Improving LLM-based Document-level Machine Translation with Multi-Knowledge Fusion

Anonymous ACL submission

Abstract

Recent studies in prompting large language model (LLM) for document-level machine translation (DMT) primarily focus on the inter-sentence context by flatting the source document into a long sequence. This approach relies solely on the sequence of sentences within the document. However, the complexity of document-level sequences is greater than that of shorter sentence-level sequences, which may limit LLM's ability in DMT when only this single-source knowledge is used. In this paper, we propose an enhanced approach by incorporating multiple sources of knowledge, including both the document summarization and entity translation, to enhance the performance of LLM-based DMT. Given a source document, we first obtain its summarization and translation of entities via LLM as the additional knowledge. We then utilize LLMs to generate two translations of the source document by fusing these two single knowledge sources, respectively. Finally, recognizing that different sources of knowledge may aid or hinder the translation of different sentences, we refine and rank the translations by leveraging a multi-knowledge fusion strategy to ensure the best results. Experimental results in eight document-level translation tasks show that our approach achieves an average improvement of 0.8, 0.6, and 0.4 COMET scores over the baseline without extra knowledge for `LLaMA3-8B-Instruct`, `Mistral-Nemo-Instruct`, and `GPT-4o-mini`, respectively. We will release our code on GitHub.

1 Introduction

Large language model (LLM) has shown impressive performance across various natural language processing (NLP) tasks (Adams et al., 2023; Dong et al., 2023). Many researchers also explore how to utilize LLM to solve the document-level machine translation (DMT), in which LLM needs to capture the inter-sentence dependency for addressing discourse issues, such as pronoun translation and word translation inconsistency. Existing approaches can be roughly categorized into 1) supervised fine-tuning (SFT) approaches (Lyu et al., 2024; Li et al., 2024; Wu et al., 2024) and 2) prompt engineering (PE) approaches (Wang et al., 2023b; Wu and Hu, 2023). The former directly leverages the document-level parallel corpus to tune LLM via some parameter-efficient methods, while the later mainly relies on the ability of LLM in in-context learning. Compared to SFT approaches, the PE approaches do not require additional computing resources to train or tune LLM, and are more resource-efficient. Therefore, in this paper we effectively explore LLM for DMT by a novel PE approach, called multi-knowledge fusion.

Despite its resource-efficient superiority, the context or knowledge utilized in the existing PE approach is limited. For example, Wang et al. (2023b) translate documents sentence by sentence, using only inter-sentence context/knowledge most relevant to the current sentence. For sentence-level machine translation, benefiting from the powerful in-context learning ability, randomly sampling bilingual parallel sentence pairs as prompts can effectively enhance the translation abilities of LLMs. However, different from sentence-level translation, the single knowledge fusion strategy may not effectively solve the tough discourse problem in DMT. Professional human translators typically explicit multi-knowledge information to ensure that its translation has stronger coherence and lexical cohesion when translating a source document, such the topics, keywords, and entity words.

When translating a document, as depicted in Figure 1, a professional translator first reads the entire text to understand its content. Additionally, the translator may highlight key entities within the document and consider their translations in advance. To mimic this behavior, we propose a

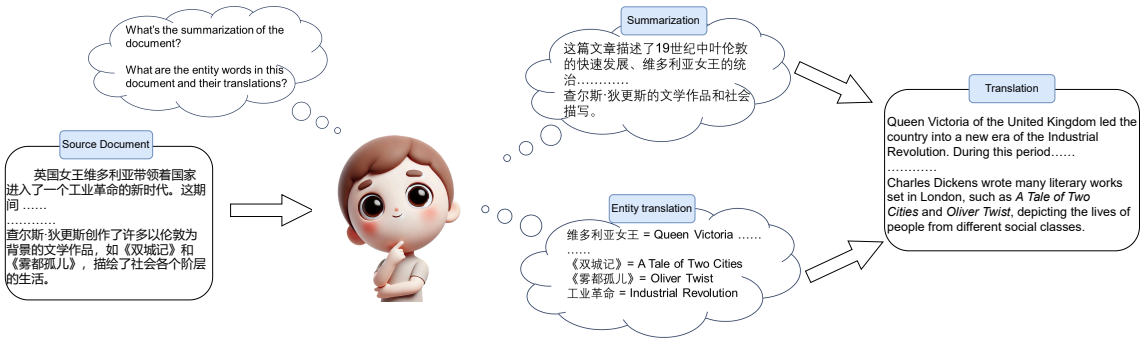

Figure 1: Illustration of a professional translator translating a document from Chinese to English.

multi-knowledge fusion approach to prompt LLMs to generate better document translation. Motivated by He et al. (2024), our approach consists of three essential steps:

- *Document-Level Knowledge Acquisition*: In this initial step, we prompt the LLM to extract two critical types of inter-sentence knowledge: **summarization** and **entity translation**. Summarization enables readers to quickly grasp the main ideas of lengthy documents, facilitating a better understanding of the key points. This improves comprehension of the overall context allows the LLM to produce more coherent translations. Additionally, knowledge of entity translation aids in organizing and structuring documents by maintaining consistency in the translation of entities, which enhances the clarity of the output.

- *Single-Knowledge Integration*: This step involves incorporating specific pieces of acquired knowledge into the process of document-level machine translation. While not every sentence in the document will need this integration, certain sentences will benefit from the added knowledge.

- *Multi-Knowledge Fusion*: In the final step, we re-evaluate and rank translations that integrate multiple facets of knowledge. This process involves merging various elements and refining the translations to ensure that the final output accurately and comprehensively represents the source document.

Overall, our main contributions in this work can be summarized as follows:

- We introduce two additional aspects of knowledge, entity word translation and summarization, guiding LLMs to generate better document translations.

- Interestingly, we observe that a single type of knowledge can improve the translation of certain sentences in a document while potentially harming some others. To address this, we propose a novel multi-knowledge fusion strategy to enhance the performance of LLM-based DMT further.

- Upon various LLMs, including `LLaMA3-8B--Instruct`, `Mixtral-Nemo-Instruct` and `GPT-4o-mini`, we demonstrate the effectiveness of the proposed approach across eight document-level machine translation directions. And additional analysis further verify the superiority of the proposed approach in addressing discourse issues.

2 Background

In this section, we briefly introduce the conventional DMT and prompting LLM for DMT.

2.1 Conventional DMT

Given a source document $\mathcal{X} = \{X_1, \cdots, X_N\}$ with N sentences, the conventional DMT model maps each sentence $X_i = \{x_1, \cdots, x_{|X_i|}\}$ with $|X_i|$ words into the target sentence Y_i by leveraging the inter-sentence context C. More specifically, the target document $\mathcal{Y}$ is generated as follows:

$$\mathcal{Y} = \arg\max P\left(\mathcal{Y}|\mathcal{X};\theta\right), \tag{1}$$

$$P(\mathcal{Y}|\mathcal{X};\theta) = \prod_{i=1}^{N}\prod_{j=1}^{|Y_i|} P(y_j^i|y_{<j}^i, X_i, C;\theta), \tag{2}$$

where θ denotes the model parameters and $|Y_i|$ is the length of sentence Y_i. C includes both the source-side and target-side inter-sentence contexts.

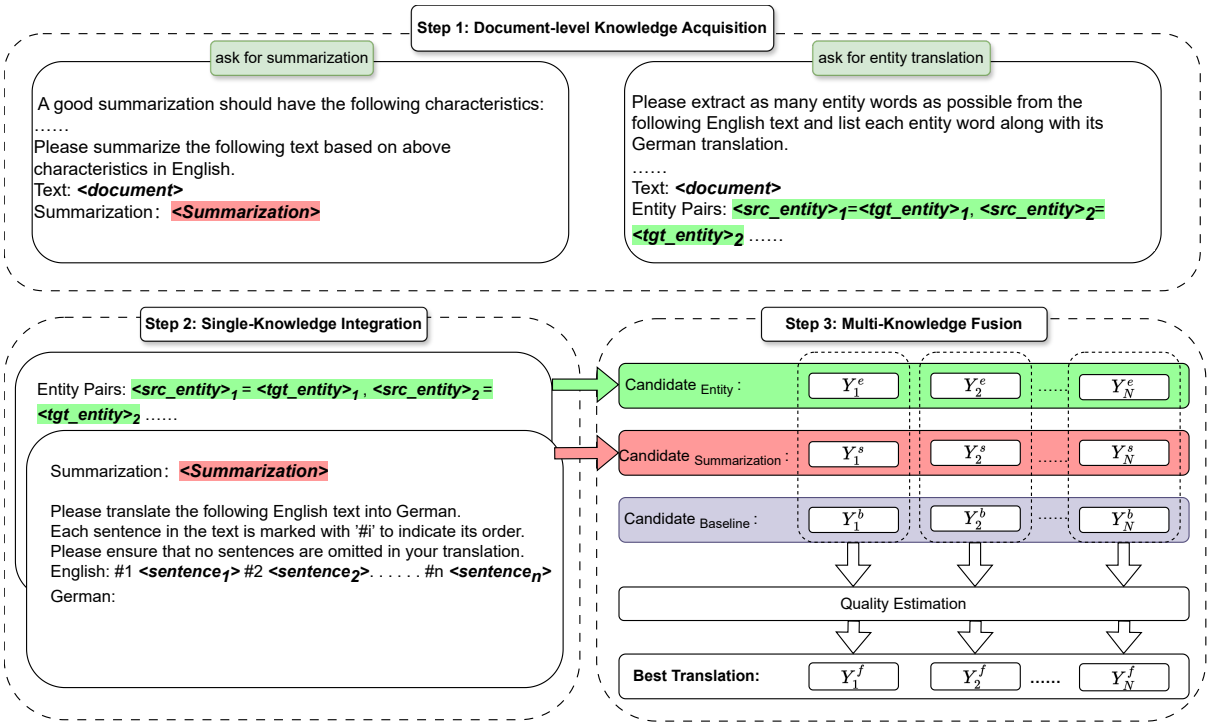

Figure 2: Illustration of our approach, which mimics the human-preferring translating process. Given a document, we first obtain its summarization and entity translation (**step 1**), then prompt LLMs to generate better document translation based on these additional knowledge (**step 2** and **step 3**).

2.2 Prompting LLM for DMT

Different from the conventional DMT, the LLM has impressive ability in instruction following. We can prompt LLM to translate the given document into the target document via concatenating a translation instruction. Similarly, LLM generates the translation of a given document as follows:

$$\mathcal{Y} = \arg\max P(\mathcal{Y}|\mathcal{X}, \mathcal{P}; \theta), \tag{3}$$

$$P\left(\mathcal{Y}|\mathcal{X}, \mathcal{P}; \theta\right) = \prod_{i=1}^{N} \prod_{j=1}^{|Y_i|} P(y_j^i | y_{<j}^i, X_i, C, \mathcal{P}; \theta), \tag{4}$$

where $\mathcal{P}$ is the instruction text.

3 Methodology

As shown in Figure 2, our approach consists of three essential steps: document-level knowledge acquisition, single-knowledge integration, and multi-knowledge fusion.

3.1 Document-Level Knowledge Acquisition

Given the source document $\mathcal{X}$, we use **summarization** and **entity translation** to provide additional context, helping the LLM produce a more accurate and fluent translation of the document.

Summarization Knowledge. A summarization provides an overall view of the document, capturing its main themes and key points. By generating a summarization, the LLM gains a clearer understanding of the document's overall context, which helps in translating complex ideas, metaphors, and culturally specific content more accurately. Research by Pu et al. (2023) and Zhang et al. (2024) demonstrate that large models often produce summarization with superior fluency and authenticity compared to humans. Thus, we first prompt the LLM to summarize the source document, with the specific prompt outlined in row #1 of Table 1.

Entity Translation Knowledge. Entity translation knowledge can enhance the translation consistency of specific terms in the document (Lyu et al., 2021). Additionally, identifying entities in the text can reduce the occurrence of untranslated segments. The entities used here include not only the general types of entities such as *name, place* and *organization* but also *events*. We prompt the LLM to fetch the entity translation knowledge with the instruction shown in row #2 of Table 1.

ID	Task	Prompt Template
#1	Summarization Acquisition	A good summarization should have the following characteristics: - Include the main points - Include key details - Be concise (no more than 3 sentences) - Remain objective Please summarize the following text based on above characteristics in English. Text: ***<document>*** Summarization:
#2	Entity Translation Acquisition	Please extract as many entity words as possible from the following ***<src_lang>*** text and list each entity word along with its ***<tgt_lang>*** translation. Entity words include but are not limited to: Person, Organization, Location, Date, Money, Work of Art, Product, Event, Occupation, Social Group, Animal, and so on. Text: ***<document>*** Entity Pairs:
#3	Prompting LLM for DMT w/o Knowledge	Please translate the following ***<src_lang>*** text into ***<tgt_lang>***. Each sentence in the text is marked with '#i' to indicate its order. Please ensure that no sentences are omitted in your translation. ***<src_lang>***: #1 ***<sentence$_1$>*** #2 ***<sentence$_2$>*** ..... ***<tgt_lang>***:
#4	Prompting LLM for DMT with Summarization	Summarization: ***<summarization>*** Please translate the following ***<src_lang>*** text into ***<tgt_lang>***. Each sentence in the text is marked with '#i' to indicate its order. Please ensure that no sentences are omitted in your translation. ***<src_lang>***: #1 ***<sentence$_1$>*** #2 ***<sentence$_2$>*** ..... ***<tgt_lang>***:
#5	Prompting LLM for DMT with Entity Translation	Entity pairs: ***<src$_{E_1}$>*** = ***<tgt$_{E_1}$>*** , ***<src$_{E_2}$>*** = ***<tgt$_{E_2}$>*** , ... Please translate the following ***<src_lang>*** text into ***<tgt_lang>***. Each sentence in the text is marked with '#i' to indicate its order. Please ensure that no sentences are omitted in your translation. ***<src_lang>***: #1 ***<sentence$_1$>*** #2 ***<sentence$_2$>*** ..... ***<tgt_lang>***:

Table 1: Prompt templates used for document-level knowledge acquisition (#1, #2 and #3) and single-knowledge integration (#4 and #5).

3.2 Single-Knowledge Integration

As long as we obtain the extracted knowledge from the given document, i.e., the summarization or entity translation, we explicitly integrate the knowledge into $\mathcal{P}$, as shown in rows #4 and #5 of Table 1, prompting LLM to generate more accurate translation by Eq. 3 and 4.

After that, we generate two different translations of the source document: $\mathcal{Y}^s$, incorporating summarization knowledge, and $\mathcal{Y}^e$, incorporating entity translation knowledge. Additionally, we produce $\mathcal{Y}^b$, a baseline translation without additional knowledge, as detailed in row #3 of Table 1.

3.3 Multi-Knowledge Fusion

Intuitively, each piece of knowledge does not always be beneficial to the translations of all sentences within the source document. The summarization knowledge can promote the translation quality of the sentences that are more related to main topic of the document. While the translation quality of sentences in which appear more entity words tend to be benefit from the entity translation knowledge. To better leverage these different types of knowledge, we fuse different knowledge to obtain a better translation by integrating $\mathcal{Y}^s$, $\mathcal{Y}^e$ and $\mathcal{Y}^b$. Specifically, we assume each translation of $\mathcal{X}$, i.e., $\mathcal{Y}^s = \{Y_1^s, \cdots, Y_N^s\}$, $\mathcal{Y}^e = \{Y_1^e, \cdots, Y_N^e\}$ and $\mathcal{Y}^b = \{Y_1^b, \cdots, Y_N^b\}$, contain N translation segments, corresponding to the translations of N sentence within $\mathcal{X}$.[1] We first select its best translation, Y_i^f, for each sentence X_i in $\mathcal{X}$ from Y_i^s, Y_i^e and Y_i^b:

$$Y_i^f = \arg\max \ S(Y, X_i), \tag{5}$$

where $Y \in \{Y_i^s, Y_i^e, Y_i^b\}$ and $S(\cdot)$ is a reference-free scoring function. Then the final translation of $\mathcal{X}$ can be formulated as $\mathcal{Y}^f = \{Y_1^f, \cdots, Y_N^f\}$.

[1] Sentence-level translations can be readily obtained using the instructions provided in row #3 of Table 15 in Appendix H. In row #1 and #2 ,we can see our instructions for obtaining stable format of summarization and entity translation.

4 Experimentation

We verify the effectiveness of our approach on three popular LLMs, including open-source and closed-source, across eight translation directions.

4.1 Experimental Settings

LLMs and Datasets. We evaluate our approach upon three LLMs, including `GPT-4o-mini` (OpenAI, 2024)[2], `LLaMA3-8B-Instruct` (Meta, 2024)[3] and `Mistral-Nemo-Instruct` (MistralAI, 2024)[4]. Our test set are extracted from WMT 2023 News Commentary v18, including English (En) $\rightleftharpoons$ {German (De), French (Fr), Spanish (Es), Russian (Ru)} eight translation directions. The test set for each translation direction contains 150 document pairs. For additional details, please refer to Table 7 in Appendix A.

Inference Settings. We run the inference of open-source LLMs, i.e., `LLaMA3-8B-Instruct` and `Mistral-Nemo-Instruct`, on a single NVIDIA V100 32GB GPU using greedy decoding strategy. For closed-source `GPT-4o-mini`, we run the inference by calling the official API. The temperature is set to 0 in the inference of all LLMs. In multi-knowledge fusion, we employ reference-free COMET[5] as the scoring function in Eq. 5.

Evaluation Metrics. Following recent studies (Wang et al., 2023b; Li et al., 2024; Wu et al., 2024), we report reference-based COMET (Rei et al., 2022) to evaluate system performance. Specifically, We use `wmt22-comet-da`[6] as our evaluation model. For translation performance in dCOMET and BLEU, please refer to Table 8 and Table 9 in Appendix B. Additionally, we report performance using the BlonDe metric (Jiang et al., 2022), which evaluates discourse phenomena based on a set of discourse-related features, with further details available in Appendix D.

4.2 Experimental Results

For better demonstrating the effect of integrating knowledge, we also build a `Reranking` system that ranks three different translation generated by `Baseline`. Table 2 presents the main experimental results, which highlight the following observations:

- Our multi-knowledge-fusion approach significantly enhances LLM performance on DMT. Specifically, our `KFMT` achieves an average improvement of 0.8, 0.6, and 0.4 COMET scores over the `Baseline` for `LLaMA3-8B-Instruct`, `Mistral-Nemo-Instruct`, and `GPT-4o-mini`, respectively. $\texttt{KFMT}_{\texttt{Oracle}}$ represents the upper bound of our approach, with an average maximum improvement of 1.2, 1.0, 0.6 COMET scores across the three LLMs.

- Due to the predominance of English data in the training of LLMs, our proposed approach shows a more pronounced improvement in En$\rightarrow X$ translation tasks compared to $X\rightarrow$En translation tasks. For instance, with `LLaMA3-8B-Instruct`, the average improvement for En$\rightarrow X$ translation tasks is 1.1, which is notably higher than the 0.5 improvement observed for $X\rightarrow$En translation tasks.

- Our approach consistently outperforms the naive `Reranking` approach, which does not incorporate any additional knowledge during reranking. This further suggests the necessity of integrating diverse knowledge sources.

- The single-knowledge fusion methods, namely `SuMT` and `EnMT`, do not always show improvements over the `Baseline`. This indicates that the benefits of different types of knowledge are most pronounced in translations of sentences closely related to that specific knowledge, rather than in all sentences within a document.

4.3 Experimental Analysis

To clarify the proposed approach, we conduct an in-depth analysis using the En$\rightleftharpoons$Ru and En$\rightleftharpoons$FR translation tasks as representatives. These analyses are carried out with the `LLaMA3-8B-Instruct` model to gain further insights. Additionally, Appendix F includes an analysis of summarization and entity translation accuracy.

Effect of Each Knowledge. Although the results presented in Table 2 highlight the overall effectiveness of our approach, the specific impact of each type of knowledge on the final translation is not entirely clear. To address this, we analyze the influence of different types of knowledge from two

[2] https://openai.com/research/gpt-4
[3] https://ai.meta.com/blog/meta-llama-3
[4] https://mistral.ai/news/mistral-nemo/
[5] https://huggingface.co/Unbabel/wmt22-cometkiwi-da
[6] https://huggingface.co/Unbabel/wmt22-comet-da

System	En→De	De→En	En→Es	Es→En	En→Ru	Ru→En	En→Fr	Fr→En	Average
LLaMA3-8B-Instruct									
Baseline	85.2	88.2	87.1	88.8	83.8	83.9	84.9	87.0	86.1
Reranking	85.7	88.4	87.4	88.9	84.5	84.2	85.3	87.2	86.5
SuMT	85.3	88.3	87.2	88.8	83.7	84.1	85.0	87.3	86.2
EnMT	85.3	88.3	86.9	88.4	83.4	83.9	84.8	86.9	86.0
KFMT	86.1	88.6	87.8	89.0	85.5	84.7	85.8	87.6	86.9
KFMT$_{\text{Oracle}}$	**86.4**	**88.8**	**88.2**	**89.4**	**86.0**	**85.1**	**86.3**	**88.0**	**87.3**
Mistral-Nemo-Instruct									
Baseline	86.5	89.0	87.3	89.4	87.0	85.2	85.9	88.0	87.3
Reranking	87.1	89.0	87.7	89.4	87.4	85.2	86.3	88.0	87.5
SuMT	86.5	88.5	87.5	89.1	87.0	84.7	86.0	87.6	87.1
EnMT	86.6	88.8	87.3	89.2	87.0	85.0	86.0	87.8	87.2
KFMT	87.6	89.3	88.2	89.6	87.9	85.4	86.7	88.1	87.9
KFMT$_{\text{Oracle}}$	**88.1**	**89.6**	**88.5**	**89.9**	**88.3**	**85.9**	**87.1**	**88.6**	**88.3**
GPT-4o-mini									
Baseline	88.5	89.3	88.9	89.6	88.7	85.5	87.3	88.1	88.2
Reranking	88.7	89.4	89.1	89.7	88.8	85.6	87.5	88.2	88.4
SuMT	88.5	89.3	88.9	89.6	88.7	85.5	87.3	88.1	88.2
EnMT	88.2	89.1	88.7	89.4	88.3	85.3	87.0	87.9	88.0
KFMT	88.9	89.6	89.3	89.9	89.2	85.8	87.7	88.4	88.6
KFMT$_{\text{Oracle}}$	**89.1**	**89.8**	**89.4**	**90.0**	**89.4**	**86.1**	**87.8**	**88.6**	**88.8**

Table 2: Performance in reference-based COMET score. **Baseline** shows results from prompting LLMs without additional knowledge. **Reranking** denotes ensemble results by reranking three translations generated by **Baseline**. **SuMT** and **EnMT** are the results for prompting LLMs via *summarization* and *entity translation* knowledge, respectively. **KFMT** and **KFMT$_{\text{Oracle}}$** are the results with the multi-knowledge fusion strategy, where **KFMT** uses a reference-free scoring function and **KFMT$_{\text{Oracle}}$** uses a reference-based one to select the best translation.

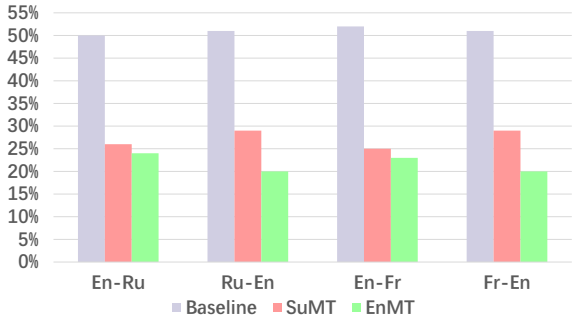

Figure 3: Visualization of the proportions of translations produced by the Baseline, SuMT, and EnMT systems relative to the total number of translations.

System	En→Ru	Ru→En	En→Fr	Fr→En
Baseline	83.8	83.9	84.9	87.0
KFMT	85.5	84.7	85.8	87.6
w/o Sum.	84.8	84.4	85.5	87.4
w/o Enti.	85.0	84.5	85.5	87.5

Table 3: Ablation study for the summarization and entity translation fusion. *w/o* Sum. and *w/o* Ent. denote that we remove the translation $\mathcal{Y}^s$ and $\mathcal{Y}^b$ when generating the final translation via Eq. 5, respectively.

perspectives. First, we visualize the proportion of sentences translated with entity ($\mathcal{Y}^e$), summarization ($\mathcal{Y}^s$), or without any additional knowledge ($\mathcal{Y}^b$) across all sentences in the test set, as shown in Figure 3.[7] Our observations reveal that, on average, 50% of the translations come from the baseline system. The remaining 50% of sentences benefit from incorporating summarization or entity translation knowledge, with summarization knowledge being utilized more frequently than entity translation knowledge. This indicates that integrating more relevant knowledge can significantly enhance translation quality, while incorporating redundant knowledge might actually impair it. Second, we perform an ablation study by individually removing either summarization or entity translation knowledge from the knowledge fusion. As shown in Table 3, the results tell that summarization knowledge plays a more critical role than entity translation knowledge in enhancing translation performance. For more details of comparing SuMT (or EnMT) with Baseline, please refer to Appendix E.

Performance in Translation Fluency. Since *KFMT* selects translations from multiple translation systems, we investigate whether this affects

[7] From Section 3.3, all sentences in our final translation, i.e., $\mathcal{Y}^f$, are selected from $\mathcal{Y}^b$, $\mathcal{Y}^e$ or $\mathcal{Y}^s$ by Eq. 5.

System	En→Ru	Ru→En	En→Fr	Fr→En
		Perplexity		
Baseline	10.7	30.8	62.7	29.9
KFMT	**7.5**	**29.5**	**57.5**	**29.6**
		Coherence		
Baseline	56.6	41.0	91.7	42.1
KFMT	56.6	41.1	**92.6**	42.2

Table 4: Translation fluency evaluation in perplexity and coherence.

System	En→Ru	Ru→En	En→Fr	Fr→En
Baseline	70.1	82.7	79.7	84.0
KFMT	**74.6**	**84.7**	**82.9**	**86.1**

Table 5: Averaged evaluation score of GPT-4o.

translation fluency. Fluency refers to how well the translated text aligns with the norms and naturalness of the target language, which in turn enhances its readability and ease of understanding. Building on previous studies (Li et al., 2023a; Kallini et al., 2024), we evaluate translation fluency using two metrics: perplexity and coherence. Specifically, we compute perplexity scores using the GPT-2 (Radford et al., 2019), and assess coherence by measuring the similarity between neighboring sentences using the SimCSE (Gao et al., 2021).

As shown in Table 4, KFMT achieves an improvement of 2.5 scores in perplexity over Baseline while it maintains stability and consistency with the Baseline in terms of the coherence metric. while maintaining stability and consistency in coherence scores. This suggests that, despite selecting translations from different systems, our multi-knowledge fusion approach either preserves or enhances the fluency of document-level machine translations.

GPT-based and Human Evaluations. In addition to the automatic evaluation metrics, we employ GPT-based and human evaluation to achieve a more comprehensive assessment of our results.

Research by Kocmi and Federmann (2023) suggests the superiority of GPT-based evaluations over traditional automatic metrics like BLEU and COMET, with reliable evaluation requiring models outperforming GPT-3.5-turbo. Consequently, we utilized GPT-4o for our evaluations. GPT-4o not only outperforms GPT-3.5-turbo but is also comparable to GPT-4. Specifically, we adopt the prompt template from Kocmi and Federmann (2023) but refined it to address issues with the original prompt,

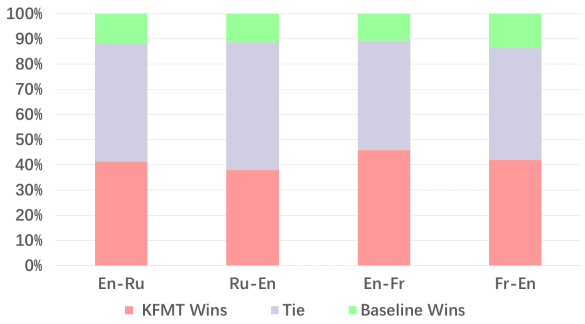

Figure 4: Human evaluation results on the test set when comparing KFMT with Baseline.

	En→Ru	Ru→En	En→Fr	Fr→En
Baseline	83.8	83.9	84.9	87.0
TFMT	84.0	84.2	85.1	87.2
KFMT	**85.5**	**84.7**	**85.8**	**87.6**

Table 6: Performance comparison of token-level knowledge fusion (TFMT) and our KFMT.

which included unnecessary explanations. This refinement resulted in a stable and consistent format with 100% reliability. In our prompt, the model is instructed to assign a score from 0 to 100 to the translation results, where 0 indicates "no retained meaning" and 100 denotes "perfect meaning and grammar." For the specific prompt used, see Table 14 in Appendix G. Table 5 presents the average scores, demonstrating that KFMT outperforms Baseline across all language pairs.

We also conduct a human evaluation on the test set. For each document, annotators receive the source document along with translations of KFMT and Baseline, presented in random order. In line with Lyu et al. (2021), annotators are asked to choose from one of three options based on fluency and correctness: (1) the first translation is better, (2) the second translation is better, or (3) both translations are of equal quality. Two annotators, one for En⇌Ru and one for En⇌Fr, are encouraged to select one of the first two options if they can identify a clear preference, rather than opting for the third. Figure 4 displays the results. On average, annotators rate 46.3% of the cases as having equal quality, while KFMT is preferred over Baseline in 41.8% of the cases compared to 11.9%, indicating a clear preference for our approach.

Comparison to Token-Level Knowledge Fusion. Our approach to multi-knowledge fusion operates at the sentence level, as it involves reranking translations on a per-sentence basis. In contrast, we in-

vestigate token-level multi-knowledge fusion. For a decoder-only model used in translation, the probability of token t at the i-th time step is given by:

$$P(t_i) = P(t_i \mid \mathcal{P}, \mathcal{X}, t_{j<i}), \quad (6)$$

where $\mathcal{P}$ represents the prompt, $\mathcal{X}$ denotes the source document, and $t_{j<i}$ indicates the previously generated tokens. Let $P_B(t_i)$,$P_S(t_i)$,$P_E(t_i)$ represent the probabilities of token at the i^{th} time step for the systems of `Baseline`, `SuMT`, and `EnMT`, respectively. Motivated by Hoang et al. (2024), we perform token-level fusion by combining these systems in an ensemble. We assign weight parameters λ_1, λ_2 and λ_3 to the respective systems, ensuring that their sum equals 1 ($\lambda_1 + \lambda_2 + \lambda_3 = 1$). Thus, the probability of i-th token t_i is:

$$P_{\text{ensemble}}(t_i) = \lambda_1 P_B(t_i) + \lambda_2 P_S(t_i) + \lambda_3 P_E(t_i). \quad (7)$$

During inference, we set the temperature to 0 and the weight to $\lambda_1 = 0.4$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.3$, respectively. Table 6 compares the performance. It shows that token-level knowledge fusion (i.e., `TFMT`) provides an average improvement of 0.2 COMET scores over `Baseline`. However, it performs less effectively compared to our proposed `KFMT`, which achieves an average improvement of 0.7 COMET scores.

5 Related Work

Conventional Document-Level Machine Translation. Conventional DMT, which are built upon the Transformer (Vaswani et al., 2017), have made significant advancements in recent years. These models generally fall into two main categories. The first category focuses on translating document sentences one by one while incorporating document-level context (Zhang et al., 2018; Maruf et al., 2019; Zheng et al., 2020). The second category extends the translation unit from a single sentence to multiple sentences (Tiedemann and Scherrer, 2017; Agrawal et al., 2018; Zhang et al., 2020) or the entire document (Junczys-Dowmunt, 2019; Liu et al., 2020; Bao et al., 2021; Li et al., 2023b).

LLMs for Document-Level Machine Translation. The adaptation of LLMs for DMT is an emerging research area with significant potential. Current research in this domain primarily explores two types of approaches: supervised fine-tuning and prompt engineering. Supervised fine-tuning aims to enhance LLMs' capabilities for document-level machine translation through targeted training. For example, Zhang et al. (2023) fine-tune the model using the Q-LORA method and compare its document translation performance with the prompt engineering approach. Wu et al. (2024) introduce a two-step fine-tuning method. Initially, LLMs are fine-tuned on monolingual data, and subsequently, they are further fine-tuned on parallel documents. Li et al. (2024) propose a hybrid approach that integrates sentence-level fine-tuning instructions into the document-level fine-tuning process, which aims to improve overall translation performance. Furthermore, Lyu et al. (2024) show that LLMs can be more effectively adapted for context-aware NMT by discriminately modeling and utilizing both inter- and intra-sentence contexts.

Prompt engineering focuses on designing prompts to optimize DMT. Wang et al. (2023b) investigate how various document translation prompts impact translation performance and assess the capabilities of different LLMs. Additionally, Cui et al. (2024) explore the use of contextual summaries to select the most relevant examples, thus enhancing sentence context for translation. Moreover, Wang et al. (2024) propose a document-level translation agent that improves the consistency and accuracy of document translation by utilizing a multi-level memory structure. Our work is aligned with prompt engineering but diverges from previous approaches by proposing the integration of various types of knowledge to enhance document translation. The most relevant work to ours is He et al. (2024) which investigates the use of knowledge for sentence-level translation. In contrast, our approach extends this research to DMT by incorporating document-level knowledge.

6 Conclusion

In this paper, we propose a multi-knowledge fusion approach that mimics human translators for document-level machine translation. Our approach explicitly combines different types of knowledge to enhance translation quality. It involves three key steps: First, it acquires two types of document-level knowledge—summarization and entity translation. Next, it integrates this knowledge to improve the translation process. Finally, recognizing that different knowledge sources may impact sentence translation differently, we optimize results using a multi-knowledge fusion strategy for refinement and ranking. Experiments across eight DMT tasks demonstrate that our approach consistently enhances performance across three different LLMs.

Limitations

Our approach has only been validated on a news dataset, and all its language pairs include English, so its effectiveness on a broader range of datasets and non-English language pairs remains uncertain. Moreover, in terms of performance improvements across the three LLMs, our approach demonstrates more significant gains on LLMs with weaker document translation performance, whereas the relative improvement is less pronounced on LLMs with stronger translation capabilities.

References

Griffin Adams, Alexander R. Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: GPT-4 summarization with chain of density prompting. *CoRR*, abs/2309.04269.

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of EAMT*, pages 11–20.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. In *Proceedings of ACL*, pages 3442–3455.

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024. Efficiently exploring large language models for document-level machine translation with in-context learning. In *Findings of ACL*, pages 10885–10897.

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2023. Self-collaboration code generation via chatgpt. *CoRR*, abs/2304.07590.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP*, pages 6894–6910.

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.

Hieu Hoang, Huda Khayrallah, and Marcin Junczys-Dowmunt. 2024. On-the-fly fusion of large language models and machine translation. In *Findings of NAACL*, pages 520–532.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of NAACL-HLT*, pages 1550–1565.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of WMT*, pages 225–233.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of ACL*, pages 14691–14714.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of EAMT*, pages 193–203.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023a. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of ACL*, pages 12286–12312.

Yachao Li, Junhui Li, Jing Jiang, Shimin Tao, Hao Yang, and Min Zhang. 2023b. P-Transformer: Towards Better Document-to-Document Neural Machine Translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3859–3870.

Yachao Li, Junhui Li, Jing Jiang, and Min Zhang. 2024. Enhancing document-level translation of large language model via translation mixed-instructions. *CoRR*, abs/2401.08088.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. 2021. Encouraging lexical translation consistency for document-level neural machine translation. In *Proceedings of EMNLP*.

Xinglin Lyu, Junhui Li, Yanqing Zhao, Min Zhang, Daimeng Wei, Shimin Tao, Hao Yang, and Min Zhang. 2024. Dempt: Decoding-enhanced multi-phase prompt tuning for making llms be better context-aware translators. In *Proceedings of EMNLP*, pages 20280–20295.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of NAACL-HLT*, pages 3092–3102.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/.

MistralAI. 2024. Mistral nemo. https://mistral.ai/news/mistral-nemo/.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/research/gpt-4.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *CoRR*, abs/2309.09558.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of WMT*, pages 578–585.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of DiscoMT*, pages 82–92.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, pages 5998–6008.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of WMT*, pages 118–128.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good NLG evaluator? A preliminary study. *CoRR*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023b. Document-level machine translation with large language models. In *Proceedings of EMNLP*, pages 16646–16661. Association for Computational Linguistics.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2024. Delta: An online document-level translation agent based on multi-level memory. *CoRR*.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *CoRR*, abs/2401.06468.

Yangjian Wu and Gang Hu. 2023. Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In *Proceedings of WMT*, pages 166–169.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of EMNLP*, pages 533–542.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2020. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of EMNLP*, pages 1081–1087.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen R. McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Trans. Assoc. Comput. Linguistics*, pages 39–57.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with qlora. In *Proceedings of WMT*, pages 468–481.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards making the most of context in neural machine translation. In *Proceedings of IJCAI*, pages 3983–3989.

A Data Statistics

Table 7 presents the detailed statistics of test sets. On average, each document across the four language pairs contains between 37 and 39 sentences.

Dataset	# Document	# Sentence
De⇌En	150	5,967
Es⇌En	150	5,815
Ru⇌En	150	5,794
Fr⇌En	150	5,619

Table 7: Data statistics on our test sets.

B Performance in dCOMET and BLEU

Unlike traditional sentence-level COMET, the document-level COMET (dCOMET) introduced by Vernikos et al. (2022) takes into account the context of previous sentences during encoding. This approach results in more accurate evaluations of document-level translations. Following the work of (Wang et al., 2024), we employed a reference-free model to obtain dCOMET scores. However, we replaced the original model, `wmt21-comet-qe-mqm`, with the latest version, `wmt22-cometkiwi-da`.The dCOMET scores, derived using the `wmt22-cometkiwi-da` model, are presented in Table 8. As shown, the performance trend of dCOMET closely follows that of sentence-level COMET, which is also presented in Table 8.

Table 9 presents the detailed performance in sentence-level BLEU (Papineni et al., 2002). From

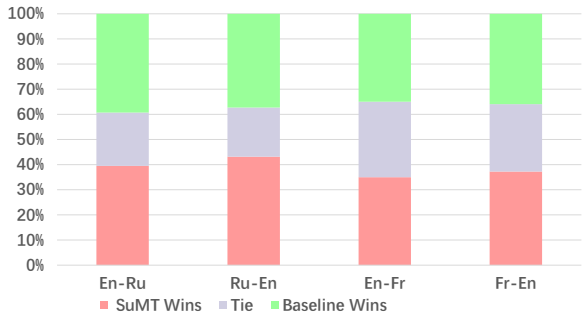

Figure 5: Comparison of SuMT and the Baseline in terms of sentence-level reference-based COMET scores.

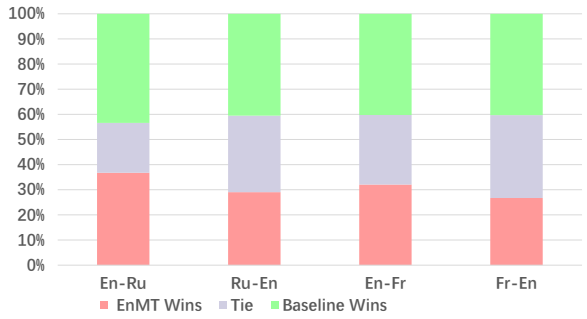

Figure 6: Comparison of EnMT and the Baseline in terms of sentence-level reference-based COMET scores.

it, we can observe that our approach achieves consistent and stable improvements in BLEU scores, although the gains in both COMET and dCOMET scores are relatively modest.

C Performance in LTCR

Following Lyu et al. (2021), we use LTCR to assess lexical translation consistency. Table 10 compares the performance of LLaMA3-8B-Instruct in the LTCR metric. The results show that our approach improves lexical consistency with an average gain of 1.62 points on the LTCR score.

D Performance in BlonDe

Table 11 compares the performance of LLaMA3-8B-Instruct in BlonDe. It shows that our approach outperforms Baseline with 4.6 BlonDe scores.

E Comparison of Baseline, SuMT, and EnMT

Incorporating a single source of knowledge benefits only certain sentences in the translation, while others may be negatively affected. Figure 5 and Figure 6 present the results of LLaMA3-8B-Instruct on the test set, comparing SuMT (or EnMT) with the Baseline at sentence-level reference-based COMET score by wmt22-comet-da. Note that we consider two translations to be tied if their COMET scores differ by no more than 0.08.

F Performance of Summarization and Entity Translation

While recent studies have demonstrated that LLMs are highly effective in generating summaries (Pu et al., 2023; Zhang et al., 2024) and translating entities (He et al., 2024), their performance on our experimental datasets remains uncertain.

The studies by Wang et al. (2023a) demonstrate that GPT evaluations achieve remarkable performance and closely align with human assessments in various NLP tasks. Therefore, we adopt and modify the reference-free scoring prompt from Wang et al. (2023a) to evaluate the quality of our summarization knowledge and entity translation knowledge. Table 12 presents the prompts used in our evaluation of the generated summarization knowledge and entity translation knowledge with GPT-4o. Table 13 provides the corresponding evaluation scores. It shows that our summarization and entity translation are of good quality.

G Prompt for GPT Evaluation

Table 14 presents the prompt template used in GPT evaluation in Section 4.3. The highlighted part demonstrates our difference from the template described in Kocmi and Federmann (2023). By specifying that the scores be returned in dictionary format, we achieve 100% consistency in format output.

H Prompt for formatting LLM's answer

Table 15 presents the prompt we used for formatting LLM's answer, including summarization, entity translation and document translation.

System	En→De	De→En	En→Es	Es→En	En→Ru	Ru→En	En→Fr	Fr→En	Average
LLaMA3-8B-Instruct									
Baseline	83.8	82.4	85.3	84.3	81.8	80.8	85.3	83.0	83.3
Reranking	84.0	82.4	85.4	84.3	82.2	81.0	85.4	83.0	83.5
SuMT	83.8	82.4	85.3	84.3	81.8	80.8	85.3	83.0	83.3
EnMT	83.8	82.4	85.3	84.2	81.7	80.8	85.3	83.0	83.3
KFMT	**84.3**	**82.7**	**85.8**	**84.6**	**82.8**	**81.4**	**85.8**	**83.5**	**83.9**
Mistral-Nemo-Instruct									
Baseline	83.7	82.6	84.8	84.4	83.0	81.1	85.3	83.4	83.5
Reranking	84.0	82.6	85.0	84.4	83.2	81.1	85.5	83.4	83.6
SuMT	83.7	82.4	84.9	84.4	83.1	80.9	85.4	83.2	83.5
EnMT	83.8	82.6	84.9	84.5	83.1	81.1	85.5	83.3	83.6
KFMT	**84.4**	**82.7**	**85.7**	**84.7**	**83.5**	**81.3**	**85.8**	**83.5**	**84.0**
GPT-4o-mini									
Baseline	84.6	82.7	86.0	84.6	83.6	81.3	85.9	83.5	84.0
Reranking	84.7	82.7	86.0	84.6	83.6	81.3	85.9	83.6	84.0
SuMT	84.6	82.7	86.0	84.6	83.6	81.3	85.9	83.5	84.0
EnMT	84.6	82.7	86.0	84.6	83.5	81.2	85.8	83.4	84.0
KFMT	**84.8**	**82.9**	**86.2**	**84.8**	**83.9**	**81.5**	**86.1**	**83.7**	**84.2**

Table 8: Performance in document-level (dCOMET) score.

System	En→De	De→En	En→Es	Es→En	En→Ru	Ru→En	En→Fr	Fr→En	Average
LLaMA3-8B-Instruct									
Baseline	31.2	39.5	42.0	45.4	27.7	31.8	35.9	37.6	36.4
Reranking	31.6	39.5	42.2	45.4	28.1	32.3	36.1	37.6	36.6
SuMT	31.0	40.0	42.0	45.7	27.4	32.4	35.7	38.4	36.6
EnMT	31.0	40.0	41.6	44.0	27.3	31.2	35.6	36.4	35.9
KFMT	**32.1**	**40.4**	**42.5**	**46.1**	**28.8**	**32.9**	**36.5**	**38.4**	**37.2**
Mistral-Nemo-Instruct									
Baseline	34.6	42.6	43.3	48.0	31.7	35.7	38.2	40.2	39.3
Rerankding	35.0	42.7	43.5	48.0	32.0	35.8	38.5	40.2	39.5
SuMT	34.2	39.5	43.2	45.4	31.2	33.4	38.5	37.8	37.9
EnMT	34.5	41.3	43.1	47.0	31.4	34.8	38.3	39.3	38.7
KFMT	**35.8**	**43.2**	**44.6**	**48.5**	**32.5**	**36.1**	**39.1**	**40.7**	**40.1**
GPT-4o-mini									
Baseline	41.2	44.0	48.6	50.1	36.0	37.0	42.0	41.4	42.5
Reranking	41.3	44.0	48.6	50.1	36.0	37.1	42.0	41.5	42.6
SuMT	41.0	44.0	48.4	50.0	35.8	36.9	41.9	41.4	42.4
EnMT	39.9	42.7	47.5	48.9	34.8	35.8	41.1	40.2	41.4
KFMT	**41.5**	**44.2**	**49.0**	**50.4**	**36.5**	**37.3**	**42.4**	**41.7**	**42.9**

Table 9: Performance in SacreBLEU score.

System	En→De	De→En	En→Es	Es→En	En→Ru	Ru→En	En→Fr	Fr→En	Average
Baseline	86.67	81.25	74.19	85.71	42.42	66.67	83.33	63.64	73.00
KFMT	**93.33**	**82.35**	**75.00**	**86.67**	**45.45**	66.67	83.33	**64.12**	**74.62**

Table 10: Lexical translation consistency evaluation in LTCR.

System	De→En	Es→En	Ru→En	Fr→En	Average
Baseline	49.4	59.4	41.1	50.1	50.0
KFMT	**58.9**	**60.5**	**46.8**	**52.0**	**54.6**

Table 11: Performance in BlonDe score.

ID	Task	Prompt Template
#1	Summarization	Score the following summarization for overall quality on a continuous scale from 0 to 100. A score of zero means "poor quality" (disjointed, hard to read, or containing significant factual inaccuracies), and a score of one hundred means "excellent quality" (fluent, coherent, and consistent with the key ideas of the original text). Overall Quality measures both: 1. Fluency: Whether the summarization is well-written, grammatically correct, and easy to understand, with smooth sentence transitions and a natural flow. 2. Consistency: Whether the summarization accurately conveys the main points, key details, and intended meaning of the original text, without introducing errors, distortions, or irrelevant information. Original Text: <Original text> Summarization: <Summarization> Score:
#2	Entity Translation	Score the following ***<src_lang>*** translation of extracted entities from original text for overall quality on a continuous scale from 0 to 100. A score of zero means "poor quality" (missing important entities or containing significant translation errors), and a score of one hundred means "excellent quality" (all key entities are extracted and translated accurately). Overall Quality measures both: 1. Correctness of Entity Extraction: Whether the extracted entities include all critical persons, locations, organizations, dates, and other relevant information explicitly mentioned in the original text. 2. Translation Accuracy: Whether the English translations are faithful to the meaning, spelling, and nuances of the original entities. Original Text: <Original text> Extracted Entities (Original): <the list of extracted entities in the original language> Extracted Entities (Translated): <the list of extracted entities in ***<src_lang>***> Score:

Table 12: Prompt Templates for evaluating summrization and entity translation.

Task	En→Ru	Ru→En	En→Fr	Fr→En	Average
Summarization	78.0	78.6	77.1	80.5	78.6
Entity Translation	87.1	87.1	91.7	83.6	87.4

Table 13: Averaged evaluation score of `GPT-4o` for Summarization and Entity Translation.

Prompt Template
Score the following translation from ***<src_lang>*** to ***<tgt_lang>*** with respect to the human reference on a continuous scale from 0 to 100, where a score of zero means "no meaning preserved" and a score of one hundred means "perfect meaning and grammar". Please return the score you gave in the dictionary format of {{"score": score}}. You only need to give the score, no additional explanation is needed. ***<src_lang>*** source: ***<src_seg>*** ***<tgt_lang>*** human reference: ***<ref_seg>*** ***<tgt_lang>*** translation: ***<tgt_seg>*** Score:

Table 14: Prompt template used in `GPT` evaluation.

ID	Task	Prompt Template
#1	Formatting summarization	I hope you can return the summarization in dictionary format. The format of the dictionary is: {\" summarization \": your summarization}
#2	Formatting entity translation	I hope you can return the entity and its translation in dictionary format. The key of the dictionary is the entity, and the value is the translation of the entity.
#3	Formatting document translation	I hope you can return your translation results in dictionary format. The keys of the dictionary should be the sentence numbers, and the values should be the translation results of the sentences. For example, if your text consists of two sentences, the format of your final translation results should be: {'#1': translation result of sentence 1, '#2': translation result of sentence 2}.

Table 15: Prompt Templates for formatting LLMs' answer.