

How different prompts affect GPT-5's Chinese-to-English translation performance of government work reports

Jingjing Feng

Shanghai International Studies University, Shanghai, China

420606919@qq.com

Abstract. In recent years, the rapid advancements of Large Language Models (LLMs) such as ChatGPT and GPT-5 have ushered machine translation into a new era. This study examines the impact of different prompts - simple prompts, complex prompts, and few-shot prompts - on GPT-5's translation performance for the 2024 Chinese Government Work Report, finding that while complex prompts yielded better results in automatic evaluation metrics, human assessment showed no substantial differences in translation quality between simple and complex prompts. The few-shot prompting approach displayed potential in adapting to the text style, but still faced common machine translation challenges, underscoring the importance of thoroughly analyzing text requirements and providing targeted prompt instructions when utilizing large language models for translation, as well as the need for future translators to master the characteristics of these models and develop the ability to identify and adjust translation issues, in order to enhance the practical effectiveness of machine translation.

Keywords: large language models, few-shot prompting, government text translation

1. Introduction

In recent years, the breakthrough advancements of Large Language Models (LLMs) such as ChatGPT and GPT-5 [1, 2] have ushered machine translation into a brand-new stage [3, 4]. These powerful LLMs have not only made remarkable progress in natural language understanding but also introduced a new translation paradigm based on prompts, which is expected to further improve the quality of machine translation [5].

Studies have shown that the translation quality of ChatGPT and GPT-5 outperforms commercial machine translation systems in many aspects, especially for language pairs with abundant corpora [6, 7]. ChatGPT performs exceptionally well even with zero-shot prompts, although it still faces limitations such as gender bias, stereotypes, and factual errors [8]. As an upgraded version of ChatGPT, GPT-5 has achieved further improvements in various natural language processing tasks, and its translation potential will be further unleashed.

Therefore, to further explore the impact of different types of prompts on GPT-5's Chinese-to-English translation performance of government texts, this study takes the 2024 Chinese Government Work Report as an example, comparing GPT-5's translation outputs under three prompt conditions: simple prompts, complex prompts, and few-shot prompts. A simple prompt refers to providing the most concise and clear instructional information to the machine translation system, such as "Please translate the following paragraph into English";

a complex prompt involves assigning a role to GPT-5 and requiring it to comply with the general translation requirements specified in the national standard GB/T 19682-2005; a few-shot prompt uses a small number of samples to pre-train GPT-5, aiming to simulate scenarios with scarce actual data and explore its performance and adaptability under limited data conditions.

In addition, this study employs two automatic evaluation metrics, BLEU [9] and chrF [10], to analyze the similarity between each translated version and the official translation, and conducts a comprehensive analysis and discussion combined with human assessment. Through this research, the author intends to provide empirical support for further improving the performance of large language models, especially GPT-5, in machine translation, and offer suggestions on prompt design for translation practitioners.

2. Related research

A major advantage of large language models lies in their high customizability. Users can supplement additional detailed information for translation tasks by providing various forms of prompts [11]. This prompt-based personalized approach is expected to play an important role in improving translation quality.

In recent years, researchers have paid increasing attention to the application potential of prompt technology in zero-shot and few-shot machine translation [12-15]. The formulation of prompts has a significant impact on the performance of large language models, and minor modifications may lead to substantial changes in results. This has inspired researchers to develop more advanced prompt strategies to fully unlock the machine translation potential of large language models.

Taking the PaLM model as an example, Vilar et al. [12] found that providing high-quality prompt examples to the model is a key factor in improving its translation performance, while the content relevance of the examples is not as important; in addition, although the machine-translated texts output by the PaLM model are highly fluent, they contain certain omissions and errors in accuracy.

This stands in partial contrast to the research results of Agrawal et al. [13], who selected examples from the context to train the model and studied their impact on translation quality. The results showed that both the quality and relevance of examples are critical—noisy examples can cause serious negative impacts, while optimized single examples or example sets selected based on similarity can significantly improve translation performance. The study also found that pre-trained language models can learn from contextual examples and imitate their styles. Garcia et al.'s research [16] also indicated that the style of text generated by large language models can be controlled through few-shot prompts.

Furthermore, the translation prompts designed by Gao et al. [5] for ChatGPT performed excellently in both high-resource language translation and multilingual translation tasks. This suggests that reasonable prompt design can effectively improve the translation performance of large language models. Deilen et al.'s study [17] found that assigning a role to ChatGPT helps enhance its translation performance.

In summary, prompt technology is a key means to give full play to the customizability of large language models. By providing the model with more contextual information and instructions, it is expected to help the model better understand and execute translation tasks, thereby promoting the further development of machine translation technology. Therefore, the author believes it is necessary to explore what kind of prompts can more effectively guide large language models to produce ideal translations. This study only takes the 2024 Chinese Government Work Report as an example for exploration.

3. Research design

3.1. Research procedures

This study randomly selected five paragraphs from the 2024 Government Work Report of the People's Republic of China, used GPT-5 as the machine translation engine to translate them into English, and compared the translations with the official English version. The specific procedures are as follows:

- a. Data Collection: Collect the original Chinese text of the government work report and the corresponding official English translation as the research dataset.
- b. Model Training: Use GPT-5 as the machine translation engine for model training under three prompt conditions: simple prompts, complex prompts, and few-shot prompts. A simple prompt refers to providing the most concise and clear instructional information to the machine translation system, such as "Please translate the following Chinese paragraphs into English"; a complex prompt involves assigning a role to GPT-5 and requiring it to comply with the general translation requirements specified in the national standard GB/T 19682-2005; a few-shot prompt selects three samples from the same text to pre-train GPT-5.
- c. Machine Translation: Use the trained GPT-5 model to translate the five selected paragraphs respectively, obtaining the corresponding English translations.
- d. Evaluation Metrics: Use two automatic evaluation metrics, BLEU and chrF, combined with human assessment to compare the machine translation quality of GPT-5 under different prompts.

3.2. Three types of prompts

This section specifically lists the three types of prompts used in the study (Table 1).

3.3. Two common automatic evaluation metrics

This study uses two common automatic evaluation metrics for machine translation, as follows:

Bilingual Evaluation Understudy (BLEU): A widely used method for evaluating machine translation quality. It calculates the score by comparing the similarity between the machine translation output and one or more reference translations. The BLEU score mainly focuses on the matching degree of n-grams (word phrases) and is usually used to evaluate the translation quality at the sentence level or the entire document level. A high BLEU score indicates that the translation output is closer to the reference text in terms of vocabulary selection and order. Generally, a BLEU score above 40 can be considered a good machine translation result.

Character n-gram F-score (chrF): Another method for evaluating machine translation quality, which mainly focuses on character-level matching. Compared with BLEU, chrF calculates not only the matching degree of words but also that of characters. chrF computes precision, recall, and their weighted harmonic mean (F-score), enabling a more comprehensive evaluation of the overall translation quality. Generally, a chrF score above 0.6 can be considered a good machine translation result.

These two evaluation metrics have their own advantages and application scenarios. BLEU is suitable for quickly evaluating the surface accuracy of translations, while chrF has more advantages in handling the complexity of language morphology. However, it should be emphasized here that automatic evaluation metrics can only be used as references, and the final judgment of translation quality needs to comprehensively consider multiple factors such as linguistic fluency and the completeness of semantic transmission in combination with human assessment.

Table 1. Three types of prompts used in this study

Simple Prompt	Please translate the following Chinese paragraphs into English.
Complex Prompt	<p>You are an expert in government text translation. Please translate the following Chinese paragraphs from the Government Work Report into English in accordance with the following requirements: 1. Be faithful to the original text, fully and accurately express the information of the original text without core semantic errors. 2. Maintain consistent terminology that conforms to the industry-specific general standards or conventions of the target language. 3. Ensure fluent writing that complies with the linguistic norms and expression habits of the target language, and is clear and understandable.¹</p> <p>Please refer to the wording and style in the following Chinese-English parallel samples to translate the Chinese paragraphs I provide into English:</p> <p>Sample 1: 传统产业加快转型升级,战略性新兴产业蓬勃发展,未来产业有序布局,先进制造业和现代服务业深度融合,一批重大产业创新成果达到国际先进水平。</p> <p>Traditional industries saw accelerated transformation and upgrading, strategic emerging industries achieved vigorous growth, and ground work was laid for developing industries of the future. Advanced manufacturing was further integrated with modern services, and a number of world-class innovations were made in major industries.</p>
Few-shot Prompt	<p>Sample 2: 积极的财政政策要适度加力,提质增效.综合考虑发展需要和财政可持续,用好财政政策空间,优化政策工具组合.</p> <p>We should appropriately enhance the intensity of our proactive fiscal policy and improve its quality and effectiveness. We should take into account both development needs and fiscal sustainability, leverage fiscal policy space, and improve the policy toolkit.</p> <p>Sample 3: 完成今年发展目标任务,必须深入贯彻习近平经济思想,集中精力推动高质量发展.强化系统观念,把握和处理好重大关系,从整体上深入谋划和推进各项工作.</p> <p>In order to fulfill the development goals and tasks for this year, we must act on Xi Jinping Thought on the Economy and devote full energy to promoting high-quality development. We should apply systems thinking, pay attention to and properly handle major relationships, and develop in-depth plans and advance all areas of work with the overall picture in mind.</p>

¹ Excerpted from the National Standard of the People's Republic of China GB/T 19682—2005 Quality Requirements for Translated Texts in Translation Services.

4. Results and evaluation

4.1. Five selected paragraphs (Tables 2-6)

Human Assessment: Whether driven by simple or complex prompts, GPT-5's translations consistently employ a subject-less structure, mirroring the original Chinese text where the subject "we" or "the Chinese government" is often omitted. This feature is typical in Chinese governmental documents but is unusual in English, where extensive use of subject-less verb-object constructions can be confusing for readers. From this perspective, minimal prompting has proven effective in guiding the translation style.

Translations under minimal prompting not only use active and passive voices but also incorporate modal verbs like "must" and "should". However, it's crucial to consider the broader context, such as planning for the upcoming year, where using "will" might be more appropriate. This emphasizes the importance of the translator's initial analysis of the original text and understanding of its features. Adding requirements about

tense and complete sentence structure in prompts could lead to translations that are closer to official reference translations.

Table 2. Selected paragraph 1

Original Chinese	坚持教育强国,科技强国,人才强国建设一体统筹推进,创新链产业链资金链人才链一体部署实施,深化教育科技人才综合改革,为现代化建设提供强大动力.
Simple Prompt Translation	Persist in the integrated planning and advancement of building a powerful nation in education, science and technology, and talent. Implement a unified deployment of the innovation chain, industrial chain, capital chain, and talent chain. Deepen comprehensive reforms in education, science and technology, and talent to provide a strong driving force for modernization efforts.
Complex Prompt Translation	Adhere to the integrated and coordinated development of building a powerful nation in education, science and technology, and talent. Implement a unified deployment of innovation chains, industrial chains, financial chains, and talent chains. Deepen comprehensive reforms in education, science and technology, and talent development to provide a strong impetus for modernization.
Few-shot Prompt Translation	We must persist in the coordinated advancement of building a strong nation in education, science and technology, and talent. The innovation chain, industrial chain, capital chain, and talent chain should be strategically implemented as a whole. By deepening comprehensive reforms in education, science, technology, and talent, we provide a powerful impetus for modernization.
Official English Translation	We will make holistic and coordinated efforts to strengthen China's education, science and technology, and workforce, integrate planning and policy implementation for innovation, industrial, capital, and talent chains, and deepen comprehensive reforms of education, science and technology, and human resources, so as to create a strong impetus for our modernization drive.

Furthermore, the differences between the three machine translations and the official translation are also evident in specific word choices and the logical relationships between sentences. For instance, the term "人才" is variably translated in the official text as "workforce", "talent", and "human resources", showcasing the adaptability and subjective input of human translators. Concerning logical relationships, the final phrase "为现代化建设提供强大动力" should serve as the goal of the previously mentioned work tasks, not just the preceding sentence. This discrepancy in logical rigor highlights a gap that needs addressing in machine translation outputs.

Human Assessment: Similar to the previous example, the machine translations generated by the three prompts all contain subject-less sentences. Moreover, as this paragraph is a summary of work completed last year in the Government Work Report, the past tense should be used. Similarly, it is necessary to add instructions about tense and sentence structure in the prompts.

In addition, regarding the logic between sentences, the three machine translations basically adopt a parallel structure. In reality, however, "高度重视稳就业" (placing high importance on stabilizing employment) is not in a parallel relationship with the subsequent two sentences ("出台政策" and "加强就业服务"); instead, it is more of a relationship between "purpose/tenet" and "specific measures", which belongs to the "implied logic" in Chinese and can only be detected and reflected in the translation by individuals familiar with Chinese and government text characteristics.

Furthermore, if the official translation has room for improvement, the last sentence should be the result of the specific measures mentioned in the previous sentences, and a connecting word such as "As a result" could be added to express the implied logic.

Table 3. Selected paragraph 2

Original Chinese	聚焦群众关切,办好民生实事.高度重视稳就业,出台支持企业稳岗拓岗政策,加强高校毕业生等重点群体就业促进服务,脱贫人口务工规模超过3300万.
Simple Prompt Translation	Focus on public concerns and address issues important to people's livelihoods. Place high importance on stabilizing employment, introduce policies to support businesses in maintaining and expanding job positions, and enhance employment promotion services for key groups such as college graduates. The scale of employment among the population lifted out of poverty has exceeded 33 million.
Complex Prompt Translation	Focus on the concerns of the public and address key livelihood issues effectively. Place high importance on stabilizing employment, introduce policies to support enterprises in maintaining and expanding job positions, and strengthen employment promotion services for key groups such as college graduates. The number of people employed from poverty-stricken backgrounds has exceeded 33 million.
Few-shot Prompt Translation	Focus on the concerns of the public and address key livelihood issues effectively. We place a high priority on stabilizing employment, introducing policies to support businesses in maintaining and expanding job opportunities, and enhancing employment promotion services for key groups such as college graduates. The number of people lifted out of poverty and engaged in employment has exceeded 33 million.
Official English Translation	We addressed the people's concerns and delivered benefits to them. We adopted, with a strong emphasis on keeping employment stable, policies to support enterprises in stabilizing and expanding employment and improved services to help college graduates and other key groups secure employment. Over 33 million people lifted out of poverty found jobs last year.

Table 4. Selected paragraph 3

Original Chinese	我国经济持续回升向好的基础还不稳固,有效需求不足,部分行业产能过剩,社会预期偏弱,风险隐患仍然较多,国内大循环存在堵点,国际循环存在干扰.
Simple Prompt Translation	The foundation for the sustained recovery and improvement of our country's economy is still not solid, with insufficient effective demand, overcapacity in some industries, weak social expectations, and numerous risks and hidden dangers. There are bottlenecks in the domestic economic cycle and disruptions in the international cycle.
Complex Prompt Translation	The foundation for the continued recovery and improvement of our country's economy is still not solid, with insufficient effective demand, overcapacity in some industries, weak social expectations, and numerous risks and hidden dangers. There are bottlenecks in the domestic major cycle and disturbances in the international cycle.
Few-shot Prompt Translation	The foundation for the sustained recovery and improvement of our economy is not yet solid, with insufficient effective demand, overcapacity in some industries, weak social expectations, and still numerous risks and hidden dangers. There are bottlenecks in the domestic major cycle and disturbances in the international cycle.
Official English Translation	The foundation for China's sustained economic recovery and growth is not solid enough, as evidenced by a lack of effective demand, overcapacity in some industries, weak public expectations, and many lingering risks and hidden dangers. Furthermore, there are blockages in domestic economic flows, and the global economy is affected by disruptions.

Human Assessment: The wording and sentence structures of the three machine translations are roughly the same, with only minor differences. One notable distinction is that the official translation specifies "我国经济"

as "China's...", while the machine translations either translate it as "our country's economy" or "our economy". In addition, regarding the translation of the Chinese politically characteristic expression "国内国际两个循环", all three machine translations literally translate it as "cycle" without explaining its specific meaning, making it difficult for English readers to understand. In contrast, the official translation translates "国内循环" and "国际循环" as "economic flows" and "the global economy" respectively, which is a specific translation based on the context, fully reflecting human translators' understanding and proficient use of the connotation of Chinese characteristic discourse.

Human Assessment: Since the original text of this paragraph provides the formal subject "我们" (we), the three machine translations avoid the subject-less issue present in the previous two examples. In terms of word choice, the three machine translations are very close to the official translation. The machine translation generated by the few-shot prompt divides the sentence into finer segments, avoiding the situation where one subject leads multiple parallel verbs—a structure that even exists in the official translation. Similar to Example 2, the first few sentences of this paragraph should be "means or measures," while the last sentence is the "goal." Neither the machine translations nor the official translation successfully reflect this implied logic.

Table 5. Selected paragraph 4

Original Chinese	我们要以更大的决心和力度深化改革开放,促进有效市场和有为政府更好结合,持续激发和增强社会活力,推动高质量发展取得新的更大成效。
Simple Prompt	We must deepen reform and opening up with greater determination and intensity, promote a better combination of an effective market and a capable government, continuously stimulate and enhance social vitality, and drive high-quality development to achieve new and greater successes.
Complex Prompt	We must deepen reform and opening up with greater determination and intensity, promote a better combination of an effective market and a capable government, continuously stimulate and enhance social vitality, and drive high-quality development to achieve new and greater success.
Few-shot Translation	We must deepen reform and opening up with greater determination and intensity, promoting a better combination of an effective market and a capable government. By continuously stimulating and enhancing social vitality, we aim to achieve new and greater successes in high-quality development.
Official English Translation	We should deepen reform and opening up with greater resolve and effort, promote better interplay between an efficient market and a well-functioning government, further stimulate and boost the vitality of our society, and strive to make greater progress in pursuing high-quality development.

Table 6. Selected paragraph 5

Original Chinese	制定支持数字经济高质量发展政策,积极推进数字产业化,产业数字化,促进数字技术和实体经济深度融合。
Simple Prompt Translation	Formulate policies to support the high-quality development of the digital economy, actively promote the digitalization of industries and the industrialization of digital technologies, and facilitate the deep integration of digital technology and the real economy.
Complex Prompt Translation	Formulate policies to support the high-quality development of the digital economy, actively promote the digitalization of industries and the industrialization of digital technologies, and encourage the deep integration of digital technology with the real economy.
Few-shot Prompt Translation	Policies to support the high-quality development of the digital economy are being formulated. We are actively promoting the digitalization of industries and the industrialization of digital technologies, fostering a deep integration of digital technology with the real economy.

Official English Translation	We will formulate policies to support high-quality development of the digital economy. We will actively develop the digital industry, transform traditional industries with digital technologies, and fully integrate digital technology into the real economy.
------------------------------	---

Human Assessment: The subject-less issue re-emerges in this example. Although the translation generated by the few-shot prompt still has certain tense issues, it uses complete sentences. In addition, regarding the Chinese characteristic expression "数字产业化,产业数字化", all three machine translations only provide literal translations and fail to offer specific explanations like the official translation. Of course, this situation can be further optimized by providing subsequent prompts to GPT-5, but it requires individuals to conduct corresponding analysis and judgment before giving appropriate prompts.

4.2. Automatic evaluation results and analysis

Table 7. BLEU and chrF scores of the three machine translations

Automatic Evaluation Metrics	Selected Paragraph 1	Selected Paragraph 2	Selected Paragraph 3	Selected Paragraph 4	Selected Paragraph 5	Average
BLEU						
Simple Prompt	62.42	69.21	65.79	73.28	62.42	66.62
Complex Prompt	68.39	72.44	68.33	75.92	68.39	70.69
Few-shot Prompt	65.71	70.83	67.42	74.65	65.71	68.86
chrF						
Simple Prompt	0.764	0.803	0.794	0.831	0.764	0.791
Complex Prompt	0.782	0.819	0.808	0.847	0.782	0.808
Few-shot Prompt	0.775	0.811	0.801	0.839	0.775	0.798

As can be seen from Table 7, Figure 1, and Figure 2, for the five selected paragraphs in this study, the results obtained by the three automatic evaluation metrics show the same trend. Horizontally, the BLEU and chrF scores of the three machine translations for Selected Paragraph 4 are the highest, while those for Selected Paragraph 1 and Selected Paragraph 5 are the lowest, which is basically consistent with the results of human assessment. Vertically, the BLEU and chrF scores of the machine translation results generally follow the pattern: Complex Prompt > Few-shot Prompt > Simple Prompt.

In addition, all BLEU scores for the five selected paragraphs exceed 60, and all chrF scores exceed 0.7, indicating that the translations of all selected paragraphs under the three prompts are of quite high-level machine translations. According to the principle that higher scores indicate that the machine translation is closer to the reference translation, the machine translation generated by the complex prompt seems to be of the highest quality, but this may not be the case. The next section will conduct a discussion combining automatic evaluation and human assessment.

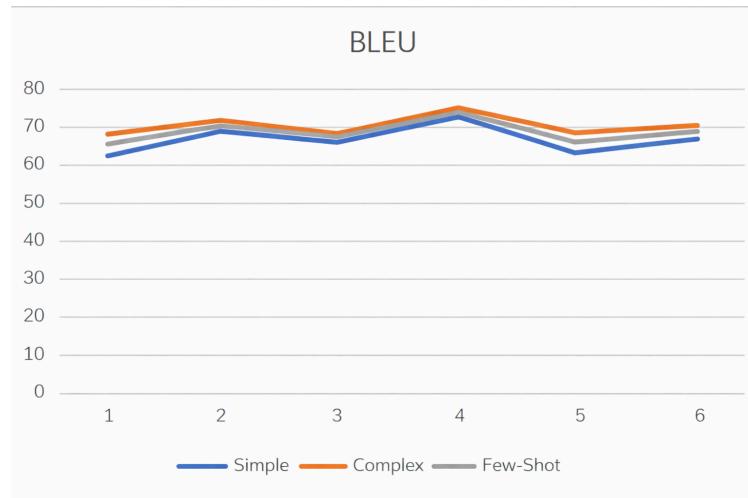


Figure 1. BLEU scores of the three machine translations

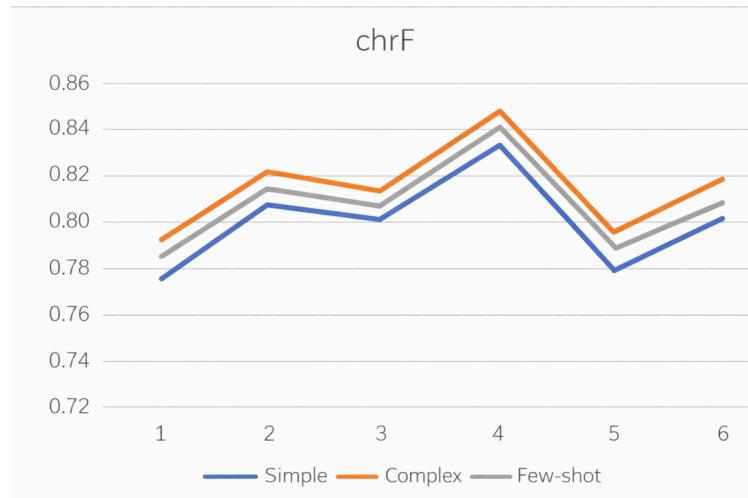


Figure 2. chrF scores of the three machine translations

4.3. Comprehensive analysis and discussion

For the five selected paragraphs in this study, there is little difference in wording and sentence structure between the machine translations generated by simple prompts and complex prompts. This indicates that GPT-5 has undergone pre-training in translation; unless human translators can provide personalized prompts based on their understanding and analysis of specific texts, general prompts will only yield similar results. In contrast, the translations generated by few-shot prompts are more in line with the needs of the selected texts—apart from using complete sentences, they also make good attempts at sentence segmentation. Overall, the performance ranking is: Few-shot Prompt > Complex Prompt \approx Simple Prompt.

A comprehensive consideration of automatic evaluation and human assessment shows that whether using BLEU or chrF scores, the relative score levels of machine translations for different paragraphs have certain reference value, which can reflect the adoptability of machine translations or the intensity of post-editing work to a certain extent. However, the scores of machine translations generated by different prompts for the same paragraph cannot reasonably reflect the true quality differences between the three translations.

In addition, the translations output by GPT-5 under the three prompts all fail to break through the inherent limitations of machine translation: 1) The translation of Chinese characteristic vocabulary is still limited to literal translation, such as "国内/国际循环" (domestic/international cycle) and "数字产业化,产业数字化" (digital industrialization and industrial digitalization); 2) The diversity of expression is insufficient, such as the discretionary handling of the term "人才" (talent); 3) The implied logic in Chinese (especially in government texts) fails to be expressed, such as causal relationships and the relationship between "measures" and "goals."

In view of this, when using GPT-5 for translation, the prompts we provide: 1) Should be based on an understanding of text characteristics and project requirements, providing specific and actionable instructions; 2) Should fully leverage GPT-5's advantage in linguistic fluency while being vigilant of its shortcomings, including issues in understanding Chinese characteristic vocabulary (or professional terminology) and grasping implied logic; 3) May not achieve the expected results in one attempt. In such cases, it is necessary to use GPT-5's functional integration [18] to provide corresponding adjustments based on the machine translation results.

In summary, the author believes that future translators may need to possess the following abilities: 1) Understand the advantages and disadvantages of GPT-5 as a machine translation engine; 2) Understand the texts they need to translate, be able to analyze text characteristics, be familiar with project requirements, and issue clear and targeted prompts to large language models; 3) After obtaining machine translation outputs, be able to quickly identify problems and conduct post-editing, or issue subsequent prompts to GPT-5 to adjust the machine translation results. These abilities will play a key role both in the future survival of translators themselves and in improving the application effect and practical value of machine translation.

5. Conclusion

Focusing on the Chinese-to-English translation task of the 2024 Government Work Report, this study explores the translation effects of GPT-5 when receiving different types of prompts (simple prompts, complex prompts, and few-shot prompts). The findings indicate that while complex prompts perform better in automatic evaluation metrics, human assessment shows no significant differences in wording and sentence structure between the translations generated by simple prompts and complex prompts. Assigning a role to GPT-5 does not bring significant quality improvement. This suggests that GPT-5's basic translation ability is already quite stable, and even complex general prompts cannot significantly improve translation quality; instead, prompts targeted at specific texts and project requirements are needed.

The translation attempts using few-shot prompts demonstrate certain potential in adapting to specific text styles but still face common machine translation challenges, such as the literal translation of culture-specific vocabulary and insufficient logical processing of complex sentence structures. Based on these findings, it is recommended that when using GPT-5 and similar large language models for translation, in-depth analysis of text characteristics and translation requirements should be conducted, and specific and detailed instructions should be provided to achieve better translation results.

In addition, future translators need to master the characteristics of large language models, be able to identify and adjust problems in translations, including conducting detailed post-editing of the outputs of large language models or adjusting prompt strategies to adapt to the needs of different translation scenarios. These abilities will play a crucial role both in the future development of translators themselves and in enhancing the practical effectiveness and value of machine translation.

Future research can attempt to expand the number of samples in few-shot prompts, such as using the previous year's government work report as samples to train GPT-5 and comparing the generated translations

with this year's government work report. Since there must be a certain correlation between the government work reports of consecutive years, it is reasonable to hypothesize that the translation quality will be improved compared to this study. On this basis, we can also extend few-shot prompts to other professional fields. In addition, although assigning a role to GPT-5 did not play a significant role in this study, further research is needed to explore whether it will have an impact on translations in certain professional fields.

References

- [1] Brown, T., et al. (2020, May 28). *Language models are few-shot learners*. arXiv. <https://arxiv.org/abs/2005.14165>
- [2] OpenAI. (2023, May 15). *GPT-5 technical report*. arXiv. <https://arxiv.org/abs/2303.08774>
- [3] Lyu, C., et al. (2023, May 2). *A paradigm shift: The future of machine translation lies with large language models*. arXiv. <https://arxiv.org/abs/2304.00207>
- [4] Hendy, A., et al. (2023, February 18). *How good are gpt models at machine translation? A comprehensive evaluation*. arXiv. <https://arxiv.org/abs/2303.16317>
- [5] Gao, Y., et al. (2023, April 5). *How to design translation prompts for ChatGPT: An empirical study*. arXiv. <https://arxiv.org/abs/2304.02182>
- [6] Wang, L., et al. (2023, April 5). *Document-level machine translation with large language models*. arXiv. <https://arxiv.org/abs/2304.02210>
- [7] Jiao, W., et al. (2023, January 20). *Is chatgpt a good translator? A preliminary study*. arXiv. <https://arxiv.org/abs/2301.08745>
- [8] Khatri, J., et al. (2023). Can you translate for me? Code-switched machine translation with large language models. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* (Volume 2: Short Papers) (pp. 83-92). Association for Computational Linguistics
- [9] Papineni, K., et al. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). Association for Computational Linguistics
- [10] Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 392-395). Association for Computational Linguistics
- [11] Siu, S. C. (2023, May 14). *ChatGPT and GPT-5 for professional translators: Exploring the potential of large language models in translation*. SSRN. <https://dx.doi.org/10.2139/ssrn.4448091>
- [12] Vilar, D., et al. (2022, November 16). *Prompting palm for translation: Assessing strategies and performance*. arXiv. <https://arxiv.org/abs/2211.09102>
- [13] Agrawal, S., et al. (2022, December 5). *In-context examples selection for machine translation*. arXiv. <https://arxiv.org/abs/2212.02437>
- [14] Zhang, B., et al. (2023, January 18). *Prompting large language model for machine translation: A case study*. arXiv. <https://arxiv.org/abs/2301.07069>
- [15] Zhu, W., et al. (2023, April 10). *Multilingual machine translation with large language models: Empirical results and analysis*. arXiv. <https://arxiv.org/abs/2304.04675>
- [16] Garcia, X., et al. (2023, February 2). *The unreasonable effectiveness of few-shot learning for machine translation*. arXiv. <https://arxiv.org/abs/2304.03656>
- [17] Deilen, S., et al. (2023, August 22). *Using ChatGPT as a CAT tool in Easy Language translation*. arXiv. <https://arxiv.org/abs/2308.11563>
- [18] Geng, F., & Hu, J. (2023). New directions in AI-assisted post-editing: A case study of translation based on ChatGPT. *Chinese Journal of Foreign Languages*, 20(3), 41-47