Yu Zhang*

Consideration of prompts as a neglected factor in research on evaluating ChatGPT's translation performance

https://doi.org/10.1515/jccall-2024-0033
Received December 31, 2024; accepted May 24, 2025; published online July 10, 2025

Abstract: The rapid development and adoption of ChatGPT have sparked increasing academic interest in its translation performance. However, little attention has been paid to the role of prompts – a key factor influencing ChatGPT's output. This study addresses this gap by analyzing the consideration of prompts in 32 articles that evaluated ChatGPT's translation performance, selected through the PRISMA framework. The findings reveal that prompt consideration is generally inadequate: nearly a third of the articles did not mention prompts, and most studies that did acknowledge prompts only addressed superficial aspects, such as merely mentioning the concept or specifying the prompts used, rather than treating prompts as a key variable. Additionally, prompt usage is often limited and lacks justification. The analysis further reveals a significant disciplinary disparity, as studies in computer science and information technology have demonstrated the most comprehensive approaches, surpassing those in translation and linguistics, as well as other disciplines. Based on these findings, this study proposes a series of implementation-focused recommendations to enhance prompt consideration in ChatGPT-related research, and underscores the importance of interdisciplinary collaboration to foster a more integrated research approach.

Keywords: ChatGPT; prompts; translation performance; evaluation

1 Introduction

ChatGPT, a generative AI chatbot developed by OpenAI, has demonstrated unprecedented capabilities in understanding and generating natural language (Che et al. 2023). It has significantly impacted numerous professional domains, including translation, even raising concerns about its potential to replace human translators (Song 2024; Zhong 2024). Researchers have increasingly applied ChatGPT to diverse

***Corresponding author: Yu Zhang**, Shanghai International Studies University, Shanghai, China, E-mail: 623436107@qq.com. https://orcid.org/0009-0005-2396-5977

translation tasks to evaluate its performance, both to explore its potential applications in translation practice and to derive insights for translation education and the industry (e.g., Hendy et al. 2023; Wang 2024; Zhang and Zhao 2024). This line of research has grown fast over the past two years both domestically and internationally, spanning diverse disciplines and yielding varied results. Despite the growing body of research on ChatGPT's translation performance, the methodological rigor of these studies warrants closer examination.

Unlike traditional translation tools, such as Google Translate and DeepL, which follow relatively fixed processes to generate translations based solely on the input of the source text, ChatGPT is capable of producing highly tailored content and making real-time adjustments according to users' instructions in natural language – referred to as "prompts" in the context of generative AI. This capability is enabled by ChatGPT's extraordinary natural language understanding and generating abilities (Che et al. 2023; Li et al. 2023; Wu et al. 2023), enhanced context awareness (Li et al. 2023; Muktadir 2023; Wu et al. 2023), and open, interactive user engagement mode (Chen and Qiu 2023; Qiu and Gu 2023; Zhao 2023). As a result, the quality of its outputs heavily depends on the effectiveness of the prompts provided by users. Crafting high-quality prompts has become a key factor in the successful use of generative AI tools like ChatGPT (Korzynski et al. 2023; Yu and Li 2024).

Given the importance of prompts, they should be a key consideration when evaluating ChatGPT's translation performance, while improper or insufficient consideration threatens the validity and reliability of research findings, and hinders the overall accumulation of knowledge about ChatGPT. However, it remains unclear whether and how prompts have been accounted for in existing studies.

Therefore, this study aims to examine how existing studies have addressed the role of prompts when evaluating ChatGPT's translation performance. Specifically, it seeks to answer three interrelated research questions (RQs):

RQ1: To what extent has existing research considered prompts in their evaluation of ChatGPT's translation performance, and is the consideration adequate?

RQ2: Does the consideration of prompts vary across different research disciplines? If so, how does it differ?

RQ3: How can the consideration of prompts in future research be improved to enhance methodological rigor and reliability?

RQ1 provides a broad overview, aiming to identify general trends and gaps in the treatment of prompts within the existing body of research. RQ2 narrows the focus to explore variations across disciplines, as the evaluation of ChatGPT's translation

performance is inherently interdisciplinary. RQ3 aims to propose practical recommendations drawing on the previous two questions for enhancing the consideration of prompts in future research.

The remainder of this paper is organized as follows. Section 2 reviews the existing research on prompts and ChatGPT-related translation studies, identifying key contributions and gaps. Section 3 outlines the methodology for the survey of 32 articles evaluating ChatGPT's translation performance, focusing on their consideration of prompts. Section 4 presents the findings, including the overall patterns of prompt consideration and variations across disciplines. Section 5 concludes with key findings and offers recommendations for future studies.

2 Literature review

2.1 Studies on prompts

With the rapid rise of ChatGPT, studies on prompts have become an emerging research area (Muktadir 2023). Researchers have been exploring prompting strategies to unlock the full potential of ChatGPT, both for general purposes and specific tasks such as translation. These studies have demonstrated that well-designed prompts can significantly influence ChatGPT's performance, especially in the context of machine translation, which is the primary focus of this study.

In terms of general prompting tactics, researchers have identified methods that can enhance ChatGPT's performance across a wide range of tasks. These include few-shot prompting (Brown et al. 2020), which improves the model's performance by providing examples that help establish context and direct the model to imitate desired styles and conventions; chain-of-thought prompting (Kojima et al. 2022; Wei et al. 2022), which boosts the model's ability to handle complex tasks by enhancing its reasoning capabilities; and role-play prompting (Njifenjou et al. 2024), which guides the model's behavior by assigning it specific roles or personas through tailored instructions. OpenAI (2023), the developer of ChatGPT, has further consolidated these findings in its *Prompt Engineering Guide*, which outlines six key strategies and detailed tactics for optimizing ChatGPT's performance. In addition, several frameworks, such as the CAST model (Jacobs and Fisher 2023), the CLEAR model (Lo 2023), and the TRUST model (Trust 2023), have been proposed to combine different prompting techniques, offering structured approaches to prompt engineering, thus further advancing the field.

In the specific context of translation, a series of studies have explored and tested various prompting methods to enhance ChatGPT's overall translation capabilities, or to address particular challenges in machine translation. Some researchers applied

standard, general prompting strategies to translation tasks, while others designed more customized prompts specifically for translation, building upon foundational prompting strategies. For example, Yamada (2023) and He (2024) investigated the effectiveness of incorporating background information about translation tasks in the prompts, including the purpose, target audience, and the personas of "translator" and "author". Gu (2023) as well as Prahallad and Mamidi (2024) successfully employed prompts with clearly articulated, step-by-step instructions to address long-standing challenges in Japanese-Chinese and English-Dravidian machine translation. In a similar vein, Jiao et al. (2023) introduced pivot prompting, a method for translating distant languages, which asks ChatGPT to translate the source sentence into a high-resource pivot language before into the target language, thereby breaking down the translation task into more manageable steps. Additionally, Jiao et al. (2024) combined various prompting strategies to propose a gradable prompting taxonomy using prompts containing expression type, translation style, part-of-speech information and explicit instructions, aiming to facilitate development of prompts tailored for various translation tasks.

These explorations not only have significantly enriched the toolkit for translation practitioners and researchers, but also have demonstrated the significant impact of prompting strategies on ChatGPT's translation performance. However, while many researchers have pointed out that prompts are insufficiently utilized in the evaluation of ChatGPT's translation capabilities (which is the premise of their own studies), they typically focused on proposing new prompting methods rather than analyzing how prompts were used in existing evaluations of ChatGPT's translation performance.

2.2 ChatGPT-related translation studies

ChatGPT, though not specifically designed for translation, has demonstrated extraordinary translation abilities and in essence has become an emerging translation tool. As a result, ChatGPT-related translation studies have gained momentum. Current research primarily falls into two categories: theoretical explorations and empirical studies.

Theoretical explorations usually take a more abstract or higher-level perspective, focusing on assessing the potential impact and challenges raised by ChatGPT or, more broadly, by generative AI and large language models on various aspects of translation, including translation studies (Hu and Li 2023; Yu and Liu 2024), translation education (Deng and Liu 2024), translation practice (Zhong 2024), translation systems (Gao and Ren 2023), translation ethics (Wu and Chen 2023; Yu and Guo 2024),

and the translation industry (Cui 2025). These studies also provide recommendations to address these challenges.

Empirical studies, on the other hand, apply ChatGPT to specific translation tasks and evaluate its performance and features, and they are the focus of this study's analysis. Some researchers examined the effectiveness of generative AI in translating different types of text, such as Chinese discourse (Wen and Tian 2024) and subtitles (Calvo-Ferrer 2023), revealing both its strengths and weaknesses. Others have explored the capabilities of ChatGPT in translating specific languages, including low-resource languages of the Belt and Road Initiative countries (Hou et al. 2024), Classical Latin, Ancient Greek, and Classical Sanskrit (Ross 2023), and Middle Polish (Klamra et al. 2023). Meanwhile, Yu (2024) analyzed the lexical diversity and syntactic complexity of ChatGPT translations using news texts, providing new insights into its translation capabilities.

In summary, while existing ChatGPT-related translation studies have examined its influence and effectiveness in various tasks, there remains a noticeable gap in systematically assessing how prompts were employed and evaluated in current research. This gap calls for further investigation into the adequacy and effectiveness of prompt usage, which this study aims to address.

3 Research design

Building on the research questions outlined in Section 1, this section presents the research design, including the procedures for paper collection and selection, the evaluation criteria, and the approach to data analysis.

3.1 Paper collection and selection

Studies focusing on evaluating the translation performance of ChatGPT, or containing such evaluation as a major part were eligible, regardless of their discipline, evaluation method or outcome. Measures were taken during the selection process to ensure the inclusion of high-quality papers, which will be detailed below.

The papers were collected following the four phases of the PRISMA methodology: (1) identification, (2) screening, (3) eligibility, and (4) inclusion (Moher et al. 2009). The entire process is documented in Figure 1. The following paragraphs provide additional clarification for each step of the process.

1) **Identification:** Relevant literature was searched across two databases – WOS and CNKI – to include studies around the world. For WOS, the search scope was the Web of Science Core Collection (all editions), using the search string

"Topic = 'ChatGPT' AND 'translat*'", with a wildcard character to capture different forms of the word "translate". For CNKI, the search scope included CSSCI, Peking University Core Journals, and CSCD databases, which are considered high-quality sources. The search string used was "Topic = 'ChatGPT' AND '翻译[1]'". Additionally, a full-text search was conducted in CNKI,[2] with the search string "Topic = '大语言模型[3]' + 'AI' AND '翻译', Full text = 'ChatGPT'" in case relevant papers used general terms like "large language model" or "AI" rather than "ChatGPT" in the topic but actually evaluated ChatGPT's translation performance. The initial search, conducted on June 12, 2024, generated a total of 232 records (CNKI: 88; WOS: 144). 25 duplicate articles were removed before screening, leaving 207 to be screened.

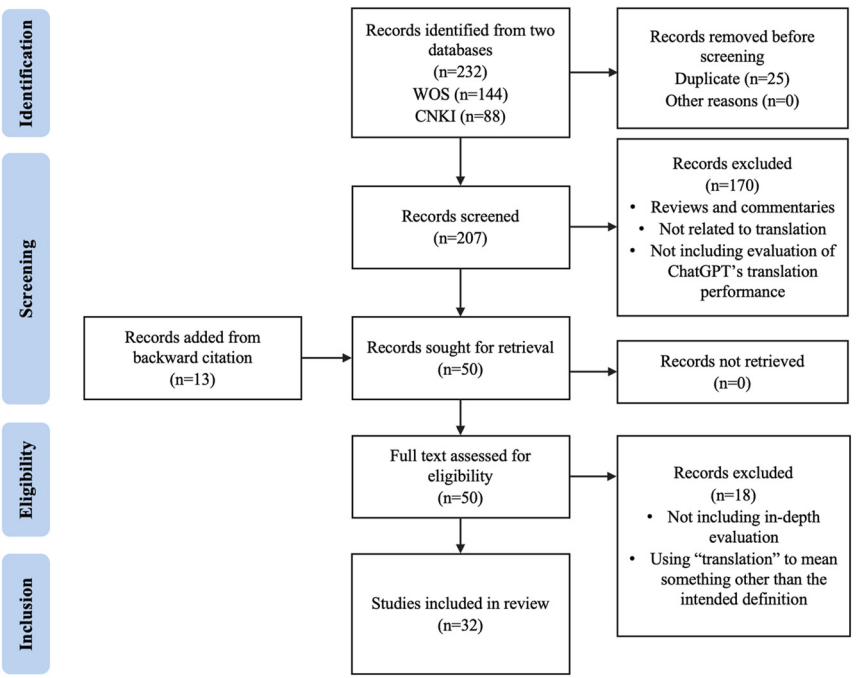


Figure 1: PRISMA flow diagram.

1 翻译 means translation/translate in Chinese.
2 WOS doesn't support full text search.
3 大语言模型 means large language model in Chinese.

2) **Screening:** Titles and abstracts of the literature were screened to assess their relevance, resulting in the removal of 170 records as shown in Figure 1. Meanwhile, an additional 13 records were identified through backward citation, as they were highly referenced by relevant and high-quality papers. This led to a total of 50 records for further analysis.
3) **Eligibility:** Full texts of the 50 records were retrieved and assessed to determine their eligibility. 18 records were excluded at this stage.
4) **Inclusion:** Ultimately, 32 studies were included for final analysis.

3.2 Evaluation criteria

From the review and analysis in previous sections, it is evident that prompts are a critical variable influencing the performance of ChatGPT. Therefore, the use of prompts in evaluation of ChatGPT's translation performance must adhere to well-established principles of experimental design to ensure meaningful, reproducible, and interpretable results. Drawing on foundational works in experimental methodology (Montgomery 2017; Seltman 2018), several criteria are proposed for assessing the consideration of prompts in existing studies:

1) **Identification and articulation of variables.** Researchers should clearly identify prompts as a key experimental variable and explicitly explain their potential influence on the research outcome. This transparency is essential for contextualizing findings and elucidating the role prompts play in shaping the results.
2) **Explicit design and justification of prompts.** Studies should provide detailed descriptions of the prompt design and justify the rationale behind their selection. This ensures the audience's understanding of the experimental process, facilitates replication, and aligns with the principle of transparency in scientific research.
3) **Alignment of prompt design with research objectives.** The chosen prompts should be suitable for the specific research objectives, with appropriate prompting strategies applied. This alignment ensures that the experimental setup effectively addresses the research questions and minimizes confounding factors.
4) **Integration of prompt influence in findings.** The influence of prompts should be explicitly analyzed and reported as part of the study's conclusions. This inclusion aligns with the principle of comprehensive reporting, which enhances the validity and reliability of the findings.

These criteria ensure the rigor and trustworthiness of an evaluation, and will serve as the foundation for the analysis in this study. To operationalize these principles, a set of assessment questions were formulated as listed in Table 1, drawing on the works of Hemkens et al. (2018) and Munkholm et al. (2020) who used standardized questions to assess whether a given factor (confounding bias) that was supposed to have impact on the research results had been adequately considered in existing studies. These questions were designed in accordance to the criteria above, addressing different aspects of prompt consideration and providing a structured

Table 1: Questions to evaluate prompt consideration.

No.	Questions
1	Did the authors make any mentions of prompts? Example for "yes": *"ChatGPT is an intelligent chatbot developed by OpenAI that builds on InstructGPT, a model designed to provide detailed responses to prompts."* (Calvo-Ferrer 2023)
2	Did the authors give any introduction to or review of the role of prompts in utilizing ChatGPT? Example for "yes": *"The style of prompts may affect the quality of translation outputs. For example, how to mention the source or target language information matters in multilingual machine translation models, which is usually solved by attaching language tokens."* (Jiao et al. 2023)
3	Did the authors examine the impact of prompts on the translation performance of ChatGPT in their evaluation? Example for "yes": *"We observe that the translation quality with few-shot in context learning can surpass that of strong encoder-decoder MT systems, especially for high-resource languages."* (Moslem et al. 2023)
4	Did the authors specify the prompts they used in the evaluation? If yes, (1) did they provide the rationale for the prompt design? If yes, how? (2) did they leverage any prompting tactics? If yes, how? Example for "yes": *"In the case of GPT-3.5, the prompt 'Traduce de español ainglés' [Translate from Spanish into English] was used for it to act as a translator without finetuning the results through a more accurate prompt – a path currently under study in the wider project where this experiment develops."* (Sanz-Valdivieso and López-Arroyo 2023) Example for "rationale given": "We consider the following three prompting strategies for GPT-3.5 that allow us to compare the model's abilities to translate with and without discourse-level context (see *Table 3 for templates and Appendix B for the exact prompts).*" (Karpinska and Iyyer 2023) Example for "prompting tactics leveraged": *"We use few-shot prompting, in which a model is provided with a prompt consisting of five demonstrations. We manually curate the five demonstrations from literary texts for each of the 18 language pairs, resulting in 90 total demonstration examples. "* (Karpinska and Iyyer 2023)
5	Did the authors reference prompts in their conclusions? Example for "yes": *"Further, we only compare GPT translations in the standard zero-shot and few-shot settings and it is quite conceivable that more specific & verbose instructions could steer the LLMs to produce translations with different characteristics."* (Raunak et al. 2023)

framework for evaluating. Each question is accompanied by illustrative examples drawn from the studies under review.

Two key points warrant further clarification:

First, Question 1 serves as a threshold to determine whether a study meets the minimum level of prompt consideration. It might seem basic but is necessary as a starting point. Naturally, studies that address the aspects covered in Question 2 to 5 will also receive an affirmative answer to Question 1, though the reverse is not necessarily true. Notably, these questions are not meant to be mutually exclusive or strictly parallel. Taken together, they provide a whole picture of prompt consideration in the studies reviewed, while each question individually serves as an important indicator in its own right.

Secondly, while Questions 1 to 5 collectively assess the overall degree of prompt consideration, Question 4 differs from the others. It specifically examines how prompts were utilized, which is a key aspect of this study, and includes 2 sub-questions: one addressing how researchers justified their selection of prompts, and the other exploring how prompting techniques were applied to meet research objectives. The analysis in the following sections will use the five main questions as indicators of overall prompt consideration, and focus on Question 4 to examine prompt usage in greater detail.

3.3 Data analysis

First, to examine how existing studies have considered prompts, each paper was systematically analyzed using the set of questions outlined in Table 1. The results were analyzed from two perspectives: (1) the overall level of prompt consideration, measured by the coverage of the five primary questions in Table 1; and (2) the specification and use of prompts, which were analyzed by focusing specifically on how Question 4, along with its two sub-questions, was addressed.

The second part of the analysis focuses on disciplinary differences in prompt consideration. For the purpose of this analysis, disciplines refer to the categories under which journals, preprint platforms, or conferences classify articles. This part starts by summarizing the distribution of examined papers across different disciplines, then compares the overall level of prompt consideration within the identified disciplines, and finally investigates how the specification and use of prompts differ across them.

Based on the data analysis, Section 5 will summarize the key findings from the analysis of the existing research and offer recommendations to improve research rigor, in response to the third research question.

4 Results

This section presents the results of the data analysis. It begins with an examination of how the selected paper, without regard to discipline, have considered prompts – both in terms of the overall level of prompt consideration, and the ways in which prompts were specified and used. This is followed by a comparison of prompt consideration across different disciplines.

4.1 Overall patterns of prompt consideration

4.1.1 Overall level of prompt consideration

Figure 2 presents the overall level of prompt consideration in the 32 articles examined through the five evaluation questions outlined in Table 1.

Figure 2: Coverage of aspects of prompt consideration.

The result indicates that the overall prompt consideration in the examined articles is relatively low and unevenly distributed across different aspects. The most frequently addressed aspect – mention of prompts – appears in only 68.75 % of the studies, meaning that 31.25 % of the articles show no indication of prompt awareness at all. Among the five aspects, only two – mention of prompts and specification of prompts used – are addressed in more than half of the studies, while the remaining three aspects are notably underrepresented, with coverage rates around 20 %–30 %. Notably, introducing or reviewing the role of prompts, assessing their impact on ChatGPT's translation performance, and integrating their influence into findings likely require more intentional effort and specialized knowledge than merely mentioning and specifying the prompts. This pattern suggests that researchers who acknowledge prompts in their studies tend to focus on more explicit and surface-level aspects, whereas deeper exploration of prompt usage remains limited, leaving a substantial gap in ChatGPT-related research.

It is also worth noting that even when an aspect is addressed, its consideration often lacks depth. For example, in *Lexical Diversity and Syntactic Complexity in ChatGPT Translation* (Yu 2024), the author mentioned in the introduction that

although ChatGPT was not specifically designed for translation, its translation abilities can be brought out with accurate prompts – an observation that is very accurate, insightful and aligned with the central idea of this study. However, this point was not further elaborated upon in the subsequent sections. Instead, the study relied on a fixed prompt, "translate from Chinese to English", for all tests, without examining how different prompts might affect ChatGPT's translation performance. This trend is particularly evident in the analysis of prompt specification and usage, as discussed in Section 4.1.2.

Figure 3 provides a more detailed view of how the examined studies addressed different aspects of prompt consideration.

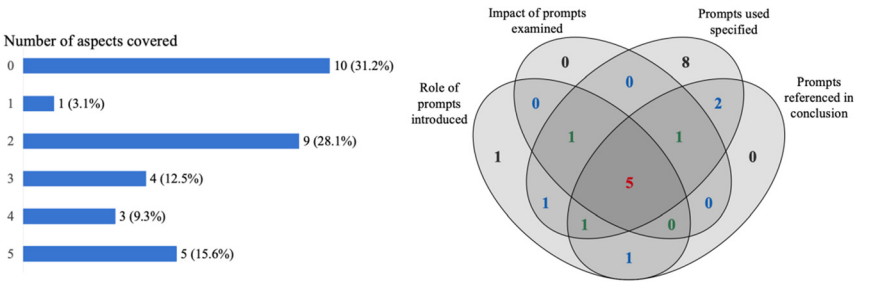


Figure 3: Distribution and overlap of articles by prompt consideration aspects.

The bar chart on the left displays the number of articles covering 0 to 5 aspects of prompt consideration, while the Venn diagram on the right illustrates the overlaps among these aspects. Each oval in the Venn diagram represents one aspect outlined by Questions 2 to 5. Question 1 is excluded as it serves as the minimal threshold for consideration and encompasses the other four aspects. The numbers in the diagram indicate how many studies share the characteristics represented by each overlapping area. Different shades and font colors are used to indicate varying levels of intersection.

Figure 3 shows that comprehensive coverage of aspects of prompt consideration is rare: only five articles (15.6 %) addressed all five aspects. Most studies only covered a subset of these aspects: three addressed four aspects, four addressed three, nine addressed two, one addressed only a single aspect, and ten did not cover any aspect at all.

Moreover, the overlaps among various aspects are limited and highly uneven. The largest area of intersection involves eight articles that both mentioned and specified prompts (some of which may have only specified prompts, as specifying is a subset of mentioning). In contrast, overlaps across other combinations are scattered.

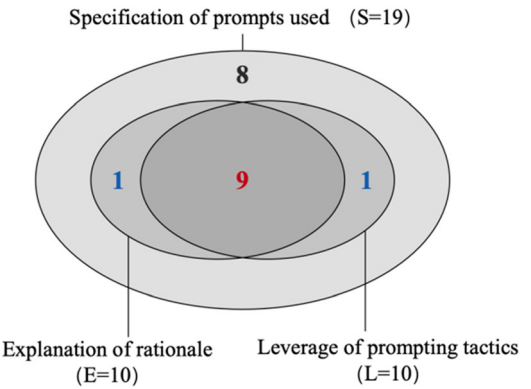
Notably, among the five papers that addressed all five aspects of prompt consideration, three are from the field of computer science and information technology, authored by researchers from institutions such as Microsoft (Hendy et al. 2023) and Tencent (Jiao et al. 2023), rather than scholars in translation studies, indicating a disciplinary disparity which will be further explored in Section 4.2.

4.1.2 Specification and use of prompts

Figures 4 and 5 visualize the specification and use of prompts in the reviewed articles. Figure 4 is a bar chart showing the number of articles that specified the prompts used, explained the rationales behind prompt selection, and leveraged prompting tactics, respectively. Figure 5 uses a Venn diagram to illustrate the overlaps among these aspects, accompanied by a summary table that explains key counts in selected intersecting areas.

Among the 19 articles that specified prompts, prompt usage appears limited and falls into two contrasting patterns. 8 out of the 19 articles neither justified their prompt selection nor employed any prompting tactics, relying solely on plain, straightforward prompts such as "translate from language A to language B". In contrast, the remaining 11 articles demonstrated greater attention to prompt usage: most of them (9 out of 11) both provided a rationale for their prompt selection and applied prompting techniques, while the other two addressed only one of these two aspects.

Figure 4: Number of articles by prompt specification and usage aspects.



Key intersection counts

E∩L	9
E∪L	11
E-(E∩L)	1
L-(E∩L)	1
S-(E∪L)	8

Figure 5: Overlap among aspects of prompt specification and usage.

Sections 4.1.2.1 and 4.1.2.2 present how the selected papers justified their selection of prompts and applied prompting tactics to meet research objectives. These are the two key aspects that reflect how prompts were utilized, corresponding to the two sub-questions of Question 4 listed in Table 1.

4.1.2.1 Rationale for prompt selection

Figures 6 and 7 indicate how examined studies used rationales for prompt selection. As shown in Figure 6, the 10 articles that gave justifications for their prompt selection drew on four types of rationale, each used with similar frequency. Figure 7 illustrates the overlaps among these rationale types. It suggests that prompt choices were often based on limited reasoning: of the 10 studies, six relied on a single justification approach, four combined two rationales, and none addressed more than two. The

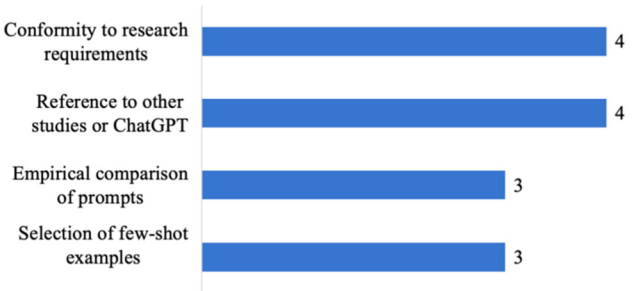


Figure 6: Number of articles by prompt selection rationale types.

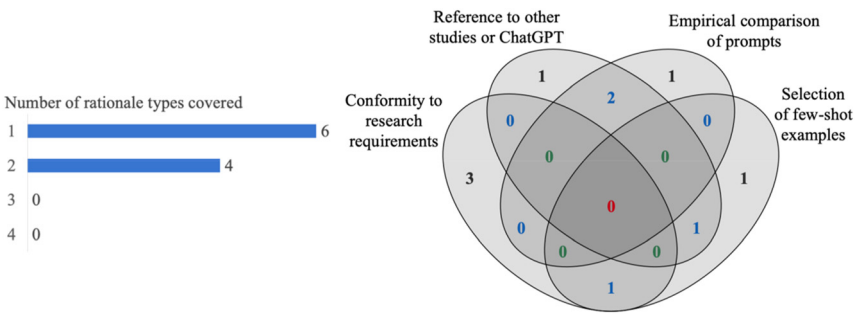


Figure 7: Distribution and overlap of articles by rationale types.

combinations of different rationale types were scattered, with no prevailing justification pattern.

Each rationale type provides a reasonable basis for prompt selection by offering methodological rigor, empirical support, or alignment with prior research. However, many justifications remain relatively simplistic and lack a strong connection to the specific characteristics of the research.

1) **Conformity to research requirements.** Four papers justified their prompt selection based on specific methodological or technical needs, ensuring alignment with research objectives. However, these connections were generally weak or indirect. For instance, Klamra et al. (2023) selected few-shot examples in their prompt to avoid format problem, as they required ChatGPT to generate responses in JSON format for easier parsing. While this rationale was valid, it was not directly related to the core evaluation task. Moreover, their justification only addressed the selection of few-shot examples, overlooking the prompt template design.
2) **Reference to other studies or ChatGPT.** Four papers relied on prompts derived from prior research or recommendations, or directly asked ChatGPT to provide prompts. This approach has the advantage of building on existing knowledge or leveraging ChatGPT's own capabilities to generate task-relevant prompts, potentially reducing bias in manual selection. However, in most cases, the researchers only briefly mentioned the source without analyzing the source itself. For example, Jiao et al. (2023) instructed ChatGPT to generate "ten concise prompts or templates that can make you translate" without specifying the task or translation requirements. Consequently, ChatGPT produced a set of generic prompts, such as "Translate this sentence from English to French", with some even being repetitive. These prompts lacked any advanced prompting techniques and were unlikely to yield significantly different results.
3) **Empirical comparison of prompts.** Three papers compared multiple prompts and selected the best-performing one. While this approach added some empirical support to their rationale, the initial selection of prompts was often questionable. For instance, Hou et al. (2024) tested three prompt templates before selecting the most effective one for a few-shot evaluation. However, the three prompts – "Translate X into Chinese", "Translate Lao sentence X into Chinese" and "Lao: X; Chinese:" – were simplistic and highly similar, with no clear explanation for why these specific variations were chosen.
4) **Selection of few-shot examples.** Three papers focused on explaining the selection of examples used in few-shot prompting. While this is a highly relevant

consideration, their discussions overlooked the design of the actual prompts themselves, leaving a critical aspect of prompt engineering unaddressed.

4.1.2.2 Leverage of prompting tactics

As noted earlier, among the 19 articles that specified prompts, only 10 employed any form of prompting tactics, while the remaining articles relied solely on plain and straightforward prompts. In terms of the use of prompting tactics, five distinct approaches with varied frequency were identified, as shown in Figure 8. Figure 9 illustrates the overlaps among these tactics, which are minimal, with non-zero counts appearing only along the peripheries.

Few-shot prompting was the most commonly used approach, appearing in five studies. In this method, researchers provided ChatGPT with examples – ranging from 1 to 15 shots – to facilitate its learning for translation tasks. Four studies adopted context provision, where researchers either supplied ChatGPT with the text to be translated within a broader context or provided background information such as the text's source and target audience, enhancing

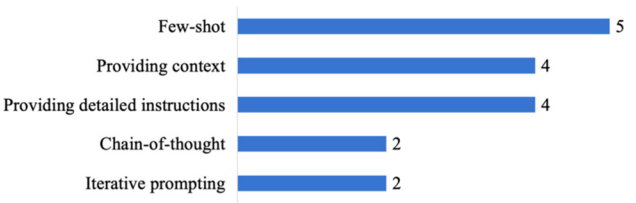


Figure 8: Number of articles by prompting tactics leveraged.

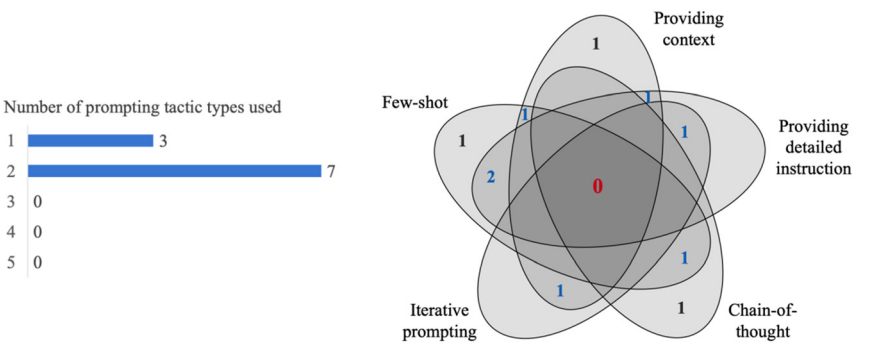


Figure 9: Distribution and overlap of articles by prompting tactics. Note: For clarity, areas in the Venn diagram without numeric labels indicate zero values and have been left unmarked to enhance visual readability.

its context awareness in translation tasks. Four studies used detailed instruction, specifying requirements such as output format, style, and terminology use. Two studies applied chain-of-thought prompting, guiding ChatGPT to break down the translation process into sequential steps. Finally, iterative prompting was utilized in two studies, involving multiple rounds of interaction with ChatGPT that provided feedback to refine its performance.

While these studies explored key prompting tactics relevant to translation tasks, their implementation appears to be somewhat simplistic and not fully optimized. Two observations can be made:

1) **Limited combination of tactics.** While most of the articles (7 out of 10) used two tactics, the combination was mostly limited to pairing few-shot prompting with another tactic. Combinations involving more than two tactics were absent. Additionally, in some cases, articles that employed two tactics used them separately in different tests, rather than integrating them organically.
2) **Rigid application of tactics.** The way each tactic was employed tends to be basic and rigid. For instance, Calvo-Ferrer (2023) employed iterative prompting by responding to ChatGPT with a fixed prompt – "It sounds awful. Please make it sound more natural in Spanish and make sure the conversation makes sense" – regardless of ChatGPT's actual output, rather than providing targeted feedback based on its responses.

4.2 Variations in prompt consideration across disciplines

4.2.1 Distribution of papers across disciplines

As shown in Table 2, the examined articles span various disciplines, indicating that academic interest in ChatGPT's translation performance extends beyond the field of translation. While translation and linguistics account for the largest share (46.88 %), a significant portion (28.12 %) of studies originate from computer science and information technology. The remaining 25 % of papers come from diverse disciplines, with no single field dominating the share, each exploring ChatGPT's translation capabilities within their respective domains.

This distribution suggests that translation and linguistics, along with computer science and information technology, are the primary disciplines investigating ChatGPT's translation performance. Meanwhile, the presence of research across other fields highlights the growing interdisciplinary engagement of diverse disciplines with ChatGPT's translation applications.

Table 2: Distribution of articles by disciplines.

Discipline	*N*	(%)
Translation and linguistics	15	(46.88)
Computer science and information technology	9	(28.12)
Others	8	(25.00)
Medicine	2	(6.25)
Multidisciplinary studies	2	(6.25)
Chemistry	1	(3.13)
Education	1	(3.13)
Environment	1	(3.13)
Libraries	1	(3.13)

4.2.2 Overall level of prompt consideration

It is evident from Figure 10 that articles in computer science and information technology exhibited the highest level of prompt consideration in most aspects, except for the specification of prompts used, where they surpass those in translation and linguistics but fall behind other disciplines by about 20 %. Furthermore, studies in computer science and information technology demonstrated a more evenly distributed consideration of prompts across different aspects, without major shortcomings, in contrast to the dramatic fluctuations observed in the other two groups. In most aspects, over half of the articles from computer science and information technology exhibited consideration.

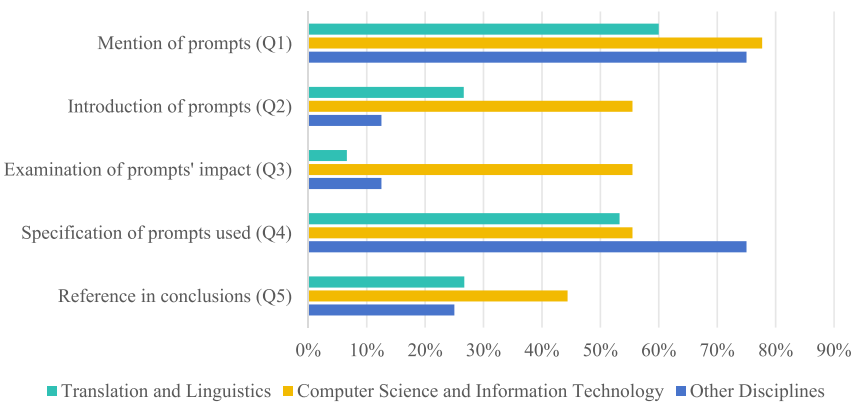


Figure 10: Comparison of overall level of prompt consideration across disciplines.

As mentioned preliminarily in Section 4.1.1, among the five papers that demonstrated the most comprehensive consideration of prompts, three are from the field of computer science and information technology, one is from translation and linguistics, and the remaining one is from the field of chemistry. Taking into account the overall distribution of articles across these disciplines, this further supports the advantage shown by articles from computer science and information technology.

When comparing articles from translation and linguistics and those from other disciplines, no definitive difference emerges, as each group excels in certain aspects. In fact, their distribution patterns closely resemble each other: apart from the similarly sharp fluctuations, they follow a comparable trend in addressing different aspects, characterized by relatively high levels of consideration in mention of prompts and specification of prompts used, and lower levels of consideration in other aspects (and very low levels of consideration in the examination of prompts' impact). This pattern contributes to the overall trend observed in the analysis of all 32 articles discussed in Section 4.1.1.

4.2.3 Specification and use of prompts

As shown in Figure 11, though a smaller proportion of articles in computer science and information technology specified the prompts used, these articles exhibited higher levels of consideration in both explanation of rationale and leverage of prompting tactics. In fact, all articles in computer science and information technology that specified prompts also provided a rationale, and employed at least one prompting tactic, resulting in identical levels of consideration across these aspects. In contrast, quite a number of articles in translation and linguistics and other

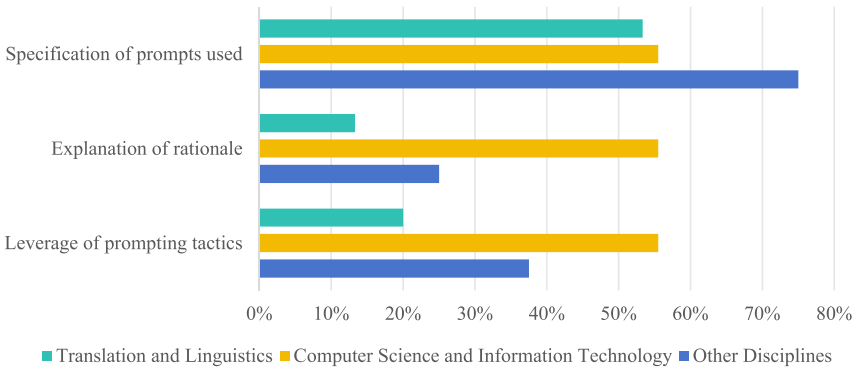


Figure 11: Comparison of specification and use of prompts across disciplines.

disciplines specified prompts but did not explain why and/or employ any prompting tactics.

A comparison between articles from translation and linguistics and those from other disciplines reveals similar patterns, with the latter exhibiting higher levels of consideration in both explanation of rationale and leverage of prompting tactics.

Furthermore, the articles from computer science and information technology not only present the most diverse rationales to justify the selection of prompts, but also cover the widest range of prompting tactics, suggesting a more intentional control over prompt use and greater familiarity with the available toolkit among researchers in this field.

5 Discussion

This section begins by identifying key observations from the data analysis, which are general under-emphasis on prompts and disparity across research disciplines, and explores their underlying causes and implications. It subsequently offers recommendations for future research based on these findings.

5.1 General under-emphasis on prompts

The survey reveals a consistent under-emphasis on prompts in the reviewed articles, both in terms of the overall prompt consideration and the way prompts were used.

First, the overall level of prompt consideration in the examined articles is relatively low. More than 30 % of the surveyed articles did not mention the concept of prompts at all. Among those that did, most focused only on a narrow subset of aspects, with very few offering a comprehensive consideration of all relevant dimensions. Even when an aspect was addressed, the discussion often lacked sufficient depth.

Second, the way prompts were used is largely limited and often lacks justification. Among the 19 articles that specified the prompts used – which is already a rather small fraction – nearly half relied solely on plain and straightforward prompts such as “Translate from language A to language B”. For the remaining studies, the justification for prompt selection often lacked a strong connection to research objectives. Additionally, the prompt usage was characterized by limited combinations and rigid applications.

From many perspectives, ChatGPT was used in these articles in much the same way as Google Translate, where users select the source and target languages, input

the text, and await the output – following a rigid, predefined workflow with limited variability. While prompts were mentioned, they often seemed to be included merely as a procedural step in the methodology rather than a focal point for critical examination. As a result, ChatGPT's potential as a generative AI remained largely underexplored.

The ways in which prompts are considered and used can be seen as a reflection of users' cognitive models, which reveals how they conceptualize generative AI and how they attempt to communicate intent to them (Shen and Yu 2025). With the rapid development of generative AI, prompt literacy – the ability to effectively craft and adopt prompts to achieve intended goals – has emerged as an important extension of digital literacy (Hwang 2023; Korzynski et al. 2023; Zhang and Jia 2024). While academic interest in prompt design and usage is rapidly growing, studies have highlighted that uses' actual prompting behaviors in real-world contexts remain underexplored, and are often insufficiently strategic in practice (Zhao et al. 2025). This aligns with the findings of the present study, where even researchers – presumably more informed users – frequently failed to make full use of ChatGPT's capabilities. These findings underscore the need for researchers to have more comprehensive consideration of prompts in ChatGPT studies, which is supported by strengthened prompt literacy, and more broadly, by a deeper understanding of AI and how to interact with them.

5.2 Disparity across research disciplines

The analysis reveals a significant imbalance in the consideration of prompts across various disciplines.

In terms of both overall prompt consideration and prompt usage, articles in computer science and information technology exhibited the highest, most balanced and most comprehensive consideration. This suggests that researchers in these fields may share a more established understanding of prompt's impact on ChatGPT's performance and how to effectively use it.

On the other hand, articles from translation and linguistics, as well as those from other disciplines – both lagging behind studies from computer science and information technology – showed no significant difference from each other. Both groups exhibited dramatic fluctuations across different aspects of prompt consideration. Specifically, they demonstrated relatively high attention to more explicit and surface-level aspects including mentioning and specifying prompts, but had considerably lower engagement in other aspects like examining prompts' impact. This disciplinary pattern largely contributes to the overall uneven distribution of prompt consideration observed in the articles surveyed.

Such pattern unfolds against the broader backdrop of generative AI increasingly blurring the boundaries between the humanities and technological fields (Chen 2024). Scholars in the humanities – including those in translation and linguistics – are actively engaging with technological tools and expanding their academic horizons by exploring new research questions and adopting innovative methods (Hu and Li 2023; Yu and Liu 2024). While these efforts are both timely and commendable, the comparison reveals differences in both methodological approaches and underlying thinking paradigms between studies conducted by researchers in computer science and information technology and those in other disciplines, even when addressing very similar research questions. To foster more meaningful interdisciplinary outcomes, it is essential for researchers in translation and linguistics to develop a more foundational and up-to-date understanding of the underlying technologies. As will be further discussed in Section 5.3.2, such understanding can be effectively enhanced through deeper interdisciplinary collaboration.

5.3 Recommendations for future research

5.3.1 More comprehensive consideration of prompts in ChatGPT studies

As outlined in Section 1, a key question is how to improve the consideration of prompts in future research. This study seeks to address this by proposing a set of practical recommendations, drawing on (1) existing research on prompt engineering, as discussed in Section 2; (2) the four criteria for prompt consideration put forth in Section 3.2; and (3) the analysis of the strengths and limitations of current studies, as examined in Section 4. These recommendations are not only intended for future evaluations of ChatGPT's performance, but may also offer broader insights for other ChatGPT-related translation studies.

To ensure practical applicability, the recommendations are structured into three phases – before, during and after the evaluation – following the typical stages of a research.

5.3.1.1 Before the evaluation

1) **Understanding ChatGPT and prompts.** Researchers should develop a solid understanding of ChatGPT's functionality, its distinctions from previous AI systems, and the role of prompts in shaping its performance.
2) **Familiarizing themselves with established research.** Researchers should review existing studies on prompt design for both general and translation-specific purposes. Keeping up with emerging research is also essential, as prompt engineering is a rapidly evolving field.

5.3.1.2 During the evaluation

1) **Clarification of the role of prompts.** Researchers need to clarify whether the study will compare ChatGPT's performance under different prompt conditions or evaluate it using a fixed prompt. If the latter, strong justifications should be provided, along with an analysis of the impact and limitations of the approach.
2) **Preliminary design of prompts.** At this stage, the general prompting strategy should be established. The design can draw from one or a combination of three sources: relevant prior studies with high applicability to the evaluation task; ChatGPT-generated suggestions based on clearly defined task requirements; and researcher-crafted prompts tailored to the study's objectives, constraints, and characteristics.

   To guide prompt selection, Table 3 outlines the application and advantages of some of the most commonly used prompting tactics in translation tasks.

Table 3: Application and strengths of some most commonly used prompting tactics in translation tasks.

Prompting tactic	Application in translation	Strengths
Providing detailed instructions	Clearly specify output requirements, such as formality, tone, terminology, and structure, etc. Example: For medical translations, instruct ChatGPT to use standardized terminology from a specific medical database	Ensures clarity, consistency, and strict adherence to terminology. Ideal for specialized texts like technical, legal and academic translations, where precision is paramount
Few-shot	Provide ChatGPT with a few high-quality and relevant translation examples Example: When translating classical poetry, provide well-crafted translations that reflect the desired style	Helps ChatGPT internalize complex styles and domain-specific conventions, especially useful when the style and tone desired cannot be explicitly defined through instructions (e.g., in literary or creative translation)
Providing context	1. Include the source text within a broader context Example: Provide the proceeding and following paragraphs along with the text to be translated 2. Provide background information such as the text's source, domain purpose, and target audience, etc.	Enhances ChatGPT's contextual awareness, reducing errors from ambiguity in isolated sentences. Particularly useful for texts with rich context
Chain-of-thought	1. Simply add "Let's think step by step" to encourage logical reasoning	Facilitates logical reasoning and deeper comprehension, which is particularly useful for culturally adaptive translations,

Table 3: (continued)

Prompting tactic	Application in translation	Strengths
	2. Specify translation steps explicitly Example: For highly flexible literary text, ask ChatGPT to first interpret the meaning, then paraphrase naturally in the target language	complex sentence restructuring, and meaning preservation. However, it may have negative influence on the translation of straightforward texts
Iterative prompting	Use multiple rounds of interaction to refine the translation Example: When translating marketing materials, review the initial output and request adjustment for greater engagement or cultural relevance	Allows progressive refinement, making it effective for creative and adaptive translation tasks. However, it may slow down the workflow in high-volume translation projects

Two key points are worth noting:
First, the use and combination of prompting tactics can be highly flexible – much like natural conversation with a human. Researchers are encouraged to explore and experiment with different approaches as long as they contribute to improved translation performance.
Second, while it is impossible to exhaust all possible prompting strategies, it's critical for researchers to make the most of the resources available, for doing so not only enhances translation quality but also advances the understanding of how to use ChatGPT effectively.

3) **Adjustment of prompts.** Even when using the same prompting tactics, the specific wording of prompts can influence ChatGPT's performance. Researchers can experiment with variations and compare the results. Additionally, they can ask ChatGPT to generate optimized versions of the prompts for further refinement.

5.3.1.3 After the evaluation

As the final product, the paper should thoroughly reflect the researchers' consideration of prompts to enhance the comparability and validity of the research. This includes clearly identifying the role of prompts; explicitly designing and justifying the prompts; ensuring that the prompts align with the research objectives; and integrating the influence into the findings.

5.3.2 More interdisciplinary collaboration to enhance exchange of insights

As highlighted in Section 5.2, there is a disparity in prompt consideration across disciplines, including between translation and linguistics, and computer science and information technology. While differences in disciplinary focus are natural, fostering more interdisciplinary collaboration is essential for enriching the exchange of insights and promoting a more effective integration of computer science and translation studies. This aligns with the recommendations put forward by Wang and Wang (2025), as well as Zong (2024). Strengthening such collaboration would benefit not only ChatGPT-related translation studies, but interdisciplinary translation studies at large. Based on this, the following two recommendations are proposed for researchers in translation studies.

5.3.2.1 Within specific studies

1) **Engage with computer science researchers early.** Discuss the study's core ideas, experimental design, and key terminology with experts in computer science and information technology. This minimizes the risk of misunderstanding during the research process.
2) **Involve computer science experts throughout the study.** Ideally, include computer science and information technology researchers as part of the research team. This provides an opportunity for them to contribute valuable technical insights while gaining a better understanding of how their technologies are applied in translation contexts.
3) **Collaborate on the final paper.** Before submission, have computer science and information technology experts review the final paper to ensure technical accuracy and clarity, addressing potential conceptual errors or misunderstandings.

5.3.2.2 Beyond specific studies

1) **Engage with relevant literature from computer science and information technology regularly.** Translation researchers should make it a habit to stay updated with key literature from computer science and information technology. This continuous engagement helps bridge the gap between technical developments in AI and the evolving needs of translation studies.
2) **Form interdisciplinary teams for long-term projects.** The formation of interdisciplinary research teams that span both fields should be encouraged, allowing for deeper collaboration on long-term projects. These teams can tackle more complex, cross-disciplinary research questions, with both translation and computer science perspectives shaping the direction of the studies.
3) **Promote sustained communication between fields.** Researchers from translation studies and computer science should establish ongoing channels for communication, such as regular meetings, conferences, or collaborative

publications. This ensures that both fields remain informed about each other's progress, leading to more effective integration of AI technologies in translation studies and continuous refinement of collaborative methodologies.

6 Conclusions

This study examines the consideration of prompts in evaluations of ChatGPT's translation performance through a survey of 32 articles. The results indicate that prompt consideration in the reviewed articles is generally inadequate. The overall level of prompt consideration across different aspects is relatively low, and the ways prompts were used are quite limited. Additionally, there is a noticeable disparity in the consideration of prompts across disciplines, with articles in computer science and information technology demonstrating the most comprehensive, balanced and thorough approaches.

Based on these findings, the study offers a set of practical recommendations to enhance prompt consideration in future studies. These include strategies to enhance consideration before, during, and after evaluations, as well as suggestions for fostering interdisciplinary collaboration to promote the exchange of insights, both within and beyond specific studies. The study addresses an underexplored gap in ChatGPT-related translation studies by highlighting the inadequate consideration of prompts, and providing targeted and actionable recommendations, thus facilitating the more comprehensive and robust development of studies in the field.

However, this study has several limitations. The primary one is its relatively small sample size of 32 papers, partly due to the short interval between ChatGPT's public release and the completion of this research. Additionally, it focuses solely on ChatGPT, as there were very few studies evaluating other emerging large language models at that time. Moreover, the articles were primarily sourced from CNKI and the WOS core collection, potentially excluding valuable papers from lower-impact journals. Expanding the scope to include more studies on different language models and from a broader range of sources will allow future research to explore prompt consideration in greater depth.

Furthermore, while this paper identifies some differences in how prompts are considered, it would be valuable to examine these differences across more dimensions and in greater detail. For example, comparing the treatment of prompts in Chinese literature versus international articles, or examining how varying consideration of prompts may lead to different conclusions in these studies, could offer deeper insights. Future studies could focus on these aspects to uncover larger patterns in the adoption of prompts, and provide more meaningful implications for academic research.

References

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33. 1877–1901.

Calvo-Ferrer, José Ramón. 2023. Can you tell the difference? A study of human vs machine-translated subtitles. *Perspectives* 32(6). 1115–1132.

Che, Wanxiang, Zhicheng Dou, Yansong Feng, Tao Gui, Xianpei Han, Baotian Hu, Minlie Huang, Xuanjing Huang, Kang Liu, Ting Liu, Zhiyuan Liu, Bing Qin, Xipeng Qiu, Xiaojun Wan, Yuxuan Wang, Jirong Wen, Rui Yan, Jiajun Zhang, Min Zhang, Qi Zhang, Jun Zhao, Xin Zhao & Yanyan Zhao. 2023. 大模型时代的自然语言处理: 挑战、机遇与发展 [Towards a comprehensive understanding of the impact of large language models on natural language processing: Challenges, opportunities and future directions]. *Scientia Sinica Informationis* 53(9). 1645–1687.

Chen, Qiuxin & Zeqi Qiu. 2023. “人机互生” 时代可供性理论的契机与危机—基于 “提示词” 现象的考察 [Opportunities and challenges of affordance theory in the age of human-machine symbiosis: An examination based on the phenomenon of prompts]. *Journal of Soochow University Philosophy & Social Sciences Edition* 44(5). 172–182.

Chen, Yuehong. 2024. 变革与创新: 人工智能的迭代发展与人文学科的未来 [Change and innovation: Iterative development of artificial intelligence and the future of the humanities]. *Frontiers* 5. 85–97.

Cui, Qiliang. 2025. 人工智能时代的语言服务行业发展趋势 [Development trends of the language services industry in the age of artificial intelligence]. *Journal of Beijing International Studies University* 47(1). 62–72.

Deng, Juntao & Wan Liu. 2024. AIGC 时代的翻译教学资源建设: 变革、挑战与对策 [Development of translation teaching resources in the AIGC era: Transformation, challenges and solutions]. *Beijing Journal of Translators* 2. 33–48.

Gao, Yuxia & Dongsheng Ren. 2023. 生成式 AI 时代翻译制度建设的挑战与对策 [Challenges and countermeasures for constructing translation institutions in the generative AI era]. *Technology Enhanced Foreign Language Education* 45(4). 9–15+114.

Gu, Wenshi. 2023. Linguistically informed ChatGPT prompts to enhance Japanese-Chinese machine translation: A case study on attributive clauses. arXiv preprint arXiv:2303.15587. https://doi.org/10.48550/arXiv.2303.15587.

He, Sui. 2024. Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. arXiv preprint arXiv:2403.00127. https://doi.org/10.48550/arXiv.2403.00127.

Hemkens, Lars, Hannah Ewald, Florian Naudet, Aviv Ladanie, Jonathan G. Shaw, Gautam Sajeev & John P. A. Ioannidis. 2018. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. *Journal of Clinical Epidemiology* 93. 94–102.

Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify & Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? A comprehensive evaluation. arXiv preprint arXiv:2302.09210. https://doi.org/10.48550/arXiv.2302.09210.

Hou, Yutao, Abudukelimu Abulizi Abulizi, Yaqing Shi, Mayilamu Musideke & Halidanmu Abudukelimu. 2024. 面向 “一带一路” 的低资源语言机器翻译研究 [Research on low-resource language machine translation for the “Belt and Road”]. *Computer Engineering* 50(4). 332–341.

Hu, Kaibao & Xiaoqian Li. 2023. 大语言模型背景下翻译研究的发展: 问题与前景 [Development of translation studies in the context of large language models: Issues and prospects]. *Chinese Translators Journal* 44(6). 64–73+192.

Hwang, Y. 2023. The emergence of generative AI and PROMPT literacy: Focusing on the use of ChatGPT and DALL-E for English education. *Journal of the Korea English Education Society* 22(2). 263–288.

Jacobs, Hayes & Michael Fisher. 2023. Prompt literacy: A key for AI-based learning. https://www.ascd.org/el/articles/prompt-literacy-a-key-for-ai-based-learning (accessed 22 May 2025).

Jiao, Hui, Bei Peng, Lu Zong, Xiaojun Zhang & Xinwei Li. 2024. Gradable ChatGPT translation evaluation. arXiv preprint arXiv:2401.09984. https://doi.org/10.48550/arXiv.2401.09984.

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi & Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv preprint arXiv:2301.08745. https://doi.org/10.48550/arXiv.2301.08745.

Karpinska, Marzena & Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. arXiv preprint arXiv:2304.03245. https://doi.org/10.48550/arXiv.2304.03245.

Klamra, Cezary, Katarzyna Kryńska & Maciej Ogrodniczuk. 2023. Evaluating the use of generative LLMs for intralingual diachronic translation of Middle-Polish texts into contemporary Polish. In *Proceedings of the international conference on Asian digital libraries*. Singapore: Springer Nature Singapore.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo & Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems* 35. 22199–22213.

Korzynski, Pawel, Grzegorz Mazurek, Pamela Krzypkowska & Artur Kurasinski. 2023. AI prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review* 11(3). 25–37.

Li, Geng, Zishuo Wang, Xiangteng He & Yuxin Peng. 2023. 从 ChatGPT 到多模态大模型: 现状与未来 [From ChatGPT to large multimodal model: Present and future]. *Scientia Sinica Informationis* 37(5). 724–734.

Lo, Leo S. 2023. The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship* 49(4). 102720.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman & PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine* 151(4). 264–269.

Montgomery, Douglas C. 2017. *Design and analysis of experiments*. Hoboken: John Wiley & Sons, Inc.

Moslem, Yasmin, Rejwanul Haque, Kelleher, John D. Kelleher & Andy Way. 2023. Adaptive machine translation with large language models. arXiv preprint arXiv:2301.13294. https://doi.org/10.48550/arXiv.2301.13294.

Muktadir, Golam Md. 2023. A brief history of prompt: Leveraging language models (through advanced prompting). arXiv preprint arXiv:2310.04438. https://doi.org/10.48550/arXiv.2310.04438.

Munkholm, Klaus, Maria Faurholt-Jepsen, John P. A. Ioannidis & Lars G. Hemkens. 2020. Consideration of confounding was suboptimal in the reporting of observational studies in psychiatry: A meta-epidemiological study. *Journal of Clinical Epidemiology* 119. 75–84.

Njifenjou, Ahmed, Virgile Sucal, Bassam Jabaian & Fabrice Lefèvre. 2024. Role-play zero-shot prompting with large language models for open-domain human-machine conversation. arXiv preprint arXiv:2406.18460. https://doi.org/10.48550/arXiv.2406.18460.

OpenAI. 2023. Prompt engineering. https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results (accessed 22 May 2025).
Prahallad, Lavanya & Radhika Mamidi. 2024. Significance of chain of thought in gender bias mitigation for English-dravidian machine translation. arXiv preprint arXiv:2405.19701. https://doi.org/10.48550/arXiv.2405.19701.
Qiu, Linan & Qianlian Gu. 2023. 从情境搭建到情境再分离: 人机传播中用户与 ChatGPT 的互动实践 [From contextual construction to contextual re-separation: User and ChatGPT interaction practices in human-machine communication]. *Chinese Editors Journal* 10. 91–96.
Raunak, Vikas, Arul Menezes, Matt Post & Hany Hassan Awadalla. 2023. Do GPTs produce less literal translations? arXiv preprint arXiv:2305.16806. https://doi.org/10.48550/arXiv.2305.16806.
Ross, Edward A. S. 2023. A new Frontier: AI and ancient language pedagogy. *Journal of Classics Teaching* 24(48). 143–161.
Sanz-Valdivieso, L. & López-Arroyo, B. 2023. Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation. In *Proceedings of the International Conference Human-informed Translation and Interpreting Technology*. Naples, Italy: HiT-IT.
Seltman, Howard J. 2018. Experimental design and analysis. https://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf (accessed 22 May 2025).
Shen, Yang & Menglong Yu. 2025. 重构智能交互范式: 基于 DeepSeek 的提示强化与人机协同 [Reconstructing a paradigm of human-AI interaction: Human-AI collaboration based on DeepSeek]. *Journal of Xinjiang Normal University (Philosophy and Social Sciences)* 46(4). 90–98.
Song, Binghui. 2024. 新技术语境下翻译文化研究何为? [Rethinking translation culture studies in the context of new technologies]. *Journal of Foreign Languages* 47(1). 10–13.
Trust, Torry. 2023. Essential considerations for addressing the possibility of AI-driven cheating, part 2. https://www.facultyfocus.com/articles/teaching-with-technology-articles/essential-considerations-for-addressing-the-possibility-of-ai-driven-cheating-part-2/ (accessed 22 May 2025).
Wang, Chutong. 2024. 人工智能时代的翻译、出版与传播: 问题与展望 [Issues and prospects of translation, publication and communication in age of AI]. *University Publishing* 4. 79–91.
Wang, Zhiguo & Shan Wang. 2025. 翻译研究的数字人文面相—以 ChatGPT 为中心的考察 [The digital humanities' landscape of translation studies—centered on ChatGPT]. *Foreign Language Research* (3). 33–40.
Wei, Jason, Xuezhi Wang, Schuurmans, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le & Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35. 24824–24837.
Wen, Xu & Yaling Tian. 2024. ChatGPT 应用于中国特色话语翻译的有效性研究 [The effectiveness of ChatGPT in translating China-specific discourse text]. *Shanghai Journal of Translators* 2. 27–34+94–95.
Wu, Junhong, Zhao Yang & Chengqing Zong. 2023. ChatGPT 能力分析与未来展望 [Analysis of ChatGPT's capabilities and future prospects]. *Bulletin of National Natural Science Foundation of China* 37(5). 735–742.
Wu, Meixuan & Hongjun Chen. 2023. 人工智能时代机器翻译的伦理问题 [Ethical Issues of Machine Translation in the Era of Artificial Intelligence]. *Foreign Language Research* (6). 13–18.
Yamada, Masaru. 2023. Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. arXiv preprint arXiv:2308.01391. https://doi.org/10.48550/arXiv.2308.01391.
Yu, Guoming & Fan Li. 2024. 生成式 AI 浪潮下平台型媒体的规则重构、价值逻辑与生态剧变 [The rule reconstruction, value logic, and ecological changes of platform media under the wave of generative AI]. *Journal of Suzhou University (Philosophy and Social Science)* 45(1). 167–175.

Yu, Hao & Yunyun Guo. 2024. 风险与超越: ChatGPT 赋能翻译的伦理分析 [Risk and beyond: An Ethical analysis of ChatGPT's empowerment in translation]. *Chinese Translators Journal* 45(4). 115–122.
Yu, Jing & Kanglong Liu. 2024. 重塑翻译研究: AI 技术影响下的范式转换与未来方向探索 [Reshaping translation studies: Paradigm shifts and future directions in the age of AI technology]. *Journal of Foreign Languages* 47(4). 72–81.
Yu, Lei. 2024. ChatGPT 翻译的词汇多样性和句法复杂度研究 [Lexical diversity and syntactic complexity in ChatGPT translation]. *Foreign Language Teaching and Research* 56(2). 297–307+321.
Zhang, Guixiang & Junzhi Jia. 2024. 生成式 AI 时代下的提示素养培育研究 [Research on prompt literacy cultivation in the era of generative AI]. *Journal of Academic Libraries* 6. 63–71.
Zhang, Wenyu & Bi Zhao. 2024. 生成式人工智能开创机器翻译的新纪元了吗? 一项质量对比研究及对翻译教育的思考 [Has generative AI opened a new era for machine translation? – A contrastive quality study and reflections on translation education]. *Journal of Beijing International Studies University* 46(1). 83–98.
Zhao, Shuang. 2023. 生产、交互与传播: 生成式人工智能的媒介可供性分析—以ChatGPT为例 [Production, interaction and communication: An analysis of media affordance in generative artificial intelligence—a case study of ChatGPT]. *Public Communication of Science & Technology* 15(14). 69–72.
Zhao, Yuxiang, Yutian Jing, Shijie Song & Wei Liu. 2025. AIGC 赋能的提示素养: 生成式 AI 时代的人智交互能力重构 [Prompt literacy for AIGC empowerment: Reconstructing human-AI interaction capabilities in the generative AI era]. *Information and Documentation Services* 46(3). 14–25.
Zhong, Houtao. 2024. 生成式人工智能给翻译实践带来的机遇与挑战 [On the AIGC's challenges for the translation and the breakthrough strategy]. *Beijing Journal of Translators* 2. 238–250.
Zong, Chengqing. 2024. 从人工智能角度谈语言类学科发展和人才培养 [On the development of language disciplines and talent cultivation from the perspective of artificial intelligence]. *Foreign Languages and Their Teaching* 5. 1–11+145.

Bionote

Yu Zhang
Shanghai International Studies University, Shanghai, China
623436107@qq.com
https://orcid.org/0009-0005-2396-5977

Yu Zhang is an incoming PhD student in Translation Studies at Shanghai International Studies University, Shanghai, China. Her areas of interest include translation practice and teaching in the context of generative AI, as well as interdisciplinary approaches to translation studies.