ChatGPT Self-Correction Outputs between Self-Provided and External Feedback in Translating Medical Texts

Hesham Abdullah Almekhlafi [1,*]
Khalil A Nagi [2]

Received: 1 November 2025
Revised: 30 November 2025
Accepted: 1 December 2025

[1] English Language (Linguistics) - English Department - Faculty of Applied Sciences and Humanities - Amran University – Yemen.
[2] English Language (Linguistics) - English Department - Faculty of Education and Sciences - Saba Region University – Yemen. E-mail: khalil.naji@usr.ac
* Corresponding author. E-mail: almekhlafihesham@gmail.com

https://doi.org/10.20428/jss.v31i10.3440

ChatGPT Self-Correction Outputs between Self-Provided and External Feedback in Translating Medical Texts

Abstract:

This study aims to investigate both external and self-provided feedback strategies and their effects on ChatGPT's self-correction for the purpose of improving its translation performance. The study evaluates the effectiveness of the proposed strategies and points out which strategy is more effective when translating English medical texts into Arabic. The researchers chose 15 English medical texts. These texts were translated into Arabic by ChatGPT using a default translation prompt. The translated texts are manually annotated and evaluated. The researchers, then, retranslate the texts, where some of the texts were retranslated using the self-provided feedback and some were retranslated using external feedback. Both feedback strategies are used to exploit ChatGPT self-correction. Manual evaluations of the initial translation and the retranslation are performed to evaluate the effectiveness of the used feedback strategies. The effectiveness of each of these feedback strategies is evaluated based on the improvement rate of the retranslation. The study reaches a conclusion regarding the most efficient feedback strategy, which is the error taxonomy feedback strategy. The study is very important for translators, post-editors, researchers, and developers of MT. It also provides recommendations for future work.

Keywords: *self-correction, self-provided, external, feedback, medical texts, ChatGPT.*

مخرجات التصحيح الذاتي لـ شات جي بي تي بين التغذية الراجعة الذاتية والتغذية الراجعة الخارجية في ترجمة النصوص الطبية

هشام عبدالله المخلافي [1,*]
خليل ناجي [2]

الملخص:

تهدف هذه الدراسة إلى التحقق من استراتيجيات التغذية الراجعة الخارجية والذاتية وتأثيراتها على التصحيح الذاتي لتشات جي بي تي (ChatGPT) بغرض تحسين أدائه في الترجمة. تقيّم الدراسة فعالية الاستراتيجيات المقترحة وتحدد الاستراتيجية الأكثر فعالية عند ترجمة النصوص الطبية باللغة الإنجليزية إلى اللغة العربية. اختار الباحثون خمسة عشر (١٥) نصًا طبيًا باللغة الإنجليزية، وتمت ترجمة هذه النصوص إلى اللغة العربية بالاعتماد على تشات جي بي تي (ChatGPT) باستخدام أمر ترجمة افتراضي. دُونت الملاحظات على النصوص المترجمة وقيِّمت يدويًا. وبعد ذلك، أعاد الباحثون ترجمة النصوص حيث أعيد ترجمة جزء من النصوص باستخدام التغذية الراجعة الذاتية، وإعادة ترجمة الجزء الآخر باستخدام التغذية الراجعة الخارجية. أستخدم كلتا استراتيجيات التغذية الراجعة لتسخير التصحيح الذاتي لتشات جي بي تي (ChatGPT). قيمت الترجمة الأولية وكذلك النصوص المُعاد ترجمتها يدوياً وتلقائياً لتقييم فعالية استراتيجيات التغذية الراجعة المستخدمة. قيمت فعالية كل من استراتيجيات التغذية الراجعة هذه بناءً على معدل التحسن في الترجمة المُعادة. وتوصلت الدراسة إلى استراتيجية التغذية الراجعة الأكثر كفاءة، وهي استراتيجية التغذية الراجعة لتصنيف الأخطاء. تعد الدراسة مهمة جداً للمترجمين، ومُدققي ما بعد التحرير، والباحثين، ومطوري الترجمة الآلية. كما أنها تقدم توصيات للأبحاث المستقبلية.

الكلمات المفتاحية ***التصحيح الذاتي، ذاتي التزويد ، خارجي، التغذية الراجعة، النصوص الطبية، تشات جي بي تي (ChatGPT)***

[1] لغة إنجليزية (لغويات) - قسم اللغة الإنجليزية - كلية العلوم التطبيقية والإنسانية - جامعة عمران – اليمن.
[2] لغة إنجليزية (لغويات) - قسم اللغة الإنجليزية - كلية العلوم التطبيقية والإنسانية - جامعة عمران – اليمن – البريد الالكتروني khalil.naji@usr.ac
[*] عنوان المراسلة: almekhlafihesham@gmail.com

Introduction

Machine translation is considered to be an inexpensive and time-saving alternative to human translation. However, it is important to get a translation output that meets the quality requirements, especially when it comes to scientific context, where the quality of translation must meet the maximum requirements.

The quality of machine translation, therefore, is a crucial aspect and, accordingly, a very interesting field of research. Investigating the performance of machine translation systems and proposing strategies to improve the quality of machine translation are the focus of the recent research literature on machine translation.

Accompanying the great advancement in the field of machine translation, there are continuous heated discussions regarding the quality of machine translation. There are proposals in the literature that state that machine translation has achieved parity with professional human translation (Hassan et al., 2018; Barrault et al., 2019). On the other hand, there are proposals that state that such parity has not been achieved yet and machine translation still falls short when it is compared to professional human translation (Läubli et al., 2018; Toral et al., 2018; Freitag et al., 2021).

Regardless of these debates, it can be stated that machine translation is advancing and that high-quality translations are provided by machine translation. However, it is also undeniable that there is still a gap between machine translation and professional human translation as proposed in the literature. Recent studies that have performed error annotation have come out with a comparatively long list of various translation errors (Popović, 2021; Kocmi, 2022).

Recently, with the advent of new large language models (LLMs), a new wave of research has started for the purpose of investigating the quality of translation provided by LLMs such as ChatGPT and Gemini. Such research is crucial since improvement of automatic translation requires more fine-grained analyses in regard to translation quality.

Recent studies on LLMs have shown that their translation outputs show numerous errors and their performance is still lacking (Zhu et al., 2023; Nagi et al., 2024). This poses a great threat, especially when it comes to medical translation. A translation error in the medical context can be critical and may lead to life-threatening situations (Karwacka, 2014; Mehandru et al., 2022).

Accordingly, careful post-editing is required. Studies have indicated that LLMs have also shown the capability of self-correction when provided with external feedback and self-provided feedback (Chen et al., 2023; Raunak et al., 2023; Feng et al., 2024; Nagi et al., 2024; Ki and Carpuat, 2024; and Hezam & Nagi, 2025). The study examines how ChatGPT's translation performance can be improved via the proposed strategies and whether it meets the quality requirements.

This study, therefore, is an effort in this area in which the translation outputs of ChatGPT are evaluated and errors are classified. The study evaluates ChatGPT's capability of self-correction using external and self-provided feedback.

1.1 Problem Statement

With ChatGPT being one of the most prominent LLMs, it is important to examine its performance in the field of translation. The model still suffers from the recurrence of translation errors, especially when translating from English to Arabic due to the wide divergence in morphological and syntactic structures. It is also because of the fact that there are not many annotated Arabic corpora. Due to that, the translation performance should be improved, especially when it comes to the translation in the medical field. Therefore, self-correction strategies should be examined to ensure high-quality translation.

1.2 Objectives of the study

The study aims at examining the effectiveness of feedback in ChatGPT's self-correction process. The study investigates two feedback strategies, namely self-provided feedback and external feedback, that induce ChatGPT's self-correction and, accordingly, the production of high-quality translation outputs. The study evaluates the effectiveness of the proposed strategies. The study evaluates the translation outputs and concludes their appropriateness to the medical context.

1.3 Research Questions

- How effective is the self-provided feedback in ChatGPT's self-correction process in the field of medical translation?
- How effective is the external feedback in ChatGPT's self-correction process in the field of medical translation?
- Do the translation outputs after self-correction meet the requirements of the medical context?

1.4 Significance of the Study

This study is the first that investigates the capacity of self-correction of ChatGPT in the case of translating from English to Arabic in the medical context. The study uses two different feedback strategies in the process. This is crucial to the development of automatic translation. Therefore, the study is a great addition to the literature of automatic translation and translation in general. To the best of the researchers' knowledge, it is the first study related to Arabic translation that investigates various feedback strategies of ChatGPT to perform self-correction on medical texts.

1.5 Limitation of the Study

The study is limited to English/Arabic translation using ChatGPT. The study is also restricted to a limited number of medical texts and reports. The concluded results,

therefore, may not apply to other languages or different contexts. This study examines two feedback strategies: external feedback and self-provided feedback.

1. Literature Review

2.1 ChatGPT

ChatGPT has recently become the most well-known artificial intelligence (AI) application whose popularity began even before its official launch, when some companies like BBC, CNN, and People's Daily announced the forthcoming AI revolution. Due to its ability to perform various tasks effectively, ChatGPT quickly rose in popularity. Generating text, classifying text, answering questions, writing code, and translating languages are some of those tasks (Siu, 2023).

Jiao et al. (2023) and Hendy et al. (2023) showed that the performance of ChatGPT in translation is on par with that of commercial translation systems such as Google Translate when it comes to the translation of high-resource European languages. However, it lags behind in translating low-resource languages. It is also indicated that ChatGPT struggles with translating biomedical abstracts or Reddit comments, but it excels at translating spoken language. According to Jiao et al. (2023), the GPT-3.5 model underperforms in specific domains when compared to its performance in the translation of spoken languages. Khoshafah (2023) mentioned that ChatGPT encounters difficulties with specialized texts such as scientific, medical, legal, or literary texts; however, it performs well with simple content. Zhu et al. (2023) also stated that while the multilingual translation capabilities of LLMs are improving, GPT-4 still has not achieved the desired level of performance when it comes to low-resource languages. Nagi et al. (2024) also asserted that ChatGPT translation outputs show a high error frequency when translating complex English sentences into Arabic.

2.2 Self-Correction

According to White et al. (2023), a prompt is referred to as "a set of instructions provided to an LLM that programs the LLM by customizing it and/or enhancing or refining its capabilities." The effectiveness of using prompts and various strategies of prompting to get a ChatGPT output with a higher quality is an interesting topic of research. It has been mentioned, in various fields, that there is a direct relationship between the output quality and the prompt quality and that using informed or specific prompts can effectively improve the output (White et al., 2023; Giray, 2023; Liu et al., 2023, among others).

In the field of translation, the effectiveness of prompt engineering on the translation output has received minimal attention. There are, however, a few interesting prompting strategies that have been implemented and confirmed to improve the translation output (Jiao et al., 2023; Gao et al., 2023; Siu, 2023). It has also been

indicated that the default prompts suggested by ChatGPT perform well with slight differences in their performance (Jiao et al., 2023).

It has been pointed out that special prompting strategies significantly enhance the quality of the translation produced by LLMs. One of these prompting strategies has been proposed by Jiao et al. (2023). It is referred to as the pivot strategy. According to this strategy, ChatGPT is required to translate source text into a high-source pivot language before translating into the target language. Jiao et al. (2023) have stated that this strategy significantly improves the translation quality.

Gao et al. (2023) have also pointed out that ChatGPT outperforms commercial translation systems when thoughtfully designed prompts are employed. In their study, they have suggested prompts that contain translation task information (both the target language and the source language are identified), context domain information (the domain of the text is identified, such as news, legal, etc.), or part-of-speech tags. It has been pointed out that the study results demonstrate that the suggested prompts significantly improve the performance of ChatGPT in translation. It has also been noted that prompts with contextual information enable ChatGPT to produce enhanced translation output (Siu, 2023).

An interesting study in this aspect has been carried out by Gu (2023), in which linguistically informed prompts have been incorporated and utilized in the translation of Japanese attributive clauses into Chinese. It has been pointed out that such prompts enhance the translation accuracy by more than 35%.

The use of a self-correction strategy is also a very interesting recent approach to informed prompting in LLM translation, where the model is asked to modify the original translation using a proposed strategy. Chen et al. (2023), Raunak et al. (2023), and Feng et al. (2024) have employed adopted strategies that rely on prompting the model to self-correct its previous translation. Nagi et al. (2024) and Hezam & Nagi (2025) have also adopted self-correction strategies to improve ChatGPT translation performance. This approach has been proven to be effective, and better translation outputs have been produced by the investigated LLMs using such an approach.

LLMs' self-correction strategies can depend on self-provided feedback, i.e., feedback that is provided by the LLM itself with prompts such as *give a better translation*. It can also depend on external feedback, where the models are provided with the translation errors or even their types.

Since the use of the self-correction approach to improving the translation output seems promising, further investigation is required. New research will help propose more effective and stable strategies. This study, therefore, aims to examine strategies that improve LLMs' translation performance within the framework of the self-correction approach. The researchers use both the self-provided feedback and the external feedback and examine the effectiveness of each method.

To the best of the researchers' knowledge, this will be the first study that examines the capacity of ChatGPT's self-correction using different feedback strategies that improve the Arabic translation of English texts.

2.3 Medical Translation

Medical translation is very significant from different points of view since it helps spread the knowledge and recent breakthroughs related to the medical field. It also helps in providing better healthcare services to foreigners or in areas where international organizations provide their services. There are major issues and challenges to medical translation. Some of these issues are medical terminology, lexical equivalence of medical texts, and readability and quality issues (Karwacka, 2015). Outstanding translation quality should be assured in medical contexts, and a mistranslation can result in substantial consequences, including life-endangering ones (Karwacka, 2014; Mehandru et al., 2022).

Recent studies have pointed out that Arabic machine translation in general faces a genuine challenge when it comes to medical context. Mahadin and Olimat (2022) have stated that due to the complexity of Arabic, rendering COVID-19 information in the Jordanian context, machine translation becomes a hard task. According to Khoshafah (2023), ChatGPT cannot perform well in dealing with translations of specialized texts, including medical texts, while it can do well in translating simple content. Alzain et al. (2024) also have concluded that ChatGPT performs poorly when translating scientific texts, including medical texts. They have stated that the translated texts show high error frequency, suggesting more training.

3 Methodology and Results

3.1 Texts and Initial Translation

The researchers select 15 English medical reports and medical texts from different essays. The texts are examined by a team of four linguist experts, including the authors, to ensure their equivalence. The texts are divided by the team into three equivalent groups of five texts each. All the texts are translated by ChatGPT using the default translation prompt. *Translate these texts from English into Arabic*.

3.2 Evaluation

Each group of the translated texts is processed differently. One of the translated text groups is left without annotation by the experts to be processed using the self-provided feedback strategy. The other two translated text groups are annotated by four professional annotators, where the errors are marked in one group and marked and classified in the other. In other words, the expert annotators point out the errors in the first group without identifying their types. They, however, point out the errors in the other group and identify their types according to the MQM framework as proposed in Lommel (2014) and later modifications. (Check https://themqm.org/the-mqm-full-typology/ for the detailed MQM taxonomy.)

To evaluate the effectiveness of the used feedback strategies, manual evaluations of the initial translation and the retranslation are performed by the four professional

annotators using a secular quality metric (Freitag et al., 2021). This metric uses a 0-6 Likert-like scale. Its ranks are as follows:

6: Perfect Meaning and Grammar: The meaning of the translation is completely consistent with the source and the surrounding context (if applicable). The grammar is also correct.

· 4: Most Meaning Preserved and Few Grammar Mistakes: The translation retains most of the meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies.

· 2: Some Meaning Preserved: The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor.

· 0: Nonsense/No meaning preserved: Nearly all information is lost between the translation and source. Grammar is irrelevant.

3.3 Self-provided Feedback Strategy

Five of the texts that are translated by ChatGPT using the default prompt are used here. ChatGPT is provided with one of the Arabic translated texts and the corresponding English source text along with a prompt that asks ChatGPT for a retranslation as indicated below.

Source Text:

HAV is a single-stranded RNA virus. Infectivity occurs primarily through fecal-oral transmission, and after ingestion and absorption, the virus replicates in the liver and is excreted in bile. HAV enters into the hepatocyte by specific receptors located on the plasma membrane. Viral RNA is uncoated after uptake and binds to ribosomes, stimulating the synthesis of viral proteins, and replication of the viral genome occurs by RNA polymerase. The virus is then secreted into the biliary tree and excreted in feces, where high concentrations of HAV are detected [12, 13]. An immunologic response occurs within the liver, leading to portal and periportal lymphocytic infiltration, and potentiates liver damage. A limited number of cases have been reported from blood transfusions and vertical transmission from mother to fetus.

Initial Translation:

فيروس التهاب الكبد A (HAV) هو فيروس RNA أحادي السلسلة. تحدث العدوى بشكل أساسي عبر الانتقال البرازي-الفموي، وبعد تناوله وامتصاصه، يتكاثر الفيروس في الكبد ثم يُفرز في الصفراء.

يدخل HAV إلى الخلايا الكبدية عبر مستقبلات محددة موجودة على الغشاء البلازمي. بعد دخوله، يتم إزالة الغلاف الفيروسي، ويرتبط الحمض النووي الريبي الفيروسي (RNA) بالريبوسومات، مما يحفز إنتاج البروتينات الفيروسية، كما تتم مضاعفة الجينوم الفيروسي بواسطة إنزيم بوليميراز.RNA

يتم بعد ذلك إفراز الفيروس في القنوات الصفراوية ومن ثم إخراجه عبر البراز، حيث يمكن اكتشاف تركيزات عالية من الفيروس فيه.

تحدث استجابة مناعية داخل الكبد، مما يؤدي إلى ارتشاح الخلايا الليمفاوية البابية وحول البابية، مما يساهم في إلحاق الضرر بالكبد.

https://doi.org/10.20428/jss.v31i10.3440

تم الإبلاغ عن عدد محدود من الحالات التي انتقلت فيها العدوى عبر نقل الدم أو الانتقال العمودي من الأم إلى الجنين.

Prompt: Improve the initial Arabic translation and provide the final Arabic translation.

Both the initial translation and the retranslation are evaluated by the team of professional annotators to measure the improvement of translation and, accordingly, the effectiveness of self-provided feedback. The results come as shown in Table 1 below.

Feedback Strategy	Original Translation	Retranslation	Improvement
Self-provided	67.50%	82.50%	15 %

3.4 Error Marking Feedback Strategy

Error marking is considered a human annotation method. In this method, the annotators mark the translation issues in MT outputs without assigning error labels. According to Kreutzer et al. (2020), this method significantly minimizes the human annotation effort. Popović (2020) has also argued that this method is more beneficial than error classification, as it is both informative and less demanding. Such an annotation method can be further leveraged to conduct distinct types of analyses (Popović, 2020). In this study, errors are marked, and the errors in one group of texts form the source of external feedback.

As with the rest of the groups, five texts are used here. After getting the initial translation, the texts are processed by the annotation experts. The errors in the translated texts are marked. At that point, each text is joined with the source text, the initial ChatGPT translation from English to Arabic, and the list of translation errors without categorization. The following format illustrates how this process is performed.

Source Text

These recommendations are based on two studies in adults. In the first, lamivudine was administered to 17 patients with fulminant HBV in a German center. Of these, 14 recovered without liver transplantation, two required transplants, and one died from cerebral herniation. An additional 20 patients from five other centers were also included in the report, of whom 15 survived without transplantation [3]. In a randomized control trial of 71 adult patients (31 lamivudine, 40 placebo) reported from India, lamivudine decreased HBV DNA level more than placebo, but there was no difference in clinical outcome [4]. There are no data regarding treatment of acute HBV in children.

Initial Translation:

تستند هذه التوصيات إلى دراستين أجريتا على البالغين. في الدراسة الأولى، تم إعطاء اللاميفودين لـ ١٧ مريضًا يعانون من التهاب كبدي B خاطف في مركز طبي ألماني. من بين هؤلاء، تعافى ١٤ مريضًا دون الحاجة إلى

زراعة كبد ، واحتاج اثنان إلى زراعة، بينما توفي مريض واحد نتيجة فتق دماغي. وقد شمل التقرير أيضًا ٢٠ مريضًا إضافيًا من خمسة مراكز أخرى، نجا منهم ١٥ دون الحاجة إلى زراعة كبد \[٣]. في تجربة عشوائية محكومة أجريت على ٧١ مريضًا بالغًا (٣١ تلقوا لاميفودين، و٤٠ تلقوا علاجًا وهميًا) في الهند ، قلل اللاميفودين من مستوى الحمض النووي لفيروس HBV أكثر من العلاج الوهمي، لكن لم يكن هناك فرق في النتيجة السريرية \[٤]. لا توجد بيانات متوفرة حول علاج التهاب الكبد B الحاد لدى الأطفال.

" Initial Translation Errors: "خاطف B التهاب كبدي،" "من بين هؤلاء، تعافى ١٤ مريضًا،" "واحتاج اثنان،" "وقد شمل التقرير أيضًا،" "نجا منهم،" "في تجربة عشوائية في الهند محكومة،" "قلل،" "لكن لم يكن هناك فرق في،" "لا توجد بيانات متوفرة حول"

Prompt: Based on the initial translation errors, please provide the final Arabic translation.

With this process of external feedback, a modified translation is given by ChatGPT. This final translation and the initial translation are evaluated by the team of professional annotators. using as a secular quality metric (Freitag et al., 2021), as mentioned earlier.

After evaluating both translations, the results are as shown in Table 1 below.

Feedback Strategy	Original Translation	Retranslation	Improvement
Error Marking	64.17%	82.92%	18.75%

3.5 Error Taxonomy Feedback Strategy

This process also includes 5 texts translated using ChatGPT's default prompt. In contrast to the prior group, the texts in this group are processed in accordance with a more organized error analysis that includes an error taxonomy. Each text here is combined with the source text, the initial ChatGPT translation from English to Arabic, and the list of translation errors with taxonomy. The types of errors made in the original Arabic translation are categorized according to the MQM error taxonomy (e.g., word order, cohesion, function words, and terminology). The process once more includes the source text, the initial translation, and the classified errors. The example below demonstrates this approach. The following format illustrates how this process is performed.

Source Text:

He received 6 cycles of induction with bortezomib-lenalidomide-dexamethasone and then commenced lenalidomide maintenance therapy. His initial disease was complicated by multilevel vertebral disease, hypercalcemia, and renal impairment.

During ongoing maintenance therapy, due to immune suppression, Fie has suffered at least 3 episodes of bacterial pneumonia requiring IV antibiotics and inpatient care. He continues to have some physical limitations due to previous vertebral involvement and therefore requires supervision and surveillance for ongoing risk of complications.

For these medical reasons, travel abroad would not be advisable.
Thank you.

Initial Translation:

تم تشخيص هذا المريض بالإصابة بالورم النقوي المتعدد في أكتوبر ٢٠١٧.
تلقى ٦ دورات من العلاج التحريضي باستخدام بورتيزوميب-لينالیدوميد-ديكساميثازون، ثم بدأ علاج الصيانة باستخدام لينالیدوميد. كانت حالته الأولية معقدة بسبب إصابته بمرض فقري متعدد المستويات، وفرط كالسيوم الدم، وضعف في وظائف الكلى.
أثناء العلاج المستمر بالصيانة، وبسبب كبت المناعة، عانى من ثلاث نوبات على الأقل من الالتهاب الرئوي الجرثومي تطلبت مضادات حيوية وريدية ورعاية داخل المستشفى. لا يزال يعاني من بعض القيود الجسدية بسبب تأثر الفقرات سابقًا، ولذلك فهو بحاجة إلى إشراف طبي ومراقبة مستمرة بسبب خطر حدوث مضاعفات.
لأسباب طبية، لا يُنصح بسفره إلى الخارج.
شكرًا لكم.

Initial Translation Errors: "تم تشخيص هذا " / word form," "بالإصابة" / word form," "المريض بالإصابة / style," "تلقى" …… / omission," "دورات / wrong term," "التحريضي" / wrong term," "الصيانة" / wrong term," "كانت / wrong term & cohesion," "المستويات" / punctuation," "الدم / punctuation," "الصيانة" / wrong term," "عانى" …… / omission," "……." "لأسباب طبية / missing function word," "يزال" ……. / omission," "خطر / wrong term," "تطلبت / style," "شكرًا لكم / style"

Prompt: Based on the initial translation errors and their types, please provide the final Arabic translation.
With this process of external feedback, a modified translation is given by ChatGPT. This final translation and the initial translation are evaluated by the team of professional annotators. using as a secular quality metric (Freitag et al., 2021), as mentioned earlier.

After evaluating both translations, the results are as shown in Table 1 below.

Feedback Strategy	Original Translation	Retranslation	Improvement
Error Taxonomy	63.33%	89.17%	25.84%

4 Discussion

The study explores the effectiveness of three different feedback strategies on the improvement of ChatGPT performance when translating medical texts from English to Arabic. The result shows that the quality of translation promisingly improved with the three feedback strategies used in the study. Despite the fact that the resulting translation is not perfect, it shows great improvement, especially in the case of the error taxonomy feedback strategy. Following this strategy, the retranslation scores

89.71% according to the experts' evaluation, with a 25.84% improvement. This improvement undeniably confirms the effectiveness of this method.

Other feedback strategies do not disappoint either. With a percentage of 18.75%, the error-marking feedback strategy also shows a great improvement. The retranslation scores a high percentage of 82.92% according to the evaluation of the experts. The self-provided feedback strategy also shows a high percentage of improvement since it achieves an improvement percentage of 15%, scoring 82.5 in terms of the total quality evaluation.

The improvement resulting from using these feedback strategies shows that both error marking and self-provided feedback strategies have achieved good improvement, and the final quality is very good. On the other hand, the error taxonomy feedback strategy has achieved a better improvement, and the final quality of translation is higher.

As mentioned earlier, the retranslation is still not perfect. However, with the high quality achieved with the error taxonomy feedback strategy, it is a step forward toward getting a translation with a higher quality and almost a perfect one, taking into consideration that the texts used in this study are medical texts.

5 Conclusion

The study investigates the effectiveness of three feedback strategies on the improvement of the quality of ChatGPT translation of English medical texts to Arabic. The study findings exhibit that feedback strategies improve the initial ChatGPT translation effectively. The error taxonomy feedback strategy is the most effective and promising strategy. Despite the fact that the translation is still not on par with professional translation, it seems with more investigation and training with various data, the translation will be more accurate.

It can be stated here that ChatGPT's response to feedback has shown its potential for development and that it is and will be a great asset to translators, particularly in all fields, including medical translators. Therefore, further future studies should be performed with the concentration on accurate error identification and prompt engineering. More investigation should focus on creating feedback strategies and how to make LLMs more adaptable to feedback. This will undoubtedly lead to more precise and accurate translation outputs.

Acknowledgement

This research received grant no. (569/2024) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

References

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., ... & Zampieri, M. (2019). *Findings of the 2019 Conference on Machine Translation (WMT19).* In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* (pp. 1–61). http://www.statmt.org/wmt19/pdf/53/WMT01.pdf

Chen, P., Guo, Z., Haddow, B., & Heafield, K. (2023). *Iterative translation refinement with large language models.* arXiv:2306.03856. https://doi.org/10.48550/arXiv.2306.03856

Feng, Z., Zhang, Y., Li, H., Liu, W., Lang, J., Feng, Y., Wu, J., & Liu, Z. (2024). *Improving LLM-based machine translation with systematic self-correction.* arXiv:2402.16379. https://doi.org/10.48550/arXiv.2402.16379

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. https://doi.org/10.1162/tacl_a_00437

Gao, Y., Wang, R., & Hou, F. (2023). *How to design translation prompts for ChatGPT: An empirical study.* arXiv:2304.02182. https://doi.org/10.48550/arXiv.2304.02182

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, 1–5. https://doi.org/10.1007/s10439-023-03272-4

Gu, W. (2023). *Linguistically informed ChatGPT prompts to enhance Japanese–Chinese machine translation: A case study on attributive clauses.* arXiv:2303.15587. https://doi.org/10.48550/arXiv.2303.15587

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., ... & Zhou, M. (2018). Achieving human parity on automatic Chinese to English news translation. arXiv:1803.05567. https://doi.org/10.48550/arXiv.1803.05567

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., ... & Awadalla, H. H. (2023). *How good are GPT models at machine translation? A comprehensive evaluation.* arXiv:2302.09210. https://doi.org/10.48550/arXiv.2302.09210

Hezam, A., & Nagi, K. A. (trans.) (2025). ChatGPT's self-correction between error marking and fine-grained taxonomy: Is taxonomy worth the effort? *Journal of Social Studies, 31*(7). https://doi.org/10.20428/jss.v31i7.3182

Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023). *Is ChatGPT a good translator? A preliminary study.* arXiv:2301.08745. https://doi.org/10.48550/arXiv.2301.08745

Karwacka, W. (2014). Quality assurance in medical translation. *The Journal of Specialised Translation, 21*, 19–34.

Karwacka, W. (2015). Medical translation. In Ł. Bogucki, S. Goźdź-Roszkowski, & P. Stalmaszczyk (Eds.), *Ways to translation* (pp. 271–298). Wydawnictwo Uniwersytetu Łódzkiego.

Khoshafah, F. (2023). *ChatGPT for Arabic–English translation: Evaluating the accuracy.* ResearchSquare. https://doi.org/10.21203/rs.3.rs-2814154/v1

Ki, D., & Carpuat, M. (2024). *Guiding large language models to post-edit machine translation with error annotations.* arXiv:2404.07851.

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., ... & Popović, M. (2022). *Findings of the 2022 Conference on Machine Translation (WMT22).* In *Proceedings of the Seventh Conference on Machine Translation* (pp. 1–45). https://aclanthology.org/2022.wmt-1.1

Kreutzer, J., Berger, N., & Riezler, S. (2020). Correct me if you can: Learning from error corrections and markings. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 135–144). https://aclanthology.org/2020.eamt-1.15

Läubli, S., Sennrich, R., & Volk, M. (2018). Has machine translation achieved human parity? A case for document-level evaluation. In *2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4791–4796). https://doi.org/10.18653/v1/D18-1512

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in NLP. *ACM Computing Surveys, 55*(9), 1–35. https://doi.org/10.1145/3560815

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica, 12*, 455–463. https://doi.org/10.5565/rev/tradumatica.77

Mahadin, D. K., & Olimat, S. N. (2022). Jordanian translators' use of machine translation and glossary of COVID-19 terminology with reference to Arabic. *New Voices in Translation Studies, 26*(1), 25–54.

Mehandru, N., Robertson, S., & Salehi, N. (2022). Reliable and safe use of machine translation in medical settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2016–2025).

Nagi, K. A., Alzain, E., & Naji, E. (2024). Informed prompts and improving ChatGPT English to Arabic translation. *Al-Andalus Journal for Humanities & Social Sciences, 98*(11). https://doi.org/10.35781/1637-000-098-007

Popović, M. (2020). Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5059–5069). https://doi.org/10.18653/v1/2020.coling-main.444

Popović, M. (2021). On nature and causes of observed MT errors. In *Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track)* (pp. 163–175). https://aclanthology.org/2021.mtsummit-research.14

Raunak, V., Sharaf, A., Wang, Y., Awadalla, H., & Menezes, A. (2023). Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 12009–12024). https://aclanthology.org/2023.findings-emnlp.804

Siu, S. C. (2023). *ChatGPT and GPT-4 for professional translators: Exploring the potential of LLMs in translation.* SSRN. http://dx.doi.org/10.2139/ssrn.4448091

Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Attaining the unattainable? Reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 113–123). https://aclanthology.org/W18-6312

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT.* arXiv:2302.11382. https://doi.org/10.48550/arXiv.2302.11382

Zhu, W., Liu, H., Dong, Q., Xu, J., Kong, L., Chen, J., ... & Huang, S. (2023). Multilingual machine translation with large language models: Empirical results and analysis. arXiv:2304.04675. https://doi.org/10.48550/arXiv.2304.04675