New Trends in Machine Translation with Large Language Models

Chenyang Lyu[1*] and **Zefeng Du**[2*] and **Jitao Xu**[3*] and **Yitao Duan**[3] and **Derek F. Wong**[2]
and **Longyue Wang**[4]
[1] Dublin City University [2] University of Macau [3] NetEase Youdao [4] Tencent AI Lab
chenyang.lyu2@mail.dcu.ie, mc05583@umac.mo, xujt01@rd.netease.com, duan@rd.netease.com,
vinnlywang@tencent.com

Abstract

Machine Translation (MT) has been significantly advanced in recent years, particularly with the emergence of Large Language Models (LLMs) like GPT-4 and ChatGPT, leading to new challenges and opportunities. In this paper, we explore various interesting directions for MT using LLMs, such as Long-Document Translation, Stylised MT, Interactive MT, Translation Memory-based MT, and new evaluation paradigms. Additionally, we address privacy concerns in MT using LLMs and propose basic privacy-preserving methods to mitigate potential risks. We present several examples to showcase the feasibility of the proposed directions and demonstrate their potentials for enhancing the quality and diversity of MT output, especially in challenging scenarios like translating long documents. Ultimately, we highlight the opportunities and challenges for future research in MT using LLMs, suggesting potential avenues for further exploration.

1 Introduction

Machine Translation (MT), especially Neural Machine Translation (NMT, [Bahdanau *et al.*, 2015; Vaswani *et al.*, 2017; Castilho *et al.*, 2017; Stahlberg, 2020; He *et al.*, 2022b; Kocmi *et al.*, 2022]) is a fundamental task in natural language processing (NLP) that aims to automatically translate texts from one language to another. Despite 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 [He *et al.*, 2022a]. Recently, the emergence of Large Language Models (LLMs), such as GPT-3 and ChatGPT [Brown *et al.*, 2020; Chen *et al.*, 2021; Ouyang *et al.*, 2022; Wei *et al.*, 2022], has significantly advanced the state-of-the-art in MT. The zero-shot MT performance of LLMs is even on par with strong fully supervised MT systems while LLMs can also be used in various scenarios beyond MT [Wei *et al.*, 2022; Jiao *et al.*, 2023b; Wang *et al.*, 2023].

However, MT using LLMs also poses new challenges and opportunities that require new directions and methodologies. In this paper, we brainstorm several interesting directions for MT using LLMs, including Stylised MT, Interactive MT, and Translation Memory (TM) based MT, as well as a potential new evaluation paradigm of translation quality using LLMs. Stylised MT [Sennrich *et al.*, 2016; Niu and Carpuat, 2020] aims to preserve the stylistic features of the source text in the translation output, such as the tone, register, formality, genre, etc. Interactive MT [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 translations retrieved from the TM to improve the MT performance. The new evaluation paradigm using LLMs aims to leverage the power of LLMs for a more accurate and efficient evaluation of MT systems from various aspects instead of only evaluating the similarity between system outputs and references.

In addition to 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 LLM-based MT is becoming increasingly important, as LLMs may inadvertently reveal sensitive information in the source text or the translation output.

To preliminarily investigate the feasibility of the interesting directions mentioned above, we present several examples using GPT-4 for MT under various scenarios, demonstrating the feasibility of the directions. Our results demonstrate 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, which can sometimes lead to translations that are devoid of context and coherence. 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 chal-

---

[*] The first three authors contributed equally to this work.

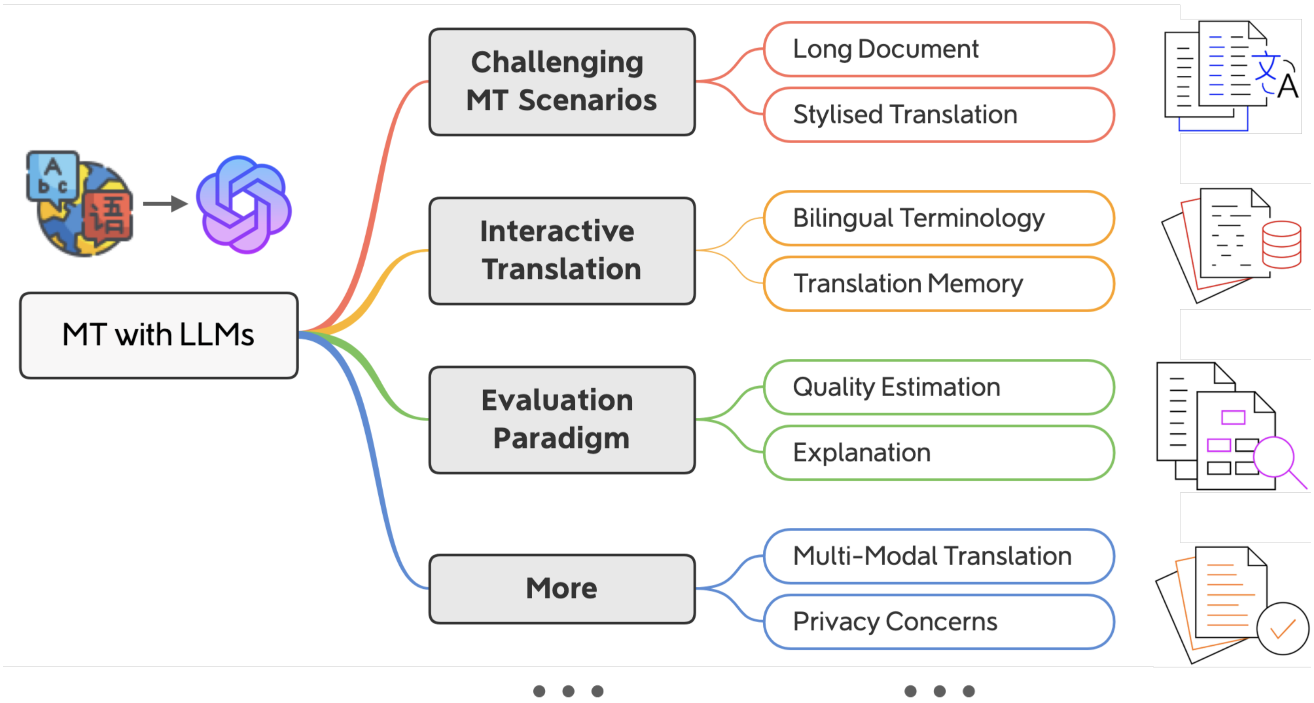


Figure 1: Interesting directions for MT using LLMs (e.g. GPT models), including challenging MT scenarios, interactive MT, new evaluation paradigm for MT using LLMs, etc.

lenges [Wang *et al.*, 2017a; Zhang *et al.*, 2022]. 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.*, 2023]. Figure 2 illustrates an example of GPT-4 translating a document.

2.2 Stylised MT

Stylised MT refers to the ability of generating translations that match a specific style or genre [Wang *et al.*, 2022], 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 MT has many potential applications, such as in marketing, literature, or cultural preservation.

However, Stylised MT is difficult to achieve before the presence of LLMs as there lacks such multi-parallel corpora for Stylised MT to fit various styles while the zero-shot ability of LLMs makes these tasks achievable. We can directly prompt LLMs to translate the text with a specific style expressed by natural language or we can first 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 that GPT-4 can handle translation with a poetic style while also keeping the semantic information of the original text, which can be hardly achieved by conventional MT systems.

Nevertheless, Stylised MT 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 MT, 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 MT

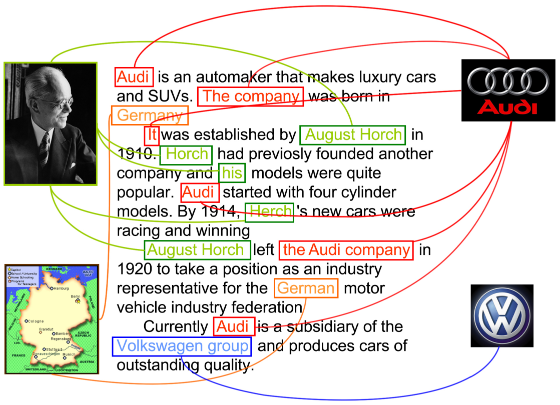
3.1 Bilingual Terminology

Interactive MT [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 with interactive user interfaces, such as chatbots or online forums, 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 MT 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 and user-friendly, yet also informative and flexible. Another challenge is how to incorporate user feedback into the trans-

GPT-4

Prompt:



Audi is an automaker that makes luxury cars and SUVs. The company was born in Germany

It was established by August Horch in 1910. Horch had previously founded another company and his models were quite popular. Audi started with four cylinder models. By 1914, Horch's new cars were racing and winning

August Horch left the Audi company in 1920 to take a position as an industry representative for the German motor vehicle industry federation

Currently Audi is a subsidiary of the Volkswagen group and produces cars of outstanding quality.

Output:
奥迪是一家生产豪华汽车和SUV的汽车制造商。该公司起源于德国。
它由奥古斯特·霍希 (August Horch) 于1910年创立。霍希此前曾创立过另一家公司，其车型相当受欢迎。奥迪起初生产四缸车型。到1914年，霍希的新车开始参加比赛并获胜。
奥古斯特·霍希于1920年离开奥迪公司，担任德国机动车行业联合会的行业代表。
目前，奥迪是大众集团的子公司，生产质量卓越的汽车。

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.

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.

lation 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.

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 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]. The use of retrieved similar sentence pairs [Pham *et al.*, 2020] seems to be a natural fit to few-shot prompting techniques when performing MT using LLMs [Vilar *et al.*, 2022; Moslem *et al.*, 2023]. LLMs have emerged with the In-Context Learning (ICL) ability that they can learn specific tasks through task examples given in the prompt.

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

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 seen during training to largely improve domain-specific translation during inference [Xu *et al.*, 2020; Xu *et al.*, 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

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.

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.

4 New Evaluation Paradigm for MT using LLMs

Evaluating the quality of MT using LLMs is a challenging task, as existing evaluation metrics may not be sufficient to capture the full range of translation quality. 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. Evaluating on these test sets cannot correctly reflect the MT performance of LLMs.

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.

A new evaluation paradigm for MT using LLMs should take into 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, we can also use extrinsic evaluation - we can use the translation output in other tasks and measure the corresponding performance instead of directly assessing the translation quality.

However, developing a new evaluation paradigm also poses several challenges. One challenge is how to balance the tradeoff 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 6. 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 trans-

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 6: 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.

lation text. Overcoming these challenges requires rigorous experimental design, statistical analysis, and transparency in reporting.

5 Privacy in MT using LLMs

As LLMs become more powerful and widely used in MT, there are growing concerns about privacy and 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 7. This is similar to methods integrating terminologies or user dictionaries into conventional NMT systems [Crego *et al.*, 2016].

However, privacy in MT using LLMs also poses 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.

6 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.*, 2017b]. 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.

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 7: 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).

Multi-modal MT [Yao and Wan, 2020; Sulubacak *et al.*, 2020] - Another promising direction is multi-modal MT, 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], can be employed to develop models that can learn from multi-modal data and generate translations that accurately convey the meaning of the input.

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.

7 Conclusion

In this paper, we explored several intriguing and promising research directions for MT in the context of using LLMs. We presented case examples for Long-Document Translation, Stylised MT, Interactive MT, TM-based MT, and new evaluation paradigms for MT using LLMs, along with examples preserving user privacy in LLM-based MT. Furthermore, we identified additional directions such as personalized MT and multi-modal translation. 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.

References

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

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Bulte and Tezcan, 2019] Bram Bulte and Arda Tezcan. 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, July 2019. Association for Computational Linguistics.

[Cai *et al.*, 2021] Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 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, August 2021. Association for Computational Linguistics.

[Castilho *et al.*, 2017] Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 2017.

[Chen *et al.*, 2021] 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. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[Crego *et al.*, 2016] 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. Systran's pure neural machine translation systems. *CoRR*, abs/1610.05540, 2016.

[Dinu *et al.*, 2019] Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 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, July 2019. Association for Computational Linguistics.

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

[He *et al.*, 2022a] Zhiwei He, Xing Wang, Zhaopeng Tu, Shuming Shi, and Rui Wang. 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), December 2022. Association for Computational Linguistics.

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

[Ji *et al.*, 2022] Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 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, May 2022. Association for Computational Linguistics.

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

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

[Knowles and Koehn, 2016] Rebecca Knowles and Philipp Koehn. Neural interactive translation prediction. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, pages 107–120, Austin, TX, USA, October 28 - November 1 2016. The Association for Machine Translation in the Americas.

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

[Kocmi *et al.*, 2022] 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. 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), December 2022. Association for Computational Linguistics.

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

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

[Moslem *et al.*, 2023] Yasmin Moslem, Rejwanul Haque, and Andy Way. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*, 2023.

[Niu and Carpuat, 2020] Xing Niu and Marine Carpuat. Controlling neural machine translation formality with synthetic supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8568–8575, Apr. 2020.

[OpenAI, 2023] OpenAI. Gpt-4 technical report, 2023.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

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

[Rabinovich *et al.*, 2017] Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 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, April 2017. Association for Computational Linguistics.

[Santy *et al.*, 2019] Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 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, November 2019. Association for Computational Linguistics.

[Sennrich *et al.*, 2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, June 2016. Association for Computational Linguistics.

[Stahlberg, 2020] Felix Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418, 2020.

[Sulubacak *et al.*, 2020] Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. Multimodal machine translation through visuals and speech. *Machine Translation*, 34:97–147, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 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., 2017.

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

[Wang *et al.*, 2017a] Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, 2017.

[Wang *et al.*, 2017b] Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 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, September 2017. Association for Computational Linguistics.

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

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

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

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

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

[Xu *et al.*, 2022] Jitao Xu, Josep Crego, and François Yvon. 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, December 2022. Association for Computational Linguistics.

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

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

[Zhang *et al.*, 2022] Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan, and Orhan Firat. 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, 2022.

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