Exploring Human-Like Translation Strategy with Large Language Models

Zhiwei He[1*]   **Tian Liang**[2*]   **Wenxiang Jiao**[3]   **Zhuosheng Zhang**[1]   **Yujiu Yang**[2]
Rui Wang[1†]   **Zhaopeng Tu**[3†]   **Shuming Shi**[3]   **Xing Wang**[3]

[1]Shanghai Jiao Tong University, China   [2]Tsinghua University, China   [3]Tencent AI Lab, China
[1]{zwhe.cs,zhangzs,wangrui12}@sjtu.edu.cn
[2]{liangt21@mails,yang.yujiu@sz}.tsinghua.edu.cn
[3]{joelwxjiao,shumingshi,zptu,brightxwang}@tencent.com

Abstract

Large language models (LLMs) have demonstrated impressive capabilities in general scenarios, exhibiting a level of aptitude that approaches, in some aspects even surpasses, human-level intelligence. Among their numerous skills, the translation abilities of LLMs have received considerable attention. Compared to typical machine translation that focuses solely on source-to-target mapping, LLM-based translation can potentially mimic the human translation process, which might take preparatory steps to ensure high-quality translation. This work explores this possibility by proposing the **MAPS** framework, which stands for **M**ulti-**A**spect **P**rompting and **S**election. Specifically, we enable LLMs first to analyze the given source sentence and induce three aspects of translation-related knowledge (keywords, topics, and relevant demonstrations) to guide the final translation process. Moreover, we employ a selection mechanism based on quality estimation to filter out noisy and unhelpful knowledge. Both automatic (3 LLMs × 11 directions × 2 automatic metrics) and human evaluation (preference study and MQM) demonstrate the effectiveness of MAPS. Further analysis shows that by mimicking the human translation process, MAPS reduces various translation errors such as hallucination, ambiguity, mistranslation, awkward style, untranslated text, and omission. Source code is available at https://github.com/zwhe99/MAPS-mt.

1 Introduction

Large language models (LLMs) have recently demonstrated remarkable general capabilities across a wide range of tasks, making substantial strides in the field of artificial general intelligence. These capabilities have led to LLMs exhibiting a certain degree of human-level intelligence, particularly in the areas of language understanding and generation (Liang et al., 2022; Bubeck et al., 2023; Wu et al., 2023; Moghaddam and Honey, 2023). Among numerous tasks, translation has emerged as a prominent area where LLMs have shown impressive capacity and competence (Jiao et al., 2023b; Agrawal et al., 2023; Zhang et al., 2023a; Vilar et al., 2022; Moslem et al., 2023; Pilault et al., 2023; Garcia et al., 2023; Hendy et al., 2023; Zhu et al., 2023b; Jiao et al., 2023a; Wang et al., 2023b; Karpinska and Iyyer, 2023; Peng et al., 2023; Lyu et al., 2023; Bawden and Yvon, 2023; Lu et al., 2023). This progress above harkens back to the long-term aspirations and dreams of earlier machine translation research in the 1960s (Bar-Hillel, 1960; Macklovitch, 1995): Can LLMs employ a translation process similar to human translators?

Figure 1 illustrates the difference between the processes of machine and human translation. While conventional machine translation is typically a direct source-to-target mapping process, professional human translators tend to take preparatory steps when working with the given source text, including gathering and meticulously analyzing information such as keywords, topics, and relevant example sentences (Baker, 2018; Koehn, 2009; Bowker, 2002; Hatim and Munday, 2004). These steps are critical for ensuring high-quality translations that accurately capture the nuances of the source material. Although recent
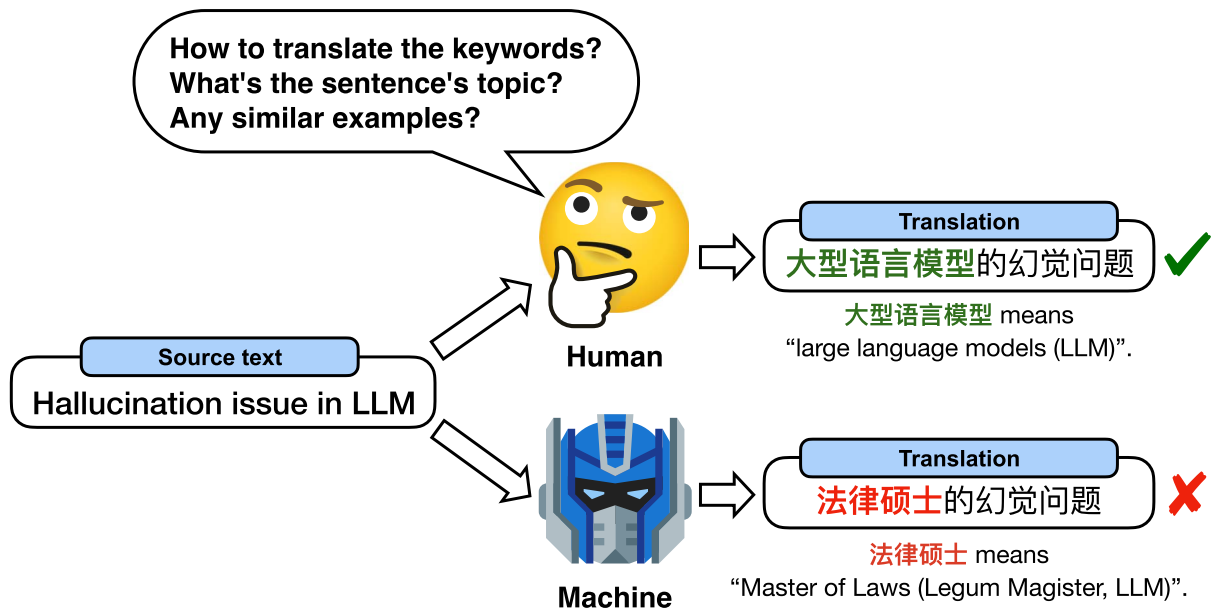
Figure 1: The difference between machine and human translation in an English→Chinese example. Typical neural machine translation is a source-to-target mapping process, while human translators can take complex steps to ensure the quality and accuracy of the translation.

advances in LLM research indicate that current LLMs are approaching human-like general intelligence (Bubeck et al., 2023; Park et al., 2023), the extent to which LLMs can emulate such strategies remains underexplored.

The primary focus of this paper is to explore whether LLMs can imitate the translation strategies employed by human translators. Specifically, we aim to investigate whether LLMs can effectively preprocess the source text and leverage the relevant knowledge to improve their translations.

To this end, we propose a method called **MAPS**, which stands for **M**ulti-**A**spect **P**rompting and **S**election. MAPS prompts the LLMs to analyze the source sentence and elicit translation-related knowledge in three aspects: **keywords**, **topics**, and **relevant demonstrations**. This knowledge then guides the LLM toward generating more accurate translations. To further enhance translation quality, we also employ a post-selection process to filter out unhelpful knowledge and select the best translation based on reference-free quality estimation (QE). We validate our approach across 11 translation directions (covering high-, medium- and low-resource language pairs) from WMT22 (Kocmi et al., 2022) and 3 LLMs (text-davinci-003, Alpaca, and Vicuna). Automatic evaluation shows that MAPS achieves significant improvement over other baselines in terms of COMET

and BLEURT. Further analysis emphasizes the importance of the extracted knowledge in resolving hallucination and ambiguity in translation. We also conduct human preference studies and Multidimensional Quality Metrics (MQM) evaluation (Burchardt, 2013) which show that MAPS produces more favorable translations by reducing mistranslation, awkward style, untranslated text, and omission errors.

In contrast to other LLM-based translation approaches, such as Dictionary-based Prompting (Ghazvininejad et al., 2023) and In-context Learning (ICL) (Agrawal et al., 2023), MAPS focuses on translating general scenarios without any preconceived assumptions about the domain of translation. As a result, MAPS does not require the preparation of any external ''datastore'', which might include a meticulously constructed glossary (Moslem et al., 2023), dictionary (Ghazvininejad et al., 2023), or sample pool (Agrawal et al., 2023), for specific language pairs and domains in advance.

In summary, the contributions of this work are detailed as follows:

- Inspired by human translation strategy, we propose the MAPS method, which mimics the human process of analyzing the source text to gather useful knowledge, ultimately leading to an accurate translation.
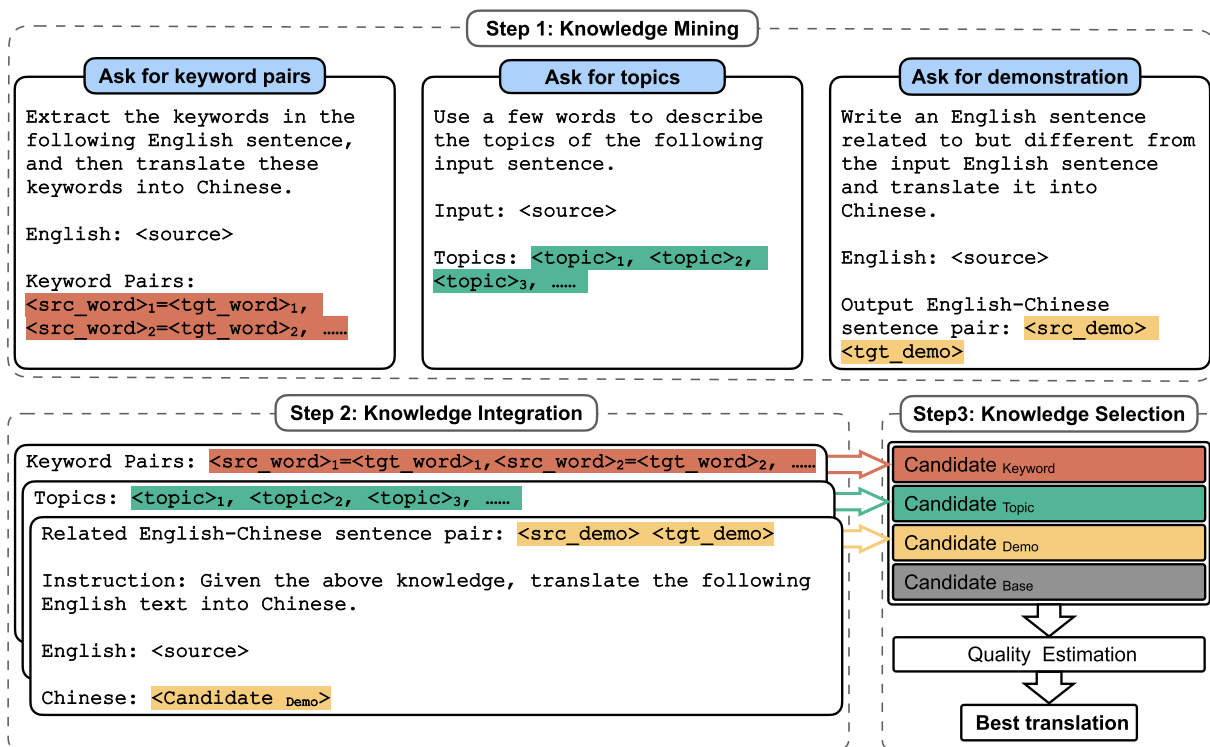
Step 1: Knowledge Mining

Ask for keyword pairs

```
Extract the keywords in the
following English sentence,
and then translate these
keywords into Chinese.

English: <source>

Keyword Pairs:
<src_word>₁=<tgt_word>₁,
<src_word>₂=<tgt_word>₂, ……
```

Ask for topics

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

Input: <source>

Topics: <topic>₁, <topic>₂,
<topic>₃, ……
```

Ask for demonstration

```
Write an English sentence
related to but different from
the input English sentence
and translate it into
Chinese.

English: <source>

Output English-Chinese
sentence pair: <src_demo>
<tgt_demo>
```

Step 2: Knowledge Integration

```
Keyword Pairs: <src_word>₁=<tgt_word>₁,<src_word>₂=<tgt_word>₂, ……

Topics: <topic>₁, <topic>₂, <topic>₃, ……

Related English-Chinese sentence pair: <src_demo> <tgt_demo>

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

English: <source>

Chinese: <Candidate Demo>
```

Step3: Knowledge Selection

Candidate Keyword
Candidate Topic
Candidate Demo
Candidate Base

⇩

Quality Estimation

⇩

Best translation

Figure 2: Framework of MAPS. On a high level, MAPS consists of three stages: *(1) Knowledge Mining*: the LLM analyzes the source sentence and generates three aspects of knowledge useful for translation: keywords, topics, and relevant demonstration; *(2) Knowledge Integration*: guided by different types of knowledge separately, the LLM generates multiple translation candidates; *(3) Knowledge Selection*: the candidate deemed best by the QE is selected as the final translation. Best viewed in color.

- We demonstrate that the three types of translation-related knowledge (keywords, topics, and relevant demonstrations) complement each other. The best translation performance can be achieved by using all three types of knowledge simultaneously.

- Our in-depth analyses of MAPS, encompassing both automatic and human evaluations, demonstrates its proficiency in resolving ambiguities and reducing hallucinations and other prevalent translation errors. Furthermore, we examined the inference time of MAPS and investigated potential acceleration techniques.

2 MAPS: Multi-Aspect Prompting and Selection

In this section, we introduce the MAPS framework. As depicted in Figure 2, MAPS consists of three steps—knowledge mining, integration, and selection. When mining the knowledge, the LLM operates in a manner akin to a human translator, analyzing the source text and generating background knowledge that is beneficial to translation purposes. The acquired knowledge is integrated as contextual guidance, enabling the LLM to produce translation candidates. However, the generated knowledge may contain noise (see §4.3 for further analysis). As a result, a filtering mechanism becomes necessary to select useful knowledge while filtering out unhelpful or noisy ones.

2.1 Knowledge Mining

Akin to the initial understanding and interpretation phase that human translators take (Gile, 2009), the knowledge mining step requires the LLM first to analyze the source text and elicit three aspects of knowledge generally beneficial to translation:

Keywords are essential words or phrases that convey the core meaning of a text and act as focal points for understanding the main idea. Accurate translation of keywords is crucial for conveying the intended meaning and ensuring

faithfulness (Baker, 2018; Koehn, 2009). Additionally, identifying and maintaining a list of keywords guarantees that specific terms are translated consistently across different parts of the text.

Topic refers to the overall subject or theme being discussed. A keen awareness of the topic helps translators sidestep potential issues arising from ambiguity, such as mistranslations or misinterpretations (Bowker, 2002). It is important to highlight that topics are generally more specific than the broader domains that have been widely discussed within the machine translation community. For example, while the news domain encompasses a wide range of subjects, subcategories like political news and entertainment news should adopt different registers and tones.

Demonstrations, or example sentences, illustrate how comparable sentences can be translated accurately. They assist the translators in identifying appropriate equivalents within the target language, enabling translators to produce natural and fluent translations to native speakers (Hatim and Munday, 2004).

As shown in Step 1 of Figure 2, given the source sentence, we prompt the LLM to elicit keyword pairs, topics, and relevant demonstrations.[1]

2.2 Knowledge Integration

Just as human translators weave their understanding of the source text into their translations (Pym, 2014), knowledge integration embeds the acquired knowledge into the context (Step 2 in Figure 2) and enables the LLM to utilize this information to generate multiple translation candidates. We obtain four candidates, which the LLM generates without guidance from any *external* knowledge.

2.3 Knowledge Selection

Knowledge selection resembles the final decision-making phase in human translation, where the best translation of the source text is chosen based on the context. Although keywords, topics, and relevant demonstrations generally benefit translation, not all the LLM-generated knowledge is helpful. For example, LLMs may generate trivial or noisy content that might distract the translation process (Shi et al., 2023; Agrawal et al., 2023). Our quantitative experiments in §4.3 support this hypothe-

sis. Therefore, we employ a filtering mechanism to select the most useful knowledge and filter out the unhelpful or noisy ones. Specifically, we adopt quality estimation (QE) to select the best candidate as the final output (Step 3 in Figure 2). The selection method is flexible, and both an externally trained QE model and the LLM itself served as QE are effective in our experiments.

3 Experiments

3.1 Experimental Setup

Models. We adopt three LLMs, encompassing both closed- and open-source models.

• text-davinci-003: A strong yet closed-source LLM developed by OpenAI, which employs advanced Reinforcement Learning with Human Feedback (RLHF) techniques (Ouyang et al., 2022). We query it via the official API.

• Alpaca (Taori et al., 2023): An open-source and instruction-following LLM fine-tuned on LLaMA model (Touvron et al., 2023a) with 52K Self-Instruct (Wang et al., 2022b) data.

• Vicuna (Chiang et al., 2023): An open-source and instruction-following LLM fine-tuned on LLaMA-2 (Touvron et al., 2023b) with user-shared conversations collected from ShareGPT (ShareGPT, 2023).

For both Alpaca and Vicuna, we use the 7B version and perform inference on a single NVIDIA V100 32GB GPU.

Comparative Methods. For a rigorous comparison, we consider several variants, including single-candidate and multi-candidate methods. Within single-candidate methods, we consider:

- **Baseline**: Standard zero-shot translation with temperature set to 0 (default value in this work).

- **5-Shot** (Hendy et al., 2023): Five high-quality labeled examples from the training data are prepended to the test input, which performs best overall in Hendy et al. (2023); meanwhile, increasing the number of examples will not result in meaningful improvement. This method requires meticulous construction of training data for each translation direction, including collecting, cleaning, and sorting by quality.

---

[1]To ensure a uniform response format, we manually constructed 5-shot exemplars for each kind of knowledge.

Within multi-candidate methods, we consider:

- **Rerank**: Using the same prompt as the Baseline, but with the temperature set to 0.3 (following Moslem et al., 2023). We randomly sample three times and add Baseline to form four candidates. The best candidate is selected through QE. It can be considered as a pure reranking method without any guidance from extracted knowledge (Fernandes et al., 2022).

- **MAPS**: Our proposed method described in Section 2. Three translation candidates are generated with guidance from three aspects of knowledge. Combined with the Baseline, the best one is selected using QE.

Knowledge Selection Methods.

- **LLM-SCQ**: Composing a single choice question (**SCQ**) that asks the LLM to choose the best candidate on its own.

- **COMET-QE**: A trained QE scorer that assigns a numerical score to each candidate. Selection is based on the highest score.

- **COMET** (oracle): A reference-based scorer that assigns a numerical score to each candidate. It can be considered as the oracle QE method, representing the upper bound of selection.

Test Data. To avoid data leakage issues (Bubeck et al., 2023; Garcia et al., 2023; Zhu et al., 2023b), we use the latest WMT22 test set, covering 11 translation directions at different resource levels (English ⇔ Chinese, English ⇔ German, English ⇔ Japanese, German ⇔ French, Ukrainian ⇔ Czech and English ⇒ Croatian). WMT22 moves away from testing only on the news domain like in previous years and shifts to focus on the general scenario covering news, social, conversational, and e-commerce (Kocmi et al., 2022).

Metrics. We adopt COMET (Rei et al., 2022a) and BLEURT (Sellam et al., 2020) as the main metrics. These neural-based learned metrics show superiority over string-based metrics like BLEU (Kocmi et al., 2021; Bawden and Yvon, 2023) and have been adopted broadly by LLM-based translation literature (Moslem et al., 2023; Hendy et al., 2023; Garcia et al., 2023; Pilault et al., 2023). We use wmt22-comet-da and BLEURT-20 checkpoints for these two metrics.

3.2 Results

For consistency, we are solely interested in comparing different methods under the same LLM. As presented in Table 1, MAPS is broadly effective and exhibits a higher upper bound. To be detailed, we have the following observations:

• **The effectiveness of MAPS has been validated across a wide range of settings.** Across 11 language pairs, 3 LLMs, and 2 metrics, MAPS consistently outperforms Rerank and Baseline. After employing MAPS $_{\text{COMET-QE}}$, text-davinci-003 surpasses the best submissions in WMT22 in 5 out of the 11 translation directions. This suggests that LLMs can enhance translation quality by emulating the human strategy of analyzing before translating.

• **MAPS outperforms Rerank consistently when the knowledge selection method is held constant.** This indicates that the improvements brought by MAPS stem from three types of translation-related knowledge: keywords, topics, and relevant demonstrations. We delve into the utilization of different types of knowledge and ablation study in §4.2.

• **Different knowledge selection methods can affect the final performance, and MAPS exhibits a higher upper bound for selection.** When using LLM-SCQ, the performance of MAPS is on par with 5-Shot (MAPS $_{\text{LLM-SCQ}}$ ≈ 5-Shot); when using COMET-QE, MAPS consistently outperforms 5-Shot (MAPS $_{\text{COMET-QE}}$ > 5-Shot). More importantly, MAPS shows higher upper bounds for selection than Rerank (MAPS $_{\text{COMET}}$ > Rerank $_{\text{COMET}}$), implying that superior knowledge selection methods like a better QE model (Rei et al., 2022b), AutoMQM (Fernandes et al., 2023) or ranking strategy (Fernandes et al., 2022) can further improve MAPS.

4 Analysis

In this section, we conduct analyses to understand the MAPS framework. If not otherwise specified, MAPS $_{\text{COMET-QE}}$, text-davinci-003, and WMT22 En-Zh are default tested method, model and language pair, respectively.

4.1 Human Evaluation

Preference Study. We perform human preference studies on En⇔Zh test sets. For each test sample, our annotators (professional translators)

Method	En-Zh	Zh-En	En-De	De-En	En-Ja	Ja-En	De-Fr	Fr-De	Cs-Uk	Uk-Cs	En-Hr
WMT22 Best | COMET											
WMT22 Best	86.8	81.0	87.4	85.0	89.3	81.6	85.7	89.5	91.6	92.2	88.4
text-davinci-003 | COMET											
Baseline	86.2	81.6	85.8	85.2	87.9	81.8	82.8	86.3	88.0	89.2	85.9
5-Shot (Hendy et al.)	87.0	81.1	86.5	85.2	88.2	82.0	83.6	86.6	—	—	—
Rerank LLM-SCQ	86.4	81.7	86.0	85.2	88.0	82.0	83.0	86.4	88.3	89.4	86.3
MAPS LLM-SCQ	86.8	**82.0**	86.4	**85.4**	**88.5**	**82.4**	83.4	**86.9**	88.8	89.9	86.5
Rerank COMET-QE	86.9	82.1	86.4	85.5	88.8	82.3	83.4	86.8	89.4	90.1	87.1
MAPS COMET-QE	**87.6**	**82.6**	**87.2**	**85.7**	**89.5**	**82.9**	**84.1**	**87.5**	90.1	91.1	88.1
⇑ **Rerank** COMET	87.5	82.6	86.9	85.8	89.3	82.3	83.4	86.8	89.9	90.7	87.7
⇑ **MAPS** COMET	**88.5**	**83.8**	**88.0**	**86.7**	**90.3**	**82.9**	**84.1**	**87.5**	90.9	92.0	89.0
text-davinci-003 | BLEURT											
Baseline	71.1	69.6	75.6	74.0	66.3	67.8	70.4	77.6	75.0	78.8	75.0
5-Shot (Hendy et al.)	72.2	69.2	76.3	74.5	67.1	68.0	70.9	78.0	—	—	—
Rerank LLM-SCQ	71.4	69.8	75.9	74.1	66.6	68.1	70.6	77.7	75.3	79.0	75.4
MAPS LLM-SCQ	72.1	**70.5**	76.3	74.4	**67.4**	**68.8**	**71.4**	78.6	76.1	80.2	76.0
Rerank COMET-QE	71.7	70.1	76.1	74.3	67.3	68.3	71.2	78.1	76.4	79.7	75.9
MAPS COMET-QE	**72.6**	**70.8**	**77.1**	**74.6**	**68.3**	**69.1**	**71.9**	**78.9**	**77.4**	**81.2**	**77.1**
⇑ **Rerank** COMET	72.4	70.6	76.5	74.6	68.0	68.8	71.8	78.6	76.8	80.2	76.4
⇑ **MAPS** COMET	**74.0**	**72.1**	**77.8**	**75.7**	**69.4**	**70.9**	**73.6**	**80.2**	**78.3**	**82.1**	**77.9**
Alpaca | COMET											
Baseline	58.9	73.1	75.5	81.9	56.6	71.8	71.7	75.4	74.1	71.1	65.9
Rerank COMET-QE	66.2	74.9	78.5	82.6	64.7	73.7	74.5	78.2	78.1	76.3	70.5
MAPS COMET-QE	**69.0**	**76.0**	**79.7**	**83.3**	**66.9**	**74.7**	**75.9**	**79.1**	**80.8**	**78.5**	**72.3**
Alpaca | BLEURT											
Baseline	42.3	58.0	62.2	69.8	31.4	55.4	52.2	63.4	52.4	54.3	53.2
Rerank COMET-QE	47.5	59.5	64.7	70.4	36.2	56.7	55.0	66.0	55.2	59.0	56.0
MAPS COMET-QE	**50.6**	**60.6**	**66.3**	**71.1**	**38.2**	**57.7**	**56.6**	**66.8**	**59.5**	**61.2**	**57.2**
Vicuna | COMET											
Baseline	81.3	78.4	79.8	82.9	82.3	77.3	75.5	77.1	74.9	72.7	69.3
Rerank COMET-QE	83.6	79.3	81.8	83.6	85.2	78.8	77.8	79.6	79.9	77.7	74.2
MAPS COMET-QE	**84.5**	**80.2**	**82.7**	**84.1**	**86.5**	**79.7**	**79.2**	**81.1**	**81.8**	**80.1**	**76.0**
Vicuna | BLEURT											
Baseline	64.9	65.3	67.4	71.0	58.7	62.8	58.8	66.0	57.8	56.6	57.7
Rerank COMET-QE	66.7	66.0	69.2	71.8	61.6	64.0	61.2	68.2	61.8	61.2	60.5
MAPS COMET-QE	**67.8**	**66.9**	**70.0**	**72.4**	**63.0**	**64.8**	**62.5**	**69.3**	**64.0**	**64.3**	**63.4**

Table 1: Translation performance on WMT22. **Bold** entries: denote statistically significant differences with $p < 0.05$ in the paired t-test compared to Baseline, 5-Shot and Rerank (with the same knowledge selection method). ⇑: indicates the upper bound of selection, using COMET, a reference-based metric, as the selection method.

were presented with a source sentence and two translations. They were then tasked with selecting the superior translation or determining that neither translation was better than the other. Figure 3 shows the results of human preference studies, and MAPS is generally more preferred by humans.

MQM Evaluation. To understand which aspects of translation that MAPS improves, we carried out MQM evaluations (Burchardt, 2013). MQM requires the annotators to identify the errors in translation and label the category and severity level for each error. Based on the weights of the different error types, the MQM ends up with a penalty score. We followed the assessment method in Freitag et al. (2021), including guidelines to annotators, error category, severity level and error weighting. We employed professional
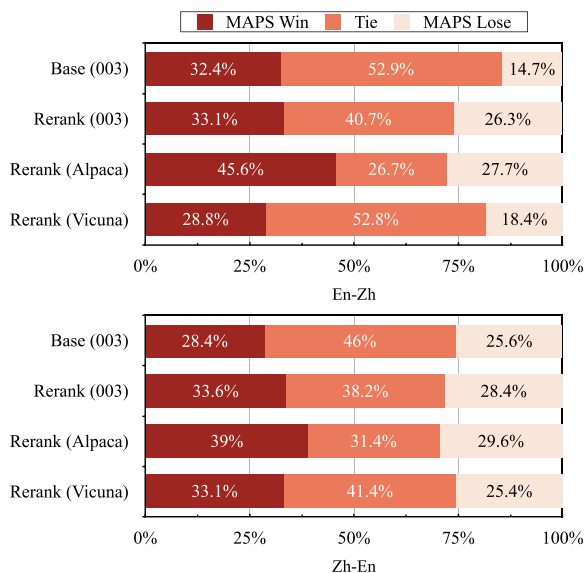
Figure 3: Human preference study, comparing MAPS with Base and Rerank. ''003'' denotes text-davinci-003.

Method	En-Zh	Zh-En
Base	1.94	2.96
Rerank	1.79	2.84
MAPS	**1.59**	**2.60**

Table 2: Averaged MQM Score (↓).

translators who had MQM experience as the annotators. We evaluated the first 1K samples on the Chinese⇔English test sets for cost reasons. Table 2 shows that MAPS outperforms Base and Rerank significantly. In terms of error categories, the improvements brought about by MAPS are mainly in the reduction of mistranslation, awkward style, untranslated text, and omission errors, as presented in Figure 4.

4.2 Utilization of Knowledge

Although Table 1 reports the overall performance of MAPS, the utilization of the three aspects of knowledge remains unclear. For instance, it is uncertain whether the majority of samples rely on relevant demonstrations rather than keywords and topics to guide the translation process. To provide further insight, we illustrate the utilization of three types of knowledge in Figure 5. We additionally present the performance differences among these three aspects of knowledge when applied to different subsets, relative to the baseline. Figure 5 reveals a relatively balanced utilization
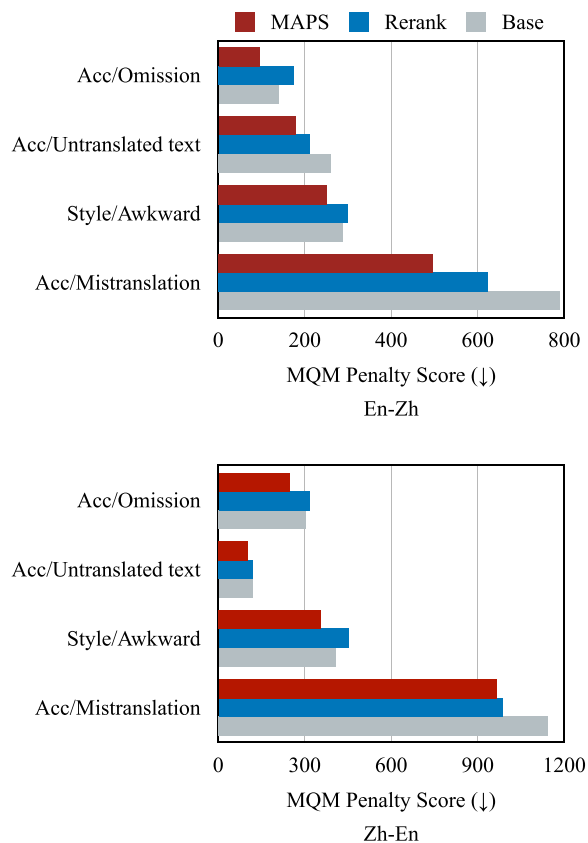

Figure 4: Selected MQM penalty scores (before average) under different error categories.
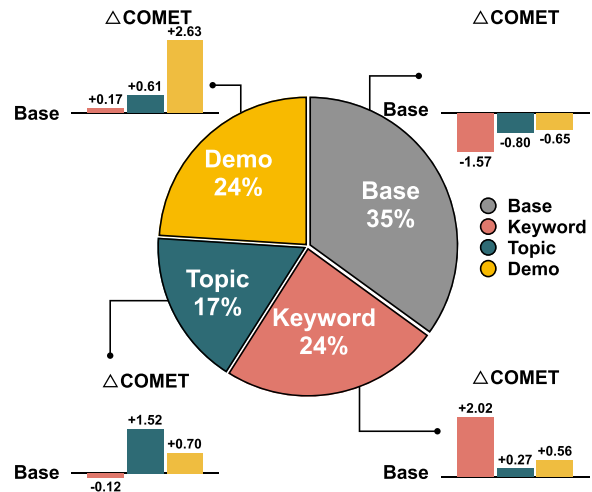

Figure 5: Utilization of keyword, topic, and relevant demonstration in MAPS.

among them. This implies that the three types of knowledge complement each other well within the MAPS framework. The ablation study presented in Table 3 further demonstrates the effectiveness of each type of knowledge. Replacing any knowledge-guided translation with random sampling leads to performance degradation.

Method	COMET	BLEURT
Rerank	87.0	71.8
MAPS	87.6	72.6
- w/o Keyword	$87.1_{\downarrow 0.5}$	$72.1_{\downarrow 0.5}$
- w/o Topic	$87.2_{\downarrow 0.4}$	$72.4_{\downarrow 0.2}$
- w/o Demo	$86.9_{\downarrow 0.7}$	$72.0_{\downarrow 0.6}$

Table 3: Ablation study. We replace the knowledge-guided translation with random sampling translation in MAPS and report average values of four experiments. "$\downarrow$": statistically significant difference with $p < 0.05$.

En-Zh			Zh-En		
P_{src}	P_{tgt}	R	P_{src}	P_{tgt}	R
98.8	55.8	97.1	99.2	41.8	89.5

Table 4: Quality of LLM-generated keyword pairs.

Method	COMET	BLEURT	Accuracy
Rerank	81.5	70.2	61.5
MAPS	**82.2**	**70.6**	**65.5**

Table 5: Results on lexical ambiguity test set.

In Figure 5, we also note that the three types of knowledge cause different degrees of performance degradation when applied to the Base subset. We conjecture that the knowledge elicited from the LLM is not always helpful and may even be noisy. This finding motivates the knowledge selection step and is discussed in detail in §4.3.

4.3 Noise in Elicited Knowledge

The quality of extracted knowledge is essential for guiding translation. We use keywords as an example to evaluate the quality of LLM-generated knowledge. We design two metrics to characterize the quality of keyword pairs. We denote $D = \{(s_i, t_i, h_i, K_i)\}$ as the set of test samples, where s_i, t_i, h_i are source, target and hypothesis guided by the keyword pairs, respectively. $K_i = \{(\text{sw}_{ij}, \text{tw}_{ij})\}$ denotes the LLM-generated keyword pairs for the ith sample, where $(\text{sw}_{ij}, \text{tw}_{ij})$ is the jth keyword pair. The precisions of LLM-generated keyword pairs concerning the source or target are defined as:[2]

$$P_{\text{src}} = \frac{\sum_i \sum_j \mathbf{1}(\text{sw}_{ij} \subseteq s_i)}{\sum_i |K_i|}, \quad (1)$$

$$P_{\text{tgt}} = \frac{\sum_i \sum_j \mathbf{1}(\text{tw}_{ij} \subseteq t_i)}{\sum_i |K_i|}, \quad (2)$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. P_{src} and P_{tgt} reflect the proportion of LLM-generated keyword pairs that do exist in the source and target, respectively. Similarly, to evaluate how well the

model follows the given keyword pairs, we define the recall of keywords in the LLM hypothesis:

$$R = \frac{\sum_i \sum_j \mathbf{1}(\text{tw}_{ij} \subseteq h_i)}{\sum_i |K_i|}. \quad (3)$$

The statistical results in Table 4 show that: (1) although most LLM-generated keywords appear in the source sentences, only about half of them appear in the target sentences (55.8% for En-Zh; 41.8% for Zh-En); (2) the LLM strictly follows the given keyword pairs when performing translations (97.1% for En-Zh; 89.5% for Zh-En).

Combining the above two observations, we can conclude that the LLM-generated knowledge contains a certain degree of noise (at least content that is not consistent with the reference), which can easily mislead the translation process. This explains why incorporating that knowledge in the ''Base'' part of Figure 5 brings negative effects. Hence, knowledge selection is a crucial step in the MAPS framework to reduce the impact of noise.

4.4 MAPS Helps Ambiguity Resolution

Ambiguity resolution has long been one of the most challenging problems in machine translation. To evaluate the ambiguity resolution capability of machine translator, He et al. (2020) provide a lexical ambiguity test set for Chinese→English. The hard part of this test set involves Chinese sentences which are difficult to translate correctly unless the translator resolves their ambiguities. Our test results in Table 5 show the superiority of MAPS in ambiguity resolution, where the ''accuracy'' indicates the percentage of successfully disambiguated sentences (evaluated by human).

---

[2]For simplicity's sake, we use subset notation to represent substring relationship.

Method	Δ% hallucinations
Baseline	–
Rerank COMET-QE	−3%
MAPS COMET-QE	−8%
⇑̄ **Rerank** COMET	−6%
⇑̄ **MAPS** COMET	−12%

Table 6: Δ% of token-level hallucinations. ⇑̄: indicates the upper bound of selection, using COMET, a reference-based metric, as the selection method.

4.5 MAPS Reduces LLMs' Hallucinations

Hallucination issue in natural language generation (NLG) refers to the phenomenon *where the content generated by the model is nonsensical or unfaithful to the provided source content* (Ji et al., 2023; Filippova, 2020; Maynez et al., 2020; Parikh et al., 2020; Zhou et al., 2021; He et al., 2022). This has been one of the key challenges in LLMs (Zhang et al., 2023c). In this section, we analyze the phenomenon of hallucination through automatic and human evaluation.

In automatic evaluation, we use the hallucination detector provided by Zhou et al. (2021) to identify token-level hallucination in Alpaca's translation on Chinese→English test set. The detector assigns a binary label to each generated token. In Table 6, MAPS outperforms Rerank and demonstrates a higher upper bound.

In human evaluation, we employed professional human translators to label the hallucination errors in both MAPS and Rerank. We sampled 500 sentences from each of the English⇔Chinese test sets and evaluated text-davinci-003, Alpaca, and Vicuna. The human annotators were required to decide whether the translation belongs to the hallucination error following the definition from Guerreiro et al. (2023b). The results from Figure 6 show that MAPS outperforms Rerank by a notable margin in resolving hallucination.

We conjecture that one of the key differences between MAPS and Rerank is that MAPS is enabled to correct the probability distribution of the next token prediction, while Rerank is not. If a hallucinatory token occupies a high probability mass (Wang et al., 2022a), it is difficult for Rerank to avoid selecting this token by diverse sampling. In contrast, MAPS, providing additional
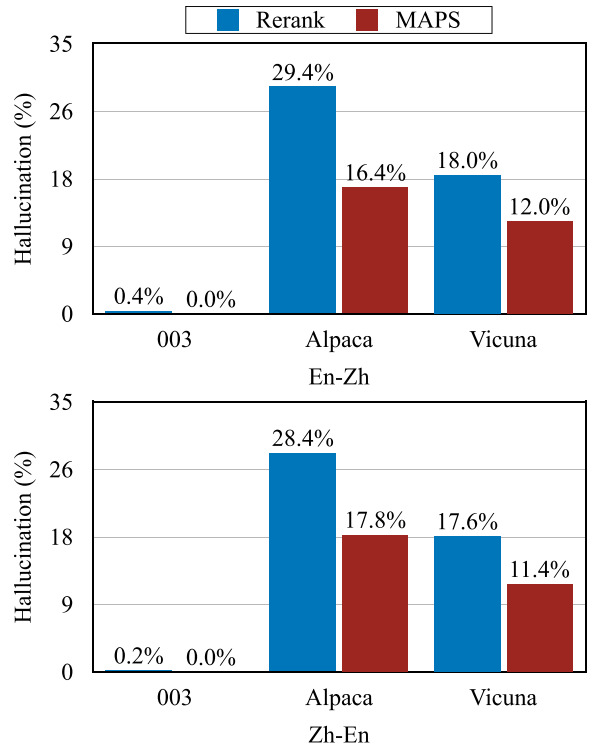

Figure 6: Ratio of hallucinations. Human annotators were tasked with labeling whether the generated translation is a hallucination error. "003" denotes text-davinci-003 and it has almost no hallucination errors.

translation-related knowledge in the prompt, enables the model to redistribute the probability of the next token, thus offering more possibilities to avoid choosing the hallucinatory token.

4.6 Three-in-One Prompting

So far, we have discussed the case where the LLM uses the three types of knowledge separately. An immediate question is how the LLM would perform if the three types of knowledge were integrated into one prompt. We call this method three-in-one prompting and present results in Table 7.

Within single-candidate methods (Baseline vs. Three-in-One), three-in-one prompting brings positive results overall, which means that the LLM can use three types of knowledge simultaneously. However, the degree of improvement varies significantly under different language pairs, with notable absence of effect in De-En translation. Regarding multi-candidate methods (MAPS COMET-QE vs. MAPS⁺ COMET-QE), incorporating three-in-one prompting into MAPS yields only marginal improvements (≤0.2). Considering that the candidate

Method	En-Zh	Zh-En	En-De	De-En
text-davinci-003 | COMET				
Baseline	86.2	81.6	85.8	85.2
Three-in-One	86.7 $_{\uparrow 0.5}$	81.7 $_{\uparrow 0.1}$	86.1 $_{\uparrow 0.3}$	85.1 $_{\downarrow 0.1}$
MAPS $_{\text{COMET-QE}}$	87.6	82.6	87.2	85.7
MAPS$^{+}_{\text{COMET-QE}}$	87.6 $_{-0.0}$	82.6 $_{-0.0}$	87.3 $_{\uparrow 0.1}$	85.7 $_{-0.0}$
text-davinci-003 | BLEURT				
Baseline	71.1	69.6	75.6	74.0
Three-in-One	71.8 $_{\uparrow 0.7}$	70.1 $_{\uparrow 0.5}$	76.1 $_{\uparrow 0.5}$	74.0 $_{-0.0}$
MAPS $_{\text{COMET-QE}}$	72.6	70.8	77.1	74.6
MAPS$^{+}_{\text{COMET-QE}}$	72.6 $_{-0.0}$	70.9 $_{\uparrow 0.1}$	77.2 $_{\uparrow 0.1}$	74.6 $_{-0.0}$

Method	En-Ja	Ja-En	De-Fr	Fr-De
text-davinci-003 | COMET				
Baseline	87.9	81.8	82.8	86.3
Three-in-One	88.4 $_{\uparrow 0.5}$	81.9 $_{\uparrow 0.1}$	83.1 $_{\uparrow 0.3}$	86.5 $_{\uparrow 0.2}$
MAPS $_{\text{COMET-QE}}$	89.5	82.9	84.1	87.5
MAPS$^{+}_{\text{COMET-QE}}$	89.6 $_{\uparrow 0.1}$	83.0 $_{\uparrow 0.1}$	84.2 $_{\uparrow 0.1}$	87.6 $_{\uparrow 0.1}$
text-davinci-003 | BLEURT				
Baseline	66.3	67.8	70.4	77.6
Three-in-One	67.3 $_{\uparrow 1.0}$	68.3 $_{\uparrow 0.5}$	70.6 $_{\uparrow 0.2}$	78.0 $_{\uparrow 0.4}$
MAPS $_{\text{COMET-QE}}$	68.3	69.1	71.9	78.9
MAPS$^{+}_{\text{COMET-QE}}$	68.5 $_{\uparrow 0.2}$	69.3 $_{\uparrow 0.2}$	71.9 $_{-0.0}$	79.0 $_{\uparrow 0.1}$

Table 7: Three-in-one prompting. Three-in-One: three types of knowledge are integrated into one prompt. MAPS$^{+}_{\text{COMET-QE}}$: adding candidate produced by Three-in-One into MAPS $_{\text{COMET-QE}}$. The subscripts indicate relative improvements from three-in-one prompting.

set generated by the three-in-one prompting overlaps significantly with the candidate sets generated individually by the three types of knowledge, this result is as expected.

5 Related Work

5.1 LLMs for Translation

Research evaluating the translation capabilities of LLMs falls into two main lines. The first line involves issues specific to LLMs, including the impact of demonstration selection in ICL (Vilar et al., 2022; Zhang et al., 2023a; Garcia et al., 2023) and prompt templates (Zhang et al., 2023a; Jiao et al., 2023b) on translation performance. The second line focuses on comprehensive evaluations of LLMs under various translation scenarios, covering multilingual (Jiao et al., 2023b; Zhu et al., 2023b; Hendy et al., 2023), document-level (Hendy et al., 2023; Wang et al., 2023b; Karpinska and Iyyer, 2023), low-resource translation (Jiao et al., 2023b; Garcia et al., 2023; Zhu et al., 2023b;

Bawden and Yvon, 2023), robustness (Jiao et al., 2023b), hallucination (Guerreiro et al., 2023a), and domain adaptation (Hendy et al., 2023; Wang et al., 2023a). Our work evaluates the translation capabilities of LLMs across eleven translation directions, varying from same-family (En⇔De), distant (En⇔Ja, En⇔Zh), and non-English-centric (De⇔Fr) and low-resource (Cs⇔Uk, En⇒Hr) language pairs. Zhu et al. (2023b) emphasizes the risk of data leakage. Therefore, we adopt the latest WMT22 test sets. Our work also quantitatively evaluates ambiguity resolution and token-/sentence-hallucination in LLM-based translation.

Jiao et al. (2023a) incorporate human evaluation into instruction data for training, resulting in translations that are preferred by humans during interactive chat sessions. In contrast, our work takes a different approach by mimicking the human translation process and achieves higher-quality translations without training.

Agrawal et al. (2023) propose an algorithm based on *n*-gram recall for demonstration selection. Given the ground-truth context, Pilault et al. (2023) introduce an interactive-chain prompting method for ambiguity resolution. Moslem et al. (2023) suggest prompting the LLMs with terminology hints extracted from the selected demonstrations or a compiled glossary for domain-specific translation such as COVID-19. Concurrently, Ghazvininejad et al. (2023) and Lu et al. (2023) use external dictionaries to augment prompts for low-resource and domain-specific translation. While our work can be viewed as a form of ''prompting strategy'', it differs from this line of research in that it does not rely on any external ''datastore'', such as sample pools, dictionaries, or ground-truth context, which should be curated carefully for specified language pairs or domains. In contrast, we consider the LLM itself as a ''datastore'' containing broad knowledge that can assist its translation process.

5.2 Chain-of-Thought Prompting

Wei et al. (2022) explore how chain-of-thought (CoT) prompting improves the ability of LLMs to perform complex reasoning such as arithmetic reasoning, commonsense reasoning, and symbolic reasoning. By guiding LLMs through generating intermediate reasoning chains prior to reaching a final solution, CoT prompting has propelled the multi-step reasoning abilities of LLMs to an

extraordinary level, as substantiated by previous research (Wei et al., 2022; Wang et al., 2023c). CoT prompting manifests through two distinct paradigms, namely, zero-shot CoT (Kojima et al., 2023; Yang et al., 2023) and few-shot CoT (Wei et al., 2023; Zhang et al., 2023b). Zero-shot CoT simply appends a trigger prompt such as *Let's think step by step* after the test question, with the motivation to harness the step-by-step reasoning capacities of LLMs in a zero-shot manner. Few-shot CoT operates by utilizing a few input-output demonstrations, each of which comprises a question, a reasoning chain, and the corresponding answer. These demonstrations are seamlessly integrated before the test question, resulting in a prompted input that is subsequently processed by an LLM to deduce the answer.

So far, most CoT prompting studies focus on complex reasoning problems. Although there are a few preliminary attempts to extend CoT prompting techniques to machine translation tasks, Peng et al. (2023) find that straightforwardly applying CoT to translation tasks resulted in word-by-word translations, which is less than satisfactory. Following this line, our work can also be viewed as a form of CoT prompting for translation as it dissects the translation process into distinct steps, which is the first successful attempt of CoT in translation tasks to the best of our knowledge. Notably, our work has successfully achieved improved translation performance by inducing three aspects of translation-related knowledge including keywords, topics, and relevant demonstrations to guide the final translation process.

5.3 Self-Prompting

Self-prompting is a line of research that utilizes the LLMs to prompt themselves and extract relevant knowledge to aid downstream tasks (Li et al., 2022; Wang et al., 2023d). Diverging from CoT prompting, which focuses on providing intermediate reasoning steps on the output side, self-prompting techniques dissect the input problem into specific sub-problems on the input side and extract the salient knowledge for the sub-problems one by one. This extracted knowledge is then utilized to deduce the ultimate solution.

Several studies exemplify the diversity of self-prompting applications. Specifically, Kim et al. (2022) and Li et al. (2022) use the LLMs to generate in-context exemplars for text classification

and open-domain question answering, respectively. Yu et al. (2023) generate diverse documents from the LLMs to improve knowledge-intensive tasks. Wang et al. (2023d) compel LLMs to first extract the core elements for news texts, such as entity, date, event, and result. Then, the extracted elements are used to generate summaries. Further innovations emerge in multimedia contexts. Zhu et al. (2023a) and Chen et al. (2023) empower LLMs to pose inquiries regarding provided images and videos to enrich the caption. Remarkably, MAPS extends the domain of self-prompting into machine translation for the first time.

6   Conclusion

This work introduces MAPS, a method that enables LLMs to mimic human translation strategy for achieving high-quality translation. MAPS allows LLMs to take preparatory steps before translation. Specifically, LLMs analyze the given source text and generate three aspects of translation-related knowledge: keywords, topics, and relevant demonstrations. Using a filtering mechanism based on quality estimation, the selected knowledge guides the LLMs' translation process. In experiments with text-davinci-003, Alpaca, and Vicuna, MAPS yields significant and consistent improvements across eleven translation directions from WMT22 and exhibits a higher upper bound of candidate selection. Human evaluations show that MAPS provides more favorable translations by reducing mistranslation, awkward style, untranslated text, and omission errors. Further analyses show that MAPS effectively resolves ambiguities and hallucinations in translation. Future work includes designing more aspects of translation-related knowledge and better filtering mechanisms to improve the translation capabilities of LLMs further. Another interesting direction is to explore the human-like translation strategy in training LLMs (e.g., instruction tuning).

7   Discussion

7.1   Inference Time

Since MAPS consists of three sequential stages, the main limitation of MAPS lies in inference time. As shown in Figure 7, when processing serially, the inference times of Three-in-One, Rerank,
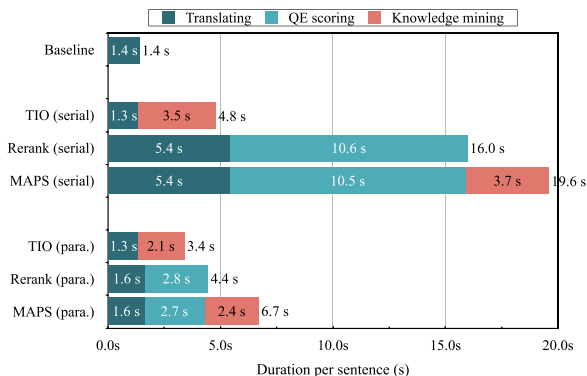
Figure 7: Durations of different processing stages for Baseline, Three-in-One prompting (TIO), Rerank, and MAPS. The results represent the average of five independent trials. Experiments were conducted using a T4-8C GPU, with each trial consisting of 50 sentences. ''Serial'' and ''Para.'' denote serial and parallel processing of multiple types of knowledge and candidates, respectively.

and MAPS are 3×, 11×, and 14× the Baseline, respectively.

Given that all three methods involve processing multiple types of knowledge or candidates without any dependencies between them, a practical approach for acceleration is to parallel processing, which drastically reduces the running times (↓29% for Three-in-One; ↓73% for Rerank; ↓66% for MAPS) to an acceptable level.

The additional overhead from MAPS is mainly in the knowledge mining phase, where the LLM generates three types of knowledge separately. One possible acceleration is to have the LLM generate three types of knowledge in a single call. By controlling the format of the output, e.g., JSON, we can extract each type of knowledge. However, the LLM is not guaranteed to output valid JSON content, which may lead to degradation of the final translation performance (see Table 8).

In addition, the running time of the QE scoring can be reduced by techniques such as model quantization or compression.

7.2 Is MAPS Overfitting Evaluation Metrics?

In this work, we rely on COMET and BLEURT for automatic evaluation for their strong alignment with human evaluation, as highlighted by Freitag et al. (2022). We also use COMET-QE as one of the knowledge selection methods, whose training data has overlap with evaluation metrics. This

Method	En-Zh			Zh-En		
	CT	BT	JSON E.	CT	BT	JSON E.
MAPS	87.6	72.6	—	82.6	70.8	—
MAPS^{JSON}	87.7	72.6	0.1%	82.1↓	70.3↓	2.0%

Table 8: MAPS^{JSON}: generating three types of knowledge in one JSON object. CT: COMET; BT: BLEURT. JSON E.: percentage of model output that is not in valid JSON format. Times taken for knowledge mining using MAPS and MAPS^{JSON} are 3.7 and 2.3 seconds per sentence, respectively.

Error type	Translation	COMET-QE (ref-free)	COMET (ref-based)	BLEURT (ref-based)
None (the reference)	The rule of drinking Red Label Whisky:	2.8	**96.2**	**101.8**
Hallucination (irrelevant content)	The rule of drinking Red Label Whisky: 1. Always drink responsibly.2. Never drink alone.3. Avoid drinking on an empty stomach.4. Set limits and stick to them.5. Drink in moderation.	**5.6**	64.1	70.2
Off-target (wrong target language)	So trinkt man Red-La-bel-Whisky:	**4.6**	72.6	53.2

Table 9: Case study of Chinese-to-English translation for ''红牌威士忌喝法:''. COMET-QE assigns higher scores to translations with hallucination and off-target errors compared to an error-free translation. Conversely, COMET and BLEURT award lower scores to these erroneous translations. Error spans are highlighted.

leads to a pertinent question: Is MAPS merely overfitting to COMET and BLEURT?

To ensure reliable evaluations, we integrated human assessments into all our experiments, including: MQM evaluation (§4.1), human preference studies (§4.1), ambiguity resolution (§4.4), and analysis of hallucination (§4.5). These evaluations substantiate MAPS's effectiveness from the viewpoint of human translators.

Furthermore, we demonstrate that MAPS remains effective even in the absence of COMET-QE. As shown in Table 1, by formulating single-choice questions, the LLM itself can select the best translation candidates (Rerank $_{LLM-SCQ}$ and MAPS $_{LLM-SCQ}$).

From a data perspective, all three models above were trained using datasets from WMT. However, they use the data in different ways. COMET-QE is reference-free and does not utilize reference data during training or inference. On the contrary, COMET and BLEURT are reference-based, with both training and inference

processes relying on reference data. This difference allows COMET and BLEURT to penalize translation errors against a reference, a function that COMET-QE lacks due to its reference-free design (see Table 9).

Overall, MAPS is widely effective by employing human strategy for translation.

Acknowledgments

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.564

Mona Baker. 2018. *In Other Words: A Coursebook on Translation*. Routledge. https://doi.org/10.4324/9781315619187

Yehoshua Bar-Hillel. 1960. A demonstration of the nonfeasibility of fully automatic high quality translation. *Advances in Computers*, 1:158–163.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: The case of bloom. *ArXiv preprint*, abs/2303.01911. https://doi.org/10.48550/arXiv.2303.01911

Lynne Bowker. 2002. *Computer-aided Translation Technology: A Practical Introduction*.

University of Ottawa Press. https://doi.org/10.1353/book6554

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv preprint*, abs/2303.12712. https://doi.org/10.48550/arXiv.2303.12712

Aljoscha Burchardt. 2013. Multidimensional quality metrics: A flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. 2023. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *ArXiv preprint*, abs/2304.04227. https://doi.org/10.48550/arXiv.2304.04227

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, Andre F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. https://doi.org/10.18653/v1/2023.wmt-1.100

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.100

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from

noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.76

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474. https://doi.org/10.1162/tacl_a_00437

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

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. *ArXiv preprint*, abs/2302.01398. https://doi.org/10.48550/arXiv.2302.01398

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *ArXiv preprint*, abs/2302.07856. https://doi.org/10.48550/arXiv.2302.07856

Daniel Gile. 2009. *Basic Concepts and Models for Interpreter and Translator Training*. Benjamins Translation Library. John Benjamins, Amsterdam. https://doi.org/10.1075/btl.8

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023a. Hallucinations in large multilingual translation models. *ArXiv preprint*, abs/2303.16104. https://doi.org/10.1162/tacl_a_00615

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023b. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.75

Basil Hatim and Jeremy Munday. 2004. *Translation: An Advanced Resource Book*. Psychology Press. https://doi.org/10.4324/9780203501887

Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.327

Zhiwei He, Xing Wang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2022. Bridging the data gap between training and inference for unsupervised neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6611–6623, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.1145/3571730

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? A comprehensive evaluation. *ArXiv preprint*, abs/2302.09210. https://doi.org/10.48550/arXiv.2302.09210

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12). https://doi.org/10.1145/3571730

Wenxiang Jiao, Jen tse Huang, Wenxuan Wang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. Parrot: Translating during chat using

large language models. In *ArXiv*, arXiv:2304.02426.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023b. Is chatgpt a good translator? A preliminary study. In *ArXiv preprint*, abs/2301.08745.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *ArXiv preprint*, abs/2304.03245. `https://doi.org/10.18653/v1/2023.wmt-1.41`

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2022. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. *ArXiv preprint*, abs/2206.08082.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Philipp Koehn. 2009. A process study of computer-aided translation. *Machine Translation*, 23(4):241–263. `https://doi.org/10.1007/s10590-010-9076-3`

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. In *Advances in Neural Infor-*

mation Processing Systems, volume 35, pages 22199–22213.

Junlong Li, Zhuosheng Zhang, and Hai Zhao. 2022. Self-prompting large language models for open-domain qa. *ArXiv preprint*, abs/2212.08635. `https://doi.org/10.48550/arXiv.2212.08635`

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. *ArXiv preprint*, abs/2211.09110. `https://doi.org/10.48550/arXiv.2211.09110`

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *ArXiv preprint*, abs/2305.06575. `https://doi.org/10.48550/arXiv.2305.06575`

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt. *ArXiv preprint*, abs/2305.01181. `https://doi.org/10.18653/v1/2023.emnlp-main.1036`

Elliott Macklovitch. 1995. The future of MT is now and Bar-Hillel was (almost entirely) right. In *Proceedings of the Fourth Bar-Ilan Symposium on the Foundations of Artificial Intelligence. url: http://rali. iro. umontreal. ca/Publications/urls/bisfai95.ps*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization.

In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.173

Shima Rahimi Moghaddam and Christopher J. Honey. 2023. Boosting theory-of-mind performance in large language models via prompting. *ArXiv preprint*, abs/2304.11490. https://doi.org/10.48550/arXiv.2304.11490

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *ArXiv preprint*, abs/2301.13294. https://doi.org/10.48550/arXiv.2301.13294

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.89

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *ArXiv preprint*, abs/2304.03442. https://doi.org/10.48550/arXiv.2304.03442

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *ArXiv preprint*, abs/2303.13780. https://doi.org/10.18653/v1/2023.findings-emnlp.373

Jonathan Pilault, Xavier Garcia, Arthur Bražinskas, and Orhan Firat. 2023. Interactive-chain-prompting: Ambiguity resolution for cross-lingual conditional generation with interaction. *ArXiv preprint*, abs/2301.10309. https://doi.org/10.48550/arXiv.2301.10309

Anthony. Pym. 2014. *Exploring Translation Theories*, 2nd edition. Routledge. https://doi.org/10.4324/9781315857633

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.52 and https://github.com/Unbabel/COMET

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. 2022b. 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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.704

ShareGPT. 2023. Sharegpt: Share your wildest chatgpt conversations with one click. Available at: https://sharegpt.com/

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. *ArXiv preprint*,

abs/2302.00093. https://doi.org/10.48550/arXiv.2302.00093

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971. https://doi.org/10.48550/arXiv.2302.13971

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288. https://doi.org/10.48550/arXiv.2307.09288

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *ArXiv*

preprint, abs/2211.09102. https://doi.org/10.18653/v1/2023.acl-long.859

Longyue Wang, Zefeng Du, DongHuai Liu, Deng Cai, Dian Yu, Haiyun Jiang, Yan Wang, Shuming Shi, and Zhaopeng Tu. 2023a. Guofeng: A discourse-aware evaluation benchmark for language understanding, translation and generation.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023b. Document-level machine translation with large language models. *ArXiv preprint*, abs/2304.02210. https://doi.org/10.18653/v1/2023.emnlp-main.1036

Wenxuan Wang, Wenxiang Jiao, Yongchang Hao, Xing Wang, Shuming Shi, Zhaopeng Tu, and Michael Lyu. 2022a. Understanding and improving sequence-to-sequence pretraining for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2591–2600, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.185

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023d. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.482

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022b. Self-instruct: Aligning language model with self generated instructions. https://doi.org/10.18653/v1/2023.acl-long.754

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani

Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, et al. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Hao Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael R. Lyu. 2023. Chatgpt or grammarly? Evaluating chatgpt on grammatical error correction benchmark. *ArXiv preprint*, abs/2303.13648. `https://doi.org/10.48550/arXiv.2303.13648`

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *ArXiv preprint*, abs/2309.03409. `https://doi.org/10.48550/arXiv.2309.03409`

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *International Conference for Learning Representation (ICLR)*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. *ArXiv preprint*, abs/2301.07069. `https://doi.org/10.48550/arXiv.2301.07069`

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023c. Multimodal chain-of-thought reasoning in lan-guage models. *ArXiv preprint*, abs/2302.00923. `https://doi.org/10.48550/arXiv.2302.00923`

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics. `https://github.com/violet-zct/fairseq-detect-hallucination`, `https://doi.org/10.18653/v1/2021.findings-acl.120`

Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. 2023a. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *ArXiv preprint*, abs/2303.06594. `https://doi.org/10.48550/arXiv.2303.06594`

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023b. Multilingual machine translation with large language models: Empirical results and analysis. *ArXiv preprint*, abs/2304.04675. `https://doi.org/10.48550/arXiv.2304.04675`

A Knowledge-specific Prompting for Rerank

Method	En-Zh		Zh-En	
	COMET	BLEURT	COMET	BLEURT
text-davinci-003				
Baseline	86.2	71.1	81.6	69.6
Rerank	86.9	71.7	82.1	70.1
Rerank $_{KEYWORD}$	86.7	71.9	81.8	70.2
Rerank $_{TOPIC}$	87.1	71.8	**82.2**	**70.5**
Rerank $_{DEMO}$	**87.4**	**72.7**	82.1	70.4

Table 10: Knowledge-specific prompting for Rerank. Four hypotheses are sampled from each knowledge-specific prompt and reranked. The subscript indicates the type of knowledge.