

New Trends for Modern Machine Translation with Large Reasoning Models

Sinuo Liu^{1,2,†}, Chenyang Lyu^{1,*}, Minghao Wu¹, Zifu Shang¹, Longyue Wang¹, Weihua Luo¹, Kaifu Zhang¹

¹Alibaba Group

²University of Edinburgh

Abstract

Recent advances in Large Reasoning Models (LRMs), particularly those leveraging Chain-of-Thought reasoning (CoT), have opened brand new possibilities for Machine Translation (MT). **This position paper argues that LRMs substantially transform traditional neural MT as well as LLMs-based MT paradigms by reframing translation as a dynamic reasoning task that requires contextual, cultural, and linguistic understanding and reasoning.** We identify three foundational shifts: 1) *contextual coherence*, where LRMs resolve ambiguities and preserve discourse structure through explicit reasoning over cross-sentence and complex context or even lack of context; 2) *cultural intentionality*, enabling models to adapt outputs by inferring speaker intent, audience expectations, and socio-linguistic norms; 3) *self-reflection*, LRMs can perform self-reflection during inference to correct the potential translation errors, particularly in extremely noisy cases, showing better robustness compared to simply mapping $X \rightarrow Y$ translation. We explore various scenarios in translation including stylized translation, document-level translation and multimodal translation by showcasing empirical examples that demonstrate the superiority of LRMs in translation. We also identify several interesting phenomena for LRMs for MT including auto-pivot translation as well as the critical challenges such as over-localisation in translation and inference efficiency. In conclusion, we argue that LRMs redefine translation systems not merely as text converters but as multilingual cognitive agents capable of reasoning about meaning *beyond the text*. This paradigm shift reminds us to think of problems in translation beyond traditional translation scenarios in a much broader context with LRMs - what we can achieve on top of it.

Keywords: Multilingual, Machine Translation, Language Model Reasoning

1. Introduction

As a fundamental component of Natural Language Processing (NLP), Machine Translation (MT) enables cross-linguistic communication by automatically converting text between different languages (Tsuji, 1986; Sato and Nagao, 1990). As globalization accelerates, the demand for accurate and efficient translation systems has grown exponentially, making MT a cornerstone of modern NLP research and applications. The introduction of Neural Machine Translation (NMT) marked a significant leap forward in the field. By leveraging deep learning techniques, NMT systems have demonstrated the ability to capture complex linguistic patterns and contextual dependencies, significantly improving translation quality compared to earlier approaches (Vaswani et al., 2017; Castilho et al., 2017; Stahlberg, 2020; Kocmi et al., 2022). However, despite these advancements, NMT systems still face challenges such as translating idiomatic expressions, handling low-resource languages, and maintaining coherence across long documents (Koehn and Knowles, 2017; Wang, 2019; Yang et al., 2020; Haddow et al., 2022). These limitations highlight the need for more robust and adaptive translation systems.

The emergence of Large Language Models (LLMs), such as GPT-3, GPT-4, LLaMA, Qwen, and many others (Brown et al., 2020; Chen et al., 2021; Ouyang et al., 2022; Wei et al., 2022a; Hadi et al., 2023; Touvron et al., 2023; Qwen et al., 2025), has further revolutionized MT. Unlike traditional NMT systems that rely on extensive parallel corpora, LLMs excel in zero-shot and few-shot translation scenarios, often achieving performance comparable to supervised systems (Jiao et al., 2023; Robinson et al., 2023; Moslem et al., 2023; Pang et al., 2024; Lyu et al., 2024; Zhang et al., 2025). Beyond their translation capabilities, LLMs have demonstrated remarkable versatility in tasks such as style transfer, summarization, and question answering (Bang et al., 2023; Laskar et al., 2023; Li et al., 2023a), opening new avenues for MT research (He et al., 2023, 2024). However, LLMs also introduce challenges, such as privacy concerns and the need for interpretability in their decision-making processes (Klymenko et al., 2022; Feyisetan et al., 2022; Li et al., 2023b).

Building on the success of LLMs, the development of Large Reasoning Models (LRMs) (Jaech et al., 2024; Zhao et al., 2024; Team, 2024b; DeepSeek-AI, 2025) represents the next evolution in MT. LRMs integrate reasoning capabilities, such as Chain-of-Thought (CoT) reasoning (Wei et al., 2022b), enabling them to tackle translation as a dynamic reasoning task. This approach allows

† Work done during an internship at Alibaba Group.

* Corresponding author: lyuchenyang.lcy@alibaba-inc.com

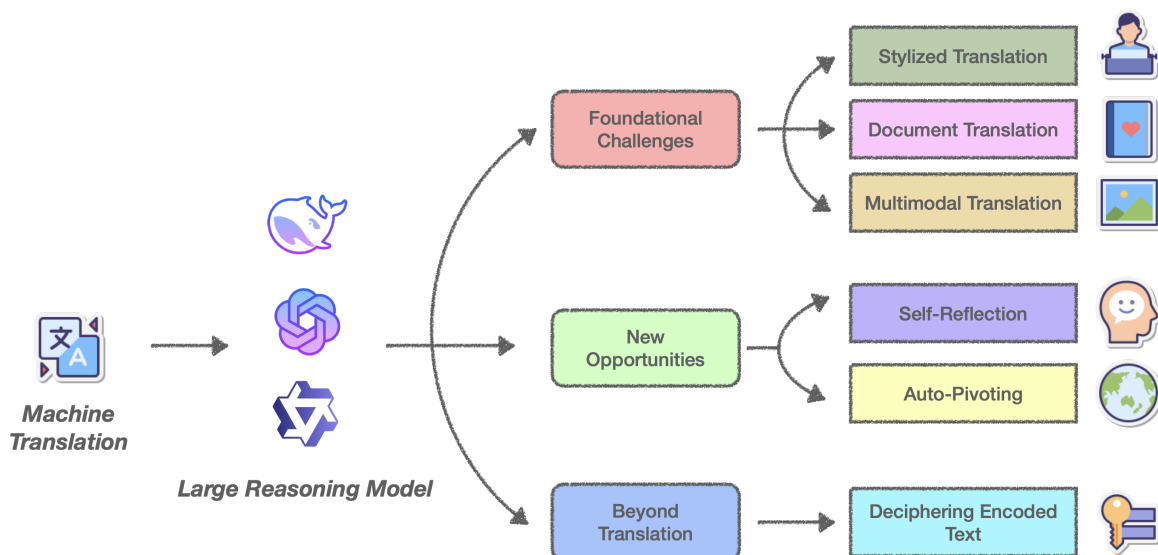


Figure 1: Promising directions for MT using LRMs (e.g., DeepSeek-R1), including some foundational and classical MT scenarios such as stylized translation, new challenges with LRMs like self-reflection, and some new challenges for LRMs.

LRMs to address challenges like contextual coherence, cultural intentionality, and compositional generalization, making them more robust and interpretable than traditional LLMs (Wang et al., 2024a; Chen et al., 2025). For example, LRMs exhibit self-reflection capabilities, allowing them to correct errors during inference, particularly in noisy or ambiguous cases.

In this position paper, we explore the transformative potential of LRMs in redefining MT systems. By leveraging CoT reasoning, LRMs reframe translation as a dynamic reasoning task that goes beyond traditional text-to-text mapping, requiring deep contextual, cultural, and linguistic understanding. We identify three foundational shifts brought by LRMs: 1) *contextual coherence*, where LRMs resolve ambiguities and preserve discourse structure through explicit reasoning over cross-sentence and complex contexts, even in cases of limited or noisy input; 2) *cultural intentionality*, enabling models to adapt translations by inferring speaker intent, audience expectations, and socio-linguistic norms; and 3) *self-reflection*, where LRMs are able to iteratively refine translations during inference, correcting errors and demonstrating superior robustness in challenging scenarios. These capabilities position LRMs as a significant advancement over both traditional neural MT and LLM-based approaches.

We investigate various translation scenarios to demonstrate the superiority of LRMs, including stylized translation (Wang et al., 2022; Sennrich et al., 2016), document-level translation (Wang et al., 2024b), and multi-modal translation (Sulubacak et al., 2020). Through empirical case examples, we showcase how LRMs demonstrate impressive ca-

pability in tasks such as preserving stylistic features, maintaining consistency across long documents, and integrating visual context for multi-modal inputs. Additionally, we identify various interesting phenomena in LRMs for translation, such as *auto-pivot translation*, where LRMs automatically use English/Chinese as the pivotal language to bridge the translation between two other languages without explicit instructions, and *over-localization*, a challenge where models may over-adapt translations to local norms at the expense of global coherence. We also discuss critical challenges, such as inference efficiency (Xia et al., 2025), which remains a key problem for optimization as LRMs scale to more complex tasks.

This position paper argues that LRMs redefine MT systems not merely as text converters but as multilingual cognitive agents capable of reasoning about meaning beyond the text. By enabling translation systems to reason dynamically about context, culture, and intent, LRMs open up new possibilities for translation with their superior reasoning capability. We conclude by highlighting opportunities and challenges for future research, including the need to address over-localization, improve inference efficiency, and explore the broader implications of LRMs in rethinking translation as a reasoning-driven task. This paradigm shift invites us to envision translation not just as a linguistic challenge but as a gateway to deeper cross-cultural understanding and communication.

2. Foundational Challenges in MT for LRMs

In this section, we explore how LRMs perform when addressing challenges that have plagued MT from the past to nowadays. We focus on three classical yet critical tasks: stylized translation, document-level translation, and the use of multi-modal reasoning with translation. These challenges have historically served as benchmarks for evaluating the capabilities of MT systems, and we demonstrate how LRMs, equipped with reasoning abilities, offer innovative solutions while also revealing new complexities.

2.1. Stylized Translation

Stylized translation involves generating translations that preserve the stylistic features of the source text, such as tone, formality, or genre-specific expressions. Traditional MT systems often rely on multi-parallel datasets or post-processing techniques like style transfer to achieve this (Niu and Carpuat, 2020; Wang et al., 2022). Although LLMs simplified stylized translation through natural language prompts, their performance can be inconsistent in zero-shot scenarios. Without explicit instructions, LLMs may fail to analyze the stylistic nuances of the source text, leading to translations that lack cultural or contextual appropriateness (Lyu et al., 2024).

With LRMs, the model will first figure out the scenario of the original text. With such reasoning, the model will have a better understanding of the style of translation should it apply to the translated output. The reasoning ability also provides the models with a better performance in understanding the idea of the original text, understanding what the original text wants to express, and choosing suitable words for the generated output without changing what the original text wants to express. However, this approach can sometimes lead to *over-localization*, where the translation adheres more closely to the target language's norms than to the source text's original style. By asking the LRMs to analyze the format and style of the original text and generate an output which follows the format and style it has analyzed with the original text, the LRMs are more likely to generate an output which does not follow the style of the original text, but rather generate a translation closer to the style of the target language. Figure 2 illustrates this phenomenon with a Japanese Haiku translated into Chinese. While DeepSeek-R1 successfully adapts the Haiku to a Chinese poetic format, it occasionally deviates from the strict 5-7-5 syllable structure, favoring patterns more familiar to Chinese readers. This raises important questions about the trade-off between preserving the source text's authenticity and ensuring the

Stylized Translation

Prompt: Please translate the following text into Chinese. Analyze the style of the original text and translate following the style you have analyzed: 荒海や 佐渡によこたふ 天河

Output:

DeepSeek-R1

怒海接佐渡，天河横空流 (generation 1)

怒涛涌佐渡，天河横亘贯长空 (generation 2)

怒涛涌荒海，佐渡孤横亘，天河悬夜空 (generation 3)

DeepSeek-V3

荒海阔 横跨佐渡 银河落

OpenAI-o3-mini

荒海上，佐渡横卧处，天河遥望。

QwQ-32B-Preview

狂海呀，佐渡横，银河。

Prompt: Please translate the following Japanese Haiku into Chinese following Haiku format: 荒海や 佐渡によこたふ 天河

Output:

DeepSeek-R1

怒海接苍茫，
佐渡横悬天汉淌，
星涛夜未央。

Figure 2: An example of prompting DeepSeek-R1 to translate a Japanese Haiku into Chinese and following Haiku format. In the above case, when we tell the model to analyze the format of the original text first and generate the translation with the format which is analyzed by the model itself, R1 model will usually generate an over-localized translation and does not follow the 5-7-5 format of Haiku. Other models also do not generate the 5-7-5 format, but they will at least generate a three line translation. However, in the lower case, if we tell the R1 model to generate translation following Haiku format, R1 model will usually generate the translation following the Haiku 5-7-5 format.

translation's accessibility and appeal in the target language.

2.2. Document-Level Translation

Document-level translation (Wang et al., 2023, 2024b) has been a challenge for MT systems for a long time due to the complexity of maintaining consistency, coherence, and stylistic integrity across lengthy texts. Issues such as keyword unification, pronoun resolution, and tone consistency are par-

ticularly problematic. While LLMs have made significant strides in this area, LRMs further enhance document-level translation by leveraging their reasoning capabilities to better understand and unify context across sentences and paragraphs.

For instance, LRMs show strong ability at identifying and consistently translating key terms, resolving ambiguous pronouns, and adapting the tone of the translation to suit the target language's conventions. We tested this capability with the translation of a scientific abstract. DeepSeek-R1 not only maintains consistency in terminology but also adapts the tone to match the formal style expected in Chinese academic writing. In contrast, non-reasoning models often overuse certain phrases (e.g., Using "我们" to translate "we", which is grammatically true, but not suitable in this situation) and fail to adhere to the target language's stylistic norms. The ability of LRMs to reason about context at the document level also enables them to handle complex narrative structures, such as those found in novels or legal documents. By understanding the relationships between sentences and paragraphs, LRMs can generate translations that preserve the logical flow and coherence of the original text. This represents a significant advancement over traditional MT systems, which often struggle with long-range dependencies and contextual ambiguities.

2.3. Multi-Modal Reasoning with Translation

The integration of multi-modal inputs, such as images, videos, or audio, has expanded the capabilities of MT systems. LLMs like GPT-4o (Hurst et al., 2024) have demonstrated the ability to process and translate text in conjunction with visual or auditory context, enabling more accurate and context-aware translations. For example, when translating a sentence like "他在看报纸," an accompanying image can help the model to determine whether the subject is reading a physical newspaper or browsing a digital one. This multi-modal approach allows LLMs to resolve ambiguities that are difficult to address with text alone. LRMs further enhance multi-modal translation by incorporating reasoning capabilities into the process (Team, 2024a). Unlike traditional LLMs, which primarily rely on pattern recognition, LRMs can infer relationships between textual and non-textual inputs, enabling deeper contextual understanding. For instance, when presented with an ambiguous sentence and an accompanying image, LRMs can reason about the visual context to generate translations that align with the intended meaning. Figure 3 illustrates how an LRM leverages visual context to disambiguate translations, demonstrating their superior ability to handle complex multi-modal scenarios.

Contextual Reasoning with Image

Prompt:



请基于图片，将以下文本翻译成英文: "捕猎的是老虎"

<think> (Too long, can't provide) </think>

Output:

QvQ-72B-preview

Final answer: Yes, the translation "The animal being hunted is a tiger" is accurate based on the photograph provided.

Prompt:



请基于图片，将以下文本翻译成英文: "捕猎的是老虎"

Output:

QvQ-72B-preview

<think> (Too long, can't provide) </think>


Final Answer: It is the tiger that is hunting.

Figure 3: An example of asking the reasoning model to translate an ambiguous sentence and providing an image as the context. It shows that the model could translate correctly with the provided image.

However, challenges remain in handling specialized multi-modal inputs, such as sign language or highly domain-specific visual content. Figure 4 shows an example where OpenAI-o3-mini fails to interpret an American Sign Language gesture from an image. This highlights the limitations of current multi-modal reasoning capabilities, particularly in areas requiring fine-grained understanding of non-textual inputs. To address these challenges, future research could explore the integration of domain-specific knowledge, such as sign language dictionaries or gesture recognition algorithms, into LRMs. Additionally, advancements in multi-modal training datasets and architectures could further enhance the ability of LRMs to process and translate complex multi-modal inputs effectively.

Sign Language Translation

Prompt:



Which number does this sign show in American Sign Language?

Output:
OpenAI-o3-mini
 The hand sign shown in the image corresponds to the number "3" in American Sign Language (ASL). This is made by extending the index, middle, and thumb fingers while keeping the other fingers folded down.
QVQ-72B-Preview
 Final Answer: 4

Figure 4: An example of prompting the OpenAI GPT-3o-mini to translate American Sign Language numbers into Arabic numbers. However, the model failed to understand the image.

3. New Challenges and Opportunities with Reasoning-Enhanced MT

As LRMs introduce reasoning capabilities to MT, they also bring new challenges and opportunities. In this section, we explore several new characteristics of LRMs in MT such as self-reflection and the use of intermediate language during translation, while also identifying areas for further improvement.

3.1. Self-Reflection

One of the key advantages of LRMs is their ability to perform self-reflection during the translation process (DeepSeek-AI, 2025). This allows them to identify and correct errors, particularly in ambiguous or noisy input scenarios like when typos existed in the input, or the input sentence has been randomly rearranged into an unreadable order. For example, when translating an ambiguous Chinese sentence like "捕获的是猎人," DeepSeek-R1 initially interprets it as "The hunter is the one who hunts." However, through self-reflection, the model revisits its reasoning and considers an alternative interpretation: "The one who captures is the hunter." This iterative process demonstrates the potential of LRMs to refine translations dynamically, though further research is needed to fully understand the scope and limitations of this capability.

Self-reflection also enables LRMs to handle noisy or incomplete input more effectively. For instance, when presented with a sentence containing typos or grammatical errors, LRMs can infer the intended

meaning and generate a coherent translation. This capability is particularly valuable in real-world applications, where input quality can vary significantly. However, the effectiveness of self-reflection depends on the model's ability to accurately assess its own reasoning process, which remains an area of active research.

3.2. Auto-Pivot Translation

An interesting behavior observed in LRMs is their automatic use of a pivot or bridge language during the translation process, even without explicit instruction (Wu and Wang, 2007; Kim et al., 2019; Dabre et al., 2021). This phenomenon, which we call *auto-pivot translation*, occurs when LRMs internally reason through a high-resource language (such as English or Chinese) to translate between less-resourced languages. For example, when translating from Irish to Chinese, an LRM might first translate the Irish text into English and then generate the final Chinese translation based on the English intermediate output. This behavior is evident in the reasoning chains of LRMs, as shown in Figure 5.

The auto-pivot translation phenomenon highlights the model's reliance on high-resource languages as a scaffolding mechanism for reasoning. This approach leverages the extensive training data and linguistic knowledge available for high-resource languages, enabling LRMs to handle low-resource language pairs more effectively. However, it also raises important questions about the transparency and efficiency of the translation process. For instance, the automatic insertion of an intermediate language step may introduce additional computational overhead and potential distortions, particularly when the pivot language lacks equivalent expressions. One of the key implications of auto-pivot translation is its impact on translation quality and cultural fidelity. While the use of a pivot language can improve fluency and coherence, it may also lead to inaccuracies or loss of meaning, especially for idiomatic expressions or culturally specific terms. For example, a proverb in Irish might lose its cultural significance when first translated into English and then into Chinese. Additionally, the choice of pivot language can influence the final output, as different high-resource languages may capture different aspects of the source text's meaning.

4. New Challenges beyond Conventional Translation

In this section, we explore unique challenges that go beyond traditional text-to-text translation tasks. Specifically, we investigate the ability of LRMs to

Auto-pivot Translation

Prompt: Please translate the following Irish sentence into Chinese: Ní leor teanga amháin

Output:

DeepSeek-R1

<think> Okay, I need to translate the Irish sentence "Ní leor teanga amháin" into Chinese. Let me start by breaking down the sentence word by word.

First, "Ní" is a negative particle in Irish, so it means "not". Then "leor" translates to "enough" or "sufficient". "Teanga" means "language", and "amháin" is "one" or "single". Putting that together, the literal translation would be "Not enough language one" or more naturally, "One language is not enough."

Now, I need to make sure that the translation into Chinese is both accurate and natural. The user probably wants a concise and correct translation. The phrase "One language is not enough" emphasizes the insufficiency of a single language. In Chinese, a common way to express this might be "掌握一种语言是不够的" which translates back to "Mastering one language is not enough." However, the original Irish sentence doesn't explicitly mention "mastering"; it's more about the language itself being insufficient.

Alternatively, a more direct translation could be "一种语言不够" which is shorter and closer to the original structure. But maybe that's too terse. Another option is "只懂一种语言是不够的" meaning "Knowing only one language is not enough." This adds "knowing" which isn't in the original, but it makes the sentence more complete in Chinese.

I should check if the verb is necessary in Chinese. Sometimes in Chinese, the subject can be implied. So maybe "一种语言不够" is acceptable, but it might feel a bit abrupt. Adding a verb like "掌握" (master) or "懂" (understand/know) makes it clearer. However, since the original sentence doesn't specify the action (like master, know, etc.), perhaps the best approach is to stick with the simplest structure that conveys the meaning without adding extra words.

Wait, the original is "Ní leor teanga amháin." "Teanga amháin" is "one language", and "ní leor" is "not enough". So the structure is "[Not enough] [one language]." In Irish, the word order is different from English or Chinese. So the literal translation is "Not enough one language," which in correct English becomes "One language is not enough."

In Chinese, the equivalent structure would be "一种语言不够" (Yī zhǒng yǔyán bùgòu), which is direct and matches the original structure. However, in Chinese, it's more natural to sometimes include a verb. But in this case, since the Irish sentence doesn't have a verb, maybe the direct translation is acceptable.

Alternatively, if the context is about language proficiency or learning, adding a verb might be better. But without context, the safest translation is the direct one. Let me confirm with some examples. If I search for similar phrases, "One language is not enough" in Chinese is often translated as "只会一种语言是不够的" (Zhǐ huì yī zhǒng yǔyán shì bùgòu de), which includes "会" (know/how to). But again, that's adding a verb not present in the original.

So, to balance between faithfulness and naturalness, maybe "一种语言是不够的" (Yī zhǒng yǔyán shì bùgòu de) – "One language is not enough." This includes "是...的" structure, making it a complete sentence in Chinese. That seems accurate and natural.

Therefore, the translation should be "一种语言是不够的" or "只一种语言不够". But the former is more standard. Let me check again. The original is "Ní leor teanga amháin." "Leor" as an adjective meaning sufficient, so the structure is "Not sufficient (is) one language." In Chinese, adjectives can function as predicates without a verb, so "一种语言不够" is correct. Adding "是" and "的" makes it more emphatic.

In summary, both translations are possible, but the more natural and commonly used one would be "一种语言是不够的" to include the "是...的" structure for emphasis and completeness in Chinese. Alternatively, if brevity is preferred, "一种语言不够" works too. But considering standard translations, the former is better.

</think>
The Irish sentence "Ní leor teanga amháin" translates to "一种语言是不够的" in Chinese.

DeepSeek-V3: "一种语言是不够的。"

OpenAI-o3-mini: "只用一种语言是不够的" (Zhǐ yòng yī zhǒng yǔyán shì bùgòu de), which means "One language is not enough."

GPT-4o: "只需要一种语言" (Zhǐ xūyào yī zhǒng yǔyán). This means "One language is not enough" or "One language alone is not enough."

Figure 5: An example of translating minor language into Chinese when the Reasoning Language is English rather than Chinese, causing the intermediate translation language problem. The Reasoning model will first tries to understood the original text into English and translating it into English first, then translating it into Chinese by using the English translation generated in the thinking step.

handle encoded or ciphered text, a task that requires not only linguistic understanding but also cryptographic reasoning. This scenario tests the limits of LRMs in deciphering and translating text that has been intentionally obfuscated, providing

insights into their reasoning capabilities and limitations.

4.1. Deciphering Encoded Text

A key strength of LRMs lies in their ability to reason through complex tasks, including the deciphering of encoded text. For example, when presented with a Caesar cipher—a simple substitution cipher where each letter is shifted by a fixed number—LRMs can often deduce the shift and decode the text without explicit instructions. Our experiment successfully shows that reasoning models can decipher a Caesar-encoded text by inferring the shift value through reasoning from " Wsqixlmrk ria erh wsqixlmrk sph" to "Something new and something old".

However, the performance of LRMs degrades significantly when faced with more complex ciphers, such as the Vigenère cipher, which uses a keyword to determine the shift for each letter. In this case, the lack of a known key increases the complexity of the task exponentially. For instance, when provided with the encoded text "Mwsimpqv pm ss" (which corresponds to "Together we go" encoded with the key "TIME"), the LRM struggles to deduce the correct key and often generates hallucinated outputs. Instead of admitting uncertainty, the model may produce an incorrect key and a fabricated decoded message, such as "The key is 'KEY' and the decoded text is 'MESSAGE TO HI'." This behavior highlights a critical limitation of LRMs: their tendency to generate plausible but incorrect answers when faced with tasks beyond their reasoning capabilities.

This phenomenon underscores the challenges of applying LRMs to tasks that require not only linguistic and contextual understanding but also advanced problem-solving skills. While LRMs excel in tasks with clear reasoning pathways, their performance in highly ambiguous or computationally intensive scenarios remains inconsistent. Future research could explore methods to improve the robustness of LRMs in such tasks, such as integrating cryptographic algorithms or enhancing their ability to recognize and handle uncertainty.

5. Discussion

The exploration of LRMs in MT reveals both their transformative potential and their inherent limitations. LRMs represent a significant advancement over traditional MT systems and even LLMs, particularly in their ability to reason about context, culture, and intent. However, their performance varies across different tasks, highlighting the need for further research and refinement.

5.1. Strengths of LRMs in MT

One of the key strengths of LRMs is their **ability to handle complex translation tasks**, such as

stylized translation and document-level translation, by leveraging reasoning capabilities. For example, LRMs can preserve stylistic features and maintain coherence across long documents, tasks that have historically challenged traditional MT systems. Additionally, their self-reflection capabilities enable them to iteratively refine translations, improving accuracy in ambiguous or noisy scenarios. These advancements position LRMs as powerful tools for applications such as low-resource language translation, interactive translation, and multi-modal translation.

5.2. Limitations and Challenges

Despite their strengths, LRMs face several limitations that hinder their widespread adoption. One major challenge is their performance in highly complex reasoning tasks, such as deciphering Vigenère ciphers without a known key. In such cases, LRMs often generate hallucinated answers rather than admitting uncertainty, highlighting a critical area for improvement. Similarly, the use of intermediate languages in translation, while beneficial for leveraging high-resource language knowledge, can introduce biases or inaccuracies, particularly when translating between less-resourced languages.

Another challenge lies in the integration of multi-modal inputs. While LRMs show promise in leveraging visual or auditory context to disambiguate translations, their performance in specialized domains, such as sign language interpretation, remains limited. This suggests the need for domain-specific training and the integration of external knowledge sources to enhance their capabilities.

5.3. Inference Efficiency and Long Chain-of-Thought Reasoning

A significant practical challenge for LRMs is their inference efficiency, particularly due to the generation of long CoT reasoning steps. While CoT reasoning enables LRMs to tackle complex tasks by breaking them down into interpretable sub-problems, it also increases computational overhead and latency (Xia et al., 2025). For example, in tasks like document-level translation or deciphering encoded text, LRMs may generate extensive reasoning chains to arrive at a solution, which can slow down inference and increase resource consumption.

This inefficiency poses a barrier to real-time applications of LRMs, such as interactive translation or live multi-modal translation. To address this issue, future research could explore methods to optimize CoT generation, such as: 1) **Pruning Redundant Reasoning Steps**: Identifying and eliminating unnecessary or repetitive reasoning steps to streamline the inference process. 2) **Model**

Method	Lexical		Contextless		Contextual	
	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT
DeepSeek-R1	84.3	73.9	84.7	73.9	84.0	73.3
DeepSeek-V3	84.7	74.2	84.4	74.1	84.1	73.2
QwQ-32B	84.1	73.0	84.0	72.8	84.1	72.9
GPT-4o	84.8	74.1	84.6	73.7	85.0	74.9
o3-mini	84.8	73.8	84.7	74.0	83.6	73.2
Gemini 2.0 Flash	83.2	72.7	84.1	72.7	71.0	72.8

Table 1: Result of commonsense translation performance on commonMT (He et al., 2020).

Compression: Applying techniques like quantization or distillation to reduce the computational load of LRMs without significantly compromising performance. Improving inference efficiency will be crucial for scaling LRMs to real-world applications, where speed and resource constraints are critical considerations.

5.4. Future Directions

To summarise, while LRMs represent a significant step forward in MT, their full potential has yet to be realized. Future research should focus on improving their robustness in complex reasoning tasks, enhancing their ability to handle uncertainty, and expanding their capabilities in specialized domains. Additionally, addressing the inference efficiency problem will be essential for enabling real-time and resource-efficient applications of LRMs. By tackling these challenges, LRMs can further redefine the boundaries of MT and enable new applications in cross-cultural communication and beyond.

6. Experiment with CommonMT - Comparing models with and without Reasoning

This section presents an experimental result and analysis the performances of various LLMs with or without reasoning ability on translating Chinese-to-English sentence with commonsense understanding (He et al., 2020). These LLMs are evaluated on BLEURT and COMET.

As shown in Table 1, we could not see significant difference between the results generated by the four models we chose for the experiment under automatic evaluation metrics. However, when we examine the MT result of each model and compare with them, we see that in some cases, a model might get a lower comet score even if the MT result is correct, but using different words which is different than the reference translation. For example, for sentence "正在采收的是果园里的果农," the reference translation is "The orchard worker in the orchard is harvesting." DeepSeek-R1 translated it as "The orchard farmers are harvesting" which

received a COMET score of 0.7748, and the translation generated by DeepSeek-V3 is "The orchard farmers are currently harvesting the fruits" which received a COMET score of 0.8039. We could see that DeepSeek-R1 generated a probably better translation than DeepSeek-V3, but the score of it is actually lower than the other model. We believe that this happens because COMET and BLEURT require a reference translation as a standard, and any translation which is close to the reference will receive a higher score. However, reasoning models could generate more diverse translations, which could be different than the reference translation, thus receiving a lower score under the metrics requiring a reference. To better scoring this situation, new automatic scoring metric are needed in the future to solve such problem.

7. Conclusion

In this paper, we have explored the transformative potential of LRMs in the field of MT. By leveraging reasoning capabilities, LRMs can tackle long-standing challenges in MT, such as stylized translation, document-level translation, and multi-modal translation, while also introducing new capabilities like self-reflection and auto-pivot language translation. However, our findings also highlight the limitations of LRMs, particularly in complex reasoning tasks and specialized domains. For example, while LRMs can decipher simple ciphers, they struggle with more complex cryptographic challenges and may generate hallucinated answers when faced with uncertainty. Similarly, their performance in multi-modal translation, such as interpreting sign language, remains limited, underscoring the need for further advancements in domain-specific reasoning. These insights provide promising directions for future research in LRM-based MT. The main areas worth for exploration include improving the robustness of LRMs in ambiguous or computationally intensive tasks, enhancing their ability to handle uncertainty, and expanding their capabilities in specialized domains. By addressing these challenges, LRMs can further redefine the boundaries of MT and enable new applications in cross-cultural

communication and beyond. In conclusion, LRMs represent a paradigm shift in MT, transforming systems from mere text converters into multilingual cognitive agents capable of reasoning about meaning beyond the text. This evolution urges us to re-think translation not just as a linguistic task but as a gateway to deeper cross-cultural understanding.

8. Bibliographical References

- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovénia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*.
- Andong Chen, Yuchen Song, Wenxin Zhu, Kehai Chen, Muyun Yang, Tiejun Zhao, et al. 2025. Evaluating o1-like llms: Unlocking reasoning for translation through comprehensive analysis. *arXiv preprint arXiv:2502.11544*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Raj Dabre, Aizhan Imankulova, Masahiro Kaneko, and Abhisek Chakrabarty. 2021. [Simultaneous multi-pivot neural machine translation](#).
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#).
- Oluwaseyi Feyisetan, Sepideh Ghanavati, Patricia Thaine, Ivan Habernal, and Fatemehsadat Mireshghallah, editors. 2022. [Proceedings of the Fourth Workshop on Privacy in Natural Language Processing](#). Association for Computational Linguistics, Seattle, United States.
- Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. [Survey of low-resource machine translation](#). *Computational Linguistics*, 48(3):673–732.
- Muhammad Usman Hadi, Qasem Al-Tashi, Rizwan Qureshi, Abbas Shah, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Mohammed Ali Al-Garadi, et al. 2023. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects.
- 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.
- Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *arXiv preprint arXiv:2305.04118*.
- Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. *arXiv preprint arXiv:2401.12873*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.
- Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. [Pivot-based transfer learning for neural machine translation between non-English languages](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computational Linguistics.

- Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. [Differential privacy in natural language processing the story so far](#). In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11, Seattle, United States. Association for Computational Linguistics.
- 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.
- Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.
- Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. [A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.
- Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. [Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation](#). *arXiv preprint arXiv:2305.15011*.
- Yansong Li, Zhixing Tan, and Yang Liu. 2023b. [Privacy-preserving prompt tuning for large language model services](#).
- Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. [A paradigm shift: The future of machine translation lies with large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.
- Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. [Adaptive machine translation with large language models](#). *arXiv preprint arXiv:2301.13294*.
- Xing Niu and Marine Carpuat. 2020. [Controlling neural machine translation formality with synthetic supervision](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8568–8575.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.
- Jianhui Pang, Fanghua Ye, Longyue Wang, Dian Yu, Derek F Wong, Shuming Shi, and Zhaopeng Tu. 2024. [Salute the classic: Revisiting challenges of machine translation in the age of large language models](#). *arXiv preprint arXiv:2401.08350*.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. [Qwen2.5 technical report](#).
- Nathaniel R Robinson, Perez Ogayo, David R Mortensen, and Graham Neubig. 2023. [Chatgpt mt: Competitive for high-\(but not low-\) resource languages](#). *arXiv preprint arXiv:2309.07423*.
- Satoshi Sato and Makoto Nagao. 1990. [Toward memory-based translation](#). In *13th International Conference on Computational Linguistics, COLING 1990, University of Helsinki, Finland, August 20-25, 1990*, pages 247–252.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Controlling politeness in neural machine translation via side constraints](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.
- Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.
- Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia,

- and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*, 34:97–147.
- Qwen Team. 2024a. [Qvq: To see the world with wisdom](#).
- Qwen Team. 2024b. [Qwq: Reflect deeply on the boundaries of the unknown](#).
- 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. 2023. [Llama: Open and efficient foundation language models](#).
- Jun'ichi Tsujii. 1986. Future directions of machine translation. In *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2024a. [Drt-o1: Optimized deep reasoning translation via long chain-of-thought](#). *arXiv preprint arXiv:2412.17498*.
- Longyue Wang. 2019. *Discourse-aware neural machine translation*. Ph.D. thesis, Dublin City University. School of Computing.
- Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, et al. 2024b. [Findings of the wmt 2024 shared task on discourse-level literary translation](#). *arXiv preprint arXiv:2412.11732*.
- Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. [Document-level machine translation with large language models](#). *arXiv preprint arXiv:2304.02210*.
- Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2022. [Controlling styles in neural machine translation with activation prompt](#). *arXiv preprint arXiv:2212.08909*.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022a. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. [Chain-of-thought prompting elicits reasoning in large language models](#). *Advances in neural information processing systems*, 35:24824–24837.
- Hua Wu and Haifeng Wang. 2007. [Pivot language approach for phrase-based statistical machine translation](#). *Machine Translation*, 21:165–181.
- Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. 2025. [Tokenskip: Controllable chain-of-thought compression in llms](#).
- Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. 2020. [A survey of deep learning techniques for neural machine translation](#).
- Hongbin Zhang, Kehai Chen, Xuefeng Bai, Xiucheng Li, Yang Xiang, and Min Zhang. 2025. [Exploring translation mechanism of large language models](#).
- Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. [Marco-o1: Towards open reasoning models for open-ended solutions](#). *arXiv preprint arXiv:2411.14405*.