ENHANCING CHINESE-ENGLISH TRANSLATION IN AI CHATBOTS: A COMPARATIVE EVALUATION OF CHATGPT-4O AND GROK-BETA USING A HEALTH SCIENCE TEXT FROM THE NEW YORK TIMES

Wang Wei[1] and Zhou Weihong[2]

[1]School of Interpreting and Translation, Beijing International Studies University, Beijing, China
[2]Department of College English Education, Beijing City University, Beijing, China

ABSTRACT

The present study examines the effectiveness of contextual prompting, utilizing a universal prompting template for translation tasks, and revision prompting in enhancing the quality of Chinese-to-English translations of scientific texts. ChatGPT-4o and Grok-beta were employed as the AI translation models. The research utilized a New York Times article on the health benefits of sweet potatoes, along with its official Chinese translation, as the source material. Translation quality was evaluated using BLEU metrics complemented by qualitative measures, including accuracy, faithfulness, fluency, genre consistency, and terminology consistency, which are critical for assessing translations in science and technology domains. Statistical analysis indicated only marginal improvements with the use of second-stage prompting, which involved commands for review and revision. These findings raise questions about the reliability of BLEU scores as a sole evaluation metric. The study highlights the potential of AI-assisted translation for specialized genres while identifying notable discrepancies in chatbot outputs. Based on the findings, the study underscores the need for refined methodologies in evaluating translation quality and advocates for integrating more robust qualitative metrics in future research to enhance the reliability and applicability of AI-assisted translation in specialized contexts.

KEYWORDS

AI-assisted translation; contextual prompting; BLEU metric; qualitative evaluation; Chinese-English translation; health science texts

1. INTRODUCTION

The emergence of AI chatbots has revolutionized the field of translation, introducing unprecedented levels of speed, accessibility, and efficiency [1-2]. These advancements have made machine translation a powerful tool across various domains. However, evaluating the performance of AI-driven translation systems, particularly for specialized genres such as health science texts, presents significant challenges. These texts often demand a high degree of accuracy, terminological precision, and stylistic consistency, which can be difficult for AI systems to achieve consistently. The current research seeks to address these challenges by assessing the translation quality of two prominent AI chatbots, ChatGPT-4o and Grok-beta, through the application of contextual and revision prompting techniques. Contextual prompting involves guiding the models with specific instructions to enhance the relevance and quality of their outputs, while revision prompting focuses on enabling the models to self-evaluate and refine their translations. To provide a comprehensive evaluation, the analysis integrates both

Bibhu Dash et al: CCITT, EDUTEC, NLDML 2025
pp. 11-25, 2025. IJCI – 2025 DOI:10.5121/ijci.2025.140102

quantitative and qualitative metrics. Quantitative assessment is conducted using BLEU (Bilingual Evaluation Understudy) scores, which measure lexical overlap with reference translations. However, recognizing the limitations of BLEU in capturing nuanced aspects of translation, the study also employs qualitative evaluation criteria, including accuracy, faithfulness, fluency, genre consistency, and terminology consistency. Through this dual approach, the research aims to identify effective practices for AI-assisted translation in specialized fields and to offer insights into the capabilities and limitations of current chatbot models. The dissemination of health-related information across linguistic barriers is crucial for global health literacy. With advancements in AI, machine translation has become an indispensable tool for translating complex scientific content. This study investigates whether AI-driven translations can accurately convey the nuanced health benefits of sweet potatoes from a translated Chinese version back to English, focusing on the effectiveness of AI models in maintaining the scientific integrity and clarity of the original content.

2. LITERATURE REVIEW

Machine translation (MT) has undergone significant advancements with the development of large language models (LLMs), such as ChatGPT and Grok. However, these advancements have also exposed key challenges in the evaluation and refinement of translations, particularly in specialized domains such as health and science. This section reviews the current state of research on translation evaluation metrics, the role of prompt engineering, and the challenges specific to translating health science texts. The BLEU metric [3], widely regarded as the gold standard for assessing MT quality, relies on n-gram overlap between machine-generated translations and reference texts. While effective for measuring lexical similarity, BLEU has notable shortcomings in evaluating idiomatic expressions, semantic nuance, and stylistic fidelity. These limitations are particularly pronounced in specialized genres like health and science, where domain-specific terminology and consistent stylistic alignment are critical. To address these gaps, alternative metrics such as METEOR, ROUGE, and TER have been developed. Although they incorporate additional aspects, such as synonymy and recall, their applicability to complex, nuanced texts remains limited. Furthermore, these metrics fail to capture qualitative dimensions like fluency, cultural appropriateness, and genre fidelity, which are essential for health-related translations. As a result, researchers have increasingly called for hybrid evaluation frameworks that combine quantitative metrics with qualitative assessments, such as accuracy, faithfulness, fluency, genre consistency, and terminology consistency [4].

Prompt engineering has proven instrumental in leveraging the potential of LLMs for a wide range of applications. As research progresses, innovations in prompt design and evaluation, combined with advancements in LLM architectures, will continue to expand the applicability and precision of these transformative technologies. Prompt engineering has become a critical technique for enhancing the performance of Generative Pretrained Transformer (GPT) models and other large language models (LLMs). By designing task-specific input prompts, users can direct these models to produce outputs that align closely with desired outcomes. The rapid evolution of prompt engineering reflects the ongoing quest to maximize the precision, adaptability, and efficiency of LLMs across various domains. Prompting strategies have undergone significant development since the inception of GPT and similar models. Early approaches primarily involved: 1) zero-shot prompting: requesting the model to perform tasks without prior examples; 2) few-shot prompting: including a limited number of examples in the prompt to guide the model's response. These foundational techniques demonstrated the adaptability of LLMs for tasks ranging from translation, post-editing, to creative writing [5],[6], [7], [8],[9]. More advanced methods, such as Chain-of-Thought (CoT) prompting, emerged to address complex reasoning tasks like mathematical problem-solving and logical inference. By structuring prompts to encourage step-by-step reasoning, CoT significantly improves the model's capacity for higher-

order reasoning [6], [10]. This highlights the role of structured prompts in enhancing task-specific performance. Recent research emphasizes the need for tailored prompts suited to specific contexts or domains. For instance, in professional translation, incorporating information about the target audience or purpose of the translation has been shown to improve output quality significantly [11]. Other techniques, such as persona matching, metadata inclusion, and temperature adjustments, enable refined outputs by aligning the model's behavior with the task's requirements [7], [11]. The concept of in-context learning has further advanced prompt engineering. By embedding task-specific examples within the prompt, models can dynamically adapt to new tasks without requiring additional fine-tuning. This technique effectively transforms the prompt into an ad hoc fine-tuning mechanism [7]. Despite its success, prompt engineering faces several challenges: 1) ambiguity in prompt design: crafting clear and task-specific prompts remains difficult, particularly for complex or nuanced tasks [6], [11]; 2) model sensitivity to phrasing: the phrasing of a prompt can significantly affect the model's output, underscoring the importance of precise wording; 3) context limitations: including extensive contexts within prompts often leads to performance degradation, necessitating more efficient input representations [7]; 4)evaluation metrics: while human evaluation is the gold standard, automating prompt evaluation using metrics like semantic similarity scores is an area of active research (Ibid). Improvements in LLM architecture and pretraining methodologies have also contributed to the effectiveness of prompt engineering. Scaling model parameters and training datasets enhances generalization, thereby increasing the efficacy of task-specific prompting techniques [6], [8], [9]. The field of prompt engineering is evolving, with ongoing research focusing on several key areas including automated tools for generating effective prompts to democratize access to this technology, robustness to reduce model sensitivity to variations in prompt phrasing, enhancing model performance across diverse applications.

Health and science texts present unique challenges for MT systems due to their high semantic and terminological demands: 1) semantic and terminological accuracy: Scientific texts rely on precise terminology to ensure clarity and avoid misinterpretation. Misaligned translations of technical terms can compromise the credibility of the text and lead to serious consequences, particularly in health communication [4]; 2) idiomatic and cultural nuances: Health information often includes idiomatic expressions and culturally specific metaphors that are challenging for MT systems to render accurately. For instance, idioms like "an apple a day keeps the doctor away" require not just linguistic translation but also cultural contextualization to maintain meaning and impact [12]; 3) genre and style consistency: Maintaining consistent terminology, tone, and style throughout the translation is crucial for readability and reliability. In journalistic and scientific genres, these features enhance the text's credibility and engagement, which is vital for specialized audiences [13]. The rise of advanced AI models like ChatGPT and Grok has provided promising solutions to many of these challenges. These systems leverage enhanced learning algorithms and vast multilingual training datasets to improve translation quality:1) enhanced learning algorithms: State-of-the-art deep learning techniques enable better pattern recognition and contextual understanding. For example, transformer-based architectures allow AI models to capture long-range dependencies in text, improving their ability to handle complex linguistic phenomena [14]; 2) large training datasets: Exposure to massive, domain-specific bilingual corpora has significantly improved the models' ability to translate health and scientific texts accurately. These datasets enhance the models' grasp of specialized terminology and nuanced expressions [15]. Prompt engineering has emerged as a pivotal technique for optimizing LLM performance in translation tasks. Early prompt engineering methods relied on simple command prompts to elicit desired outputs. However, as translation tasks have become more complex, more sophisticated strategies have been developed: 1) contextual prompting involves crafting prompts that provide detailed instructions or additional context to guide the model's output. For instance, specifying genre requirements (e.g., journalistic tone or scientific rigor) or emphasizing the importance of accuracy and fluency can significantly improve translation quality [16]. Contextual prompting

has proven especially effective in aligning the output with the expectations of specialized audiences; 2) while contextual prompting focuses on the initial output, revision prompting encourages models to self-assess and refine their translations. This iterative process aims to enhance aspects such as semantic accuracy, idiomaticity, and genre fidelity. Despite its potential, the efficacy of revision prompting remains underexplored, particularly for texts requiring precise terminology and cultural adaptation.

Although previous studies have demonstrated the efficacy of contextual prompting in improving MT performance [14], [16], the potential of revision prompting as a complementary strategy has received little attention. Additionally, while advanced AI models show promise in handling domain-specific challenges, inconsistencies in their outputs and limitations in evaluation metrics continue to hinder their application in critical fields like health and science. This study builds on these findings by examining the interplay between contextual and revision prompting in enhancing translation quality. It also explores the limitations of BLEU as a standalone evaluation metric and advocates for the integration of qualitative measures. By focusing on health science texts, the research contributes to a deeper understanding of the strengths and weaknesses of AI-assisted translation in specialized genres.

3. METHODOLOGY

The methodology employed in this study was designed to assess the translation performance of ChatGPT-4o and Grok-beta on a specialized text, combining quantitative and qualitative metrics to ensure a comprehensive and objective evaluation. The following steps outline the process in detail:

3.1. Source Text Selection

To ensure linguistic complexity and the inclusion of domain-specific terminology, the current research has utilized a Chinese text (“紅薯有多健康？”) detailing the health benefits of sweet potatoes. This report was itself a translated version of the original article titled “How Healthy Are Sweet Potatoes?”, published in *The New York Times* [17]. The dual-language origin of the text presented a unique opportunity to evaluate translation models on material that required careful handling of nuanced expressions, idiomatic phrases, and specialized terms. By selecting this text, the study aimed to simulate a challenging and authentic scenario typical of health and science communication, providing a rigorous and contextually rich test case for evaluating the performance of the translation models.

3.2. Translation Models and Prompting Strategies

Two state-of-the-art AI translation models—ChatGPT-4o and Grok-beta—were utilized to translate the Chinese text back into English, employing a two-stage prompting strategy designed to assess and enhance translation quality. The process included the following steps:

Each model was provided with a carefully crafted contextual prompt explicitly instructing them to prioritize accuracy, fluency, and genre fidelity. This prompt aimed to capture the models’ ability to generate high-quality initial translations by emphasizing the need for precise terminology, seamless readability, and adherence to the stylistic conventions of the health and science genre. In light of text typology theory, we have introduced a standardized template for specific translation task by fusing constant and fixed commands (accuracy, faithfulness, naturalness, idiomaticity) and variable commands (which is determined by text types and language types). Thus the BASE TEMPLATE is composed like this:

Prompt {Translate the following [TEXT TYPE] text into accurate, faithful, fluent, natural, and comprehensible [X (Language)]. [ADDITIONAL GUIDELINES RELATED TO TEXT TYPE].} Considering the source text (which actually is a Chinese version of the English report) is a health report, the initial prompting design was composed like this:

CONTEXTUAL PROMPT {We now have a report on the health benefits of sweet potatoes in Mandarin Chinese. Please analyze the source text and translate it into ACCURATE, FAITHFUL, NATURAL, IDIOMATIC, JOURNALISTIC and SCIENTIFIC English, paying special attention to genre and terminology consistency.}

After completing the initial translation, a second prompt leveraged the AI chatbots' memory capabilities to encourage self-assessment and iterative refinement of their outputs. This step required the models to critically analyze their translations, focusing on identifying and addressing potential issues such as semantic inaccuracies, misinterpretation of idiomatic expressions, and inconsistencies in stylistic tone or genre alignment. The revision prompt emphasized improving alignment with the domain-specific requirements of the health and science genre, ensuring clarity, precision, and fluency in the revised translations. The second prompting design was briefly composed like this:

PROMPT {Now please compare the English version you just generated and the source text, revise the version to make it more ACCURATE, FAITHFUL and EUIVALENT to the Chinese source text.}

This two-step prompting framework provided a structured approach to evaluate the interplay between contextual and revision prompts in enhancing translation quality. By incorporating a dedicated revision phase, the study not only tested the models' ability to generate high-quality initial translations but also their capacity to refine and improve their outputs through iterative processing. This approach offered valuable insights into the effectiveness of revision-based strategies, demonstrating the potential of AI models to produce polished, accurate, and genre-consistent translations in specialized domains.

3.3. Evaluation Metrics

To comprehensively assess the performance of the AI translation models, the study employed both quantitative and qualitative evaluation metrics, ensuring a balanced and multidimensional analysis of translation quality.

The Bilingual Evaluation Understudy (BLEU) metric was employed as a quantitative measure to evaluate the similarity between AI-generated translations and the original English text. This metric calculates n-gram overlap, capturing both lexical and structural correspondence, and was implemented using Python-based algorithms for precision. BLEU scores provided an objective, numerical assessment of translation quality, offering a straightforward method to gauge alignment between the source and target texts. However, BLEU's reliance on surface-level matching has well-documented limitations. It struggles to capture deeper aspects of translation quality, such as semantic nuance, contextual accuracy, idiomatic expressions, and adherence to stylistic conventions. Despite its utility, BLEU has notable limitations: BLEU's reliance on exact word or phrase matches means it often misses. BLEU does not capture how well the translation conveys the intended meaning or nuances of the original text. BLEU overlooks the importance of context, where words or phrases might need to be translated differently based on their surrounding text. The BLEU metric struggles with idioms and expressions that do not have direct translations. It does not evaluate for stylistic adherence, which can be crucial for the translation's readability and cultural relevance. In fields like health and science, where

terminology and precision are critical, BLEU's shortcomings become more pronounced. Here, translations need to be not just accurate but also clear, understandable, and culturally appropriate. These shortcomings are particularly pronounced in specialized texts, such as health and science articles, where precise terminology and natural fluency are critical. Recognizing these constraints, the current research has incorporated complementary qualitative evaluation metrics to provide a more holistic and nuanced assessment of translation performance. This integrated approach ensured that both the measurable accuracy of the translations and their contextual, semantic, and stylistic fidelity were rigorously analyzed.

A rigorous framework of qualitative metrics was designed to evaluate critical aspects of translation quality, with a focus on both linguistic precision and contextual fidelity. These metrics were tailored to address the nuanced demands of translating health and science texts, emphasizing the following dimensions:

Table 1. Qualitative metrics.

Dimensions	**Explanation**
Accuracy	The extent to which the translation preserved the exact meaning of the source text, ensuring that no content was omitted, distorted, or misrepresented. Accuracy was especially crucial given the factual and technical nature of the content.
Faithfulness	The ability of the translation to maintain the original text's intent, central themes, and subtle nuances. In health and science domains, where precision and clarity are paramount, preserving the integrity of the message was a key focus.
Fluency	The naturalness, coherence, and grammatical correctness of the translated text, evaluated to ensure it met the standards of human-authored content. This criterion assessed whether the output was easily readable and free from linguistic errors.
Genre consistency	Adherence to the stylistic and tonal conventions characteristic of health journalism. This included maintaining a professional, authoritative voice while aligning with the conventions of journalistic reporting in the health domain.
Terminology consistency	The accuracy, precision, and uniformity in the translation of domain-specific terms. Special emphasis was placed on health and scientific terminology to prevent ambiguity or miscommunication, which could undermine the text's credibility.

To enhance the objectivity and reliability of these qualitative evaluations, 19 independent AI models (either through the direct link to the official websites: https://chatgpt.com/; https://x.com/i/grok?focus=1; or AI chatbot hubs: https://lmarena.ai/) were utilized to independently score the translations across the defined dimensions. This diverse panel of evaluators provided a broad spectrum of assessments, mitigating potential biases inherent in single-model evaluations. The mean scores for each criterion were calculated to capture aggregated performance trends, ensuring consistency and robustness in the evaluation process. By combining the objective measurement capabilities of BLEU scores with the nuanced insights from qualitative evaluations, the study adopted an integrative approach to translation assessment. BLEU scores quantified lexical and structural alignment, offering a numerical baseline for performance. Meanwhile, qualitative metrics provided a more comprehensive analysis, capturing the contextual, stylistic, and semantic fidelity of the translations. This combined methodology not only facilitated a thorough evaluation of the AI models' ability to translate complex, domain-specific texts but also illuminated areas where quantitative and qualitative measures intersected or diverged. Ultimately, this approach allowed for a more holistic understanding of the strengths and limitations of AI-assisted translation in the health and science domains.

3.4. Statistical Analysis

The current study has employed a rigorous and multifaceted statistical methodology to evaluate and compare the performance of AI translation models, ensuring robustness, precision, and replicability. The primary statistical methods applied include the following:

The BLEU (Bilingual Evaluation Understudy) metric served as a key tool to evaluate the lexical and structural alignment between AI-generated translations and the reference English text. By quantifying similarities in word choice and phrase structures, BLEU provided an objective measure of translation quality. To assess the influence of the revision prompt, BLEU scores were calculated for both the initial and revised translations across different AI models. Comparing these scores allowed for a systematic analysis of how effectively the self-assessment phase enhanced translation accuracy and fluency. This quantitative benchmarking was instrumental in determining whether incorporating a revision step led to measurable improvements in the translation outputs.

To complement BLEU's focus on n-gram correspondence, qualitative metrics were employed to provide a more holistic evaluation of translation performance. These metrics encompassed key dimensions such as ''accuracy'' (the degree to which the meaning aligns with the source text), ''faithfulness'' (adherence to the source text's content without unwarranted additions or omissions), ''fluency'' (naturalness and grammatical correctness in the target language), ''genre consistency'' (alignment with stylistic norms of specific text types), and ''terminology precision'' (correct use of domain-specific terms). Scores for each criterion were generated by 19 independent AI models, which offered diverse perspectives and minimized individual bias. The mean scores across all models were calculated for both initial and revised translations, allowing for a robust comparison. This aggregated analysis revealed patterns of improvement, particularly in areas such as fluency and genre consistency, and provided insights into how self-assessment and contextual prompts impacted linguistic and contextual fidelity across multiple datasets.

To explore the relationship between quantitative and qualitative metrics, a correlation analysis was conducted between BLEU scores and the aggregated qualitative metrics. This analysis investigated the degree to which BLEU scores aligned with the human-like assessments provided by qualitative measures. The results offered insights into the strengths and limitations of BLEU as a standalone metric, particularly for nuanced or specialized translations, and emphasized the need for multidimensional evaluation frameworks.

All translations, prompts, and evaluation results were carefully recorded and stored in a structured dataset. This documentation ensured full transparency and reproducibility of the research, allowing other scholars to replicate the study or apply its methodology to different genres, languages, or AI translation models. The dataset included annotated translations, evaluation scores, and metadata describing the prompts and evaluation process. The dual-language source text, along with the comprehensive evaluation framework, serves as a replicable template for future research. By providing detailed documentation and clear methodologies, this study facilitates the extension of its approach to other specialized domains, such as legal, technical, or medical translations. The inclusion of both contextual and revision prompting strategies further enriches the methodology, enabling researchers to explore iterative translation processes in a variety of contexts. This methodological framework integrates quantitative rigor with qualitative depth, offering a balanced and nuanced approach to evaluating AI translation systems. The combination of BLEU scores with rich qualitative metrics captures both measurable accuracy and contextual, stylistic, and semantic fidelity. Furthermore, the analysis of iterative prompting strategies underscores the potential for AI models to refine their outputs dynamically, a feature of increasing importance for specialized, high-stakes domains such as health and

science. By combining these advanced evaluation techniques, the study provides a robust foundation for assessing the capabilities of AI-assisted translation, paving the way for future advancements in translation technologies and their application to complex linguistic tasks.

4. TRAINING DATA ANALYSIS

4.1. BLEU Scores Analysis

The dataset for this study comprised three key components: the original New York Times (NYT) report on the health benefits of sweet potatoes, its official Chinese translation, and the English translations generated by ChatGPT-4o and Grok-beta under two prompting conditions. The first condition utilized a contextual translation prompt to guide the models in producing the initial translations. The second condition employed a revision prompt, encouraging the models to self-assess and refine their outputs. This two-step prompting strategy allowed for an examination of how iterative prompting impacts translation quality. Both quantitative (BLEU scores) and qualitative evaluations (as illustrated in Table 2) were used to assess the translations. To enhance reliability, multiple independent AI evaluators (19 models involved, as illustrated in Table 3) contributed to scoring the translations across various dimensions, and their evaluations were aggregated to minimize bias and ensure robustness.

Table 2. BLEU Scores vs. Qualitative Scores.

Models	**Versions**	**BLEU Scores**	**Qualitative scores**
ChatGPT-4o	1	32.38%	91.43%
ChatGPT-4o	2	35.12%	91.63%
Grok-beta	1	39.12%	87.01%
Grok-beta	2	33.28%	87.90%

BLEU (Bilingual Evaluation Understudy) scores were calculated to measure the lexical and structural similarity between the AI-generated translations and the original English text. BLEU scores for both the initial (Version 1) and revised (Version 2) translations were compared for each model: 1) ChatGPT-4o's marginal improvement (2.74% increase) suggests that the revision prompt helped refine the translation by addressing n-gram inconsistencies or errors. 2) While for Grok-beta, the performance decline (5.84% decrease) indicates potential limitations in Grok-beta's ability to leverage revision prompts effectively. These results demonstrate that the revision prompt benefited ChatGPT-4o slightly, whereas Grok-beta exhibited diminished performance, highlighting divergent capabilities in responding to iterative prompting strategies.

4.2. Qualitative Scores Analysis

Each dataset row contains the scores for ChatGPT-4o (Versions 1 and 2) and Grok-beta (Versions 1 and 2) for direct comparison. The datasets are organized by name and version, allowing an easy scan for performance trends. ChatGPT-4o generally shows consistent or improved performance in V2 compared to V1, except in a few datasets like "ChatGPT-4o-2024-05-13"and "llama-3.1-8b-instruct." Grok-beta exhibits mixed performance, with some datasets (e.g., "ChatGPT-4o") showing significant improvements in V2, while others (e.g., "ChatGPT-4o-2024-08-16") show declines. "Gemini-test" and "Gemini-exp-1114" exhibit the highest scores for both models, indicating a strong alignment with the task's requirements. "c4ai-aya-expanse-32b" has the lowest scores, particularly for Grok-beta V1 (78%) and Grok-beta V2 (81%), which suggests challenges with this dataset. Some datasets, like "anonymous-chatbot2" and "jade-1,"

reveal high consistency and minimal performance variation, making them benchmarks for further testing.

Table 3. Qualitative Scores Calculated by 19 AI Models.

Models	ChatGPT-4o		Grok-beta	
	Version 1	Version 2	Version 1	Version 2
ChatGPT-4o	84%	90%	82%	86%
ChatGPT-4o mini-2024-07-18	92%	91.8%	87.6%	88.2%
ChatGPT-preview-01-mini	92%	92%	87%	87%
ChatGPT-4o-2024-05-13	90.2%	88.4%	86.4%	88.4%
ChatGPT-4o-2024-08-16	88%	88%	82%	80%
Grok-beta	85%	88.6%	92%	92%
Grok-2-mini-2024-08-13	89%	90.8%	84.6%	87%
Gemini-test	96.8%	97.8%	94.6%	96.2%
Gemini-exp-1114	96%	100%	92%	96%
llama-3.1-70b-instruct	85.5%	87%	83%	86.5%
llama-3.1-8b-instruct	88.4%	86.4%	83.2%	80.8%
claude-3-opus-20240229	93%	94%	88%	88%
claude-3-5-sonnet-20241022	94.4%	96.4%	89%	88.8%
folson-03	93%	93%	88%	88%
anonymous-chatbot1	94%	89%	85%	85%
anonymous-chatbot2	96.8%	95.8%	90.8%	91.2%
jade-1	94%	95%	94%	94%
secret-chatbot	94%	90%	86%	86%
c4ai-aya-expanse-32b	91%	87%	78%	81%

Qualitative assessments focused on dimensions such as accuracy, faithfulness, fluency, genre consistency, and terminology consistency. The mean scores across these criteria were calculated for both prompting conditions. Minimal improvement (0.20% increase) indicates that while the revision prompt resulted in slightly improved performance, the gains were relatively insignificant. Although Grok-beta demonstrated modest improvements in qualitative metrics, they were insufficient to offset the decline in BLEU scores, suggesting mixed performance outcomes. The qualitative evaluation reveals that while both models benefited slightly from the revision prompt in terms of contextual and linguistic quality, the effect size was minimal. This highlights the limited impact of revision prompting in significantly enhancing translation quality, particularly for complex, domain-specific texts. The statistical findings underscore a divergence in the models' responses to the prompting strategy: 1) ChatGPT-4o exhibited modest but consistent improvements in both BLEU scores and qualitative metrics, suggesting that its architecture and training data may be better suited to iterative refinement. 2) Grok-beta, in contrast, showed a reduction in BLEU scores after the revision phase, despite minor gains in qualitative dimensions. This decline points to potential difficulties in leveraging revision prompts to enhance structural or

lexical accuracy. Additionally, the weak improvements in qualitative scores across both models suggest that while revision prompting may address minor errors, its ability to generate substantial advancements in overall translation quality remains limited. The findings call for further research into more effective prompting strategies and model-specific adaptations to optimize iterative translation processes. The results of this study have several implications: 1) Iterative Prompting Effectiveness: The marginal improvements highlight that while revision prompting can refine translations, it is not a panacea for resolving deeper linguistic or contextual challenges in AI-assisted translation. 2) Model-Specific Behavior: The contrasting performance trends of ChatGPT-4o and Grok-beta demonstrate the need to tailor prompting strategies to individual model architectures to maximize their strengths. 3) Evaluation Framework: The combination of BLEU scores and qualitative metrics provides a balanced and multidimensional framework for evaluating translation quality, addressing the limitations of using a single metric. This analysis paves the way for future research into optimizing prompting techniques and refining evaluation methodologies for domain-specific AI translations, particularly in health and science communication.

4.3. Correlation Analysis

The correlation coefficient between the BLEU scores and the qualitative evaluations is approximately -0.576 (as illustrated in Figure 1). This suggests a moderate negative correlation, indicating that as BLEU scores increase, qualitative evaluation scores tend to decrease, and vice versa. The chart above visualizes this relationship, highlighting the data points for each model version and their respective BLEU and qualitative evaluation scores.

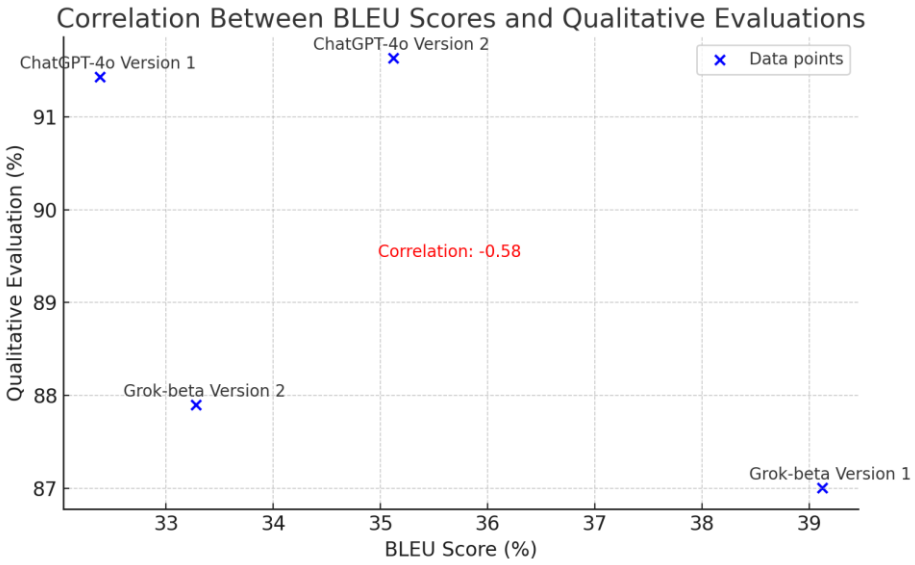

Figure 1. Coefficient between BLEU Scores and Qualitative Scores

For ChatGPT-4o, there's a perfect positive correlation, suggesting that as BLEU scores increase, so do the qualitative scores. This might indicate that for this model, the automatic metrics align well with human judgment. For Grok-beta, there's a perfect negative correlation, which is unusual. This might imply that improvements in BLEU score correspond with a decrease in qualitative scores, possibly suggesting a misalignment between automatic metrics and human preferences for this particular model's translations.

5. DISCUSSION

5.1. The Limitations of BLEU Scores in Translation Quality Evaluation

BLEU (Bilingual Evaluation Understudy) scores, while a prevalent metric in machine translation (MT) evaluation, exhibit significant limitations when assessing translation quality comprehensively. BLEU primarily focuses on n-gram overlap between the machine-generated translation and a reference text. However, this approach inherently biases translations towards surface-level similarity, often neglecting deeper nuances such as context, genre, and idiomatic expressions. In our experiment, the BLEU scores of ChatGPT-4o and Grok-beta varied across their initial and revised versions, but these differences were minimal and insufficiently reflective of substantial qualitative improvements. For example, while Grok-beta's first version had the highest BLEU score (39.12%), its qualitative assessment indicated less fluency and idiomatic accuracy than the second version, which scored lower on BLEU but improved on journalistic tone and terminological precision. This discrepancy underscores a critical issue: BLEU cannot adequately capture non-literal translations that are contextually and stylistically faithful to the source text. Scientific and journalistic texts, such as our study's sweet potato report, require fidelity not only to factual content but also to stylistic conventions that enhance readability and audience engagement. Consequently, we argue that BLEU scores are insufficient as a standalone metric, necessitating the integration of additional qualitative criteria for a holistic evaluation.

5.2. Efficacy of Contextual Prompting

The results of our study reveal that contextual prompting significantly improves translation quality for scientific and journalistic genres. By explicitly instructing the AI models to prioritize accuracy, faithfulness, fluency, and genre consistency, both ChatGPT-4o and Grok-beta produced translations that aligned more closely with the standards of journalistic and scientific writing. These improvements were reflected in higher qualitative scores in areas such as idiomaticity and terminological accuracy. For instance, the second versions generated by both models exhibited better alignment with the original NYT text's nuanced expression, even though the BLEU scores showed only marginal improvement. This finding aligns with prior research emphasizing the importance of task-specific prompting in natural language processing (NLP). Contextual prompts guide AI models to adapt their generative outputs to the stylistic and functional requirements of the target genre, mitigating common issues such as lexical awkwardness and terminological inconsistency. For scientific texts, this approach proves particularly effective, as precise and accessible language is paramount. However, the relatively small margin of improvement observed in BLEU and qualitative scores suggests diminishing returns for repeated prompting, highlighting the need for more refined prompt engineering techniques or hybrid evaluation frameworks.

5.3. Variability in Qualitative Scores among AI Models

One intriguing observation from our study was the variability in qualitative scores among different AI models. Despite identical prompts and evaluation criteria, the scores assigned to translations varied, with some models producing inconsistent or inflated evaluations. For example, ChatGPT-4o Version 1 received an average score of 84% from one evaluation model and 96.8% from another, raising concerns about the objectivity and reliability of AI-assisted scoring mechanisms. This variability points to potential issues such as divergent interpretation of evaluation metrics or algorithmic bias inherent in specific models. Even more concerning is the possibility of data plagiarism among AI chatbots. Instances where models appear to replicate evaluation data from mainstream AI systems undermine the credibility of autonomous

evaluations. This emphasizes the need for transparent and standardized evaluation protocols to ensure that assessments are both independent and reliable. Future research should explore the development of cross-model evaluation frameworks that minimize biases and enforce consistent standards.

The spider web radar analogy is apt in conceptualizing how AI systems work. Each "node" of the spider web represents a cluster of knowledge or patterns embedded in the AI model. When a prompt is issued, it activates certain "nodes" or areas, influenced by the following factors:The prompt's specific words or phrases can trigger different areas of the model's training. Neural networks prioritize certain patterns over others based on probabilities learned during training. Random sampling mechanisms determine the exact path taken within the "web," adding variability to outputs. Imagine asking for a description of "translation." A direct prompt might activate "nodes" related to linguistics, while the same query framed within a cultural context might activate nodes related to anthropology or media studies. The variability in AI chatbot responses reflects the nuanced and probabilistic nature of their design. The spiderweb radar analogy helps illustrate how different areas of the AI model's "knowledge network" are activated based on input. While randomness introduces variability, it also adds richness and flexibility, allowing AI systems to produce diverse and contextually adaptive outputs. Fine-tuning response parameters, such as temperature or prompt design, can help mitigate or leverage this variability depending on user goals.

5.4. Implications for Translation Practices and Future Research

The findings from our study have several implications for AI-assisted translation practices and research. Firstly, the marginal improvement observed in the second prompting suggests that while revision prompting can enhance quality, its utility may be limited for well-designed initial prompts. This highlights the importance of crafting comprehensive and precise initial prompts, particularly for high-stakes genres such as scientific and journalistic texts. Secondly, our results reinforce the necessity of multidimensional evaluation metrics that incorporate qualitative aspects such as fluency, idiomaticity, and genre alignment alongside quantitative scores like BLEU. For practitioners, these insights suggest that contextual prompting can effectively guide AI models to produce translations tailored to specific audiences and purposes, reducing the need for extensive post-editing. However, the potential for score variability and data manipulation in AI evaluation tools calls for increased vigilance and the adoption of robust, transparent evaluation frameworks.

6. CONCLUSION

In conclusion, contextual prompting has shown considerable promise in enhancing the quality of AI-generated translations, improving dimensions such as accuracy, fluency, and adherence to genre-specific conventions. However, its benefits are subject to diminishing returns as prompts become increasingly intricate, raising questions about scalability and efficiency. This underscores the necessity of adopting a comprehensive, multidimensional approach to evaluate its impact effectively. The limitations of BLEU scores as a standalone metric, coupled with inconsistencies in qualitative assessments among AI models, highlight the inadequacy of current evaluation paradigms. This study reaffirms the potential of contextual prompting, particularly for domains such as scientific and journalistic texts, where precision and clarity are paramount. Nonetheless, challenges remain in achieving consistent, objective, and scalable evaluation frameworks, revealing critical opportunities for further investigation and methodological refinement. A more robust foundation for future work should focus on the following key areas:

6.1. Development of hybrid evaluation metrics

Current evaluation frameworks often rely heavily on quantitative metrics like BLEU, which, while computationally efficient, fail to capture subtleties in semantic fidelity, cultural appropriateness, or user satisfaction. Future efforts should explore hybrid methodologies that integrate traditional metrics with qualitative assessments, such as human evaluation panels or linguistic profiling. Advances in automated metrics that leverage neural quality estimation tools and attention mechanisms may also provide more nuanced insights into translation performance.

6.2. Refinement of prompt engineering techniques

Contextual prompting strategies must evolve to accommodate increasingly complex linguistic phenomena, such as idiomatic expressions, ambiguous syntactic structures, and culturally nuanced meanings. This requires identifying optimal levels of specificity that maximize model responsiveness without compromising flexibility. Research should also address the unique challenges posed by low-resource languages, where limited data exacerbates variability in output quality. Techniques such as dynamic prompt generation, few-shot learning, and meta-prompting could play a critical role in overcoming these barriers.

6.3. Mitigation of model biases and systemic inconsistencies

Variability in evaluation results—arising from differences in training data, model architecture, or prompt design—poses a significant challenge to fair comparisons across systems. Future work should explore mechanisms for diagnosing and reducing biases, such as the over-reliance on memorized data or preferential treatment of high-frequency linguistic patterns. Safeguards against score inflation and redundancy in generated translations will be vital to ensure that evaluation outcomes reflect true quality improvements rather than artifacts of dataset overlap.

6.4. Benchmarking across diverse genres and use cases

While contextual prompting has proven effective in structured genres such as scientific and journalistic texts, its applicability to other domains remains underexplored. Expanding experimental frameworks to include diverse text types—such as literary, legal, or technical documents—will be essential for testing the generalizability of these techniques. These domains pose unique challenges, including the need to preserve stylistic elements, adhere to domain-specific terminologies, and balance creative and factual accuracy.

6.5. Exploration of user-centric and multilingual paradigms

Beyond genre-specific applications, future research should investigate user-centric evaluation methodologies, emphasizing the subjective experience of end-users. Studies on multilingual translation contexts—especially for underrepresented languages—could yield insights into how contextual prompting adapts across varying linguistic and cultural landscapes. Additionally, leveraging contextual prompts to enable collaborative workflows, where human translators and AI systems jointly refine outputs, represents a promising avenue for practical application.

By addressing these research priorities, the field can advance toward the development of AI translation systems that are not only more accurate and contextually aware but also equitable, transparent, and adaptable to diverse real-world scenarios. These efforts will be instrumental in shaping a future where AI-driven translation serves as a reliable bridge for global communication and cross-cultural understanding.

REFERENCES

[1] OpenAI, (2023) *GPT-4 Technical Report*. OpenAI.

[2] OpenAI, (2024) *ChatGPT-4o mini Technical Overview*. OpenAI.

[3] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu, (2002) "Bleu: a Method for Automatic Evaluation of Machine Translation". In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[4] Ghassemi, Marzyeh, Tristan Naumann, Peter Schulam, Andrew L. Beam, Irene Y. Chen, Rajesh Ranganath, (2020) "A Review of Challenges and Opportunities in Machine Learning for Health." *AMIA Jt Summits Transl Sci Proc*. 2020 May 30;2020:191-200. PMID: 32477638; PMCID: PMC7233077.

[5] Yamada, Masaru., (2023) "Optimizing Machine Translation through prompt engineering: An Investigation into ChatGPT's Customizability." *Proceedings of Machine Translation Summit XIX (September 4–8, 2023, Macau SAR, China.)*, Vol. 2: Users Track, pages 195–204.

[6] Patil, Rajvardhan & Venkat Gudivada, (2024) "A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs)." *Applied Sciences*, 2024, 14, 2074. https://doi.org/10.3390/app14052074

[7] Chen, I-Sheng, Wang, Danyang, Xu Luyi, Cao Chen, Fang Xiao, Lin Jionghao, (2024) "A Systematic Review on Prompt Engineering in Large Language Models for K-12 STEM Education." https://arxiv.org/pdf/2410.11123

[8] Wang, Wei & Zhou Weihong, (2023) "Assessment of the Translation and Post-editing of Machine Translation (MT) With Special Reference to Chinese-English Translation." *Cross-Cultural Communication*, Vol. 19, No. 4, 2023, pp. 1-9. DOI:10.3968/13178

[9] Wang, Wei & Zhou Weihong, (2024) "Optimizing Prompt Engineering in Translation Practice: A Comparative Study of ChatGPT-4.0 and ChatGPT-4o Mini" *Canadian Social Science*, Vol. 20, No. 5, 2024, pp. 12-29. DOI:10.3968/13573

[10] Wei, Jason, Wang Xuezhi, Schuurmans Dale, Bosma Maarten, Ichter Brian, Xia Fei, Ed Chi, Le Quoc, Denny Zhou, (2022) "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." arXiv preprint arXiv:2201.11903, 2022.

[11] Castaldo, Antonio & Johanna Monti (2024) "Prompting Large Language Models for Idiomatic Translation." *Proceedings of the 1st Workshop on Creative-text Translation and Technology: 32–39.* https://aclanthology.org/2024.ctt-1.4.pdf

[12] Wang, Ping, (2018) "Cultural Characteristics of Idiomatic Expressions and Their Approaches to Translation." *Journal of Literature and Art Studies*, 2018, Vol. 8, No. 2, 295-300. DOI: 10.17265/2159-5836/2018.02.016

[13] Yaseen, Heba Shaji Sa'adeh., (2013) *Terminological Inconsistency in Medical Translation from English into Arabic*. An-Najah National University, MA thesis.

[14] Mohamed, Yasir Abdelgadir, Akbar Kanna, Mohamed Bashir, Abdul Hakim Mohamed, Mousab A. E. Adiel, Muawia A. Elsadig, (2022) "The Impact of Artificial Intelligence on Language Translation: A review." IEEE Access: DOI 10.1109/ACCESS.2024.3366802

[15] Shen, Li, Sun Yan, Yu Zhiyuan, Ding Liang, Tian Xinmei, Tao Dacheng, (2023) "On Efficient Training of Large-Scale Deep Learning Models: A Literature Review." https://arxiv.org/pdf/2304.03589

[16] Brown, Tom B., Benjamin Mann, Nick Ryder, et al., (2020) "Language Models are Few-Shot Learners", arXiv:2005.14165

[17] Whitcomb, Isobel, (2024) "How Healthy Are Sweet Potatoes?". *The New York Times*. https://cn.nytimes.com/health/20241119/sweet-potatoes-health-benefits-recipes/zh-hant/dual/

AUTHORS

Wang Wei an associate professor at Beijing International Studies University, specializes in translation studies. He earned his PhD from Shanghai Jiao Tong University in 2008. His academic interests encompass historical linguistics, corpus linguistics, and contrastive linguistics, reflecting a strong focus on the interplay between language history,

computational analysis of language data, and comparative studies of linguistic structures across languages.

Weihong Zhou a senior lecturer at Beijing City University, specializes in applied linguistics and data analysis. She earned his MA from Beijing International Studies University in 2004. Her academic interests encompass applied linguistics and translation studies.