JOURNAL OF
EMERGING INVESTIGATORS

Article

Large Language Models are Good Translators

Zhaohan Zeng[1], Zhibin Liang[2]
[1]Fontbonne Academy, Boston, Massachusetts
[2] School of Mathematical Sciences, Capital Normal University, Beijing, China

SUMMARY

Machine translation, which uses computers to translate one language into another, is one of the most challenging tasks in artificial intelligence. During the last decade, neural machine translation (NMT), which builds translation models based on deep neural networks, has achieved significant improvement. However, NMT still faces several challenges. For example, the translation quality of an NMT system greatly depends on the amount of bilingual training data, which is expensive to acquire. Furthermore, it is difficult to incorporate external knowledge into an NMT system to obtain further improvement for a specific domain. Recently, large language models (LLMs) have demonstrated remarkable capabilities in language understanding and generation. This raises interesting questions about whether LLMs can be good translators and whether it is easy to adapt LLMs to new domains or to meet specific requirements. In this study, we hypothesized that LLMs can be adapted to perform translation by using prompts or fine-tuning and these adapted LLMs would outperform the conventional NMT model in four aspects: translation quality, interactive ability, knowledge incorporation ability, and domain adaptation. We compared GPT-4 and Google Translate, the representative LLM and NMT models, respectively, on the WMT 2019 (Fourth conference on machine translation) dataset. Experimental results showed that GPT-4 outperformed Google Translate in the above four aspects by exploiting appropriate prompts. Further experiments on Llama, an open-source LLM developed by Meta, showed that the translation quality of LLMs can be further improved by fine-tuning on limited language-related bilingual corpus, demonstrating strong adaptation abilities of LLMs.

INTRODUCTION

Machine translation (MT), which aims to translate from one language (the source language) to another (the target language), is one of the most challenging tasks in artificial intelligence. Across more than 70 years of development, MT has made great progress (1). With the advantages of low cost and high efficiency, MT systems are widely used for applications that require cross-language communication, such as traveling, e-commerce, and foreign language studying.

The last decade has witnessed the rapid development of neural machine translation (NMT), an end-to-end method that builds MT models based on deep neural networks such as recurrent neural networks, convolutional neural networks, and transformer networks (2-6). Typically, an NMT model consists of an encoder that maps source language text into a high dimensional vector and a decoder that generates target language text based on the vector. NMT systems directly learn translation knowledge from bilingual training corpora, consisting of sentence pairs of source and target languages. Generally, the larger the training corpus size, the better the translation quality. To perform translation, the NMT models first "read" the source text as a whole and then generate the target text word-by-word based on an understanding of the source text. This process is analogous to human translators, producing high-quality translation in terms of both accuracy and fluency. These advantages have made NMT a breakthrough technology in the history of MT (1).

However, NMT models still face challenges (7, 8). Firstly, their performance significantly depends on the amount of annotated bilingual training corpus. However, the collection of large training data is very expensive. Secondly, incorporating external knowledge, such as domain knowledge, named entities and terminologies, into NMT models is a considerable task. To improve translation quality for a specific domain, the NMT models typically need to be retrained or fine-tuned using in-domain data or to leverage external memory to store translation knowledge of terminologies (9, 10). As mentioned, it is expensive to collect an annotated corpus, especially for specific domains. Thirdly, conventional NMT models are usually specialized for a particular translation task. Thus, additional training would be needed if we wanted to refine the outputs. To solve the above problems, a model must have more comprehensive abilities in language understanding and generation.

In this study, we investigated the translation ability of large language models (LLMs). An LLM is a pre-trained model that uses an unsupervised machine learning method to train a deep neural network, usually with hundreds of billions of parameters, on large unannotated data. Leveraging big data, deep neural networks, and numerous parameters, LLMs show good language understanding and generation abilities in many natural language processing tasks, such as text generation, human-machine dialogue, and question and answering (11-13).

Since MT requires both good understanding and generation abilities, we hypothesized that LLMs can be easily adapted for translation tasks via prompts or fine-tuning, and they can outperform conventional NMT models in four aspects, including translation quality, interactive ability, knowledge incorporation ability, and domain adaptation. We compared

https://doi.org/10.59720/24-020

Methods	Chinese-English		English-Chinese	
	COMET scores	Improvement	COMET scores	Improvement
Google Translate	0.3194	N/A	0.2828	N/A
GPT-4-Prompt-base	0.3381**	1.87%	0.2519	-3.09%
GPT-4-Prompt-refine	**0.3800****	**6.03%**	**0.2901****	**0.73%**
Llama-2-13B-chat	0.3248*	0.54%	0.1046	-17.82%

Table 1: COMET scores of Google Translate, GPT-4, and Llama-2-13B-chat for Chinese-English and English-Chinese translations. "Improvement" means the improvements of LLMs over Google Translate. The highest scores are shown in bold. When we used a prompt to refine the initial translation, GPT-4 outperformed Google Translate in both translation directions. Asterisks (*) and (**) indicate results that are are significantly better than Google Translate with $p<0.05$ and $p<0.001$, respectively. The prompts are defined in Table 4.

GPT-4 and Google Translate, which are representatives of LLM and NMT systems, respectively (14, 15). We conducted the experiments on Chinese-English translation and German-English translation. The experimental results showed that GPT-4 outperforms Google Translate in terms of both translation quality and adaptability. Furthermore, GPT-4 offers a user-friendly interactive approach through prompts. We obtained further improvements by polishing the initial translation, integrating external knowledge, and transferring to diverse domains via minor modification of the prompts. In addition, we conducted experiments on an open-source LLM, Llama-2, which was mainly trained with English corpora, to show that the translation abilities of LLMs can be easily improved with other language corpora (16). We have shown that the English-Chinese translation quality of Llama-2 is greatly improved with Chinese-related corpora, demonstrating strong adaptation. Our study confirmed the hypothesis that LLMs are good at performing translation and outperformed the traditional NMT model by using prompts or fine-tuning on related corpora.

RESULTS

We tested four competencies of LLMs related to MT including translation quality, interactive ability, knowledge incorporation ability, and domain adaptation, by designing appropriate prompts. We used Google Translate, GPT-4, and Llama-2-13B-chat, which is a version of Llama-2 (17). Besides using prompts, we also carried out further experiments on Llama-2-13B-chat by fine-tuning on additional corpora. We evaluated the translation quality using Crosslingual Optimized Metric for Evaluation of Translation (COMET) scores (18). COMET scores are calculated with a neural framework, which captures the semantic similarity between source texts and target texts, achieving a high correlation with human judgments (18). COMET scores range from 0 to 1, with higher scores indicating higher translation quality. The experimental results showed that the LLMs (GPT-4 and Llama-2-13B-chat) outperformed the NMT system (Google Translate) in the four mentioned test areas.

We initially tested the translation quality, a crucial factor in evaluating the translation ability of translation systems. We used the WMT 2019 test sets as our test data and performed translations between Chinese and English in both directions (19). The prompts we used for GPT-4 and Llama-2-13B-chat were "Please translate [SOURCE] into [TARGET]" and "[src]:<SRC>\n [tgt]:<TGT>", respectively. The result showed that GPT-4 outperformed Google Translate on Chinese-English translation, with an improvement of 1.87 percentage points in the COMET score, yet underperformed Google Translate on English-Chinese translation with a decrease of 3.09 percentage points (Google Translate and GPT-4-Prompt-base, **Table 1**). Both the improvement and the decrease in the Chinese-English direction are statistically significant ($p<0.001$), using the built-in function of the COMET tool.

Although Llama-2-13B-chat has smaller parameters than GPT-4, it also outperformed Google Translate on Chinese-English translation. However, because the training corpus for Llama-2-13B-chat contains little Chinese data, it significantly underperformed Google Translate in the English-Chinese direction: the decrease in terms of COMET score was 17.82 percentage points (**Table 1**). To study the language adaptation ability of Llama, we fine-tuned the Llama-2-13B pre-trained model with English-Chinese binlingual corpus by selecting variable numbers (ranging from 1,000 to 400,000) of sentence pairs from the WMT training corpus. We observed that the COMET scores were steadily improved by adding more English-Chinese training data (**Figure 1**), which indicated the potential ability of Llama for language adaptation.

To further improve translation quality, human translators usually review and polish the draft translation. Although it is difficult to do this in traditional MT systems, LLMs provide an efficient solution. To assess the ability of an LLM to improve a translation, we first asked GPT-4 to refine the outputs

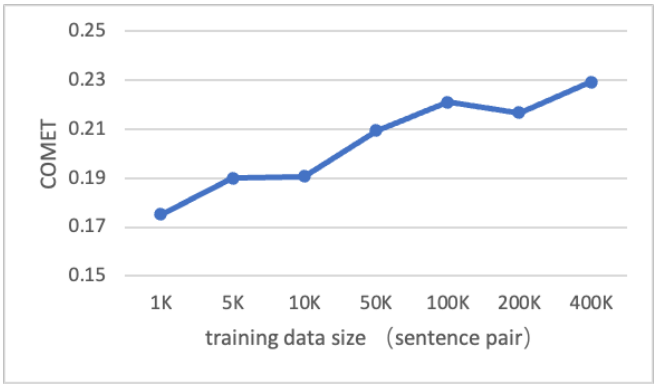

Figure 1: COMET scores during the fine-tuning of Llama-2-13B with increasing training data size for English-Chinese translation. The COMET scores indicate the translation quality with different numbers of fine-tuning sentence pairs. The translation quality steadily improved when more Chinese training data was used for fine-tuning.

https://doi.org/10.59720/24-020

Source	宁夏银川市曾有一个"尴尬"的称号：沿黄城市中唯一一个"**守着**黄河缺水喝"的城市。
Google Translate	Yinchuan City in Ningxia once had an "embarrassing" title: it was the only city along the Yellow River that "**guarded** the Yellow River for lack of water."
GPT-4-Prompt-base	The city of Yinchuan in Ningxia once had an "awkward" title: it is the only city **along** the Yellow River that is "short of water to drink despite being by the Yellow River."
GPT-4-Prompt-refine	The city of Yinchuan in Ningxia once bore a somewhat "awkward" distinction: it was the only city **along** the Yellow River that faced a shortage of drinking water, despite its proximity to the river.

Table 2: Chinese-English Translation results of Google Translate and GPT-4. For the Chinese words "守着"(in bold), GPT-4 produced the correct translation, while Google Translate used another meaning of the source word, which is not appropriate here. Furthermore, with a prompt asking the model to refine the translation, GPT-4-Prompt-refine delivered a better output. This indicates that LLMs have better language understanding and generation capabilities. The prompts are defined in Table 4.

generated by GPT-4-Prompt-base using the prompt "Please review and polish the translation result". We observed that the translation quality was further improved (GPT-4-Prompt-refine, **Table 1)** with improvements in the Chinese-English and English-Chinese translations of 4.19 and 3.82 percentage points, respectively. With this prompt, GPT-4 newly outperformed Google Translate in the English-Chinese translation direction.

We manually analyzed the translation results of GPT-4 and Google Translate. Table 2 shows an example selected from the Chinese-English test set. The Chinese word "

(shou zhe)" has two meanings, "to protect somebody/something" or "being near somebody/something in distance". In this case, Google Translate incorrectly translated it to "guarded", indicating its misunderstanding of the intended meaning. GPT-4-Prompt-base, on the other hand, produced a correct translation for the word, albeit with a tense error. With a refinement prompt, GPT-4-Prompt-refine significantly improved the translation to deliver a much more accurate result (**Table 2**).

We next tested the ability of LLMs to integrate external knowledge. We analyzed the translation results and the translation errors of named entities (NE), and then asked GPT-4 to retranslate the sentences by providing the correct translations for these entities. We used the prompt named Prompt-NE "Please retranslate the sentence, and note that the [NE in source] should be translated as [NE in target]". Since there are no publicly available test sets with translation errors of named entities, we collected and tested 20 sentences with named entity (NE) translation errors. Experimental results showed that all these errors were correctly translated. As we know, incorporating external knowledge into traditional NMT models is a challenging task. However, we can easily achieve this with GPT-4 via prompts. Google Translate and GPT-4-Prompt-base failed to translate the company's name "FangDD". By using the prompt "Please retranslate the sentence, note that '房多多' should be translated as 'FangDD'", GPT-4-Prompt-NE successfully generated the correct translation (**Table 3**).

Domain adaptation, which adjusts an MT model to perform well on a target data whose domain is different from that of training data, is important for a translation system to achieve desirable performances in real-world applications. We examined the ability of the LLMs to perform domain adaptation on the OPUS multi-domain dataset, which contains law and medical domain data (7). Since there is no domain adaptation data for Chinese-English in this domain adaptation test set, we chose the German-English language pair to conduct this experiment. To incorporate the domain information, we used three kinds of prompts. For GPT-4-Prompt-domain, we used the prompt "You are an expert in [DOMAIN], please translate the sentence from [SOURCE] into [TARGET]". For GPT-4-Prompt-src, we used the prompt "You will be first provided 5 source sentences in [DOMAIN], and then took these sentences as examples to perform the translation in this domain. The sample sentences are: [Sent 1, …, Sent 5]. Please translate the sentences from [SOURCE] into [TARGET]". For GPT-4-Prompt-tgt, we used an analogous prompt to GPT-4-Prompt-src except that we used 5 target sentences instead of source sentences (**Tables 4, 5**).

The experimental results showed that GPT-4 outperformed Google Translate in both law and medical domains (**Table 6**). In addition, with domain information (either domain name or

Source	公开资料显示，**房多多**于 2019 年 11 月登陆纳斯达克
Google Translate	Public information shows that **Fangduoduo** was listed on Nasdaq in November 2019.
GPT-4-Prompt-base	Public information shows that **Fangduoduo** landed on NASDAQ in November 2019,
GPT-4-Prompt-NE	Publicly available information indicates that **FangDD** made its debut on the NASDAQ in November 2019.

Table 3: An example to illustrate the translation errors of a named entity. For the company name "房多多" (in bold), both Google Translate and GPT-4-Prompt-base produced an incorrect translation. By providing the translation of the company name in the prompt, GPT-4-Prompt-NE generated the correct translation, which indicates LLMs can integrate specific knowledge. The prompts are defined in Table 4.

https://doi.org/10.59720/24-020

Tested Abilities	Prompt Names	Prompts
Translation	Prompt-base	Please translate [SOURCE] into [TARGET].
Refinement	Prompt-refine	Please review and polish the translation result.
Knowledge integration	Prompt-NE	Please retranslate the sentence, note that the [NE in source] should be translated as [NE in target].
Domain adaptation	Prompt-domain	You are an expert in [DOMAIN], please translate the sentence from [SOURCE] into [TARGET].
Few-shot learning	Prompt-src	You will be first provided 5 source sentences in [DOMAIN], and then take these sentences as references to perform the translation in this domain. The sample sentences are: [Sent 1, …, Sent 5]. Please translate the sentences from [SOURCE] into [TARGET].
	Prompt-tgt	Similar to Prompt-src, except that the examples are 5 target sentences in a specific domain.

Table 4: Prompts we used to test different abilities of GPT-4. SOURCE means the language being translated from, and TARGET means the language being translated to. NE is an abbreviation of Named Entity.

monolingual sample sentences), GPT-4 produced further improvements. For the law domain, the prompt with target samples (GPT-4-Prompt-tgt) achieved the highest COMET score, with an improvement of 1.31 percentage points; and for the medical domain, the prompt with domain information (GPT-4-Prompt-domain) performed the best, with an improvement of 1.51 percentage points. Both improvements were statistically significant ($p<0.05$).

As shown in Table 7, with domain information or sample sentences, GPT-4 accurately translated "chronic renal insufficiency" from the source phrase "chronischer Niereninsuffizienz", which refers to the early stages of kidney disease. The translation "chronic kidney failure" produced by Google Translate implies a more advanced stage. The distinction between the two target phrases is critical. In addition, the word "renal" is more clinical or technical than "kidney" in the medical field (**Table 7**).

DISCUSSION

In this study, we investigated the translation ability of large language models (LLMs), and compared the performances of GPT-4, Llama-2-13B-chat and Google translate. The translation quality was automatically evaluated with COMET scores. We found that LLMs are good at performing translation and outperformed the traditional NMT model by using prompts or fine-tuning on related corpora, which confirmed our hypothesis.

The performance of GPT-4 was likely possible because it has a comprehensive capability of language understanding and generation, which is achieved via training on a large number of corpora (11-12). Furthermore, GPT-4 has a promising ability to refine its own output according to the refine prompt. In our experiments, when GPT-4 received the prompt "Please review and polish the translation result", it reconsidered the initial translation given the source text by checking whether the word or phrase translations are accurate and correcting possible mistakes. During the process of refinement, it also improved the translation fluency. This process is analogous to that of human translators, who iteratively review and refine the translations. Impressively, such improvements can be achieved by designing appropriate prompts, with no need to write source codes. This kind of refinement ability was further verified in Table 3, where the NE translations could be modified by inputing the desired translations in the prompts.
GPT-4 also demonstrated excellent potential for domain adaptation. While traditional NMT models required bilingual sentences from a specific domain for fine-tuning to enhance performance, GPT-4 can adapt to a particular domain by merely being fed with a few domain-specific monolingual samples, a process known as few-shot learning. Few-shot learning can improve the performances of LLM with a few samples rather than large-scale domain-specific training data.

Furthermore, the performace of LLMs can be improved via fine-tuning. It is worth noting that the performance of LLMs mainly depends on three factors: the model size, the data size, and the amount of compute used for training. According to the scaling law, the performance of LLMs has a power-law relationship with each of these factors (20). In our experiments, we also observed that the translation quality was steadily improved with the increased data sizes.

Both automatic evaluation and manual review showed the powerful translation ability of GPT-4. Furthermore, the translation results can be easily revised, pre-edited or post-edited by using instructions in natural language or by fine-tuning. However, LLMs also face challenges. Although LLMs take advantage of prompts for smart interaction, they may cause inconsistency. In our experiments, we found that different prompts could yield different outputs when provided the same initial sentence to translate. Thus, users should carefully design prompts to achieve optimal results (21). Although LLMs can generate diverse translations, in some real applications such as translating legal documents, users expect the model to have a stable output given the same

Few-shot learning prompt for medical domain
You will be first provided 5 target sentences in medical domain, and then take these sentences as references to perform the translation in this domain. The sample sentences are: Sent-1: the effect of Fabrazyme treatment on the kidney function was limited in some patients with advanced renal disease. Sent-2: Wash hands carefully before and after applying the cream. Sent-3: Symptoms of overdose are low blood pressure, increased heartbeat, possibly decreased heartbeat. Sent-4: As with other insulins, NovoMix may cause hypoglycaemia (low blood glucose). Sent-5: It is a medicine that decreases the inflammation process of these diseases. Please translate the sentence from German into English: Dementsprechend wird eine Kontrolle der Nierenfunktion in regelmäßigen Abständen bei Patientinnen mit chronischer Niereninsuffizienz empfohlen.

Table 5: An example of few-shot learning prompt for German-English medical translation. Few-shot learning allows a model to learn from a few samples rather than a large amount of training data. In the prompt (Prompt-tgt), only examples in the target language are used.

https://doi.org/10.59720/24-020

Methods	Law domain		Medical domain	
	COMET scores	Improvement	COMET scores	Improvement
Google Translate	0.3578	N/A	0.5304	N/A
GPT-4-Prompt-base	0.3601	0.23%	0.5381	0.77%
GPT-4-Prompt-domain	0.3624	0.46%	**0.5455***	1.51%
GPT-4-Prompt-src	0.3697*	1.19%	0.5428*	1.24%
GPT-4-Prompt-tgt	**0.3709***	1.31%	0.5406*	1.03%

Table 6: COMET scores of GPT-4 and Google Translate on domain test sets of German-English translation. "Improvement" means the improvements of LLMs over Google Translate. We calculated the COMET scores for the law domain and the medical domain. All three methods that used domain-specific prompts (the lastthree rows) improved the translation quality. The highest scores are shown in bold. An asterisk (*) indicates results that are significantly better than Google Translate ($p<0.05$). The prompts are defined in Table 4.

source text. In the future, we will focus on prompt engineering and further explore the impact of different prompts on the outputs.

For automatic MT evaluation, although COMET score is widely used, it has some limitations. For example, since COMET is a neural network framework, its performance depends on both the size and the quality of its training data. Large-scale and high-quality training data can improve the consistency between the COMET scores and human evaluations.

The ability to communicate across languages is crucial to allow understanding between people globally. MT provides a convenient way to automate this process, increasing our ability to connect the world. In this study, we found that the LLMs excelled in delivering high-quality translations effectively and conveniently, which may enable LLMs to become an indispensable aid in cross-language communication.

MATERIALS AND METHODS

For NMT, we chose Google Translate as it has been widely used and showed good performance. Google Translator uses the Transformer network, which is built on attention mechanisms (4, 6). The attention mechanism describes the relationship between the input and output, playing a crucial role in modeling neural networks.

For the pre-trained model, we used GPT-4. GPT-4 allows users to input their instructions via prompts. The prompts we used are summarized in Table 4. Table 5 shows a sample for few-shot learning in the medical domain.

As GPT-4 is a closed model, we could not carry out fine-tuning experiments on it. Therefore, we employed an open source LLM, Llama, to study the language enhancement ability. Llama-1 was released by Meta in February 2023; the name Llama is an abbreviation for "Large Language Model Meta AI". Llama-2, the upgraded version, was trained on 2 trillion tokens and has twice the context length (4096 tokens) compared to Llama-1 (17). The Llama-2 family contains 3 distinct model size: 7B, 13B, and 70B (B denotes one billion tokens). In this study, we carried out experiments on the medium size model Llama-2-13B-chat ("chat" means it is a fine-tuned model rather than a pre-trained model). We downloaded the model and further fine-tuned it on an NVIDIA A100 under default settings with randomly selected English-Chinese parallel sentences from WMT training corpus. We used greedy search for decoding (22-23).

The test data we used were subsets of WMT, which are collections of Chinese and English news of 500 sentences, as well as the subsets of the OPUS multi-domain test sets, which include the law and medical domains of German to English translation, consisting of 462 and 403 sentences, respectively (7, 19). To test the few-shot learning ability for domain adaptation, we randomly selected 5 monolingual sentences from the training corpus in the domain and provided these sentences to GPT-4 as examples.

Source	Dementsprechend wird eine Kontrolle der Nierenfunktion in regelmäßigen Abständen bei Patientinnen mit **chronischer Niereninsuffizienz** empfohlen.
Google Translate	Accordingly, monitoring kidney function at regular intervals is recommended in patients with **chronic kidney failure**.
GPT-4-Prompt-base	Accordingly, regular monitoring of kidney function is recommended for patients with **chronic kidney insufficiency.**
GPT-4-Prompt-domain	Accordingly, regular kidney function monitoring is recommended for patients with **chronic renal insufficiency**.
GPT-4-Prompt-src	Accordingly, regular monitoring of kidney function is recommended in patients with **chronic renal insufficiency**.
GPT-4-Prompt-tgt	Accordingly, regular monitoring of kidney function is recommended in patients with **chronic renal insufficiency**.

Table 7: An example of German-English domain translation to show the adaptation ability of GPT-4. For the phrase "chronischer Niereninsuffizienz", Google Translate produced the output "chronic kidney failure", implying an advanced stage of kidney disease. GPT-4 produced the more accurate translation of "chronic rental insufficiency", which refers to the early stage of the disease.

To evaluate the translation quality, we used the COMET score. Traditionally, translation quality has been evaluated using BLEU, which relies on string matching and is somewhat inflexible in assessing semantic nuances (24). In contrast, COMET is a neural-based method that leverages pre-trained models and evaluates translations at the semantic level, displaying a strong correlation with human judgment. In recent years, during WMT evaluations, COMET has been officially adopted as a key metric for translation quality assessment. In our experiments, we used the wmt20-comet-qe-da model to calculate the COMET scores (25).

Received: December 18, 2023
Accepted: April 14, 2024
Published: October 16, 2024

REFERENCES

1. Wang, Haifeng, et al. "Progress in Machine Translation". *Engineering*, vol. 18, Nov. 2022, pp. 143-153, https://doi.org/10.1016/j.eng.2021.03.023.
2. Bahdanau, Dzmitry, et al. "Neural Machine Translation by Jointly Learning to Align and Translate." *The 3rd International Conference on Learning Representations (ICLR)*, 2015.
3. Sutskever, Ilya, et al. "Sequence to Sequence Learning with Neural Networks." *Advances in Neural Information Processing Systems,* 2014, pp. 3104–3112.
4. Wu, Yonghui, et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." *ArXiv.org, Cornell University Arxiv,* Oct. 8, 2016, arxiv.org/abs/1609.08144.
5. Gehring, Jonas, et al. "Convolutional Sequence to Sequence Learning." *The 34th International Conference on Machine Learning*, 2017, pp.1243-1252.
6. Vaswani, Ashish, et al. "Attention is All You Need." *The 31st Conference on Neural Information Processing*

https://doi.org/10.59720/24-020

Systems, 2017, pp. 6000–6010.
7. Koehn, Philipp, et al. "Six Challenges for Neural Machine Translation." *The First Workshop on Neural Machine Translation*, 2017, pp. 28-30, https://doi.org/10.18653/v1/W17-3204.
8. Zhang Jiajun, et al. "Neural Machine Translation: Challenges, Progress and Future." *ArXiv.org, Cornell University Arxiv,* Apr. 13, 2020, arxiv.org/abs/2004.05809.
9. Chu, Chenhui, et al. "A survey of domain adaptation for neural machine translation." *The 27th International Conference on Computational Linguistics,* 2018, pp. 1304–1319.
10. Feng, Yang, et al. "Memory-augmented Neural Machine Translation." *The 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1390-1399, https://doi.org/10.18653/v1/D17-1146.
11. Radford, Alec, et al. "Improving Language Understanding by Generative Pre-Training." *Technical Report*, 2018, openai.com/research/language-unsupervised.
12. Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 2019, pp. 4171–4186.
13. OpenAI. "GPT-4 Technical Report." *ArXiv.org, Cornell University Arxiv,* Mar. 27, 2023, arxiv.org/abs/2303.08774.
14. "GPT-4 is OpenAI's most advanced system, producing safer and more useful responses." OpenAI, openai.com/gpt-4. Accessed Mar. 16, 2023.
15. "Google Translate." Google, translate.google.com/.
16. Touvron, Hugo, et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models." *ArXiv.org, Cornell University Arxiv,* July 18, 2023, arxiv.org/abs/2307.09288.
17. "Llama 2: open source, free for research and commercial use." Meta, ai.meta.com/resources/models-and-libraries/llama/.
18. Rei, Ricardo, et al. "COMET: A Neural Framework for MT Evaluation." *The 2020 Conference on Empirical Methods in Natural Language Processing,* 2020, pp. 2685–2702, https://doi.org/10.18653/v1/2020.emnlp-main.213.
19. "Shared Task: Machine Translation of News." *ACL 2019 Fourth Conference on Machine Translation*. Accessed August 1, 2019.
20. Kaplan, Jared, et al. "Scaling Laws for Neural Language Models." *ArXiv.org, Cornell University Arxiv,* 2020, arxiv.org/abs/2304.02182.
21. Gao, Yuan, et al. "How to Design Translation Prompts for ChatGPT: An Empirical Study." *ArXiv.org, Cornell University Arxiv*, 2023, arxiv.org/abs/2304.02182.
22. Gu,Jiatao, et al. "Trainable Greedy Decoding for Neural Machine Translation." *The 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1968–1978, https://doi.org/10.18653/v1/D17-1210.
23. Xia, Heming, et al. "Unlocking Efficiency in Large Language Model Inference: A Comprehensive Survey of Speculative Decoding." *ArXiv.org, Cornell University Arxiv*, 2024, arxiv.org/abs/2401.07851.
24. Papineni, Kishore, et al. "BLEU: a Method for Automatic Evaluation of Machine Translation." *The 40th Annual Meeting of the Association for Computational Linguistics (ACL),* 2002, pp. 311-318, https://doi.org/10.3115/1073083.1073135.
25. "COMET Metrics." unbabel.github.io/COMET/html/models.html#available-evaluation-models.