This is a preprint of the following article: Siu, S.C., "Employing ChatGPT in the Evaluation of Translation Quality" (2024).

Employing ChatGPT in the Evaluation of Translation Quality

Sai Cheong Siu

Abstract This paper aims to discuss four potential areas in which ChatGPT can assist with computer-assisted evaluation of translation: (1) checking key facts and numbers, (2) assessing grammar and naturalness, (3) identifying terminological issues, and (4) checking consistent style. It is hoped that this work will provide insights into how users can capitalise on the chatbot's natural language processing capabilities for the assessment of translation quality, while also paying attention to its limitations and considering possible future work to address such issues.

1 Introduction

Effective evaluation of translation plays a critical role in professional translation, translation teaching, and translation technology development. This includes assessing the performance of translation students for the attainment of teaching and learning outcomes, evaluating translators for quality assurance and the maintenance of professional standards, and analysing translation engines for the identification of strengths, weaknesses and areas for continuous improvement. For more insights about translation quality evaluation in various settings, refer to Orlando (2011), Drugan (2013), Joss, Sheila, Gaspari, and Stephen (2018), and Chauhan & Daniel (2023).

In the context of machine translation evaluation, common approaches include human evaluation and automatic evaluation (Chatzikoumi, 2020). Human evaluation involves human raters who assign scores to translations by referring to pre-defined criteria or rank different translations of the same target text according to their quality (Freitag et al., 2021). Automatic evaluation computes performance scores automatically by using evaluation metrics, such as comparing the similarity between machine translation outputs and the source text and/or reference translations. Examples of such metrics include BLEU Papineni et al. (2002), CHRF (Popović, 2015), and BERTScore (Zhang et al., 2019).

An emerging alternative approach is the use of large language models with chatbot features, such as ChatGPT. Released by OpenAI in November 2022, ChatGPT is a decoder-based model featuring the use of a multilayer deep neural network with billions of parameters pre-trained on a large amount of text data followed by instruction fine-tuning and reinforcement learning

---

S. C. Siu
School of Translation and Foreign Languages
The Hang Seng University of Hong Kong, Hong Kong, China
e-mail: scsiu@hsu.edu.hk

Electronic copy available at: https://ssrn.com/abstract=4842272

with human feedback, which enables it to follow users' instructions (see OpenAI, 2022; Ouyang et al., 2022). The platform demonstrates promising abilities in a wide range of areas, including natural language processing tasks such as machine translation (Jiao et al., 2024; Siu, 2023a) and automatic evaluation (Araújo & Aguiar, 2023; Kocmi & Federmann, 2023; Leiter et al., 2024; Lu et al., 2023a).

It is against this background that the present work explores how ChatGPT can be used for the assessment of translation quality. Section 2 highlights the possibility of using ChatGPT as an evaluation tool. Sections 3-6 explore four key areas of evaluation that may benefit from it, with examples illustrating how the platform can help discover translation issues. Section 7 discusses potential issues that deserve users' attention and suggests possible future work to improve or better harness the tool. Section 8 presents the concluding remarks. It is hoped that this work will enhance understanding of the strengths and weaknesses of the chatbot as a tool for computer-assisted evaluation of translation.

2 ChatGPT as a Tool for Computer-aided Evaluation of Translation

A prominent feature of large language model-driven chatbots such as ChatGPT is their ability to generate versatile and coherent natural language output based on the input provided by the user in the form of prompts (see examples in OpenAI (2022, 2023b)). Leveraging this for machine translation evaluation, given the source text and translation to be assessed, the model can help us go beyond score computation and generate diverse types of comments on the translation in natural language, enabling the identification of multifaceted issues in machine translation.

It is important to note that ChatGPT is not a single model; instead there are multiple models, which can arguably be called "a family of ChatGPT models". Following its initial release in 2022, OpenAI has made available different models with varying capabilities and context lengths. According to the developer's official website (OpenAI, 2024d), as of May 2024, there are three flagship models: (1) GPT-3.5-turbo, (2) GPT-4-turbo, and (3) GPT-4o. The GPT-3.5-turbo model (OpenAI, 2024b), first released in March 2023, is largely based on the initial ChatGPT model. GPT-4-turbo (OpenAI, 2023c), released in November 2023, is based on the GPT-4 model (OpenAI, 2023a) released in March 2023, and GPT-4o (Open-AI, 2024a, 2024c) is the latest model released in May 2024, with multimodal capabilities.

Studies have shown ChatGPT's promising potential in translation quality evaluation. Araújo and Aguiar (2023) found that ChatGPT exhibited greater consistency in its evaluations compared to human evaluators across all criteria, including fluency, appropriateness, accuracy, and overall assessment. Kocmi and Federmann (2023) proposed a metric that uses GPT models to score translation quality, and it demonstrated strong performance in evaluation with and without a reference translation. Leiter et al. (2024), stressing the significance of explainability of evaluation metrics, suggested that recent large language models such as ChatGPT could offer plausible results for generating natural language explanations. Huang et al. (2024) stated that large language models have shown impressive results in evaluating machine translation quality, and they investigated how these models leverage source and reference information when evaluating translations, with ChatGPT being one of the models considered.

Electronic copy available at: https://ssrn.com/abstract=4842272

Given the above, different from existing works which examine the early releases of ChatGPT models (such as early GPT-3.5 and GPT-4 models in 2023), this work focuses on the latest model in the ChatGPT family, GPT-4o, in an attempt to gain a better understanding of the latest capability of ChatGPT in the context of translation evaluation.

As shown in the examples from Section 3 to Section 6 below, with suitable prompts, users can capitalise on ChatGPT's complex task-solving capabilities and flexible natural language output. Subsequently, ChatGPT can provide explanations of the evaluation results and even suggest revision methods (Araújo & Aguiar, 2023). This enhances the transparency of assessment, offering more insights for users compared with conventional automatic evaluation approaches that assign scores with reference to a sample translation and may provide rather limited information about translation errors or the validity/rationale of the scores. As argued by Guerreiro et al. (2023), evaluation metrics that predict translation scores by regressing on human scores offer limited insight into translation errors.

Unless otherwise specified, a general prompt "You are a chatbot designed to evaluate translation quality." is used across the examples (together with task-specific prompts as presented below), with the temperature set at 0.00 to avoid varying responses by the chatbot given the same input. Translations evaluated in the prompts are based on Google Translate (accessed in May 2024).

3 Area 1: Key Facts and Numbers

The first use case is checking the key facts and numbers in the target text with the assistance of ChatGPT. The following shows a possible prompt template, where the user provides the source text as well as the translation to be evaluated. The prompt asks the chatbot to verify whether the numbers and key facts are translated correctly, and it also states that the results should comprise the source text segment, the original translation, and comments indicating whether the corresponding segment is correct or incorrect. If any errors are found, the chatbot should explain the necessary changes. Finally, the chatbot is expected to provide a revised translation where appropriate.

> Consider the following source text and translation:
>
> **Source text:**
> {Put your source text here.}
>
> **Translation:**
> {Put your translation here.}
>
> Extract key facts and numbers from the source text provided. For each item extracted, check whether they are translated correctly, and present results in the following format:
>
> **Key information extracted:** (Describe the information extracted)
> **Source text:** (Show the information extracted together with neighbouring expressions)
> **Translation:** (Show the translation provided)

Electronic copy available at: https://ssrn.com/abstract=4842272

Comments: (Indicate whether the translation is "Correct" or "Incorrect", with changes explained).

Provide a revised translation if necessary.

A sample output based on the above template is shown in Table 1. The source text is an excerpt in Chinese from a press release by the Hong Kong Monetary Authority regarding the results of the credit card loan survey for the first quarter of 2024 (HKSAR Government, 2024d), with the official reference text in English available at HKSAR Government (2024c). A machine translation in English is provided for evaluation.

Table 1 Checking key facts and numbers in translation using ChatGPT

Role	Prompt/response
User	Consider the following source text and translation: **Source text:** 香港金融管理局今日（五月十六日）公布二○二四年第一季的信用卡貸款調查結果。 信用卡應收帳款總額在上季由節日消費及持卡人繳付薪俸稅所帶動上升8.5%後，隨着有關影響減退，於二○二四年第一季下跌5.3%至三月底的1,515億港元。 **Translation:** The Hong Kong Monetary Authority today (May 16) announced the results of the credit card loan survey for the first quarter of 2024. After rising by 8.5% in the previous quarter, driven by holiday consumption and salary taxes paid by cardholders, total credit card receivables fell by 5.3% in the first quarter of 2024 to 1,515 at the end of March as the impact receded. billion Hong Kong dollars. Extract key facts and numbers from the source text provided. For each item extracted, check whether they are translated correctly, and present results in the following format: **Key information extracted:** (Describe the information extracted) **Source text:** (Show the information extracted together with neighbouring expressions) **Translation:** (Show the translation provided) **Comments:** (Indicate whether the translation is "Correct" or "Incorrect", with changes explained). Provide a revised translation if necessary.
ChatGPT (GPT-4o)	**Key information extracted:** Date of announcement **Source text:** 香港金融管理局今日（五月十六日）公布 **Translation:** The Hong Kong Monetary Authority today (May 16) announced **Comments:** Correct **Key information extracted:** Survey period **Source text:** 二○二四年第一季的信用卡貸款調查結果

Electronic copy available at: https://ssrn.com/abstract=4842272

Translation: the results of the credit card loan survey for the first quarter of 2024
Comments: Correct

Key information extracted: Previous quarter's increase in credit card receivables
Source text: 信用卡應收帳款總額在上季由節日消費及持卡人繳付薪俸稅所帶動上升8.5%
Translation: After rising by 8.5% in the previous quarter, driven by holiday consumption and salary taxes paid by cardholders
Comments: Correct

Key information extracted: Decrease in credit card receivables in the first quarter of 2024
Source text: 於二○二四年第一季下跌5.3%
Translation: total credit card receivables fell by 5.3% in the first quarter of 2024
Comments: Correct

Key information extracted: Total credit card receivables at the end of March
Source text: 至三月底的1,515億港元
Translation: to 1,515 at the end of March as the impact receded. billion Hong Kong dollars.
Comments: Incorrect. The translation has a formatting error. It should be "1,515 billion Hong Kong dollars" or "HKD 151.5 billion."

Revised translation:
The Hong Kong Monetary Authority today (May 16) announced the results of the credit card loan survey for the first quarter of 2024.
After rising by 8.5% in the previous quarter, driven by holiday consumption and salary taxes paid by cardholders, total credit card receivables fell by 5.3% in the first quarter of 2024 to HKD 151.5 billion at the end of March as the impact receded.

In the above example, ChatGPT identifies five key facts for assessment and rightly points out the wrong translation of the number “1,515億港元” (151.5 billion) as “1,515 billion Hong Kong dollars—refer to the fifth point in the chatbot’s response. It confirms the other four key facts in the translation are correct and suggests a revised translation to fix the problem.

4 Area 2: Grammar and Fluency

ChatGPT can also be used to check the grammar and fluency of the target text. For example, Coyne et al. (2023) found that GPT-3.5 and GPT-4 demonstrate strong performance in sentence-level grammatical error correction, particularly in a revision task setting.

The following is a prompt template. The user provides the source text and the translation for evaluation, and the prompt specifies what issues in the translation should be paid attention to (such as spelling, grammar, and word choices) on a sentence-by-sentence basis. For each sentence, the chatbot is expected to provide a list of identified issues and suggested revisions.

> Consider the following text in Chinese and its translation in English:
>
> **Source text**:
> {Put your source text here.}

Electronic copy available at: https://ssrn.com/abstract=4842272

Translation:
{Put your target text here.}

For each sentence in the translation, check for errors/problems in spelling, grammar, sentence structure (e.g., fragments), word choices, collocation, usage, formality, tone, flow, cohesion, coherence, and overall fluency and naturalness. Present the results for each sentence in the following format:

Sentence [Number]
Translation: (the current sentence in the target language)
Issues identified: (if any, with explanation in point form)
Suggested revision: (if applicable; add a blank line before this section starts)

Table 2 shows a sample conversation. The source text is another press release by the Hong Kong Monetary Authority in Chinese, which is about mutual access between Mainland China and Hong Kong interest rate swap markets (HKSAR Government, 2024l), with the official reference text in English accessible at HKSAR Government (2024k). A machine translation output is used to illustrate how the chatbot identifies the issues and offers suggestions.

Table 2 Checking grammar and fluency using ChatGPT

Role	Prompt/response
User	Consider the following text in Chinese and its translation in English: **Source text**: 截至二○二四年四月末，20家境內報價商與58家境外投資者，累計達成人民幣利率互換交易3600多筆，名義本金總額約1.77萬億元人民幣，日均成交名義本金約76億元人民幣，按月計算的日均成交名義本金增長近三倍，從上線首月日均約30億元人民幣增至二○二四年四月日均120億元人民幣以上，為境內外投資者開展人民幣資產配置提供了便利、高效的風險管理工具。 **Translation**: As of the end of April 2024, 20 domestic quotations companies and 58 overseas investors had concluded more than 3,600 RMB interest rate swap transactions, with a total nominal principal of approximately RMB 1.77 trillion, and an average daily nominal principal of approximately RMB 1.77 trillion. 7.6 billion yuan, the monthly average daily transaction nominal principal has increased nearly three times, from an average daily average of about 3 billion yuan in the first month of launch to an average daily average of more than 12 billion yuan in April 2024, which is the largest domestic It provides foreign investors with a convenient and efficient risk management tool for RMB asset allocation. For each sentence in the translation, check for errors/problems in spelling, grammar, sentence structure (e.g., fragments), word choices, collocation, usage, formality, tone, flow, cohesion, coherence, and overall fluency and naturalness. Present the results for each sentence in the following format: **Sentence [Number]** **Translation:** (the current sentence in the target language) **Issues identified:** (if any, with explanation in point form) **Suggested revision:** (if applicable; add a blank line before this section starts)

Electronic copy available at: https://ssrn.com/abstract=4842272

ChatGPT (GPT-4o)	**Sentence 1** **Translation:** As of the end of April 2024, 20 domestic quotations companies and 58 overseas investors had concluded more than 3,600 RMB interest rate swap transactions, with a total nominal principal of approximately RMB 1.77 trillion, and an average daily nominal principal of approximately RMB 1.77 trillion. **Issues identified:** • Repetition of "RMB 1.77 trillion" in the context of total nominal principal and average daily nominal principal is incorrect. • "Quotations companies" should be "quote providers" or "quoting companies" for better clarity. • The sentence is too long and could be broken down for better readability. **Suggested revision:** As of the end of April 2024, 20 domestic quoting companies and 58 overseas investors had concluded more than 3,600 RMB interest rate swap transactions, with a total nominal principal of approximately RMB 1.77 trillion. The average daily nominal principal was approximately RMB 7.6 billion. **Sentence 2** **Translation:** 7.6 billion yuan, the monthly average daily transaction nominal principal has increased nearly three times, from an average daily average of about 3 billion yuan in the first month of launch to an average daily average of more than 12 billion yuan in April 2024, which is the largest domestic It provides foreign investors with a convenient and efficient risk management tool for RMB asset allocation. **Issues identified:** • The sentence is fragmented and lacks coherence. • Repetition of "average daily average" is redundant. • "The largest domestic It provides" is grammatically incorrect and unclear. • The sentence should be split for clarity and readability. **Suggested revision:** The monthly average daily transaction nominal principal has increased nearly three times, from about RMB 3 billion in the first month of launch to more than RMB 12 billion in April 2024. This provides both domestic and foreign investors with a convenient and efficient risk management tool for RMB asset allocation.

From Table 2, it can be seen that ChatGPT manages to identify different issues in the target text, especially the structural problem in the second sentence. This issue may have arisen from the lengthy source language sentence leading to segmentation issues during the translation process. This results in the duplication of the number “RMB 1.77 trillion” and the syntactically problematic expression “which is the largest domestic It provides” in the sentence. Both are detected by ChatGPT, with relevant revision suggestions presented in its response.

5 Area 3: Terminological Issues

Terminology is at the core of specialised translation, such as financial translation and medical translation, among others. Conventionally, terminology databases can be used to help ensure that terms are translated properly, often through integration into quality assurance tools (e.g., Phrase (n.d.)). ChatGPT offers a new possibility: users can incorporate bilingual terminological information into prompts so that the chatbot can check the target text against it. Here is an example:

Electronic copy available at: https://ssrn.com/abstract=4842272

> Consider the following bilingual glossary:
>
> {Put your glossary here.}
>
> Given the following source text and translation, assess the quality of the translation of terminology. Determine whether the terms are aligned with the glossary above and whether the terms are translated consistently (i.e., verify that the same term, if used multiple times, is translated in the same way throughout the text).
>
> **Source Text:**
> {Put your source text here.}
>
> **Translation:**
> {Put your translation here.}
>
> Present your evaluation results in the following format:
>
> **Term detected:** (Term in Source Language – Term in Target Language)
> **Source:** (Actual term used in the source text)
> **Translation:** (Actual term used in the target text)
> **Accuracy and consistency:** (Correct / Partially correct (in bold) / Incorrect (in bold): Explain if the term is translated incorrectly or inconsistently in the target text.)

The user first provides a bilingual glossary, which is expected to be used by the chatbot to check the appropriateness of the translated terms and confirm whether they are translated consistently. Table 3 shows an example of using this template. The user supplies a small glossary with 10 entries related to finance, along with the source text (a press release in Chinese by the Hong Kong Monetary Authority announcing e-CNY, a pilot for digital cross-border payments; HKSAR Government, 2024f), with the English version accessible at HKSAR Government (2024e).

Table 3 Checking terminology using ChatGPT

Role	**Prompt/response**
User	Consider the following bilingual glossary: 大灣區 - Guangdong-Hong Kong-Macao Greater Bay Area (GBA) 數字貨幣研究所 - Digital Currency Institute (DCI) 數字人民幣 - e-CNY 數字人民幣錢包e-CNY wallets G20跨境支付藍圖 - G20 Roadmap for enhancing cross-border payments 香港金融管理局 - Hong Kong Monetary Authority (HKMA) 中國人民銀行 - People’s Bank of China (PBoC) 三聯通、三便利 - three connection, three facilitation 零售銀行 - retail banks 轉數快 - Faster Payment System (FPS) Given the following source text and translation, assess the quality of the translation of terminology. Determine whether the terms are aligned with the glossary above and

Electronic copy available at: https://ssrn.com/abstract=4842272

	whether the terms are translated consistently (i.e., verify that the same term, if used multiple times, is translated in the same way throughout the text). **Source Text:** 香港金融管理局（金管局）及中國人民銀行（人民銀行）就數字人民幣跨境支付試點的合作取得進一步成果，擴大數字人民幣在香港的試點範圍，便利香港居民開立和使用數字人民幣錢包（數幣錢包），並透過「轉數快」為數幣錢包增值。與此同時，「轉數快」與人民銀行數字貨幣研究所（數研所）營運的數字人民幣央行端系統的互通，也是世界首個快速支付系統與央行數字貨幣系統的連接，為G20跨境支付藍圖所強調的互通性提供了創新的用例。 **Translation:** The Hong Kong Monetary Authority (HKMA) and the People's Bank of China (PBOC) have achieved further results in their cooperation on the digital renminbi cross-border payment pilot, expanding the scope of the digital renminbi pilot in Hong Kong and facilitating Hong Kong residents to open and use digital renminbi wallets (digital renminbi wallets). cryptocurrency wallet), and add value to the cryptocurrency wallet through "FPS". At the same time, the interoperability between "FPS" and the digital renminbi central bank system operated by the Digital Currency Research Institute of the People's Bank of China (the Digital Currency Research Institute) is also the world's first connection between a fast payment system and a central bank's digital currency system, providing cross-border support for the G20 The interoperability emphasized by the Payments Blueprint provides innovative use cases. Present your evaluation results in the following format: **Term detected:** (Term in Source Language – Term in Target Language) **Source:** (Actual term used in the source text) **Translation:** (Actual term used in the target text) **Accuracy and consistency:** (Correct / Partially correct (in bold) / Incorrect (in bold): Explain if the term is translated incorrectly or inconsistently in the target text.)
ChatGPT (GPT-4o)	**Term detected:** 香港金融管理局 - Hong Kong Monetary Authority (HKMA) **Source:** 香港金融管理局（金管局） **Translation:** Hong Kong Monetary Authority (HKMA) **Accuracy and consistency:** Correct **Term detected:** 中國人民銀行 - People's Bank of China (PBoC) **Source:** 中國人民銀行（人民銀行） **Translation:** People's Bank of China (PBOC) **Accuracy and consistency: Incorrect**: The abbreviation should be PBoC, not PBOC. **Term detected:** 數字人民幣 - e-CNY **Source:** 數字人民幣 **Translation:** digital renminbi **Accuracy and consistency: Incorrect**: The term should be translated as e-CNY according to the glossary. **Term detected:** 數字人民幣錢包 - e-CNY wallets **Source:** 數字人民幣錢包（數幣錢包） **Translation:** digital renminbi wallets (digital renminbi wallets) **Accuracy and consistency: Incorrect**: The term should be translated as e-CNY

Electronic copy available at: https://ssrn.com/abstract=4842272

wallets according to the glossary. Additionally, the parenthetical translation is redundant and incorrect.

Term detected: 轉數快 - Faster Payment System (FPS)
Source: 轉數快
Translation: FPS
Accuracy and consistency: Partially correct: The term FPS is correct, but the full term Faster Payment System should be mentioned at least once for clarity.

Term detected: 數字貨幣研究所 - Digital Currency Institute (DCI)
Source: 數字貨幣研究所（數研所）
Translation: Digital Currency Research Institute of the People's Bank of China (the Digital Currency Research Institute)
Accuracy and consistency: Incorrect: The term should be Digital Currency Institute (DCI) according to the glossary.

Term detected: G20跨境支付藍圖 - G20 Roadmap for enhancing cross-border payments
Source: G20跨境支付藍圖
Translation: G20 Payments Blueprint
Accuracy and consistency: Incorrect: The term should be G20 Roadmap for enhancing cross-border payments according to the glossary.

Term detected: 數字人民幣央行端系統 - (Not provided in the glossary)
Source: 數字人民幣央行端系統
Translation: digital renminbi central bank system
Accuracy and consistency: Partially correct: The term is not provided in the glossary, but the translation is understandable and consistent with the context.

Overall Evaluation:
The translation of terminology in the provided text is inconsistent with the glossary. Several terms are translated incorrectly, and some terms are not translated consistently. The translator should adhere more closely to the provided glossary to ensure accuracy and consistency.

In ChatGPT's response, terms in the target text are extracted, and the chatbot is able to state that certain terms are not consistent with the ones given in the glossary. For example, "數字貨幣研究所" (Digital Currency Institute (DCI)) is translated as "Digital Currency Research Institute", and "數字人民幣" (e-CNY) is translated as "digital renminbi". It also notes a slight discrepancy in the absence of the full form of "轉數快" (Faster Payment System) in the target text, as well as the abbreviation of the People's Bank of China ("PBoC" in the glossary, as opposed to "PBOC" in the machine translation).

6 Area 4: Stylistic Consistency

For the translation of annual/monthly reports or specialised/technical documents with standardised style and/or formatting (e.g., company announcements by listed companies for the Hong Kong Stock Exchange), it is important for the target text to adopt a translation style consistent with the other documents (e.g., previously published annual reports or other

Electronic copy available at: https://ssrn.com/abstract=4842272

company announcements of the same kind). ChatGPT can help check the style of a translation and see whether it aligns with the style of others. One possible method is to include a sample bilingual document in the prompt for the chatbot's reference. Here is a prompt template:

> Consider the following bilingual document. Pay attention to the writing style in Chinese.
>
> **English text:**
> {Put your reference text in English here.}
>
> **Chinese text:**
> {Put your reference text in Chinese here.}
>
> Now, given the following new text, comment on whether its Chinese translation is consistent with the style of writing of the above text. List any inconsistencies, and provide a revised version of the translation with a consistent style of writing:
>
> **New Text:**
> {Put your source text here.}
>
> **Translation:**
> {Put your translation here.}
>
> List any issues in stylistic consistency and provide a revised translation if applicable. Focus on differences in style.

The prompt template first provides a bilingual document so that the chatbot has a reference point for what the target style should be like. After that, the user gives the source text together with its translation, the style of which is to be checked. The prompt states that the chatbot is expected to comment on the consistency of the writing style in the target language, provide a list of inconsistent stylistic elements, and offer a revised version.

Table 4 illustrates the use of the above prompt by considering an automatic translation of the first two paragraphs of "Provisional Statistics of Retail Sales for March 2024" in English, with the official versions of the English and Chinese texts retrieved from HKSAR Government (2024i) and HKSAR Government (2024j). Note that a bilingual excerpt in English and Chinese for the monthly report for the previous month (i.e., February 2024) is first provided to inform the chatbot of the reference translation style (see HKSAR Government (2024g, 2024h) for the English and Chinese versions).

Table 4 Checking stylistic consistency using ChatGPT and a previous reference document

Role	**Prompt/response**
User	Consider the following bilingual document. Pay attention to the writing style in Chinese. **English text:** Provisional Statistics of Retail Sales for February 2024 The Census and Statistics Department (C&SD) released the latest figures on retail sales today (April 3).

Electronic copy available at: https://ssrn.com/abstract=4842272

The value of total retail sales in February 2024, provisionally estimated at $33.8 billion, increased by 1.9% compared with the same month in 2023. The revised estimate of the value of total retail sales in January 2024 increased by 0.9% compared with a year earlier. For the first two months of 2024 taken together, it was provisionally estimated that the value of total retail sales increased by 1.4% compared with the same period in 2023.

Chinese text:

二零二四年二月份零售業銷貨額臨時統計數字

政府統計處今日（四月三日）發表最新的零售業銷貨額數字。

二零二四年二月的零售業總銷貨價值的臨時估計為338億元，較二零二三年同月上升1.9%。二零二四年一月的零售業總銷貨價值的修訂估計較二零二三年同月上升0.9%。與二零二三年同期比較，二零二四年首兩個月合計的零售業總銷貨價值的臨時估計上升1.4%。

Now, given the following new text, comment on whether its Chinese translation is consistent with the style of writing of the above text. List any inconsistencies, and provide a revised version of the translation with a consistent style of writing:

New Text:

Provisional Statistics of Retail Sales for March 2024

The Census and Statistics Department (C&SD) released the latest figures on retail sales today (May 3).

The value of total retail sales in March 2024, provisionally estimated at $31.2 billion, decreased by 7.0% compared with the same month in 2023. The revised estimate of the combined value of total retail sales in January and February 2024 increased by 1.4% compared with the same period a year earlier. For the first quarter of 2024, it was provisionally estimated that the value of total retail sales decreased by 1.3% compared with the same period in 2023.

Translation:

2024年3月社會消費品零售總額臨時統計

政府統計處今日（5月3日）公佈最新零售業銷貨數字。

2024年3月零售業銷貨總額暫估為312億元，較2023年同月下跌7.0%。同一時期。 初步估計2024年第一季社會消費品零售總額較2023年同期下降1.3%。

List any issues in stylistic consistency and provide a revised translation if applicable. Focus on differences in style.

ChatGPT (GPT-4o)

The provided Chinese translation of the new text has several stylistic inconsistencies when compared to the original Chinese text. Here are the issues and a revised version:

Electronic copy available at: https://ssrn.com/abstract=4842272

Issues in Stylistic Consistency:

1. **Title Consistency**:
    - Original: "二零二四年二月份零售業銷貨額臨時統計數字"
    - New: "2024 年 3 月社會消費品零售總額臨時統計"
    - Issue: The title format is inconsistent. The original uses full Chinese characters for the year and month, while the new text uses Arabic numerals.

2. **Terminology Consistency**:
    - Original: "零售業銷貨額"
    - New: "社會消費品零售總額"
    - Issue: The terminology for "retail sales" is different. The original uses "零售業銷貨額," while the new text uses "社會消費品零售總額."

3. **Date Format**:
    - Original: "今日（四月三日）"
    - New: "今日（5 月 3 日）"
    - Issue: The original uses full Chinese characters for the date, while the new text uses Arabic numerals.

4. **Numerical Consistency**:
    - Original: "338 億元"
    - New: "312 億元"
    - Issue: The original uses Chinese characters for the numbers, while the new text uses Arabic numerals.

5. **Missing Information**:
    - The new text omits the revised estimate for January and February 2024.

Revised Translation:

二零二四年三月份零售業銷貨額臨時統計數字

政府統計處今日（五月三日）發表最新的零售業銷貨額數字。

二零二四年三月的零售業總銷貨價值的臨時估計為三百一十二億元，較二零二三年同月下跌7.0%。二零二四年一月和二月的零售業總銷貨價值的修訂估計較二零二三年同期上升1.4%。與二零二三年同期比較，二零二四年第一季合計的零售業總銷貨價值的臨時估計下跌1.3%。

This revised translation maintains consistency in style, terminology, and numerical representation with the original Chinese text.

As shown in Table 4, the chatbot highlights a few stylistic differences between the machine translation output and reference translation, including the use of Arabic numbers (e.g., “2024年3月” for March 2024) instead of Chinese characters (“二零二四年三月”) in terms of date formatting, and title inconsistency (the use of “社會消費品零售總額臨時統計” instead of “零售業銷貨額臨時統計數字” for the title “Provisional Statistics of Retail Sales”). A revised translation is provided.

Electronic copy available at: https://ssrn.com/abstract=4842272

7 Limitations and Future Work

While ChatGPT has demonstrated capabilities in the automatic assessment of translation quality as shown in the examples discussed in the previous sections, there are two limitations users should consider and areas where future work is needed.

7.1 Hallucinations

The first limitation is hallucinations, where chatbots generate false information, a well-known issue among large language models (OpenAI, 2023b; Bubeck et al., 2023). Hallucinations related to natural language processing and translation have been documented (e.g., see Siu (2023a, 2023b) for examples of hallucinations in translation). This could be attributed to an inadequate understanding of relevant knowledge or non-English languages, among other factors. For example, when dealing with content beyond the cut-off date of training data (e.g., a model trained on data prior to 2021 being asked about events in 2022 or later), the chatbot may fabricate information based on what it learnt about the world before that date from the training data (see Siu (2023a)).

In the case of automatic evaluation, hallucinations may happen if the chatbot is asked to handle linguistic issues (e.g., complex issues about the use of simplified Chinese and traditional Chinese) that may not be well-represented in the training data, and it may make things up based on its knowledge about the Chinese language, potentially leading to inaccurate or contradictory responses.

To address such issues, researchers may consider fine-tuning the model by exposing it to translation evaluation tasks and texts in relevant languages to enhance its knowledge and see if this could help reduce hallucinations. Another approach worth exploring is the use of few-shot prompts, where examples of how an assessment task should be completed are incorporated into the prompt, or the provision of additional background information to offer more context, so as to help the chatbot get ready for the task. By referring to such information in the prompt, the chatbot may have fewer hallucinations and provide more accurate results.

Examples of Hallucinations:

For the illustration of the issue of hallucinations, two ChatGPT models, namely GPT-3.5-Turbo and GPT-4o, are asked to consider the following Chinese text, an excerpt from the press release "CE meets Secretary of CPC Guangxi Zhuang Autonomous Region Committee (with photo)" (see HKSAR Government (2024a) and HKSAR Government (2024b), for the official English and Chinese versions, respectively), and its machine translation in Chinese.

> Official Text in English:
> The Chief Executive, Mr John Lee, met with the Secretary of the CPC Guangxi Zhuang Autonomous Region Committee, Mr Liu Ning, today (May 21) to exchange views on further promoting co-operation between Hong Kong and Guangxi.
>
> Official Text in Chinese:

Electronic copy available at: https://ssrn.com/abstract=4842272

行政長官李家超今日（五月二十一日）與廣西壯族自治區黨委書記劉寧會面，就進一步推動港桂合作交換意見。

Machine Translation in Chinese:
行政長官李家港今日（5月21日）與廣西壯族自治區委書記劉寧會面，就進一步推動香港與廣西合作交換意見。

The two models are asked to comment on (1) whether the names and titles of the HKSAR government officials are correctly translated and (2) whether simplified Chinese characters are properly used. They are not provided with the official text in Chinese, as the objective here is to test how they conduct evaluation based on what they learnt from the training data. Table 5 shows the responses provided by the two chatbots.

Table 5 Illustration of hallucination issues

Role	Prompt/response
ChatGPT (GPT-3.5-Turbo)	1. Upon examining the machine translation, it appears that the name of the Chief Executive, Mr John Lee, has been correctly translated as "行政長官李家港." The translation accurately captures the name and maintains its integrity. 2. In terms of the use of simplified Chinese characters, the machine translation seems to have properly utilized them. Simplified Chinese characters are commonly used in mainland China, including the Guangxi Zhuang Autonomous Region. The translation employs simplified characters throughout, which aligns with the standard practice in mainland China.
ChatGPT (GPT-4o)	1. The name of the Chief Executive is not correctly translated. The source text mentions "Mr John Lee," but the machine translation incorrectly translates it as "李家港" (Li Jia Gang). The correct translation should be "李家超" (Li Jiachao). 2. The machine translation uses traditional Chinese characters, not simplified Chinese characters. For example, "行政長官" (Chief Executive) and "會面" (meet) are in traditional Chinese. Simplified Chinese characters would be "行政长官" and "会面" respectively.

The chatbots' responses suggest that the latest model, GPT-4o, outperforms its predecessor GPT-3.5-turbo. GPT-3.5-Turbo hallucinates that the name of the Chief Executive is correct and that simplified Chinese characters are used in the target text. With more up-to-date training data (with a data cut-off date of October 2023 for GPT-4o as opposed to September 2021 for GPT-3.5-turbo), multilingual support, and other developments (OpenAI, 2024a, 2024c, 2024d), GPT-4o can detect the errors correctly—suggesting that the correct name should be "李家超" (Mr John Lee) and indicating that traditional Chinese characters are used.

7.2 Variability in Assessment Results

The second limitation is the possible variability in assessment results (Araujo & Aguiar, 2023; Lu et al., 2023a). For the same prompt, the chatbot may give different responses or have varying ways to present its findings. While the responses may be similar in terms of assessment results, there could be minor deviations. This variability could stem from the randomness in the process of generating the results.

Electronic copy available at: https://ssrn.com/abstract=4842272

As a decoder-based language model, ChatGPT works by predicting the next token given the prompt and the words generated so far, considering the probabilities of candidates for the next token (Siu, 2023b). For general text generation tasks, to enhance the diversity of the output, the model can use random sampling with a higher temperature setting to introduce randomness in the selection of the candidate for the next token, which is not necessarily the one with the highest probability. However, this randomness may not be desirable for translation evaluation because it could lead to varying results and affect reliability.

In this regard, when using ChatGPT as a translation evaluation tool, users may need to consider lowering the temperature setting, say close to zero (Lu et al., 2023a), to reduce variability and achieve more consistent results, especially if they observe significant variations or even conflicting responses. If a higher temperature is still preferred to encourage diverse findings of translation issues, users might consider generating multiple responses instead of one and asking the model to consolidate them to give the final output.

Example of Variability:

For the illustration of the issue of variability, in this example, with reference to the texts in the above example about hallucinations, a higher temperature setting of 1.00 is applied to GPT-4o, instead of 0.00 in all the previous examples, and it is asked to generate a response again. While the input is the same, we get the following different response (Table 6). The answer is still valid, but the actual wording is different. Although this can be addressed by lowering the temperature setting if possible, if the source text and the target text are different, even with the same translation issues present, whether they can still be consistently identified remains to be examined because of the probabilistic nature of the model—the prediction of tokens in the result generation process is dependent on the actual tokens in the prompts and preceding tokens.

Table 6. Illustration of variability in chatbot responses

Role	Prompt/response
ChatGPT (GPT-4o)	1. The name of the Chief Executive is not correctly translated. The source text mentions "Mr John Lee," but the machine translation uses "李家港," which is incorrect. The correct translation should be "李家超." 2. The provided machine translation uses traditional Chinese characters, not simplified Chinese characters. For example, "行政長官" should be "行政长官" in simplified Chinese, and "書記" should be "书记." Therefore, the machine translation does not use simplified Chinese characters.

7.3 *Other Future Work*

Given these limitations, it is advisable to use ChatGPT as assistive tool for evaluation rather than relying solely on it, as also stated in Araujo & Aguiar (2023). In addition, there are other issues related to the use of ChatGPT for translation evaluation that remain to be explored. Three noteworthy areas are proposed as follows:

First, the application of prompt engineering strategies should be further investigated. The importance of prompt engineering—devising and refining instructions sent to a large language

Electronic copy available at: https://ssrn.com/abstract=4842272

model—has been emphasised in the context of translation evaluation (Jiao et al., 2024) and grammatical error correction (Coyne et al., 2023). Researchers may explore whether it would be beneficial to consider advanced prompting approaches. For example, Jiao et al. (2024) showed that a multiple-level taxonomy for prompting (with the inclusion of information such as expression type, translation style, and part-of-speech information) could improve the translation quality of ChatGPT—would this be applicable to translation evaluation as well? Lu et al. (2023a) also went beyond zero-shot prompting by combining (i) chain-of-thought prompting, which highlights the generation of intermediate reasoning steps (Wei et al., 2022), and (ii) error analysis (Lu et al., 2023b). This method first prompts ChatGPT to identify major and minor errors in the translation and then asks it to score the translation based on the errors identified, showing improvements in the accuracy of evaluation. Researchers may also investigate the use of retrieval-based methods with the incorporation of information provided by third-party resources such as search engines in the hope of increasing the reliability of assessment results. For example, this could be useful for terminology checking in the absence of a user-provided glossary; the chatbot can use data retrieved from an external terminology database for evaluation. The same applies to the assessment of consistent style where, if a reference translation is not available, the chatbot powered by a retrieval-augmented generation (RAG) mechanism (Lewis et al., 2020) can conduct internet or database searches to see if reference style information can be used to augment the evaluation process.

Second, systematic analyses of the efficiency of different prompts for translation evaluation should be conducted. Researchers may explore how prompts for translation quality evaluation should be structured (e.g., should the instruction be stated first, or should the source text, translation and/or reference translation be provided first?). They may also investigate whether it would be beneficial to provide the official reference translation if available. As discussed by Huang et al. (2024), a reference translation can improve the accuracy of machine translation evaluation, while providing source information could sometimes be counterproductive. Another area to consider is whether to include a system prompt (e.g., explicitly telling the chatbot that it plays the role of a linguist or translation expert). Additionally, researchers should examine how to adjust prompting strategies in light of system updates. It is expected that more new models will be developed in the future, which may have a larger context length. While the latest GPT-4o model is shipped with a context length of 128k tokens (OpenAI, 2024c), whether different prompting strategies are needed is worth further exploration if models with larger contexts (e.g., 200k as in Claude (Anthropic, 2023) and 1 million as in Gemini (Google, 2024)) are released. A larger context length means longer prompts are possible, providing the chatbot with more additional information *in-situ* (e.g., information retrieved from third-party sources) to potentially help it deal with more complex situations unseen in training data. Researchers may also investigate whether designated prompts are needed for the evaluation of translation quality for specialised domains or specific language pairs, and if so, how they should be designed. These are all important questions for a better understanding of whether different ways of formulating evaluation prompts could have an impact on the quality of assessment, and if so, how they differ, and if not, why not.

Third, researchers may explore how ChatGPT could go beyond common areas of evaluation, such as accuracy in terms of terminology, numbers, grammar, and spelling, and take a step forward by considering the appropriateness of a translation in terms of the fulfilment of the purpose of translation and even cultural suitability. This could be achieved by providing relevant background information about a translation project in the prompt. The chatbot can then be asked to deduce the target readership and purpose of the translation, based on which it assesses suitability. The "reasoning" process using such information for determining

Electronic copy available at: https://ssrn.com/abstract=4842272

appropriateness can be requested in the chatbot's response. These areas are not easy to assess using conventional automatic scoring methods that pay attention to the similarity between the translation and the source text and/or reference translation. Exploring these aspects could lead to a more comprehensive and nuanced evaluation of translation quality.

8 Concluding Remarks

This paper explores the possibilities of using ChatGPT as a tool to assist in the evaluation of translation quality, with the following four areas explained: key facts and numbers, grammar and fluency, terminology, and stylistic consistency with other documents. Sample prompt templates, together with examples of user input and chatbot output, are presented, with a focus on the use of GPT-4o, the latest model in the ChatGPT family at the time of writing, for the illustration of state-of-the-art capabilities in the context of translation evaluation. While the platform demonstrates impressive potential beyond conventional automatic evaluation methods that tend to stress the computation of similarity between a translation and a reference translation and/or the source text, there are two areas that users should be aware of: possible hallucinations and variability in assessment results. Adjustments in chatbot configuration, the fine-tuning of models, and the exploration of advanced prompting strategies are areas of future work that may help mitigate these issues and enhance the applicability of the chatbot for translation evaluation.

Acknowledgements

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. UGC/IDS(R)14/23).

References

Anthropic. (2023). Introducing Claude 2.1. Retrieved in May 2024, from https://www.anthropic.com/news/claude-2-1

Araújo, S., & Aguiar, M. (2023). Comparing ChatGPT's and human evaluation of scientific texts' translations from English to Portuguese using popular automated translators. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)* (pp. 2908-2917).

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering, 26*(2), 137-161.

Chauhan, S., & Daniel, P. (2023). A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Processing Letters, 55*(9), 12663-12717.

Coyne, S., Sakaguchi, K., Galvan-Sosa, D., Zock, M., & Inui, K. (2023). Analyzing the performance of GPT-3.5 and GPT-4 in grammatical error correction. *arXiv preprint arXiv:2303.14342*.

Drugan, J. (2013). *Quality in professional translation: Assessment and improvement* (Vol. 9). A&C Black.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics, 9*, 1460-1474.

Google. (2024). Our next-generation model: Gemini 1.5. Retrieved in May 2024, from https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

Google Translate. (2024). Retrieved in May 2024, from https://translate.google.com/

Electronic copy available at: https://ssrn.com/abstract=4842272

Guerreiro, N. M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., & Martins, A. F. (2023). xCOMET: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*.

HKSAR Government. (2024a). CE meets Secretary of CPC Guangxi Zhuang Autonomous Region Committee (with photo). Retrieved in May 2024, from https://www.info.gov.hk/gia/general/202405/21/P2024052100528.htm

HKSAR Government. (2024b). CE meets Secretary of CPC Guangxi Zhuang Autonomous Region Committee (with photo) (Chinese Version: 行政長官與廣西壯族自治區黨委書記會面（附圖）). Retrieved in May 2024, from https://www.info.gov.hk/gia/general/202405/21/P2024052100488.htm

HKSAR Government. (2024c). Credit card lending survey results for the first quarter 2024. Retrieved in May 2024, from https://www.info.gov.hk/gia/general/202405/16/P2024051600338.htm

HKSAR Government. (2024d). Credit card lending survey results for the first quarter 2024 (Chinese Version: 二○二四年第一季信用卡貸款調查結果). Retrieved in May 2024, from https://www.info.gov.hk/gia/general/202405/16/P2024051600339.htm

HKSAR Government. (2024e). Expanding cross-boundary e-CNY pilot in Hong Kong. Retrieved in May 2024, from https://www.info.gov.hk/gia/general/202405/17/P2024051600533.htm

HKSAR Government. (2024f). Expanding cross-boundary e-CNY pilot in Hong Kong (Chinese Version: 擴大數字人民幣在香港的跨境試點). Retrieved in May 2024, from https://www.info.gov.hk/gia/general/202405/17/P2024051600553.htm

HKSAR Government. (2024g). Provisional statistics of retail sales for February 2024. Retrieved in May 2024, from https://www.info.gov.hk/gia/general/202404/03/P2024040300253.htm

HKSAR Government. (2024h). Provisional statistics of retail sales for February 2024 (Chinese Version: 二零二四年二月份零售業銷貨額臨時統計數字). Retrieved in May 2024, from https://www.info.gov.hk/gia/general/202404/03/P2024040300285.htm

HKSAR Government. (2024i). Provisional statistics of retail sales for March 2024. Retrieved in May 2024, from https://www.info.gov.hk/gia/general/202405/03/P2024050300238.htm

HKSAR Government. (2024j). Provisional statistics of retail sales for March 2024 (Chinese Version: 二零二四年三月份零售業銷貨額臨時統計數字). Retrieved in May 2024, from https://www.info.gov.hk/gia/general/202405/03/P2024050300230.htm

HKSAR Government. (2024k). Swap Connect enhancements to advance high-quality opening-up of China's financial markets. Retrieved in May 2024, from https://www.info.gov.hk/gia/general/202405/13/P2024051300430.htm

HKSAR Government. (2024l). Swap Connect enhancements to advance high-quality opening-up of China's financial markets (Chinese Version: 優化「互換通」機制安排　促進中國金融市場高水平對外開放). Retrieved in May 2024, from https://www.info.gov.hk/gia/general/202405/13/P2024051300434.htm

Huang, X., Zhang, Z., Geng, X., Du, Y., Chen, J., & Huang, S. (2024). Lost in the source language: How large language models evaluate the quality of machine translation. *arXiv preprint arXiv:2401.06568*.

Jiao, H., Peng, B., Zong, L., Zhang, X., & Li, X. (2024). Gradable ChatGPT translation evaluation. *arXiv preprint arXiv:2401.09984*.

Joss, M., Sheila, C., Gaspari, F., & Stephen, D. (2018). *Translation quality assessment: From principles to practice*.

Kocmi, T., & Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Leiter, C., Lertvittayakumjorn, P., Fomicheva, M., Zhao, W., Gao, Y., & Eger, S. (2024). Towards explainable evaluation metrics for machine translation. *Journal of Machine Learning Research, 25*(75), 1-49.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459-9474.

Lu, Q., Qiu, B., Ding, L., Xie, L., & Tao, D. (2023a). Error analysis prompting enables human-like translation evaluation in large language models: A case study on ChatGPT. *arXiv preprint arXiv:2303.13809*.

Lu, Q., Ding, L., Xie, L., Zhang, K., Wong, D. F., & Tao, D. (2023b). Toward human-like evaluation for natural language generation with error analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5892-5907).

OpenAI. (2022). Introducing ChatGPT. Retrieved in May 2024, from https://openai.com/blog/chatgpt

OpenAI. (2023a). GPT-4. Retrieved in May 2024, from https://openai.com/index/gpt-4-research/

OpenAI. (2023b). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. (2023c). New models and developer products announced at DevDay. Retrieved in May 2024, from https://openai.com/index/new-models-and-developer-products-announced-at-devday/

OpenAI. (2024a). Hello GPT-4o. Retrieved in May 2024, from https://openai.com/index/hello-gpt-4o/

Electronic copy available at: https://ssrn.com/abstract=4842272

OpenAI. (2024b). Introducing APIs for GPT-3.5 Turbo and Whisper. Retrieved in May 2024, from https://openai.com/index/introducing-chatgpt-and-whisper-apis/

OpenAI. (2024c). Introducing GPT-4o and more tools to ChatGPT free users. Retrieved in May 2024, from https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/

OpenAI. (2024d). Models. Retrieved in May 2024, from https://platform.openai.com/docs/models

Orlando, M. (2011). Evaluation of translations in the training of professional translators: At the crossroads between theoretical, professional and pedagogical practices. *The Interpreter and Translator Trainer, 5*(2), 293-308.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems, 35*, 27730-27744.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318).

Phrase. (n.d.). *Quality assurance (QA) - TMS*. Retrieved in May 2024, from https://support.phrase.com/hc/en-us/articles/5709703799324-Quality-Assurance-QA-TMS

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 392-395).

Siu, S. C. (2023a). ChatGPT and GPT-4 for professional translators: Exploring the potential of large language models in translation. Available at SSRN 4448091.

Siu, S. C. (2023b). Revolutionising translation with AI: Unravelling neural machine translation and generative pre-trained large language models. Available at SSRN 4499768.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., … & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824-24837.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Electronic copy available at: https://ssrn.com/abstract=4842272