A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models

Chenyang Lyu[1†], Zefeng Du[2†], Jitao Xu[3†], Yitao Duan[3], Minghao Wu[4], Teresa Lynn[1], Alham Fikri Aji[1], Derek F. Wong[2], Siyou Liu[2], Longyue Wang[5]

1. MBZUAI
2. University of Macau
3. NetEase Youdao
4. Monash University
5. Tencent AI Lab

chenyang.lyu@mbzuai.ac.ae, mc05583@umac.mo, xujt01@rd.netease.com, duan@rd.netease.com, vinnlywang@tencent.com

Abstract

Machine Translation (MT) has greatly advanced over the years due to the developments in deep neural networks. However, the emergence of Large Language Models (LLMs) like GPT-4 and ChatGPT is introducing a new phase in the MT domain. In this context, we believe that the future of MT is intricately tied to the capabilities of LLMs. These models not only offer vast linguistic understandings but also bring innovative methodologies, such as prompt-based techniques, that have the potential to further elevate MT. In this paper, we provide an overview of the significant enhancements in MT that are influenced by LLMs and advocate for their pivotal role in upcoming MT research and implementations. We highlight several new MT directions, emphasizing the benefits of LLMs in scenarios such as Long-Document Translation, Stylized Translation, and Interactive Translation. Additionally, we address the important concern of privacy in LLM-driven MT and suggest essential privacy-preserving strategies. By showcasing practical instances, we aim to demonstrate the advantages that LLMs offer, particularly in tasks like translating extended documents. We conclude by emphasizing the critical role of LLMs in guiding the future evolution of MT and offer a roadmap for future exploration in the sector.

Keywords: Large Language Models, Machine Translation, New Trends

1. Introduction

Machine Translation (MT) is a fundamental task in Natural Language Processing (NLP) that aims to automatically translate texts from one language to another (Tsujii, 1986; Sato and Nagao, 1990). The performance and quality of MT systems have been significantly advanced from Statistical Machine Translation (SMT) (Zens et al., 2002; Koehn et al., 2007) to Neural Machine Translation (NMT) (Cho et al., 2014; Bahdanau et al., 2015) with the employment of machine learning techniques (Vaswani et al., 2017; Castilho et al., 2017; Stahlberg, 2020; Kocmi et al., 2022). Despite these advancements achieved within decades of research, MT still faces many challenges, such as dealing with idiomatic expressions, low-resource translation, handling rare words, and maintaining coherence and fluency in the translation (Koehn and Knowles, 2017; Wang, 2019; Yang et al., 2020; Haddow et al., 2022a).

Recently, the emergence of Large Language Models (LLMs), such as GPT-3 and GPT-4 (Brown et al., 2020; Chen et al., 2021; Ouyang et al., 2022; Wei et al., 2022; Hadi et al., 2023), has substantially reshaped the paradigm for MT from multiple dimensions. The zero-shot performance of LLMs on translation is even on par with strong fully supervised MT systems (Jiao et al., 2023b; Robinson et al., 2023; Moslem et al., 2023; Pang et al., 2024). More importantly, LLMs can also be used in various scenarios beyond MT such as question answering and style transfer (Bang et al., 2023; Laskar et al., 2023; Li et al., 2023a), which enables novel scenarios and provide rooms for exploration for MT. Besides these opportunities, LLMs-based MT poses new challenges such as privacy-related issues that require new directions and methodologies to be addressed.

In this paper, with the aim of extending the scope of MT with the incorporation of the superior capability of LLMs, we present discussions on several potentially interesting and promising novel directions for LLMs-based MT, including challenging translation scenarios such as long-document translation and stylised translation, interactive translation, and Translation Memory (TM) based MT, potential new evaluation paradigms of translation quality using LLMs, as well as privacy concerns for LLMs-based MT and some other interesting directions.

Among the aforementioned directions, challenging translation scenarios include long-document translation (Maruf et al., 2021) which requires the translation of long documents with thousands of

† Equal contribution.

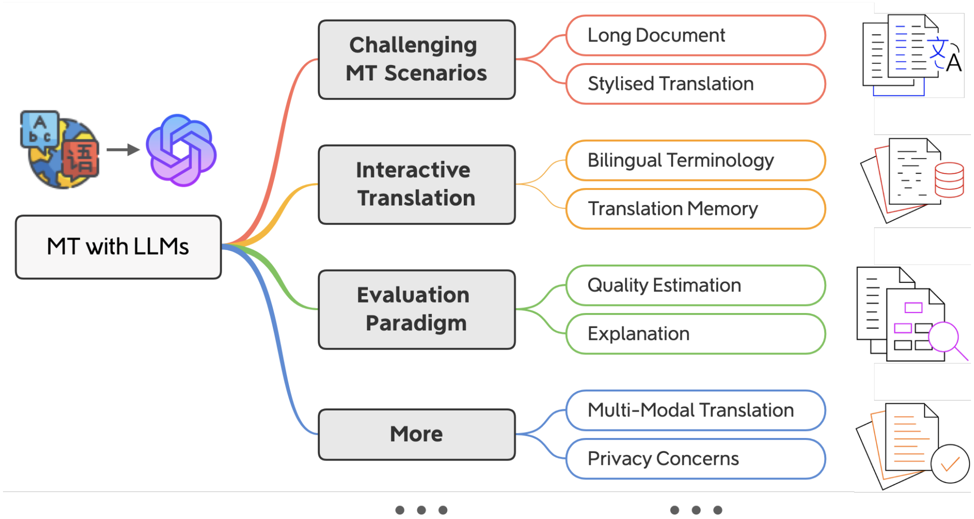

Figure 1: Interesting directions for MT using LLMs (e.g. GPT models), including challenging MT scenarios, interactive MT, new evaluation paradigm for MT using LLMs, etc.

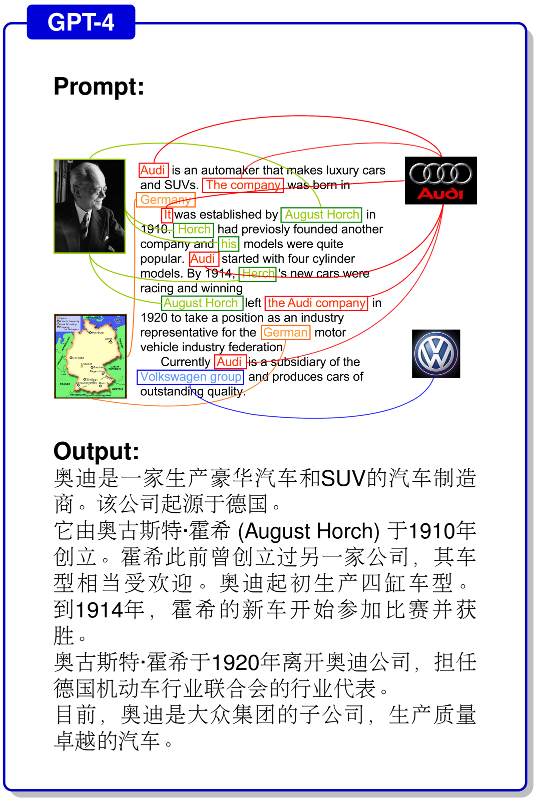

Figure 2: An example of translating a document-level text from English to Chinese using GPT-4. We highlight the discourse phenomena using figures and lines, which are invisible to GPT-4.

words or even longer and stylised translation that aims to preserve the stylistic features of the source text or inject specific language styles in the translation output, such as the tone, register, formality, genre, etc (Sennrich et al., 2016; Niu and Carpuat, 2020). Interactive translation (Knowles and Koehn, 2016; Santy et al., 2019) aims to facilitate the collaboration and feedback between human translators and MT systems, such as through chatbots or question-answering systems. TM-based MT (Bulte and Tezcan, 2019; Xu et al., 2020) tends to make use of similar or relevant translations to improve the quality of the translation output, which is increasingly important in the era of LLMs with the use of in-context learning that can inform LLMs the necessary knowledge in the demonstration examples (Moslem et al., 2023). Moreover, we explore multi-modal translation with LLMs which translates source texts with images as extra context (Sulubacak et al., 2020; Yao and Wan, 2020). Furthermore, we discuss the potential new evaluation paradigm for LLMs-based MT that aims for a more accurate and efficient evaluation of MT systems from various aspects instead of only evaluating the similarity between system outputs and references (Kocmi and Federmann, 2023; Liu et al., 2023). Besides the new directions and methodologies, we also discuss the privacy concerns in MT using LLMs and propose basic privacy-preserving methods to mitigate the risks. Privacy in NLP and LLMs is becoming increasingly important (Klymenko et al., 2022; Feyisetan et al., 2022; Li et al., 2023b), as LLMs may inadvertently reveal sensitive information in the source text or the translation output when using LLMs for translation. In addition, we present discussions on some other new scenarios for LLMs-based MT such as personalized MT and low-resource language translation using LLMs.

To preliminarily investigate the feasibility of the interesting directions mentioned above, we present some empirical evidence in Section 8 and examples using various LLMs such as LLaMA and GPT-4 for MT under various scenarios, demonstrating the feasibility of the directions. This position paper demonstrates the potentials of the prospective new directions and methodologies for enhancing the quality and diversity of MT output, as well as the importance and challenges of privacy-preserving in MT using LLMs. We conclude by highlighting the opportunities and challenges for future research in MT using LLMs and suggesting potential directions for further exploration.

2. Challenging MT Scenarios

2.1. Long-Document Translation

The majority of MT applications have traditionally concentrated on sentence-level translation (Post and Junczys-Dowmunt, 2023), which can sometimes lead to translations that are devoid of context and coherence when translating long-documents that might contain thousands of words with complex structure. Recent years have seen a growing interest in document-level translation, a task of critical importance that involves the translation of entire documents, but also presents unique challenges (Wang et al., 2017; Voita et al., 2019; Wang et al., 2018; Zhang et al., 2022; Jiang et al., 2022; Wu et al., 2023; Wang et al., 2023a; Wu et al., 2024b). Surprisingly, LLMs have shown potentials in modeling exceptionally long texts with complex discourse structures, suggesting that they could be instrumental in advancing the field of document-level translation (Wang et al., 2023a; He et al., 2023; Wu et al., 2024a). Figure 2 illustrates an example of GPT-4 translating a document, where the discourse phenomena such as pronouns must be well translated in order to maintain the document structure. Furthermore, with the introduction of larger context window for LLMs (Tworkowski et al., 2023; Xiong et al., 2023), the translation of longer documents such as novels and books with LLMs become more and more feasible.

2.2. Stylised Translation

Stylised Translation refers to the ability of generating translations that match a specific style or genre (Wang et al., 2022b), such as formal or informal expression (Sennrich et al., 2016), poetry or prose, different dialects or registers, etc. This can be achieved by training MT systems on multi-parallel data that contain translations in different styles or genres, or by using style transfer techniques (Yang et al., 2018) that can transform a given translation into a desired style. Stylised Translation has many potential applications, such as in marketing, literature, or cultural preservation.

However, Stylised Translation is difficult to achieve before the presence of LLMs as there lacks such multi-parallel corpora for Stylised Translation to fit various styles while the zero-shot ability of LLMs makes these tasks seamlessly achievable. It can be achieved by directly prompting LLMs to translate the text with a specific style expressed by natural language or firstly let LLMs translate the original text and then stylise the translation output. We present an example of translating an introduction for the Olympic Games from Wikipedia from English to Chinese while following a poetic style in Figure 3. This example shows GPT-4 can handle translation with a poetic style while keeping the semantic information of the original text, which can be hardly achieved by traditional MT systems.

Nevertheless, Stylised Translation also presents a variety of obstacles. Among these challenges, one notable issue is to determine the best approach to systematically define and quantify various styles or genres on a larger scale. Another challenge is how to evaluate the quality of Stylised Translation, as traditional evaluation metrics may not be sufficient to capture the diversity of stylistic variations. Overcoming these challenges requires interdisciplinary collaboration between linguists, literary scholars, and computer scientists.

3. Interactive Translation

3.1. Bilingual Terminology

Interactive Translation (Santy et al., 2019; Jiao et al., 2023a) allows users to actively participate in the translation process, either by correcting or refining automatic translations or by providing feedback on the translation quality. This can be achieved by integrating MT systems based on LLMs that has superior conversational capability for language understanding and generation with interactive user interfaces, such as chatbots, that allow users to engage with the translation process in real-time to provide feedback and more specific requirements such as specific translations of certain terminologies. Interactive Translation can help to improve the accuracy and fluency of the translations, especially in cases where the source language is ambiguous or the domain knowledge is limited.

However, interactive MT also raises several challenges. One challenge is how to design user interfaces that are intuitive, user-friendly, informative and flexible. Another challenge is how to incorporate user feedback into the translation process in a principled and effective way. Overcoming these challenges requires insights from human-computer interaction, NLP, and user experience design. Figure 4 illustrates an example of prompting GPT-4 in an interactive way for MT, where the specific translation requirement of named entities is provided.

Previous studies on conventional TM-based MT has also shown that conventional Transformer-based NMT system already shows the ability to make use of new TMs that the model has never

GPT-4

Prompt: Please translated the text below to Chinese following poetic style: "The Ancient Olympic Games were a religious and athletic festival that took place every four years at the sanctuary of Zeus in Olympia, Greece. Participants came from various city-states and kingdoms of Ancient Greece, competing in mainly athletic sports, but also combat sports such as wrestling and the pankration, as well as horse and chariot racing events. It is believed that during the Games, all conflicts among participating city-states were suspended until the end of the festival, which was known as the Olympic peace or truce. However, this idea is considered a modern myth as the Greeks never actually suspended their wars. Nevertheless, the truce did provide safe passage for religious pilgrims traveling to Olympia, who were protected by Zeus."

Output:
古奥林匹克运动会，四年一度，希腊奥林匹亚，
宙斯圣地，宗教盛典，竞技同场。
诸城邦国，齐聚一堂，力争上游，
田径角逐，搏击角力，马车竞速。
奥运和平，节庆期间，战争暂停，
虽为现代，神话传说，实则不然。
宙斯庇佑，朝圣者行，安全抵达，
奥林匹克，和平永驻，共祈祝愿。

Figure 3: An example of prompting GPT-4 to translate texts from English to Chinese using poetic style. GPT-4 can generate translations in a poetic style, and its format is also more in line with poetry, while the semantic information of the original English text is also preserved in the output translation.

seen during training to largely improve domain-specific translation during inference (Xu et al., 2020, 2022). This indicates that conventional NMT systems learn to understand the relationship between a given source sentence and a similar translation and to select useful information from the given similar translation, rather than remember sentences seen during training. This ability is, to some extent, similar to the ICL ability of LLMs. However, to the best of our knowledge, there does not exist research works focusing on finding the relationships between these two abilities.

3.2. Translation Memory-based MT

TM has been used for decades to help human translators in basic Computer-Aided Translation systems. The general process of using TM in MT is, for a sentence to be translated, to first search for similar translations in TM using, for instance, fuzzy matching techniques, then revised or edit

GPT-4

Prompt: Please translate the text below to Chinese: "Jordan played college basketball for three seasons under coach Dean Smith with the North Carolina Tar Heels. As a freshman, he was a member of the Tar Heels' national championship team in 1982. Jordan joined the Bulls in 1984 as the third overall draft pick, and quickly emerged as a league star, entertaining crowds with his prolific scoring while gaining a reputation as one of the game's best defensive players. His leaping ability, demonstrated by performing slam dunks from the freethrow line in Slam Dunk Contests, earned him the nicknames "Air Jordan" and "His Airness"."

Output: 乔丹在北卡罗来纳州塔尔希尔队的教练迪恩·史密斯指导下，为期三个赛季的大学篮球生涯。作为一名新生，他在1982年成为了塔尔希尔队全国冠军团队的一员。1984年，乔丹作为第三顺位选秀被公牛队选中，并迅速成为联盟明星，以高效得分娱乐观众，同时也被认为是比赛中最好的防守球员之一。他的跳跃能力，通过在扣篮大赛中从罚球线起跳完成扣篮而展示出来，为他赢得了“飞人乔丹”和“空中魔术师”的绰号。

Prompt:Please translate it to Chinese while keep the people's name (such as Jordan) and organisation's name (such as Chicago Bull) in English.

Output:乔丹在北卡罗来纳州Tar Heels队的教练Dean Smith指导下，为期三个赛季的大学篮球生涯。作为一名新生，他在1982年成为了Tar Heels队全国冠军团队的一员。1984年，乔丹作为第三顺位选秀被Chicago Bulls队选中，并迅速成为联盟明星，以高效得分娱乐观众，同时也被认为是比赛中最好的防守球员之一。他的跳跃能力，通过在扣篮大赛中从罚球线起跳完成扣篮而展示出来，为他赢得了“Air Jordan”和“His Airness”的绰号。

Figure 4: An example of using GPT-4 in an interactive way for MT. GPT-4 can adjust the translation results according to the user's input. Named entities (in red) ajusted by GPT-4 are marked in blue.

the retrieved similar translation in order to obtain a high-quality translation. TM-based MT has already been integrated into conventional NMT systems (Bulte and Tezcan, 2019; Xu et al., 2020; Cai et al., 2021). Since LLMs have emerged with the In-Context Learning (ICL) ability that they can learn specific tasks through task examples given in the prompt. The use of retrieved similar or relevant sentence pairs (Pham et al., 2020) seems to be a natural fit to few-shot prompting techniques

GPT-4

Prompt: Please translate the following sentence into French. Here is a similar translation you can use:
En cas d'atteinte sévère ou si le patient est dans l'incapacité d'avaler les comprimés (par ex. patients alimentés par sonde), il est recommandé de débuter le traitement par une administration intraveineuse de ciprofloxacine jusqu'à ce qu'un relais par voie orale soit possible.

In severe cases or if the patient is unable to take oral suspension (e.g- patients on enteral nutrition), it is recommended to commence therapy with intravenous ciprofloxacin until a switch to oral administration is possible.

Output *without TM*: Dans les cas graves ou si le patient ne peut pas prendre de suspension orale (par exemple, les patients sous nutrition entérale), il est recommandé de commencer le traitement par ciprofloxacine intraveineuse jusqu'à ce qu'un passage à l'administration orale soit possible.

Output: Dans les cas graves ou si le patient est incapable de prendre une suspension orale (par ex. les patients en nutrition entérale), il est recommandé de commencer le traitement par une administration intraveineuse de ciprofloxacine jusqu'à ce qu'un passage à l'administration orale soit possible.

Figure 5: An example of prompting GPT-4 to translate text from English to French using a similar translation retrieved from TM. Benefits directly taken by GPT-4 from the given similar translation is marked in blue, in contrast to the translation in red without using TMs.

when performing MT using LLMs (Vilar et al., 2022; Moslem et al., 2023).

However, existing works so far have mostly used randomly selected translation examples as prompts and suggest that using semantically similar examples does not significantly further improve the translation performance (Vilar et al., 2022; Zhu et al., 2023). Most of these works used sentence-level embedding built by an external language model to retrieve similar examples via an embedding similarity search. On the contrary, other studies using lexical fuzzy matches to retrieve similar translations have shown significant improvements (Moslem et al., 2023). Therefore, the conclusion about the effectiveness of using similar translations in MT using LLMs still remains unclear. Since TMs can provide useful domain and style information that can directly help LLMs to generate translations that better meet the translation requirement, it is a promising direction to further study how to better integrate TMs into LLMs for MT. Fig-

GPT-4-Vision

Prompt:

Please translated the text below to English while considering the context of the image: "欢乐的春节舞龙巡游，充满了喜庆和活力"

Output: The text translates to:
"Joyful Spring Festival loong dance parade, full of festivity and vitality."

Figure 6: An example of propmpting GPT-4-Vision for translating a sentence in Chinese to English while considering an associated image which shows a scene of celebrating new year.

GPT-4-Vision

Prompt:

Please translate the text in the image that I saw then driving a car into English:

Output: The text in the image contains Chinese characters. Here are the translations:

"前方施工" translates to "Construction ahead."
"车辆绕行" translates to "Vehicles detour."
"禁止停车" translates to "No parking."
"土石工程" translates to "Earthwork project."

Figure 7: An example of a Chinese roadside construction sign translated by GPT-4-Vision: highlighting warnings for upcoming construction, a detour for vehicles, and a no-parking zone.

GPT-4

Prompt: Please translate this tweet into Chinese: "For the next 12 months, Twitter will keep none of the money.
You will receive whatever money we receive, so that'$ 70% for subscriptions on ios & Android (they charge 30%) and -92% on the web (could be better, depending on payment processor)
After the first year. ios & Android fees drop to 15% and we will add a small amount on top of that, depending on volume.
We will also help promote your work. Our goal is to maximize creator prosperity.
At any point, you can leave our platform and take your work with you. Easy in, easy out."

Output: 在接下来的12个月里，Twitter不会保留任何钱。
我们收到的钱你们将全部获得，所以在iOS和Android上的订阅收入是70%（它们收取30%的费用），在网页上是92%（取决于支付处理器，可能会更好）
第一年过后，iOS和Android的费用降低到15%，我们会在此基础上增加一小部分，具体取决于交易量。
我们还将帮助推广你的作品。我们的目标是最大化创作者的繁荣。
在任何时候，你都可以离开我们的平台并带走你的作品。轻松进入，轻松退出。

Prompt: How good is the translation above with some specific aspects you can think of?

Output: The translation above is generally accurate and conveys the main ideas of the original tweet. However, there are a few aspects that could be improved:
1. Consistency in terms: In the original tweet, "money" is used consistently, while in the translation, both "钱" and "收入" are used. It would be better to stick with one term, such as "收入" for consistency.
2. Clarity: The part "-92% on web" in the original tweet might be a typo and should be "92% on web." The translation follows the original text, but it could cause confusion for readers. It would be better to clarify this point in the translation.
3. Punctuation: The original tweet uses a mix of punctuation marks, such as "$" and "%." In the translation, it would be better to use standard Chinese punctuation marks, like "%" for percentages.
4. Formatting: The translation is presented as a single paragraph, while the original tweet has multiple lines and breaks. To improve readability, the translation could be formatted similarly to the original tweet.
Overall, the translation is good but could benefit from some minor adjustments to improve consistency, clarity, punctuation, and formatting.

Figure 8: An example of using GPT-4 for evaluating translation output. We let GPT-4 generate a translation text, and then ask it to introduce how to evaluate a translation text. GPT-4 provides standards and evaluates the translation text accordingly.

GPT-4

Prompt: Please translate this text into Chinese: "John Smith, the CEO of Boeing Corporation, has offered to pay $10 million for the acquisition of Alphabet Corporation. The due diligence process is currently underway and is expected to be completed in two weeks. Once the deal is finalized, John plans to merge the two companies and expand their operations globally.

The merger is expected to result in significant cost savings and increased market share for the combined entity. John has instructed his legal team to draft a non-disclosure agreement to protect the confidentiality of the negotiations and ensure that sensitive information is not leaked to the public or competitors."

Output: 约翰·史密斯（John Smith），波音公司（Boeing Corporation）的首席执行官，已提议支付1000万美元收购Alphabet Corporation。尽职调查过程目前正在进行中，预计将在两周内完成。一旦交易敲定，约翰计划将两家公司合并，并将其业务扩展到全球。

这次合并预计将为合并后的实体带来显著的成本节约和市场份额的增长。约翰已指示他的法律团队起草一份保密协议，以保护谈判的机密性，并确保敏感信息不会泄露给公众或竞争对手。

Prompt: Please translate this text into Chinese: "#P1#, the CEO of #ORG1#, has offered to pay $#NUM1# million for the acquisition of #ORG2#. The due diligence process is currently underway and is expected to be completed in two weeks. Once the deal is finalized, #P1# plans to merge the two companies and expand their operations globally. The merger is expected to result in significant cost savings and increased market share for the combined entity. #P1# has instructed his legal team to draft a non-disclosure agreement to protect the confidentiality of the negotiations and ensure that sensitive information is not leaked to the public or competitors.

Output: #P1#，#ORG1#的首席执行官，已提议支付#NUM1#百万美元收购#ORG2#。尽职调查过程目前正在进行中，预计将在两周内完成。一旦交易达成，#P1#计划将两家公司合并并将其业务扩展至全球。预计合并将为合并后的实体带来显著的成本节约和市场份额增长。#P1#已指示其法律团队起草一份保密协议，以保护谈判的保密性，确保敏感信息不会泄露给公众或竞争对手。

Figure 9: An example of privacy issue using GPT-4 for MT. The above one is the example where the input is not anonymized, thus containing name information, business data, etc (in red). The bottom one is the example where the sensitive information in the input is anonymized (in blue).

ure 5 illustrates an example of prompting LLM both without and with TMs where using TMs directly improves the translation quality.

4. Multi-modal Translation

Another promising direction is multi-modal MT (Yao and Wan, 2020; Sulubacak et al., 2020), which involves integrating visual, audio, or other non-textual information into the translation process. This approach can enhance the quality and accuracy of translations in various settings, such as image or video captioning, automatic speech recognition, and sign language translation. LLMs, such as GPT-4 (OpenAI, 2023; Wang et al., 2023b), can be employed to develop models that can learn from multi-modal data and generate translations that accurately convey the meaning of the input. We demonstrate an example of multimodal translation using GPT-4-Vision [1] in Figure 6 in which the context of image must be taken into consideration when translating the sentence.

However, multi-modal MT poses several challenges, such as data heterogeneity, unbalanced datasets, and domain specificity. Overcoming these challenges would require developing novel algorithms that can learn from multi-modal data and generalize well across different modalities and domains. Leveraging the multilingual translation prowess of LLMs and combining them with models of diverse modalities unlocks the potential for remarkable applications. For instance, LLMs can be employed for video localization purposes. This tool's primary objective is to seamlessly translate video content into a desired target language while simultaneously replicating the video creator's voice using voice cloning technology for narration. Such an approach is perfectly suited for global product promotions, enabling the creation of a single video that can be effortlessly transcribed into multiple languages, catering to audiences across the world.

5. New Evaluation Paradigm for LLMs-based MT

Evaluating the quality of LLMs-based MT is a challenging task, as existing evaluation metrics may not be sufficient to capture the full range of translation quality (White and O'Connell, 1993; Isabelle et al., 2017). In addition, existing open-access test sets may suffer from the data contamination problem as they are possibly used during the training process of LLMs (Bang et al., 2023). Evaluating on these test sets cannot correctly reflect the MT performance of LLMs. A new evaluation paradigm for LLMs-based MT should take into

[1] https://openai.com/research/gpt-4v-system-card

GPT-4

Prompt:
Please translate the following text to English with emphasis on molecular biology terminology, considering the user's background and their fluency in French:
"La protéine XYZ joue un rôle crucial dans la signalisation cellulaire des plantes."

Output:
The XYZ protein plays a pivotal role in plant cellular signaling pathways.

Figure 10: An illustration of personalized MT, adapting to user's domain-specific knowledge and language proficiency.

account the unique characteristics of LLM-based MT, such as the ability to generate fluent but inaccurate translations or the sensitivity to domain-specific knowledge. Possible approaches to a new evaluation paradigm include using specifically-designed human evaluations (Graham et al., 2020; Ji et al., 2022) for such systems, or even directly employ LLMs to evaluate the translation output from LLMs (Kocmi and Federmann, 2023) - although studies show that LLMs would prefer the translation output from LLMs instead of other systems (Liu et al., 2023). Besides, using extrinsic evaluation is also feasible - the translation output can be used in other tasks and measure the corresponding performance instead of directly assessing the translation quality (Moghe et al., 2023).

However, developing a new evaluation paradigm also poses several challenges. One challenge is how to balance the trade-off between evaluation efficiency and evaluation quality, as human evaluations can be time-consuming and expensive, and LLM-based evaluation can be biased. Another challenge is how to ensure the reliability and validity of the evaluation results, as different evaluators may have different subjective judgments or biases. An example of using GPT-4 to evaluate the translation output for a tweet from Elon Musk is shown in Figure 8. Although GPT-4 can analyze the text based on the standards it lists, there is a certain hallucination phenomenon, which means pointing out errors that do not exist in the translation text. Overcoming these challenges requires rigorous experimental design, statistical analysis, and transparency in reporting.

6. Privacy in MT using LLMs

As LLMs become more powerful and widely used in MT, there are growing concerns about privacy and

GPT-4

Prompt:
Translate the following Irish text to English:
"Bhí mé ag siúl sa choill nuair a chuala mé ceol álainn ón adharc."

Output (With LLM Integration):
I was walking in the forest when I heard beautiful music from the horn.
Output (Traditional MT):
I was in the wood when I heard nice sound from the horn.

Figure 11: Comparison of LLM-based MT with traditional MT for translating a nuanced Irish sentence, illustrating the finesse of LLMs in capturing intricate linguistic details.

security (Xie et al., 2023). In particular, LLMs may inadvertently reveal sensitive information in the source text or the translation output, such as personally identifiable information, confidential business data, or political opinions. Privacy in MT using LLMs aims to mitigate these risks by developing privacy-preserving methods that can protect the confidentiality and integrity of the translation process. One basic approach to preserve privacy in MT using LLMs is to anonymize sensitive information in the textual input and then pass it to LLMs and get the output, which is then de-anonymized. An example of such an issue using GPT-4 is shown in Figure 9. This is similar to methods integrating terminologies or user dictionaries into conventional NMT systems (Crego et al., 2016).

However, privacy-preserving methods in LLMs-based MT also pose several challenges. One challenge is how to balance the trade-off between privacy and accuracy, as privacy-preserving methods may introduce additional noise or distortion to the translation output (Dinu et al., 2019). Another challenge is how to ensure the interoperability and compatibility of privacy-preserving methods across different languages, models, and platforms. Overcoming these challenges requires collaboration between experts in cryptography, privacy, and MT, as well as adherence to ethical and legal standards.

7. Discussion

Personalized MT (Mirkin and Meunier, 2015; Rabinovich et al., 2017) - With the advancements in LLM-based MT, the focus can be shifted towards personalized MT. This approach can enable the provision of customized translations that are tailored to each user's preferences and needs. It can include translations that are adapted to the user's language proficiency, domain-specific terminology, or cultural references. One possible approach to perform personalized MT is to prompt LLMs with user-specific preferences or metadata, such as the search histories or social media posts of the users. In other words, this aims to incorporate more contexts when translating text(Wang et al., 2017). A practical illustration of this can be seen in a scenario where a user's domain expertise and language fluency guide the LLM's translation output, as demonstrated in Figure 10. The zero-shot ability of LLMs makes the above tasks feasible, which are difficult to achieve in previous MT systems because such data is usually unavailable and also difficult to integrate into NMT system even when it is available. However, personalized MT still raises several challenges. One of such is how to collect and store user-specific data in a privacy-preserving manner. Another critical challenge is how to measure the effectiveness of personalized MT, as traditional evaluation metrics may not capture the nuances of user preferences and needs. Overcoming these challenges requires careful consideration of ethical, legal, and technical issues.

Low-resource MT (Zhang et al., 2021; Haddow et al., 2022b; He et al., 2022) - Translation for languages with limited resources has long been a topic of interest and concern within the NLP community. The primary obstacle has been the lack of substantial parallel corpora, which is essential for training effective translation models. The advent of LLMs has brought renewed hope to this domain, given their comprehensive training on a vast array of textual data and their proven capabilities in diverse linguistic tasks. However, a deeper investigation reveals some limitations in leveraging LLMs like ChatGPT for low-resource MT. Studies, including those by (Bang et al., 2023), have highlighted performance inconsistencies of LLMs for non-English languages. This is further supported by (Jiao et al., 2023b), which show that the translation accuracy of LLMs suffers when addressing languages that are either low in resources or significantly divergent from English. This is largely attributed to the over-representation of English in the datasets used to train these models. Nevertheless, the extensive knowledge base and generalization abilities of LLMs present several opportunities. They can be employed in generating synthetic parallel data, thereby potentially compensating for the lack of genuine training data. An example that encapsulates this potential can be observed when translating sentences from languages like Irish, as highlighted in Figure 11. Improving MT for low-resource languages is essential for preserving linguistic diversity and ensuring equitable access to information. By advancing translation capabilities for these languages, we can democratize knowledge dissemination, promote cultural understanding, and foster economic and social inclusivity on a global scale.

8. Experiments and Analysis on Translation Performance of LLMs

This section presents some experimental results and analysis of various LLMs performances in translation tasks, focusing on Chinese-to-English translations. These LLMs are evaluated based on multiple criteria, including BLEU (Papineni et al., 2002), ChrF++ (Popović, 2017), TER (Snover et al., 2006), d-BLEU (Liu et al., 2020) and detailed comparison across various datasets. Furthermore, an error rate analysis provides insights into specific challenges faced by LLMs.

8.1. Chinese-to-English Translation Performance

Table 1 is adopted from Jiao et al. (2023b), showcasing the performance of different systems in Chinese-to-English translation tasks, evaluated by BLEU, ChrF++, and TER metrics. The results have demonstrated that the early version of ChatGPT (Ouyang et al., 2022) can achieve comparable performance compared to various specialized MT systems while it is a general-purpose dialogue system, confirming the translation capability of LLMs and the prospect of utilizing LLMs on MT.

System	BLEU↑	ChrF++↑	TER↓
Google	31.66	57.09	56.21
DeepL	31.22	56.74	57.84
Tencent	29.69	56.24	57.16
GPT-3.5	24.73	53.71	62.84

Table 1: Comparative analysis of Chinese-to-English translation performance. The results are from Jiao et al. (2023b).

8.2. Document-level Translation Performance

Further analysis from Wang et al. (2023a) on various datasets reveals how commercial MT systems and LLM applications perform on document-level Chinese-to-English translation datasets including mZPRT (Wang et al., 2022a) and WMT2022 (Kocmi et al., 2022) covering domains such as news, social media, web fiction, and Q&A forum. The results shown in Table 2 demonstrate that advanced LLMs such as GPT3.5 and GPT-4 obtained strong performance on various domains and even surpass some specialized MT systems on some domains, showcasing the competitive capability of LLMs on document-level translation.

Model	News	Social	Fiction	Q&A	Avg
Google	27.7	35.4	16.0	12.0	22.8
DeepL	30.3	33.4	16.1	11.9	22.9
Tencent	29.3	38.8	20.7	15.0	26.0
GPT-3.5	29.1	35.5	17.4	17.4	24.9
GPT-4	29.7	34.4	18.8	19.0	25.5

Table 2: Document-level Translation performance on mZPRT (Wang et al., 2022a) and WMT2022 (Kocmi et al., 2022). The results are from Wang et al. (2023a).

8.3. Error Rate Analysis

An in-depth error rate analysis from Wu et al. (2024a) on provides a clear view of the challenges LLMs face for translation, which is crucial for future development of LLMs on MT. The results in Table 3 show that although the the finetuned LLaMA-7B (Touvron et al., 2023a,b) still exhibit various types of translation erro such as mis-translation and grammar issues, it achieves lower error rate on under-translation, omission and inconsistent style, etc. This analysis has demonstrated the potential of LLMs for translation.

Error Type	LLaMA	Google
Mistranslation	2,002	1,356
Overtranslation	836	715
Undertranslation	977	1,358
Addition	586	840
Omission	417	484
Grammar	474	266
Unclear reference	206	102
Cohesion	933	704
Coherence	717	565
Inconsistent style	85	771
Multiple terms in translation	386	339

Table 3: Error rate analysis between LLaMA-7B-finetune and Google Translate. The results are from Wu et al. (2024a).

9. Conclusion

In this paper, we explored several intriguing and promising research directions for MT in the context of using LLMs. We presented discussions and case examples for long-document translation, stylised translation, interactive translation, TM-based translation, multi-modal translation and new evaluation paradigms for MT using LLMs, along with examples preserving user privacy in LLMs-based MT. Furthermore, we identified additional directions such as personalized MT . Our aim is to inspire further research in the area of leveraging LLMs for MT and to advance the state-of-the-art in this rapidly evolving field.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable suggestions and comments. Jitao Xu is supported by the China Postdoctoral Science Foundation under Grant Number 2023TQ0245. Derek F. Wong is supported by the Science and Technology Development Fund, Macau SAR (Grant Nos. FDCT/060/2022/AFJ, FDCT/0070/2022/AMJ), National Natural Science Foundation of China (Grant No. 62261160648), Ministry of Science and Technology of China (Grant No. 2022YFE0204900), and the Multi-year Research Grant from the University of Macau (Grant No. MYRG-GRG2023-00006-FST-UMDF).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, 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.

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

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

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*.

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*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Josep Maria Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurélien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. 2016. Systran's pure neural machine translation systems. *CoRR*, abs/1610.05540.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

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.

Yvette Graham, Christian Federmann, Maria Eskevich, and Barry Haddow. 2020. Assessing human-parity in machine translation on the segment level. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4199–4207, Online. Association for Computational Linguistics.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022a. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022b. 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.

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, Zhaopeng Tu, Shuming Shi, and Rui Wang. 2022. Tencent AI lab - shanghai jiao tong university low-resource translation system for the WMT22 translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 260–267, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving reliable human assessment of open-domain dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6416–6437, Dublin, Ireland. Association for Computational Linguistics.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. Parrot: Translating during chat using large language models. *arXiv preprint arXiv:2304.02426*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023b. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

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.

Rebecca Knowles and Philipp Koehn. 2016. Neural interactive translation prediction. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, pages 107–120, Austin, TX, USA. The Association for Machine Translation in the Americas.

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

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180. 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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

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

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbon, Portugal. Association for Computational Linguistics.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, Toronto, Canada. Association for Computational Linguistics.

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.

OpenAI. 2023. Gpt-4 technical report.

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*.

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

Minh Quang Pham, Jitao Xu, Josep Crego, François Yvon, and Jean Senellart. 2020. Priming neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 462–473, Online. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.

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*.

Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. INMT: Interactive neural machine translation prediction. 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): System Demonstrations*, pages 103–108, Hong Kong, China. Association for Computational Linguistics.

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.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, and ohters. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Jun'ichi Tsujii. 1986. Future directions of machine translation. In *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2023. Focused transformer: Contrastive training for context scaling.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz 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.

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

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Longyue Wang. 2019. *Discourse-aware neural machine translation*. Ph.D. thesis, Dublin City University. School of Computing.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018. Translating pro-drop languages with reconstruction models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Longyue Wang, Mingzhou Xu, Derek F. Wong, Hongye Liu, Linfeng Song, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2022a. GuoFeng: A benchmark for zero pronoun recovery and translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11266–11278.

Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2022b. Controlling styles in neural machine translation with activation prompt. *arXiv preprint arXiv:2212.08909*.

Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. 2023b. Gpt4video: A unified multimodal large language model for Instruction-followed understanding and safety-aware generation. *arXiv preprint arXiv:2311.16511*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

John S. White and Theresa A. O'Connell. 1993. Evaluation of machine translation. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document flattening: Beyond concatenating context for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 448–462, Dubrovnik, Croatia. Association for Computational Linguistics.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024a. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024b. Importance-aware data augmentation for document-level neural machine translation. *arXiv preprint arXiv:2401.15360*.

Shangyu Xie, Wei Dai, Esha Ghosh, Sambuddha Roy, Dan Schwartz, and Kim Laine. 2023. Does prompt-tuning language model ensure privacy? *arXiv preprint arXiv:2304.03472*.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2023. Effective long-context scaling of foundation models.

Jitao Xu, Josep Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online. Association for Computational Linguistics.

Jitao Xu, Josep Crego, and François Yvon. 2022. Bilingual synchronization: Restoring translational relationships with editing operations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8016–8030, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. 2020. A survey of deep learning techniques for neural machine translation.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7298–7309.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *KI 2002: Advances in Artificial Intelligence: 25th Annual German Conference on AI, KI 2002 Aachen, Germany, September 16–20, 2002 Proceedings 25*, pages 18–32. Springer.

Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan, and Orhan Firat. 2022. Multilingual document-level translation enables zero-shot transfer from sentences to documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4176–4192.

Meng Zhang, Minghao Wu, Pengfei Li, Liangyou Li, and Qun Liu. 2021. Noahnmt at wmt 2021: Dual transfer for very low resource supervised machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1009–1013.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis.