

# The Double-Edged Sword of Reasoning LLMs in Translation: Disambiguation, Hallucination, and Efficiency

Anonymous ACL submission

## Abstract

The Reasoning LLMs are transforming AI by simulating human cognitive processes, but their performance in multilingual machine translation (MMT) remains underexplored. This study examines: (1) how Reasoning LLMs perform in MMT tasks and (2) what factors influence their translation quality. We evaluate multiple Reasoning LLMs and compare them with traditional LLMs like ChatGPT and GPT-4o. Results show that Reasoning LLMs establish new multilingual translation benchmarks. They demonstrate strengths in historical and cultural translation but exhibit a tendency for **rambling issues** in more challenging scenarios. Further analysis reveals three key insights: (1) High inference costs and slower processing speeds make complex translation tasks more resource-intensive. (2) Translation quality improves with model size, enhancing commonsense reasoning and cultural translation. (3) The temperature parameter significantly impacts output quality—lower temperatures yield more stable and accurate translations, while higher temperatures reduce coherence and precision.<sup>1</sup>

## 1 Introduction

Large Language Models (LLMs) have demonstrated significant potential in the field of Machine Translation (MT) (Zhu et al., 2024; Peng et al., 2023; Jiao et al., 2023a; Hendy et al., 2023a; Tyen et al., 2023; Liang et al., 2023; Guerreiro et al., 2023; Ranaldi et al., 2023; Chen et al., 2024c). Currently, Reasoning LLMs—such as OpenAI’s o1 model (OpenAI, 2024) and its replicated versions like QwQ (Qwen, 2024), Marco-o1 (Zhao et al., 2024), and DeepSeek-R1 (Guo et al., 2025) are profoundly transforming the AI landscape, enabling models to handle increasingly complex tasks. These models are renowned for their

strong reasoning capabilities, exhibiting human-like deep thinking in extended tests, exploring diverse reasoning strategies, and refining answers through decision reflection and iterative refinement. This allows them to simulate human cognitive processes in problem-solving.

Consequently, how do these models perform in Multilingual Machine Translation (MMT) tasks? Currently, the real capabilities of reasoning LLMs in MMT still require systematic investigation. MMT is a highly challenging task (Wang et al., 2024b; Huang et al., 2024a; Gao et al., 2024; Wu et al., 2024; Chen et al., 2024a), requiring models not only to achieve semantic alignment across languages but also to ensure translation accuracy in aspects such as commonsense reasoning, historical and cultural context, and terminology.

This study focuses on the performance of Reasoning LLMs in Multilingual Machine Translation (MMT) tasks, addressing two key research questions: 1) Whether Reasoning LLMs have **positive** potential to become stronger translation models. 2) In which translation tasks these models show **neutral** performance or **negative** limitations? 3) Which model characteristics influence their translation quality?

To answer the above questions, we evaluate multiple mainstream Reasoning LLMs. To provide a more comprehensive comparison between Reasoning LLMs and traditional LLMs in translation performance, we introduce ChatGPT (Ouyang et al., 2022), GPT-4o (Achiam et al., 2023), and DeepSeek-v3 (DeepSeek-AI et al., 2024) as baseline models. We conducted a systematic analysis across five types of translation tasks, including: Flores-200 (Goyal et al., 2022) for evaluating multilingual capability, Commonsense MT (He et al., 2020) for testing commonsense reasoning in translation, RTT (Zhang et al., 2023) for terminology constraint translation, and CultureMT (Yao et al., 2023) for culturally

<sup>1</sup>Our code will be made available at <https://github.com/anonymous>.

contextualized translation.

For the first question, the results show that, compared with traditional LLMs, reasoning LLMs achieve clear gains in general multilingual translation and perform even better on reasoning-oriented Commonsense MT tasks, indicating their positive potential for translation. For the second question, our analysis shows several limitations. Reasoning LLMs do not have inherent advantages in historical-cultural or terminology-constrained translation. They also have higher reasoning costs and slower speed, requiring more time and computation for complex tasks. Large-scale testing further reveals typical failure modes, such as *Instruction Following Issue* and *Rambling Issue*. We infer that the rambling issue arises from a misactivation of the reasoning paradigm when translation tasks are mistakenly treated as reasoning problems.

For the third question, our results show that translation quality generally improves as model size increases. We also find that the temperature parameter has a substantial impact on translation performance: in complex linguistic settings, a lower temperature typically produces more stable and accurate outputs. These findings reveal key characteristics of reasoning LLMs in multilingual translation and offer actionable directions for further optimizing their translation performance. The main contributions of this study are as follows:

- We evaluate multiple mainstream reasoning LLMs on four benchmark tests that cover different dimensions of multilingual machine translation, offering a comprehensive analysis of their translation abilities.
- We systematically compare reasoning LLMs with traditional LLM translation systems, revealing their positive behaviors, neutral behaviors, and negative behaviors, and analyzing the root causes of these negative behaviors.
- We identify and summarize new performance patterns of reasoning LLMs in MMT tasks, and empirically examine how their reasoning features correlate with translation quality, providing directions for future improvement.

## 2 Related Work

**Reasoning LLMs** Recently, Reasoning LLMs have shown exceptional performance in reasoning tasks, especially in mathematics and coding.

Following OpenAI’s O1 model (OpenAI, 2024), significant efforts have been made to replicate its success. Qin et al., 2024 introduced journey learning, a training paradigm that enhances long-term reasoning with only 327 training samples. Huang et al., 2024b demonstrated the effectiveness of data distillation from existing reasoning LLMs. Zhao et al., 2024 proposed the Marco-o1 model, combining Chain-of-Thought (CoT) fine-tuning, Monte Carlo Tree Search (MCTS), and reflection mechanisms to tackle open-ended problems. Guo et al., 2025 introduced the DeepSeek-R1 model, enhancing reasoning through multi-stage training and reinforcement learning (RL). QwQ (Qwen, 2024) model based on the Qwen architecture excel in mathematics and coding tasks but facing challenges like language mixing and circular reasoning. Wang et al., 2024a proposed the DRT-o1 model, applying long CoT to MT, showing superior translation capabilities, especially with literature texts involving metaphors and similes.

**Machine Translation with Large Language Models (LLM-MT)**. Large language models, such as ChatGPT (Ouyang et al., 2022), have shown significant effectiveness in machine translation across various language pairs (Chen et al., 2024d; Hendy et al., 2023b; Jiao et al., 2023b; Le Scao et al., 2023; Iyer et al., 2023; Zeng et al., 2023; Karpinska and Iyyer, 2023; Moslem et al., 2023; Wang et al., 2023; Iyer et al., 2023; Farinhas et al., 2023; Chen et al., 2024b). Recent research has explored the performance of LLMs in machine translation, including control over formality in translation outputs (Garcia and Firat, 2022), in-context translation abilities during pre-training (Shin et al., 2022), and the impact of LLM-based machine translation on culturally sensitive texts (Yao et al., 2023). Additionally, studies have examined the bilingual capabilities of LLMs to enhance translation performance (Huang et al., 2024a). For translation tasks requiring reasoning, multi-agent debates can effectively enhance the reasoning abilities of LLM-MT (Liang et al., 2023). These investigations further validate the research value of LLM-MT, offering diverse research directions for scholars.

## 3 Experimental Setting

### 3.1 Evaluation Setting

**Dataset:** We conduct experiments on four MT benchmarks: Flores-200, RTT test data,

Commonsense MT, and Culture MT. The Flores-200 dataset (Goyal et al., 2022)<sup>2</sup> is a multilingual translation benchmark that covers 200 languages. It provides a comprehensive evaluation of translation quality across a wide range of languages. The dataset includes parallel sentences for each language pair, allowing for a thorough assessment of translation performance. Commonsense MT (He et al., 2020)<sup>3</sup> is a commonsense reasoning MT test set that evaluates the ability of translation systems to handle sentences requiring commonsense knowledge. It includes sentences with ambiguous or context-dependent meanings, making it a rigorous benchmark for assessing the ability of translation systems to generate accurate and contextually appropriate translations. The cultural MT dataset (Yao et al., 2023)<sup>4</sup> provides a culturally relevant parallel corpus with annotations of cultural-specific items. It covers 6 language pairs: En↔Es/Fr/Hi/Ta/Te/Zh, including over 7,000 cultural-specific items from 18 concept categories across more than 140 countries and regions. RTT test data (Zhang et al., 2023) is a challenging test set specifically designed for terminology MT. It increases the average constraint count per sentence from 1.1~1.7 to 6.1 and the length per target constraint from 1.1~1.2 words to 3.4 words, making it a rigorous benchmark for evaluating the handling of specialized terms in translation.

**Evaluation Metrics:** In evaluating our translation methodology, we initially employ COMET<sup>5</sup> (Rei et al., 2022) and BLEURT (Sellam et al., 2020)<sup>6</sup> as automatic metrics, aligning with the established standards in LLM-based translation (He et al., 2023; Huang et al., 2024a). For traditional translation evaluation, we use BLEU (Post, 2018)<sup>7</sup>.

### 3.2 Models Selection

We evaluate translation performance on 6 reasoning LLMs. OpenAI-o1 (OpenAI, 2024) is a close-source LLM that improves the reasoning quality by increasing the length of CoT for the first time. OpenAI-o3-mini is a small reasoning model that excels at programming, math, and science problems. DeepSeek-R1 (Guo et al.,

<sup>2</sup><https://github.com/facebookresearch/flores>

<sup>3</sup><https://github.com/tjunlp-lab/commonmt>

<sup>4</sup><https://github.com/BigBinnie/Benchmarking-LLM-based-Machine-Translation-on-Cultural-Awareness>

<sup>5</sup><https://huggingface.co/Unbabel/wmt22-comet-da>

<sup>6</sup><https://github.com/lucadiliello/bleurt-pytorch>

<sup>7</sup><https://github.com/mjpost/sacrebleu>

2025) is a 671B model which is first fine-tuned using a large number of CoT examples, after which it was trained with Reinforcement Learning. Marco-o1-7B (Zhao et al., 2024) is trained based on Qwen2-7B-Instruct with multiple CoT and reasoning datasets. DRT-o1-14B (Wang et al., 2024a) is a reasoning translation model, focusing on English-to-Chinese translation task. QwQ-32B (Qwen, 2024) performs well on math and reasoning problems.

For comparison, we selected three models. ChatGPT (Ouyang et al., 2022) is a non-Reasoning model which can interact with users in a conversational manner. GPT-4o (Achiam et al., 2023) is a large multimodal model which demonstrates excellent ability on various QA benchmarks. DeepSeek-v3 (DeepSeek-AI et al., 2024) is a Mixture-of-Experts(MoE) model which performs well on multiple benchmarks.

### 3.3 Inference Settings

We utilized APIs to access the models from DeepSeek and OpenAI<sup>8</sup>. For the open-source models, we employed vLLM<sup>9</sup> for local deployment and conducted tests on 12 A100 GPUs, each with 80GB of memory. The prompt template utilized in our study is presented as follows:

```

{"role": "user", "content": "Source:  $s$ , translate the following sentences from  $L_s$  to  $L_t$ ."}

```

In the template,  $L_s$  and  $L_t$  denote the names of source language and the target language, respectively, while  $s$  is the source sentence which is to be translated.

## 4 Main Experiment

### 4.1 Positive: Evaluation of Multilingual Translation Performance

We evaluate general multilingual translation performance on the Flores-200 dataset. Closed-source Reasoning LLMs achieve the best overall results, with OpenAI-o1 reaching a maximum +34.5 BLEU and an average +11.14 BLEU gain over standard LLMs. Among open-source models, DeepSeek-R1 performs strongest, improving by +16.92 BLEU on average. Smaller models such as Marco-o1 and DRT-o1 also narrow

<sup>8</sup>The DeepSeek and OpenAI models used in this study are accessed through the DeepSeek-R1, gpt-3.5-turbo, gpt-4o-2024-11-20, o1-preview-2024-09-12, and o3-mini-2025-01-31 APIs, respectively.

<sup>9</sup><https://github.com/vllm-project/vllm>

Methods	En→De			En→Ro			Zh→Ro			En→Zh		
	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT
<b>Non-Reasoning LLMs</b>												
GPT-4o	21.74	68.35	54.53	24.91	74.11	68.08	20.77	81.40	71.41	40.02	81.24	64.50
ChatGPT	36.32	83.65	70.47	38.38	88.57	79.22	20.76	85.09	73.60	41.99	86.00	70.29
DeepSeek-v3	21.67	70.16	57.76	27.69	79.99	73.63	22.84	86.79	76.95	41.78	86.71	71.81
Qwen3	32.82	82.46	68.68	34.74	86.54	78.27	21.71	86.17	75.78	41.68	86.35	71.11
<b>Close-source Reasoning LLMs</b>												
OpenAI-o1	<b>43.90</b>	85.85	72.96	<b>41.92</b>	<b>90.31</b>	80.60	<b>25.40</b>	<b>87.95</b>	<b>78.26</b>	43.93	88.51	<b>74.07</b>
OpenAI-o3-mini	40.31	<b>86.11</b>	<b>73.35</b>	41.49	90.29	<b>81.41</b>	21.72	86.86	76.73	<b>44.76</b>	<b>89.16</b>	74.06
<b>Open-source Reasoning LLMs</b>												
Marco-o1	11.73	80.87	71.83	15.33	77.14	67.08	5.34	66.77	69.78	40.13	74.53	59.84
DRT-o1-14B	9.40	65.97	51.17	19.91	80.08	69.37	2.42	60.93	50.25	35.89	57.05	40.20
DeepSeek-R1	39.44	85.32	72.55	37.95	86.97	77.62	21.33	83.55	73.40	41.23	87.44	72.43
QwQ	9.98	80.02	66.23	12.40	81.50	70.30	11.73	78.60	66.56	39.63	80.37	64.63
<b>De→En</b>												
<b>Non-Reasoning LLMs</b>												
GPT-4o	<b>43.97</b>	87.20	78.00	41.60	87.45	77.07	26.78	64.27	48.17	20.77	81.40	71.41
ChatGPT	43.73	87.03	77.55	41.59	87.78	77.00	31.42	78.17	61.36	27.33	85.40	72.82
DeepSeek-v3	38.87	85.44	76.39	37.61	86.37	76.10	29.38	83.72	68.92	27.56	86.23	74.55
Qwen3	39.54	85.96	76.39	39.78	86.98	76.42	29.24	84.89	69.20	28.71	86.42	74.07
<b>Close-source Reasoning LLMs</b>												
OpenAI-o1	43.55	<b>87.54</b>	<b>78.70</b>	<b>42.82</b>	<b>88.79</b>	<b>78.64</b>	39.13	85.59	71.46	<b>28.96</b>	<b>86.57</b>	<b>75.25</b>
OpenAI-o3-mini	42.70	87.34	78.03	42.12	88.35	77.79	<b>39.42</b>	<b>86.76</b>	<b>71.50</b>	27.48	86.28	74.41
<b>Open-source Reasoning LLMs</b>												
Marco-o1	38.45	86.22	77.30	27.15	85.80	75.37	25.68	77.08	64.83	18.98	77.46	63.27
DRT-o1-14B	18.40	76.81	66.66	22.95	81.95	70.96	19.43	57.72	43.46	3.54	60.33	46.09
DeepSeek-R1	40.91	84.77	74.71	40.60	86.05	75.00	38.13	79.76	64.16	27.05	85.56	73.43
QwQ	13.09	83.07	72.65	16.55	84.06	72.81	19.94	75.75	64.75	15.91	82.32	69.33

Table 1: The main results from the Flores-200 dataset are presented. The bold indicates the highest values that are statistically significant, with p-values less than 0.01 in the paired t-test against all compared methods.

the gap to closed-source systems on COMET and BLEURT. Notably, reasoning LLMs show larger gains on COMET/BLEURT than on BLEU, reflecting their tendency to produce semantically faithful yet lexically varied outputs. Overall, the results indicate that Reasoning LLMs consistently outperform traditional LLMs in multilingual translation.

#### 4.2 Positive: Evaluation of Reasoning Tasks in Translation

To evaluate the performance of Reasoning LLMs on commonsense reasoning translation tasks, we conducted experiments using the Commonsense MT lexical subset. The results, shown in Table 2, indicate that Reasoning LLMs maintain clear advantages on tasks requiring explicit reasoning. In the Lexical setting, OpenAI-o1 achieved the strongest results, surpassing GPT-4o by 2.00 COMET and 3.89 BLEURT. More importantly, on the ACC metric, which directly measures reasoning consistency and correctness, Reasoning LLMs also achieved higher

Methods	BLEU	COMET	BLEURT	ACC
<b>Non-Reasoning LLMs</b>				
GPT-4o	<b>30.57</b>	80.72	68.14	70.4
ChatGPT	24.42	78.72	68.61	64.2
DeepSeek-v3	28.53	81.17	69.94	69.4
<b>Close-source Reasoning LLMs</b>				
OpenAI-o1	26.45	<b>82.72</b>	<b>72.03</b>	90.6
OpenAI-o3-mini	27.10	81.13	70.08	88.4
<b>Open-source Reasoning LLMs</b>				
Marco-o1	13.44	78.00	65.70	85.3
DRT-o1-14B	5.84	69.28	50.67	76.2
DeepSeek-R1	27.27	81.81	70.94	71.5
QwQ	1.89	73.50	61.41	62.8

Table 2: Lexical-only results from the Commonsense MT dataset.

scores than traditional LLMs. These results suggest that the structured reasoning process helps Reasoning LLMs better capture and resolve implicit commonsense relations in translation.

#### 4.3 Neutral: Analyzing Culture-Specific Translation

To evaluate models on culturally specific translation, we used the Culture MT dataset. As shown

Methods	En→Es			En→Fr			En→Hi			En→Ta			En→Te			En→Zh		
	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT
<b>Non-Reasoning LLMs</b>																		
GPT-4o	27.80	72.00	56.76	19.64	68.12	43.80	16.34	66.71	53.64	4.23	70.32	60.57	5.71	71.63	57.10	31.26	76.56	56.02
ChatGPT	36.39	79.05	64.78	24.60	73.26	49.66	13.64	68.47	56.93	2.98	64.16	53.28	3.54	65.51	49.64	28.30	78.07	59.01
DeepSeek-v3	23.45	68.70	54.99	16.23	64.32	39.46	9.68	58.28	46.44	2.83	63.07	58.20	2.20	58.05	49.69	28.60	74.24	56.81
<b>Close-source Reasoning LLMs</b>																		
OpenAI-o1	41.42	78.30	62.71	24.64	72.94	48.34	<b>20.72</b>	<b>73.67</b>	<b>62.03</b>	<b>8.48</b>	<b>79.79</b>	67.12	<b>8.06</b>	<b>79.82</b>	<b>61.12</b>	32.10	82.60	63.82
OpenAI-o3-mini	<b>41.87</b>	<b>81.15</b>	66.51	<b>29.15</b>	<b>76.18</b>	53.02	17.29	73.17	61.56	5.74	78.61	<b>67.14</b>	5.95	74.30	58.47	<b>33.22</b>	<b>83.41</b>	64.67
<b>Open-source Reasoning LLMs</b>																		
Marco-o1	30.97	80.09	<b>67.06</b>	24.57	74.89	<b>54.92</b>	0.20	33.86	22.06	0.51	38.76	33.31	0.25	36.32	29.15	19.03	72.98	56.08
DeepSeek-R1	38.09	79.42	64.63	26.67	73.77	50.17	17.57	69.07	57.28	5.96	75.29	60.28	5.06	65.94	48.97	30.18	82.57	<b>64.98</b>
QwQ	10.68	74.93	59.23	4.55	69.91	45.56	7.14	64.76	52.75	0.26	55.44	47.79	0.35	53.81	39.87	13.24	70.78	51.08
Methods	Es→En			Fr→En			Hi→En			Ta→En			Te→En			Zh→En		
	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT
<b>Non-Reasoning LLMs</b>																		
GPT-4o	41.52	78.69	63.38	29.83	74.04	55.55	<b>34.80</b>	<b>82.08</b>	<b>64.96</b>	23.48	75.73	56.53	<b>25.40</b>	<b>75.73</b>	<b>56.53</b>	<b>30.94</b>	80.47	65.12
ChatGPT	40.79	79.21	64.18	<b>31.88</b>	74.15	55.73	29.75	81.21	62.70	18.12	73.80	54.16	18.12	73.80	54.16	30.01	78.77	61.52
DeepSeek-v3	34.49	77.17	62.32	26.62	72.49	54.76	25.94	79.90	62.55	13.28	71.14	54.35	13.28	71.14	54.35	26.34	78.89	63.78
<b>Close-source Reasoning LLMs</b>																		
OpenAI-o1	<b>46.86</b>	79.64	63.81	29.77	73.93	57.14	30.41	81.85	64.03	21.45	<b>79.58</b>	<b>64.00</b>	19.72	73.45	53.85	23.05	<b>82.10</b>	<b>69.97</b>
OpenAI-o3-mini	43.99	<b>80.53</b>	65.71	31.03	<b>75.27</b>	57.21	30.60	81.12	63.08	<b>24.22</b>	75.51	56.24	24.22	75.51	56.24	28.52	80.39	64.29
<b>Open-source Reasoning LLMs</b>																		
Marco-o1	37.67	80.32	<b>67.51</b>	28.41	75.40	<b>60.14</b>	2.29	45.63	13.99	1.14	45.75	15.27	1.60	45.54	17.99	0.12	70.84	53.91
DeepSeek-R1	39.21	77.53	62.26	26.49	73.06	55.29	30.53	80.09	62.19	20.61	76.00	56.70	22.98	75.58	56.21	28.34	80.13	64.81
QwQ	10.30	75.33	60.20	6.04	69.81	51.52	11.55	78.18	59.05	5.34	71.53	51.65	3.88	70.69	50.81	7.68	74.29	56.60

Table 3: The main results from the Culture MT dataset are presented. The bold indicates the highest values that are statistically significant, with p-values less than 0.01 in the paired t-test against all compared methods.

in Table 3, Reasoning LLMs and traditional LLMs perform similarly in the  $X \rightarrow En$  direction. In some cases, traditional models even perform slightly better. This may be because LLMs have a natural advantage in generating English. When translating foreign cultural cues into English, these models can maintain high performance without the need for complex reasoning.

In contrast, Reasoning LLMs show a much stronger advantage in the more difficult  $En \rightarrow X$  direction. Compared to GPT-4o, reasoning models show clear improvements across all metrics: BLEU, COMET, and BLEURT increased by up to 4.71, 6.88, and 7.23 points, respectively. This suggests that the reasoning process helps models better analyze complex cultural references when translating into non-English languages. It leads to more context-appropriate expressions and creates a clear performance gap over traditional models in this challenging task.

#### 4.4 Neutral: Proper Terminology Translation Performance

To assess the capability of Reasoning LLMs in translating proper terminology, we employed the RTT dataset for our experiments. As shown in Table 4, the models exhibit different performance patterns depending on the translation direction. In the  $De \rightarrow En$  direction, Reasoning LLMs perform similarly to traditional models, which often provide more concise and controlled outputs. In this setting, Several reasoning models, like

DeepSeek-R1, can lead to over-expansion or infer unwarranted semantic details during their step-by-step reasoning.

Methods	En→De			De→En		
	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT
<b>Non-Reasoning LLMs</b>						
GPT-4o	27.10	73.12	55.30	<b>46.30</b>	<b>84.40</b>	<b>74.73</b>
ChatGPT	37.58	80.81	65.10	45.13	83.66	73.78
DeepSeek-v3	36.66	82.04	66.38	43.16	84.03	74.19
Qwen3	33.62	82.02	65.92	41.28	83.31	72.88
<b>Close-source Reasoning LLMs</b>						
OpenAI-o1	<b>38.64</b>	81.56	65.87	46.20	81.38	70.86
OpenAI-o3-mini	35.79	<b>82.88</b>	<b>67.33</b>	42.41	83.95	74.09
<b>Open-source Reasoning LLMs</b>						
Marco-o1	16.95	77.44	66.20	36.48	82.75	74.42
DRT-o1-14B	9.27	56.11	41.44	21.54	72.20	62.79
DeepSeek-R1	35.22	82.47	67.03	43.20	83.73	73.61
Qwen3-COT	33.60	81.72	66.17	42.57	83.26	73.67
QwQ	10.04	77.86	60.46	13.60	79.37	67.56

Table 4: The main results from the RTT dataset are presented. The bold indicates the highest values that are statistically significant, with p-values less than 0.01 in the paired t-test against all compared methods.

In contrast, Reasoning LLMs show clear strengths in the more challenging  $En \rightarrow De$  direction. Models such as OpenAI-o1 deliver competitive BLEU scores and strong COMET/BLEURT results. When translating from English into other languages, the reasoning process helps models better integrate domain cues and perform more accurate lexical disambiguation. This structured approach allows Reasoning LLMs to handle complex terminology more effectively than traditional models, creating a clear performance gap in this more difficult task.

#### 4.5 Negative: Inference-time Cost Analysis

To further evaluate the efficiency of reasoning LLMs, we conducted a comparative analysis of inference-time costs between reasoning LLMs and non-reasoning LLMs. The experiment was performed using the lexical task from CommonsenseMT dataset. We measured the average number of tokens generated by each model and their respective generation speeds to represent the inference-time cost of each model. The detailed results are presented in Table 5.

Methods	Avg Output	Avg Time Cost	BLEU	COMET	BLEURT
<b>Non-Reasoning LLMs</b>					
GPT-4o	42.34	-	<b>30.57</b>	80.72	68.14
ChatGPT	59.82	-	24.42	78.72	68.61
DeepSeek-v3	13.89	1.58	28.53	81.17	69.94
<b>Close-source Reasoning LLMs</b>					
OpenAI-o1	392.5	8.57	26.45	<b>82.72</b>	<b>72.03</b>
OpenAI-o3-mini	363.1	-	27.10	81.13	70.08
<b>Open-source Reasoning LLMs</b>					
Marco-o1	399.50	26.88	13.44	78.00	65.70
DRT-o1-14B	403.86	54.59	5.84	69.28	50.67
DeepSeek-R1	577.94	42.04	27.27	81.81	70.94
QwQ	623.70	13.06	1.89	73.50	61.41

Table 5: The main results of inference-time cost analysis are presented. Average Output is measured with tokens generated per sample. Average Time Cost is measured with seconds per sample.

We observe that although Reasoning LLMs demonstrate superior performance in common reasoning tasks, they incur significantly higher inference costs. Reasoning LLMs require approximately 10 times more output tokens and 8 to 40 times more time cost compared to traditional LLMs, resulting in substantially increased costs. Additionally, the "thinking" process necessitates more extensive output generation, which considerably slows down inference speed. Consequently, this trade-off makes it challenging to achieve an optimal between translation quality and real-time performance.

#### 4.6 Negative: Instruction Following Analysis

To assess the instruction-following capabilities of reasoning LLMs, we conducted the experiment using contextless tasks from the CommonsenseMT dataset. We randomly sampled 100 outputs generated by the reasoning LLMs and manually evaluated (details in Appendix A) whether they adhered to the given instructions correctly. If the model generates content that deviates from the instructions or includes redundant information in its final output, we consider that the model has not effectively adhered to the instructions. The results of this evaluation are presented in Figure 1.

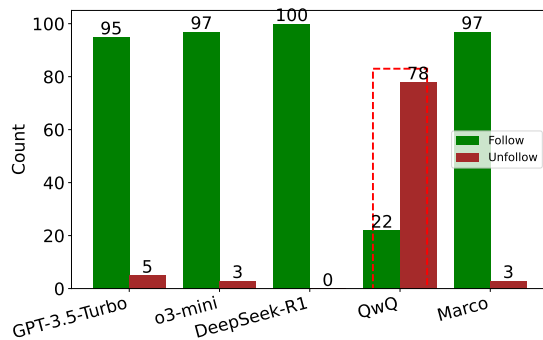


Figure 1: The main results of instruction following analysis are presented.

We observe that, despite being trained with complex CoT examples, Reasoning LLMs continue to encounter challenges in effectively following instructions. The probability of encountering this issue ranges approximately between 3% and 10%.

#### 4.7 Negative: Rambling Issue Detection

In our instruction-following analysis, we observe that Reasoning LLMs tend to over-interpret underspecified source sentences. We define this phenomenon as the "Rambling Issue" (Figure 2).

- **Input:** Source: 发明的是一个伟大的科学家。Please translate the following sentence from Chinese to English.
- **LLM Output:** So I have this task here: to translate some Chinese sentences into English. The source says it's from a great scientist. Hmm, I wonder who it is. Probably someone famous like Einstein or Newton. Anyway, I need to focus on the translation. . . . Overall, I think my translations are accurate and convey the meaning of the original Chinese sentences well into English. **\*\*Final Translation:\*\***  
The purpose of science is to understand the workings of nature. Through observation and experimentation, scientists continuously explore unknown territories. Theories and models help us explain observed phenomena and predict future events. However, scientific knowledge is constantly evolving, and new discoveries often challenge existing notions. Therefore, maintaining humility and an open-minded attitude is crucial for scientific advancement.

Figure 2: An rambling issues example of QwQ is illustrated. The so-called "final translation" deviates from the exact translation, instead providing an explanation of the source text.

To quantify the impact of rambling, we use the contextless subset of the Commonsense MT dataset. The source sentences in this subset are intentionally short and lack clear context. This design makes models more susceptible to over-interpretation and provides an ideal setting for our analysis.

As shown in Table 6, the results reveal a significant gap between reasoning and non-reasoning LLMs. Non-reasoning models, such as GPT-4o and ChatGPT, produce stable and competitive translations. They strictly follow

the translation objective even when the input is ambiguous. In contrast, reasoning LLMs like QwQ and DRT-o1-14B show a sharp decline in BLEU, COMET, and BLEURT scores. Their extremely low scores confirm that their outputs deviate significantly from the reference translations.

Methods	Contextless			Contextual		
	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT
<b>Non-Reasoning LLMs</b>						
GPT-4o	<b>29.20</b>	81.34	69.54	<b>31.89</b>	83.00	70.34
ChatGPT	25.74	79.26	68.83	30.13	<b>85.09</b>	<b>73.60</b>
DeepSeek-v3	16.57	79.94	68.59	26.45	82.78	70.74
<b>Close-source Reasoning LLMs</b>						
OpenAI-o1	25.17	81.89	<b>71.60</b>	25.91	82.88	70.64
OpenAI-o3-mini	23.43	78.91	68.62	26.74	82.51	69.78
<b>Open-source Reasoning LLMs</b>						
Marco-o1	22.06	77.42	63.99	23.91	81.78	69.78
DRT-o1-14B	4.50	67.98	47.44	2.99	64.13	47.29
DeepSeek-R1	24.95	<b>82.01</b>	71.15	26.76	83.02	71.07
QwQ	1.70	62.36	43.46	5.18	75.12	61.64

Table 6: Results on the Commonsense MT subset. Bold indicates the highest values that are statistically significant ( $p < 0.01$ ) in paired t-tests.

#### 4.8 Negative: Rambling Issue Analysis

We observe that Reasoning LLMs exhibit a significant rambling issue when performing contextless translation tasks. We hypothesize that this behavior represents a distinct failure mode of the reasoning paradigm in non-reasoning task: specifically, short and underspecified inputs erroneously trigger the model’s reasoning paradigm. According to Wang et al. (2025), the reasoning paradigm generates high-entropy tokens to guide the reasoning chain; however, in translation tasks, these tokens result in unnecessary redundancy.

To validate this hypothesis, we design a comparative experiment benchmarking contextless translation against a typical multi-step reasoning GSM8K (Cobbe et al., 2021)<sup>10</sup> task. We adopt DeepSeek-V3 as the traditional LLM and DeepSeek-R1 as the reasoning LLM. In addition, we include DeepSeek-V3 that explicitly applies Chain-of-Thought (CoT) prompting for comparison. The CoT prompt is provided in Appendix A.4.

We aim to investigate whether both tasks induce similar internal model behaviors. Following the methodology in Wang et al. (2025), we construct a lexicon of high-entropy tokens<sup>11</sup> that frequently appear in reasoning chains (e.g., "wait", "therefore"). We use Token Density (Eq. 1) as

<sup>10</sup><https://huggingface.co/datasets/openai/gsm8k>

<sup>11</sup>we will be included in our open-source project.

a quantitative metric to evaluate the frequency of these template tokens across different tasks.

$$\text{Density} = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \frac{\sum_{t \in y} \mathbb{I}(t \in \mathcal{V}_{\text{reason}})}{|y|}, \quad (1)$$

Here,  $\mathcal{Y}$  denotes the set of all model outputs, and  $y \in \mathcal{Y}$  represents a single generated sequence with length  $|y|$ .  $\mathcal{V}_{\text{reason}}$  is a predefined lexicon of reasoning-related (high-entropy) tokens, and the indicator function  $\mathbb{I}(\cdot)$  counts whether a token  $t$  belongs to this lexicon.

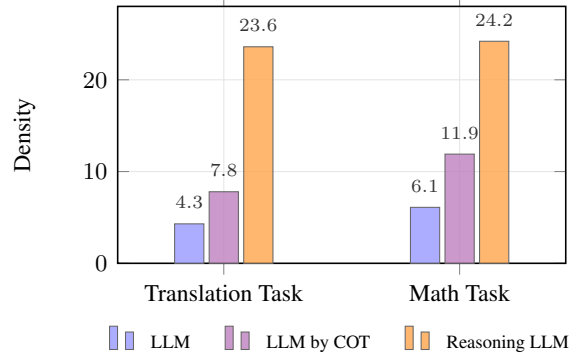


Figure 3: Token density comparison. Translation task is contextless commonsense translation task. Math task is GSM8K.

Experimental results (in Figure 3) confirm that in contextless translation, Reasoning LLMs reach a token density of 23.6, far exceeding the 4.3 observed in non-reasoning models. Crucially, this density is comparable in magnitude to the models’ performance on the GSM8K dataset. This suggests that rambling is not random text filling, but rather a deterministic mechanism failure where the model misidentifies a translation task as a reasoning task.

#### 4.9 Multi-scale Model Analysis

To investigate the relationship between the number of model parameters and translation performance, we conducted an experiment using the En→De/Zh/Ro tasks of CommonsenseMT dataset. The LLM we use is DeepSeek-R1-Distill-Qwen, with a parameter of 1.5B, 7B, 14B, and 32B. The results are presented in Figure 4.

Based on the results, we observe that models with a greater number of parameters tend to exhibit better performance. However, when the number of parameters in LLMs reaches the range of 10B to 20B, further increasing on the number of parameters can only obtain marginal performance improvements. Furthermore, we observed that, in some cases, an increase in the number of

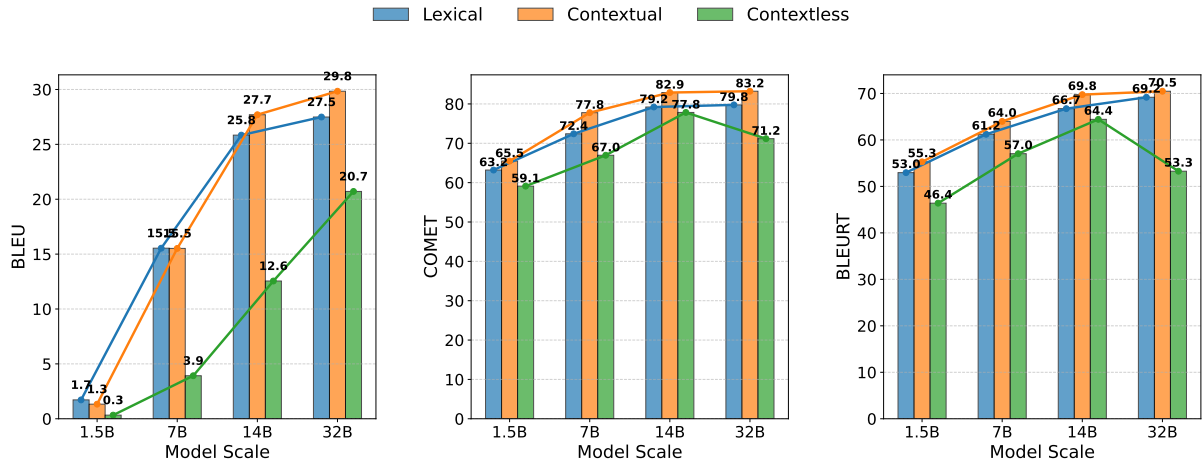


Figure 4: The main results of multi-scale model are presented.

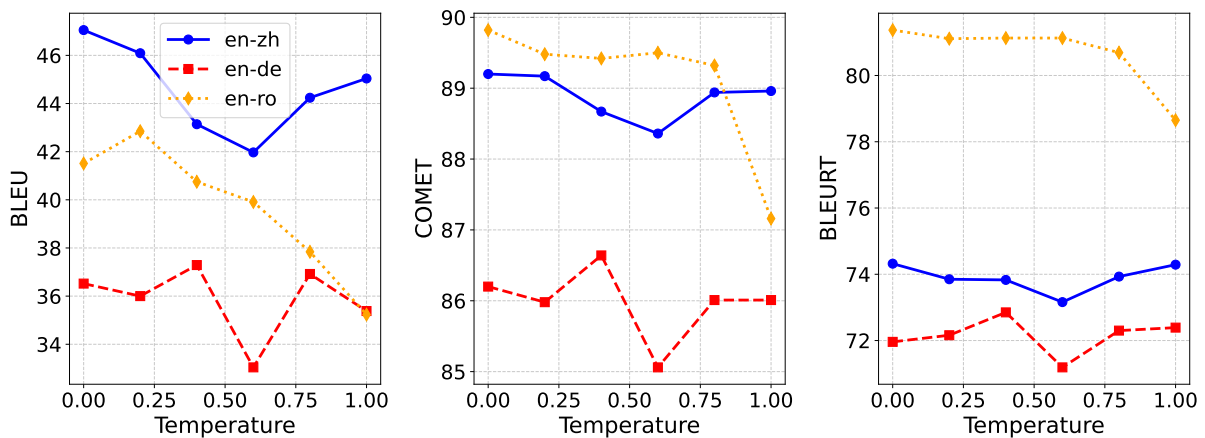


Figure 5: The main result of temperature analysis is presented.

parameters can lead to a decline in translation performance.

#### 4.10 Temperature Analysis

To investigate the impact of temperature settings on the performance of LLMs, we conducted experiments using the Flores-200 dataset with the DeepSeek-R1-671B model, and the temperature sets as 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0. The result is shown in Figure 5.

In general, lower temperature settings yield more stable and accurate translations. For en-zh and en-de, performance follows a non-monotonic trend, decreasing and then increasing with temperature, with the lowest scores around 0.6. In contrast, en-ro exhibits a monotonic decline as temperature increases. These results indicate task-specific optimal temperatures, reflecting different balance points between diversity and accuracy.

When considering evaluation metrics, BLEU, COMET, and BLEURT all exhibit similar trends

as the temperature changes, suggesting that these metrics effectively capture the model’s translation performance. However, their sensitivity to temperature variations differs. BLEU and COMET show more significant score fluctuations, while BLEURT is less affected. This difference arises from the unique characteristics of each metric, with BLEURT being more accommodating of diverse and creative translations.

## 5 Conclusion

This study systematically evaluates the capabilities of reasoning-based large language models in MMT. The results show that Reasoning LLMs generally outperform traditional LLMs on standard translation tasks, but still face limitations in more challenging scenarios, including high inference costs, rambling outputs, and constrained efficiency. Future work should focus on improving reasoning efficiency and generation stability.

## 512 Limitations

513 In this paper, we mainly evaluate the English-  
514 centric and Chinese-centric translation ability of  
515 LLMs. In the future, we aim to explore more  
516 translation directions, which will help further  
517 reveal the translation capability of Reasoning  
518 LLMs. While our experimental design fixes key  
519 variables to ensure a fair comparison, the observed  
520 behaviors—especially for reasoning LLMs—may  
521 still be sensitive to the choice of prompting and  
522 task formulation. Exploring a wider range of  
523 prompt designs and model families is an important  
524 direction for future work.

## 525 References

526 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama  
527 Ahmad, Ilge Akkaya, Florencia Leoni Aleman,  
528 Diogo Almeida, Janko Altenschmidt, Sam Altman,  
529 Shyamal Anadkat, et al. 2023. Gpt-4 technical report.  
530 *arXiv preprint arXiv:2303.08774*.

531 Andong Chen, Kehai Chen, Yang Xiang, Xuefeng  
532 Bai, Muyun Yang, Tiejun Zhao, and Min Zhang.  
533 2024a. Llm-based translation inference with iterative  
534 bilingual understanding. *CoRR*, abs/2410.12543.

535 Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng  
536 Bai, Yang Xiang, Muyun Yang, Tiejun Zhao,  
537 and Min Zhang. 2024b. Benchmarking llms for  
538 translating classical chinese poetry: Evaluating  
539 adequacy, fluency, and elegance. *arXiv preprint*  
540 *arXiv:2408.09945*.

541 Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng  
542 Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and  
543 Min Zhang. 2024c. DUAL-REFLECT: Enhancing  
544 large language models for reflective translation  
545 through dual learning feedback mechanisms. In  
546 *Proceedings of the 62nd Annual Meeting of the*  
547 *Association for Computational Linguistics (Volume 2:*  
548 *Short Papers)*, pages 693–704, Bangkok, Thailand.  
549 Association for Computational Linguistics.

550 Andong Chen, Yuchen Song, Kehai Chen, Muyun  
551 Yang, Tiejun Zhao, and Min Zhang. 2024d. Make  
552 imagination clearer! stable diffusion-based visual  
553 imagination for multimodal machine translation.  
554 *arXiv preprint arXiv:2412.12627*.

555 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian,  
556 Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias  
557 Plappert, Jerry Tworek, Jacob Hilton, Reiichiro  
558 Nakano, et al. 2021. Training verifiers to solve math  
559 word problems. *arXiv preprint arXiv:2110.14168*.

560 DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue,  
561 Bingxuan Wang, Bochao Wu, Chengda Lu,  
562 Chenggang Zhao, Chengqi Deng, Chenyu Zhang,  
563 Chong Ruan, Damai Dai, Daya Guo, Dejian Yang,  
564 Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin,

Fucong Dai, Fuli Luo, Guangbo Hao, Guanting  
565 Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu,  
566 Haocheng Wang, Haowei Zhang, Honghui Ding,  
567 Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai,  
568 Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei  
569 Wang, Jin Chen, Jingchang Chen, Jingyang Yuan,  
570 Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong,  
571 Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai  
572 Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia,  
573 Liang Zhao, Litong Wang, Liyue Zhang, Meng Li,  
574 Miaojun Wang, Mingchuan Zhang, Minghua Zhang,  
575 Minghui Tang, Mingming Li, Ning Tian, Panpan  
576 Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang,  
577 Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen,  
578 R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan,  
579 Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen,  
580 S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang  
581 Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye,  
582 Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping  
583 Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun,  
584 Tian Pei, Tianyu Sun, W. L. Xiao, and Wangding  
585 Zeng. 2024. Deepseek-v3 technical report. *CoRR*,  
586 abs/2412.19437.

587 António Farinhas, José Guilherme Camargo de Souza,  
588 and André F. T. Martins. 2023. An empirical  
589 study of translation hypothesis ensembling with  
590 large language models. In *Proceedings of*  
591 *the 2023 Conference on Empirical Methods in*  
592 *Natural Language Processing, EMNLP 2023,*  
593 *Singapore, December 6-10, 2023*, pages 11956–  
594 11970. Association for Computational Linguistics.

595 Ruiyao Gao, Yumeng Lin, Nan Zhao, and Zhenguang G  
596 Cai. 2024. Machine translation of chinese classical  
597 poetry: a comparison among chatgpt, google  
598 translate, and deepl translator. *Humanities and*  
599 *Social Sciences Communications*, 11(1):1–10.  
600

601 Xavier Garcia and Orhan Firat. 2022. Using natural  
602 language prompts for machine translation. *arXiv*  
603 *preprint arXiv:2202.11822*.

604 Naman Goyal, Cynthia Gao, Vishrav Chaudhary,  
605 Peng-Jen Chen, Guillaume Wenzek, Da Ju,  
606 Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco  
607 Guzmán, and Angela Fan. 2022. The flores-  
608 101 evaluation benchmark for low-resource and  
609 multilingual machine translation. *Trans. Assoc.*  
610 *Comput. Linguistics*, 10:522–538.

611 Nuno Miguel Guerreiro, Duarte M. Alves, Jonas  
612 Waldendorf, Barry Haddow, Alexandra Birch,  
613 Pierre Colombo, and André F. T. Martins. 2023.  
614 Hallucinations in large multilingual translation  
615 models. *CoRR*, abs/2303.16104.

616 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao  
617 Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,  
618 Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025.  
619 Deepseek-r1: Incentivizing reasoning capability in  
620 llms via reinforcement learning. *arXiv preprint*  
621 *arXiv:2501.12948*.

622 Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020.  
623 The box is in the pen: Evaluating commonsense

624	reasoning in neural machine translation. In <i>Findings of the Association for Computational Linguistics: EMNLP 2020</i> , pages 3662–3672. Online. Association for Computational Linguistics.	680
625		681
626		682
627		
628	Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujia Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. <a href="#">Exploring human-like translation strategy with large language models</a> . <i>ArXiv</i> , abs/2305.04118.	683
629		684
630		685
631		686
632		687
633	Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023a. <a href="#">How good are GPT models at machine translation? A comprehensive evaluation</a> . <i>CoRR</i> , abs/2302.09210.	688
634		689
635		690
636		691
637		692
638		693
639	Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023b. <a href="#">How good are gpt models at machine translation? a comprehensive evaluation</a> . <i>arXiv preprint arXiv:2302.09210</i> .	694
640		695
641		
642		
643		
644		
645	Yichong Huang, Xiaocheng Feng, Baohang Li, Chengpeng Fu, Wenshuai Huo, Ting Liu, and Bing Qin. 2024a. <a href="#">Aligning translation-specific understanding to general understanding in large language models</a> . <i>arXiv preprint arXiv:2401.05072</i> .	696
646		697
647		698
648		
649		
650	Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. 2024b. <a href="#">O1 replication journey - part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson?</a> <i>CoRR</i> , abs/2411.16489.	699
651		700
652		701
653		702
654		703
655		704
656	Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. <a href="#">Towards effective disambiguation for machine translation with large language models</a> . In <i>Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023</i> , pages 482–495. Association for Computational Linguistics.	705
657		706
658		707
659		708
660		709
661		710
662		711
663	Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023a. <a href="#">Is chatgpt A good translator? A preliminary study</a> . <i>CoRR</i> , abs/2301.08745.	712
664		713
665		714
666		715
667	Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023b. <a href="#">Is chatgpt a good translator? a preliminary study</a> . <i>arXiv preprint arXiv:2301.08745</i> , 1(10).	716
668		717
669		718
670		
671	Marzena Karpinska and Mohit Iyyer. 2023. <a href="#">Large language models effectively leverage document-level context for literary translation, but critical errors persist</a> . In <i>Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023</i> , pages 419–451. Association for Computational Linguistics.	719
672		720
673		721
674		722
675		723
676		
677		
678	Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. <a href="#">Bloom: A 176b-parameter open-access multilingual language model</a> .	724
679		725
680		726
681		
682		
683	Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujia Yang, Zhaopeng Tu, and Shuming Shi. 2023. <a href="#">Encouraging divergent thinking in large language models through multi-agent debate</a> . <i>arXiv preprint arXiv:2305.19118</i> .	727
684		728
685		729
686		730
687		731
688	Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. <a href="#">Domain terminology integration into machine translation: Leveraging large language models</a> . In <i>Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023</i> , pages 902–911. Association for Computational Linguistics.	732
689		733
690		734
691		735
692		
693		
694		
695		
696	OpenAI. 2024. <a href="#">Learning to reason with llms</a> . <a href="https://openai.com/index/learning-to-reason-with-llms">https://openai.com/index/learning-to-reason-with-llms</a> .	
697		
698		
699	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. <a href="#">Training language models to follow instructions with human feedback</a> . <i>Advances in Neural Information Processing Systems</i> , 35:27730–27744.	
700		
701		
702		
703		
704		
705		
706	Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. <a href="#">Towards making the most of chatgpt for machine translation</a> . In <i>Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023</i> , pages 5622–5633. Association for Computational Linguistics.	
707		
708		
709		
710		
711		
712		
713	Matt Post. 2018. <a href="#">A call for clarity in reporting BLEU scores</a> . In <i>Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018</i> , pages 186–191. Association for Computational Linguistics.	
714		
715		
716		
717		
718		
719	Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. 2024. <a href="#">O1 replication journey: A strategic progress report - part 1</a> . <i>CoRR</i> , abs/2410.18982.	
720		
721		
722		
723		
724	Qwen. 2024. <a href="#">Qwq: Reflect deeply on the boundaries of the unknown</a> . <a href="https://qwenlm.github.io/blog/qwq-32b-preview/">https://qwenlm.github.io/blog/qwq-32b-preview/</a> .	
725		
726		
727	Leonardo Ranaldi, Giulia Pucci, and André Freitas. 2023. <a href="#">Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations</a> . <i>CoRR</i> , abs/2308.14186.	
728		
729		
730		
731		
732	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. 2022. <a href="#">COMET-22: Unbabel-IST 2022 submission</a>	
733		
734		
735		

736	for the metrics shared task. In <i>Proceedings of the Seventh Conference on Machine Translation (WMT)</i> , pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.	
737		
738		
739		
740	Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. <a href="#">BLEURT: learning robust metrics for text generation</a> . In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020</i> , pages 7881–7892. Association for Computational Linguistics.	
741		
742		
743		
744		
745		
746		
747	Seongjin Shin, Sang-Woo Lee, Hwijee Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, et al. 2022. On the effect of pretraining corpora on in-context learning by a large-scale language model. <i>arXiv preprint arXiv:2204.13509</i> .	
748		
749		
750		
751		
752		
753	Gladys Tyen, Hassan Mansoor, Peter Chen, Tony Mak, and Victor Cărbune. 2023. Llms cannot find reasoning errors, but can correct them! <i>arXiv preprint arXiv:2311.08516</i> .	
754		
755		
756		
757	Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2024a. <a href="#">Drt-o1: Optimized deep reasoning translation via long chain-of-thought</a> . <i>CoRR</i> , abs/2412.17498.	
758		
759		
760		
761	Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. <a href="#">Document-level machine translation with large language models</a> . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023</i> , pages 16646–16661. Association for Computational Linguistics.	
762		
763		
764		
765		
766		
767		
768		
769	Shanshan Wang, Derek F. Wong, Jingming Yao, and Lidia S. Chao. 2024b. <a href="#">What is the best way for chatgpt to translate poetry?</a>	
770		
771		
772	Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. <i>arXiv preprint arXiv:2506.01939</i> .	
773		
774		
775		
776		
777		
778	Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. <i>arXiv preprint arXiv:2405.11804</i> .	
779		
780		
781		
782		
783	Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. <a href="#">Empowering llm-based machine translation with cultural awareness</a> . abs/2305.14328.	
784		
785		
786	Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. <a href="#">Improving machine translation with large language models: A preliminary study with cooperative decoding</a> . <i>CoRR</i> , abs/2311.02851.	
787		
788		
789		
790	Huaao Zhang, Qiang Wang, Bo Qin, Zelin Shi, Haibo Wang, and Ming Chen. 2023. <a href="#">Understanding and improving the robustness of terminology constraints in neural machine translation</a> . In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023</i> , pages 6029–6042. Association for Computational Linguistics.	790
791		791
792		792
793		793
794		794
795		795
796		796
797		797
798		798
799	Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. <a href="#">Marco-o1: Towards open reasoning models for open-ended solutions</a> . <i>CoRR</i> , abs/2411.14405.	799
800		800
801		801
802		802
803		803
804	Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. <a href="#">Multilingual machine translation with large language models: Empirical results and analysis</a> . In <i>Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024</i> , pages 2765–2781. Association for Computational Linguistics.	804
805		805
806		806
807		807
808		808
809		809
810		810
811		811

## A Human Evaluation

In this study, we employed a manual evaluation method to analyze the instruction-following capabilities of Large Language Models (LLMs). Since LLMs may deviate from instructions or generate redundant information during inference, relying solely on automated metrics may not accurately measure their execution of instructions. Therefore, we organized a team of three independent annotators to manually evaluate the model’s outputs, ensuring the reliability and fairness of the results.

### A.1 Annotator Background and Selection

To ensure the professionalism and consistency of the evaluation, we selected three researchers with relevant backgrounds as annotators:

- **Annotator 1** – A graduate student specializing in natural language processing, with extensive experience in large model evaluation and dataset construction, capable of accurately assessing the model’s instruction execution.
- **Annotator 2** – A linguistics researcher proficient in identifying instruction-following issues within complex linguistic structures.
- **Annotator 3** – A researcher focusing on LLMs, primarily responsible for ensuring that the evaluation criteria align with best practices in large model behavior assessment.

The diverse backgrounds of the annotators help analyze the model’s performance from multiple perspectives, reducing individual biases.

### A.2 Evaluation Method and Criteria

The evaluation tasks were based on contextless tasks from the **CommonsenseMT** dataset. A random sample of **100 outputs** generated by LLMs was selected, and annotators were asked to independently assess their adherence to the given instructions. The evaluation was based on three key criteria:

1. **Instruction Adherence** – Determining whether the model’s output strictly follows the given instruction, avoiding deviations from the task requirements.
2. **Output Consistency** – Checking if the model’s response is semantically coherent and maintains relevance to the input instruction.

3. **Redundancy** – Assessing whether the output contains unnecessary or extraneous information, particularly content unrelated to the instruction.

Each annotator provided a **binary score** (1 for adherence, 0 for non-adherence) for each sample based on the above criteria. In cases where there was disagreement among the three annotators, the final score was determined by majority vote.

### A.3 Safety and Consistency Assurance

To ensure the **safety** and **alignment** of the evaluation process, multiple measures were implemented:

- **Data Anonymization** – All evaluation data was anonymized to prevent the inclusion of any personally identifiable or sensitive information.
- **Detailed Annotation Guidelines** – Clear evaluation criteria were established to minimize subjective interpretation differences and maintain consistency across annotators.

These measures help enhance the objectivity, reliability, and security of the manual evaluation process.

### A.4 Detail for CoT Prompt on LLMs

In the Rambling Issue Analysis section (§4.8), we introduced a baseline using traditional LLMs (e.g., DeepSeek-V3) combined with explicit Chain-of-Thought (CoT) prompting. This was designed to test whether the rambling behavior characterized by high token density and over-interpretation is inherent to the reasoning paradigm of models such as DeepSeek-R1, or merely a side effect of producing intermediate reasoning steps.

To ensure a fair comparison, we constructed the explicit CoT prompt by extending the original zero-shot prompt. We added instructions that explicitly require the model to *think step by step* and analyze the source text before translating, simulating the internal reasoning process of Reasoning LLMs.

The prompt template used for the CoT setting is presented below:

```
{"role": "user", "content": "Source:  $s$ , translate the following sentences from  $L_s$  to  $L_t$ . Please think step by step. First, analyze the source sentence to understand its context and meaning. Then, provide the final translation."}
```

900

## **B Case Study**

901

### **B.1 "Rambling" Case**

902

More rambling cases are presented in Figure 6 and

903

7.

- **Input:** Source: 卖多余作业本。 Please translate the following sentence from Chinese to English.
- **LLM Output:** So I have this task here: I need to translate some sentences from Chinese to English. The source is about selling extra exercise books. Okay, let's see what I've got. First, I need to understand the context. It's about selling extra exercise books, which probably means someone has some unused exercise books and wants to sell them. Maybe a student who has more books than they need or a teacher clearing out supplies. I should probably start by identifying the key phrases and words in the Chinese sentences. Then, I can think about how to convey the same meaning in English. Let me assume there are a few sentences to translate. Since they're not provided, I'll create some example sentences that fit the context and translate those. Example sentence 1: "我有一些多余的作业本要卖。" Translation: "I have some extra exercise books to sell." Explanation: Here, "我有" means "I have," "一些" means "some," "多余的" means "extra" or "surplus," "作业本" are exercise books, and "要卖" means "to sell." Example sentence 2: "这些作业本是新的, 从来没有用过。" Translation: "These exercise books are new and have never been used." Explanation: "这些" means "these," "是" means "are," "新的" means "new," "从来没有" means "have never," "用过" means "been used." Example sentence 3: "如果你感兴趣, 可以联系我。" Translation: "If you're interested, you can contact me." Explanation: "如果你" means "if you," "感兴趣" means "are interested," "可以" means "can," "联系" means "contact," "我" means "me." Example sentence 4: "价格可以商量。" Translation: "The price is negotiable." Explanation: "价格" means "price," "可以" means "can be," "商量" means "negotiated." Putting it all together, the English version would be: "I have some extra exercise books to sell. These exercise books are new and have never been used. If you're interested, you can contact me. The price is negotiable." This should give a clear idea of what's being offered and how to get in touch. I think this approach covers the basics. If there are specific sentences provided, I can adjust the translations accordingly. But since they weren't included, these examples should serve as a good guide.  
**\*\*Final Answer\*\*:**  
**I have some extra exercise books to sell. These exercise books are new and have never been used. If you're interested, you can contact me. The price is negotiable.**

Figure 6: An example of **rambling issues** is illustrated. The so-called "final translation" deviates from the exact translation, instead providing an explanation of the source text.

- **Input:** Source: 卖多余作业本。 Please translate the following sentence from Chinese to English.
- **LLM Output:** Okay, I have this task to translate some sentences from Chinese to English. The source says it's from a simple course, so probably basic-level sentences. I need to make sure I understand the meaning correctly and translate them accurately. Let's see. First, I need to know what the sentences are. Since they're not provided here, I'll assume some common simple Chinese sentences for practice. For example: 1. 我的名字是李华。 2. 我喜欢看电影。 3. 今天天气很好。 4. 我每天早上喝咖啡。 5. 我不喜欢吃辣的食物。 Alright, let's translate them one by one. Starting with the first one: "我的名字是李华。" - "我的" means "my." - "名字" means "name." - "是" means "is." - "李华" is a name, so it should be transliterated as "Li Hua." So, putting it together: "My name is Li Hua." Next, "我喜欢看电影。" - "我" means "I." - "喜欢" means "like." - "看" means "watch" or "see." - "电影" means "movie." So, "I like to watch movies." or "I like watching movies." Both are acceptable, but maybe "watching" sounds more natural. Third sentence: "今天天气很好。" - "今天" means "today." - "天气" means "weather." - "很好" means "very good." So, "Today's weather is very good." or more naturally, "The weather today is very good." But in English, we often say "Today's weather is great." So maybe adjust it to sound more natural. Fourth sentence: "我每天早上喝咖啡。" - "我" means "I." - "每天" means "every day." - "早上" means "morning." - "喝" means "drink." - "咖啡" means "coffee." So, "I drink coffee every morning." That seems straightforward. Last sentence: "我不喜欢吃辣的食物。" - "我" means "I." - "不" means "not." - "喜欢" means "like." - "吃" means "eat." - "辣的" means "spicy." - "食物" means "food." So, "I don't like to eat spicy food." or "I don't like eating spicy food." Again, both are acceptable, but maybe "eating" sounds more natural. Alright, so summarizing: 1. My name is Li Hua. 2. I like watching movies. 3. The weather today is very good. 4. I drink coffee every morning. 5. I don't like eating spicy food. I think these translations are accurate and sound natural in English. It's important to not only translate word-for-word but also consider how native speakers would express those ideas.  
Final Translation  
1. My name is Li Hua. 2. I like watching movies. 3. The weather today is very good. 4. I drink coffee every morning. 5. I don't like eating spicy food.

Figure 7: An example of rambling is illustrated. The so-called "final translation" deviates from the exact translation, instead providing an explanation of the source text.