# Informed Prompts and Improving ChatGPT English to Arabic Translation

# الموجهات المستنيرة ودورها في تحسين ترجمة ChatGPT من الإنجليزية إلى العربية

10.35781/1637-000-098-007

**Khalil A Nagi**[(1)]
**Elham Alzain** [(2)]
**Ebrahim Naji** [(3)]

1) University of Saba Region
khalil.naji@usr.ac- https://orcid.org/0009-0002-1229-108X
2) King Faisal University
elhamalzain@gmail.com- https://orcid.org/0000-0002-6330-3100
3) Trine University
najie@trine.edu- https://orcid.org/0009-0007-5736-3477

## Abstract:

The aim of the study is to investigate the quality of ChatGPT translation and the effectiveness of using informed prompts to improve it. The research team built a dataset composed of various English complex sentence types (150complex sentences) that are selected from various news sites. The sentences were translated into Arabic using a default ChatGPT translation prompt (Translate the following sentences into Arabic). The translated sentences were annotated by three professional annotators. An error taxonomy was performed based on the Multidimensional Quality Metrics (MQM). The results of the error taxonomy showed a high error frequency that amounts to 2.73 errors per sentence which indicates that ChatGPT falls short when translating English complex sentences into Arabic and that it still needs to be trained effectively. The sentences whose translation outputs had the most errors were translated again using informed prompts that require the model to correct the original translation. Both the original and the new translation outputs were evaluated manually by the professional annotators and automatically using the BLEU metric. The study, therefore, identifies the effectiveness of the adopted prompt strategies in improving translation quality and recommends further research in the area of informed prompts.

**Keywords:** ChatGPT, prompts, error taxonomy, English-Arabic, translation output, complex sentences.

ISSN : 2410-1818

# الموجهات المستنيرة ودورها في تحسين ترجمة ChatGPT من الإنجليزية إلى العربية

د. خليل عبدالسلام خالد ناجي
د. إلهام دحان علي الزين
د. إبراهيم عبدالجليل خالد ناجي

الملخص

**هدفت هذه الدراسة إلى التحقق من جودة ترجمة ChatGPT وفعالية استخدام الموجهات المُستنيرة لتحسينها. قام فريق البحث ببناء مجموعة بيانات تتكون من أنواع مختلفة من الجمل الإنجليزية المركبة (150 جملة مركبة) تم اختيارها من مواقع إخبارية متنوعة. تمت ترجمة الجمل إلى اللغة العربية باستخدام موجه ترجمة افتراضي (ترجم الجمل التالية إلى العربية). ثم تم تحرير الجمل المترجمة من قبل ثلاثة محررين محترفين، وتم إجراء تصنيف للأخطاء بناءً على معايير الجودة متعددة الأبعاد .(MQM) تُظهر نتائج تصنيف الأخطاء معدل خطأ مرتفع يبلغ 2.73 خطأ لكل جملة، مما يدل على أنChatGPT لا يزال فيها قصور كبير عند ترجمة الجمل الإنجليزية المركبة إلى اللغة العربية ولا يزال بحاجة إلى التدريب بشكل فعال. تم بعد ذلك ترجمة الجمل التي تحتوي ترجمتها على أكبر عدد من الأخطاء مرة أخرى باستخدام موجهات مُستنيرة تطلب من ChatGPTتصحيح الترجمة الأصلية. كما تم تقييم كل من مخرجات الترجمة الأصلية والجديدة يدوياً من قبل المحررين المحترفين. وبالتالي، خلصت الدراسة إلى أن استراتيجيات المطالبات المستخدمة كانت فعالة في تحسين جودة الترجمة وأوصت بمزيد من البحث في مجال الموجهات المُستنيرة.**

**الكلمات المفتاحيةChatGPT :، الموجهات، تصنيف الأخطاء، اللغة الإنجليزية- اللغة العربية، الترجمة، الجمل المركبة**

## 1.Introduction

The quality of machine translation is a very interesting field of research. Accompanying the great advancement in this field, there are continuous heated discussions regarding the quality of machine translation. In the literature, there are proposals 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 which state that such parity has not been achieved yet (Läubli et al, 2018; Toral et al 2018; Freitag et al, 2021).

Regardless of these debates, there is no doubt that machine translation is advancing and that high-quality translations are performed by machine translation. However, it is also undeniable that there is still a gap between machine translation and professional human translation. Recent studies that have performed error analysis have come out with a comparatively long list of errors (Popović, 2021; Kocmi et al., 2022).

ISSN : 2410-1818

However, 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 Bard.

The improvement of automatic translation requires more fine-grained analyses in regard to translation quality. It is, therefore, an interesting topic to scrutinize areas of translation that form a problem for NMT systems and evaluate the quality of translation of both NMT systems and LLMs. It is also very important to come out with strategies that improve the translation quality. This study will be an effort in this area.

The study aims to investigate the effectiveness of informed prompts in improving ChatGPT English to Arabic translation. The research team performs an error taxonomy and translate the sentences with the most erroneous translation outputs using informed prompts. The performance of ChatGPT in English to Arabic translation with default is compared with that of informed prompts to check the effectiveness of the latter in improving the translation.

The study provides an error taxonomy of ChatGPT translation of English complex sentences into Arabic. It also investigates the effectiveness of using informed prompts in improving such translation. Both ChatGPT and the effectiveness of informed prompts in improving translation are either under-investigated. Complex sentences, on the other hand, form a challenge to machine translation. The study, therefore, is of great significance since it will contribute to the literature of MT, especially LLMs, and to improving automatic translation.

A brief introduction of ChatGPT, prompts, and complex sentences is presented in Section 2. Section 3 presents the methodology and the research results. Section 4 presents the discussion and conclusion.

## 2.Literature Review

This section generally introduces ChatGPT and briefly presents prompts and how they are utilized to produce better outputs. It also presents a concise discussion on complex sentences and some of the differences between complex sentences in English and Arabic.

### 2.1 ChatGPT

ChatGPT is the most famous large language model (LLM) nowadays. It has gained its fame even before its advent, as people anticipated its usage in daily life when some companies, such as BBC, CNN, and People's Daily, announced the upcoming AI revolution. The reason for the availability of ChatGPT is due to the technology development which started by using Natural Language Processing (NLP) in 1950s, an information processing technology based on natural language understanding and natural language generation, and the usage of Recurrent Neural

ISSN : 2410-1818

Network (RNN) and Long-Term Short-Term Memory (LTSM) models in recent years. Chat-GPT is an enhanced variant, improved specifically for conversational purposes, of the previous GPT models (text-davinci-002 and text-davinci-003). Text-davinci-002 is a large language model which was trained by using Reinforcement Learning with reward models. The reward models were trained based on human comparisons of different text outputs. Text-davinci-003 is an enhanced version of text-davinci-002 (Hendy, et al, 2023). Through Chat-GPT people can perform multiple tasks, after having access to it in November 2022, such as office work, digital images, coding, academic research, and translation more efficiently and quickly (Li, 2023). According to Siu, (2023), Chat-GPT has become very popular quickly because it can perform many tasks such as generating texts, answering questions, classifying texts, generating codes, and translating languages very well. The reason behind this is that Chat-GPT employs many methods like Natural Language Processing (NLP), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Transformer and Reinforcement Learning from Human Feedback model (RLHF).

According to Jiao et al., (2023), Chat-GPT is as good as commercial translation systems like Google Translate when translating high-resource European languages, but it falls behind at translating low-resource or distant languages. In addition, Chat-GPT is not as good as commercial translation systems at translating biomedical abstracts or Reddit comments, but it is good at translating spoken language. However, the launch of the GPT-4 engine on March 2023 has significantly improved Chat-GPT's translation performance, making it comparable to commercial translation systems even for distant languages.

## 2.2 ChatGPT and Prompts

A prompt is identified as “a set of instructions provided to an LLM that programs the LLM by customizing it and/or enhancing or refining its capabilities” (White et al., 2023).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. In various fields, it has been stated that the output quality has a direct relationship with 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, there has been little investigation on the effectiveness of prompt engineering on the translation output. There are, however, a few interesting prompting strategies that have been used and proven effective in improving the translation output (Jiao et al., 2023; Gao et al, 2023; Siu, 2023). It has been also indicated that the default prompts suggested by ChatGPT perform well with slight differences in their performance (Jiao et al., 2023).

ISSN : 2410-1818

It has been pointed out that special prompting strategies contribute greatly to the quality of the translation produced by LLMs. One of these prompting strategies has been suggested by Jiao et al. (2023). It is dubbed the pivot strategy. According to this strategy, ChatGPT is required to translate a 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 surpasses commercial translation system when properly designed prompts are used. In their study, they have proposed 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 indicated that the results of the study have shown that the proposed prompts significantly enhance the performance of ChatGPT in translation.It has also been indicated that prompts with contextual information enable ChatGPT to produce improved translation output (Siu, 2023).

An interesting study in this aspect has been conducted by Gu (2023) in which linguistically informed prompts have been introduced and used in the translation of Japanese attributive clauses into Chinese. It has been stated that such prompts improve the translation accuracy by more than 35%. There is also a very interesting recent approach to informed prompting in LLM translation which is the use of self-correction strategy 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 used adopted strategies that depend on prompting the modal to self-correct its previous translation.

It seems that the use of self-correction approach of prompting is effective in improving the translation output greatly. This requires further investigation with different language pairs, especially when a low-resource language, like Arabic, is involved. New research also need to propose more effective prompting strategies. This study is an endeavor in examining a new method of using prompts in accordance with the self-correcting approach.

## 2.3 Complex sentences and MT challenges

Complex sentences are made up of a main clause and one or more subordinate clauses, with the main clause being the primary focus. Subordinate clauses are of different types. The major types of subordinate clauses are complement clauses, relative clauses, and adverbial clauses (Miller, 2002; Diessel, 2004). These types of subordinate clauses can be identified as follows:

- Complement clauses fill in slots in the main clause that can also be occupied by noun phrases. These clauses complete the syntax of a verb in the main clause, either by following or preceding it.

ISSN : 2410-1818

- Relative clauses modify nouns and usually follow them, unlike adjectives that precede nouns. They usually provide additional information about the noun they modify

- Adverbial clauses, on the other hand, modify entire clauses and are classified by their meaning, such as reason, time, concession, manner, or condition. They are adjuncts, meaning they are optional in sentences. (Miller, 2002)

This type of sentences is proven to be interesting in the various fields of linguistic research. In the field of machine translation, it has proven that complex sentences form a challenge for machine translation, especially when it involves a morphologically rich language (Qasmi et al., 2020; Turganbayeva et al, 2022). Certain strategies have been proposed to deal with the issue, the most prominent of which is sentence / text simplification (Hasler et al., 2017; Štajner & Popović, 2018; Sulem et al., 2018; Lu et al., 2021). In Arabic MT, it is also recognized that complex sentences form a serious problem for machine translation (Nagi, 2023).

## 2.4 Variation in Arabic complex sentence structure

There are various structural differences between English and Arabic complex clauses. Some of these differences are pointed out as follows:

- In English sentences, dependent clauses mostly come before independent clauses. In Arabic, however, it is preferable to start with the independent clause.
- A referring expression in Arabic usually follows the antecedent. In English, however, whether the referring expression precedes or follows the antecedents are equally acceptable.
- Conjunctions like “wa-” (and) and “fa-” are used at the beginning of the second clause to affirm its cohesion with the preceding clause in Arabic due to the fact that Arabic has an extremely syndetic discourse (Farghal, 2017). Accordingly, Arabic complex sentences will have more conjunctions than the English ones.
- Complex sentences in Arabic require more agreement patterns than simple sentences such as agreement related to relative pronouns. English, on the other hand, has a poor agreement system compared to Arabic

All these variations add extra problems to machine translation which add a great deal to the significance of this study.

## 3. Methodology and Results

This section of the study presents the methodology and the results. It introduces the datasets used, how errors are analyzed and the results of error annotation. In addition, it introduced the prompting strategy, the evaluation of the

ISSN : 2410-1818

original translation, and the evaluation of the retranslation after applying the suggested prompting strategy.

### 3.1 The Main Dataset

The research team built a dataset with 150 complex sentences. Complex sentences are used in this study since they are confirmed to form a challenge for machine translation as indicated earlier. The sentences are selected from recent news essays to ensure that they are not included in the ChatGPT training data. The selected sentences are translated by ChatGPT-4 using the default prompt Translate these sentences into [TL].

### 3.2 Error Taxonomy

The sentences are, then, annotated by 3 professional annotators. The classification of errors in this study follows the error taxonomy provided by Multidimensional Quality Metrics (MQM), introduced in Lommel et al. (2014). The taxonomy provided by MQM divided translation errors into eight dimensions: terminology, accuracy (adequacy), linguistic conventions (fluency), style, locale conventions, audience appropriateness, design and markup, and custom. Such dimensions are defined and classified further.

The annotated errors fall under the terminology, accuracy (adequacy), linguistic conventions (fluency), and style MQM dimensions. They are classified further into different category as represented in Table 1 below. However, before presenting the number and classification of annotated errors, let us briefly explain each type of error and provide clarifying examples.

**Terminology Errors:** The errors of this dimension refers to the ones where a term in the translation does not represent the term in the source text correctly. The following categories represent the annotated errors under this dimension:

- **Wrong Term:** *According to this,* a term that a professional translator does not usually use in a certain context is used in the target text. It can also refer to the use of a certain term which can cause conceptual mismatch. The translation of the English term "provoke" as "استفزاز" instead of "إثارة" is an example of this error. The latter is more correct and suitable to the context since the former has a negative reference.

- **Inconsistent use of terminology:** This kind of error refers to the case where multiple terms are used in the translated text as equivalents of the same term in the source text where consistency is required. The translation of the English term "cape" as in "Cape Grim" as both "كيب" and "رأس" in the same sentence is an example of this error.

ISSN : 2410-1818

**Accuracy Errors (Adequacy Errors):**Errors that make the content of the target text an inaccurate match of the propositional content of the source text fall under this dimension. There can be various reasons for this such as additions, omissions, or distortions of the original message. The annotated errors in the sentences under study that fall under this dimension are subcategorize as follows.

- **Ambiguous Target Content***:* This category of errors indicates that the translated text or a part of it can be interpreted in more than one way. The Arabic word "يتذكر" can be understood as "remember" instead of the intended "be remembered". To avoid the confusion, it should be either translated as "يتم تذكر" or diacritical marks should be used.

- **Ambiguous Source Content***:* This type of error pertains to the case where the source text or a part of it can be interpreted in more than one way. For example, the term "space" is considered to be a homonym, and therefore, translating it into "فضاء" where "مساحة" is required is an example of this error.

- **Untranslated:** This type of error refers to a part of the target text that was left untranslated. Translating "***magna cum laude***" as "**ماغنا كم لاودي**" is an example of this. The English phrase here is transcribed in Arabic letters and not translated.

- **Omission:** This type of error refers to the case where a content of the source text is not present in the translated text. Omitting the Arabic word "الفائدة" and translating "rate cuts" simply as "خفض الأسعار" instead of "خفض أسعار الفائدة"is an example of this error which cause a loss in meaning.

- **Overly Literal:** The word for word translation fall under this category of errors when an idiomatic translation is required due to the idiomatic nature of the source text. Translating the following English sentence to the Arabic sentence below it is an example of this.

Only around 5% of women tend to give birth on their due dates, research shows.

تظهر الأبحاث أن حوالي 5٪ فقط من النساء يميلن إلى الولادة في مواعيد استحقاقهن.

The phrase "مواعيد استحقاقهن" is a literal translation of its English counterpart and it is not used like that in Arabic.

**Linguistic Convention Errors (Fluency Errors):**Under this dimension, errors are linked to the linguistic well-formedness of the target text. The annotated errors in the texts under study that belong to this dimension are classified into the following:

ISSN : 2410-1818

- **Incorrect Function Word:** This type of error is related to the incorrect use of function words. That is to say, an article or a preposition, for example, is not used correctly and the use of another article or preposition is more correct. For example, using "حتى" instead of"بل" is an error when translating "even" in "and even praised".

- **Missing Function Word:** This type of error is also related to the use of function words. That is to say, an article or a preposition, for example, is required in the target text but it is not present. For example, the translation of "what is something we can do" as "ما هو شيء يمكننا القيام به" is not correct. An article and a pronoun are needed and the translation should be " ما هو الشيء الذي يمكننا القيام به" to be more acceptable.

- **Extraneous Function Word***:* This type of error is also related to the use of function words. However, as opposed to the "missing function word", this type of error refers to the use of unnecessary function word, such as an article or a preposition, in the target text. The use of "معها" in "أنها ستصطحب طفلها الجديد معها" is unnecessary as a translation of "that she'd have her new baby in tow". The meaning of the preposition and its complement is included in the verb meaning.

- **Word Form:** This type of error is represented by the case where an incorrect morphological variant of a word is chosen in the target text. It includes agreement, tense, part of speech, etc. The translation "ولا يمتلك المدينة أياً منهما" clearly has the wrong form of the verb. The verb should be "تمتلك" instead of "يمتلك" to have a correct gender subject-verb agreement.

- **Cohesion:** This refers to the issue where a part of the target text is required to be connected to the context. Check the following English sentences and the Arabic translated one.

Souers text: 'As expectant mothers often realize, newborn babies don't always arrive on schedule.'

Target text: ' كما غالبًا ما تدرك الأمهات الحوامل، لا تصل الأطفال حديثي الولادة دائمًا وفقًا للجدول الزمني.'

Ignoring the other errors, it is apparent that the two Arabic clauses are not cohesive despite the fact that the English ones are. That is due to fact that Arabic is a language with a highly syndetic discourse.

- **Word Order:** This type of error simply is related to the syntactic word order of a translated sentence, clause or phrase. A structure in the target text may follow the rules of the target language. It may simply copy the structure of the source language when there are structural differences between the two languages. Check the following English sentence and its Arabic translation.

ISSN : 2410-1818

Source Text: As she campaigned in Keene, New Hampshire, Haley referenced Trump's speech the night before.

Target Text: أثناء حملتها الانتخابية في كين، نيو هامبشاير، أشارت هالي إلى خطاب ترامب في الليلة السابقة.

In the Arabic translation above, the dependent clause is used before the independent one which in undesirable in Arabic. Moreover, the dependent clause contains an anaphora that precedes the antecedent. This is acceptable in English. However, an Arabic sentence where such case occurs is considered to be flawed.

- **Punctuation:** The use of punctuation marks in this case is considered to be incorrect and does not follow the rules of the target language. In the following translated Arabic sentence the comma is not needed.

ظهر صغيرها هادئًا ومرتاحًا بينما تسلمت سزيمتشاك شهادتها بابتسامة، كما أظهر فيديو مباشر للحفل.

**Style Errors:** According to the classification of style errors, the target text is grammatically correct according to the rules of the target language. However, an awkward language style, inappropriate register, or a deviation from target language style is detected. Check the translation of the following sentence.

Source Text: Szymchack, who has an associate degree in early childhood development at Northwestern Michigan College, says she always planned on teaching.

Target Text: تقول سيمتشاك، التي تحمل درجة الزمالة في تطوير الطفولة المبكرة في كلية ميشيغان الشمالية الغربية، إنها كانت دائمًا تخطط للتدريس.

Ignoring the unneeded commas, the Arabic sentence is correct. However, it will be more natural if we add "و" (and) before the relative pronoun.

The table below shows the types and number of annotated errors in ChatGPT original translation of the complex sentences in the dataset under investigation.

ISSN : 2410-1818

**Table 1: Types and Number of Annotated Errors in ChatGPT Original Translation**

| Types of Errors & Clauses | | Adverb Clauses | Relative Clauses | Complement Clauses | Total No. of Errors |
|---|---|---|---|---|---|
| **Terminology** | Wrong Term | 22 | 19 | 20 | **61** |
| | Inconsistent use of terminology | 1 | 0 | 0 | **1** |
| **Terminology Errors** | | **23** | **19** | **20** | **62** |
| **Accuracy** | Ambiguous Target Content | 2 | 2 | 2 | **6** |
| | Ambiguous Source Content | 0 | 2 | 0 | **2** |
| | Untranslated | 6 | 8 | 7 | **21** |
| | Omission | 8 | 5 | 6 | **19** |
| | Overly Literal | 11 | 9 | 9 | **29** |
| **Accuracy Errors** | | **27** | **26** | **24** | **75** |
| | Incorrect FW | 4 | 8 | 9 | **21** |
| | Missing FW | 35 | 47 | 29 | **111** |
| | Extraneous FW | 3 | 11 | 10 | **24** |
| | Cataphoric issues | 3 | 1 | 10 | **14** |
| | Word form | 12 | 10 | 10 | **32** |
| | Cohesion | 12 | 2 | 5 | **19** |
| | Word Order | 9 | 18 | 13 | **40** |
| | Punctuation | 15 | 16 | 10 | **41** |
| **Fluency Errors** | | **93** | **113** | **96** | **302** |
| **Style Errors** | | **9** | **15** | **4** | **28** |
| **Total No. of Errors** | | **152** | **173** | **144** | **469** |

### 3.3 Error Distribution

The error taxonomy shows that ChatGPT translation outputs are still riddled with errors of various types. Since the test suite is composed of 150 sentences, the error frequency, then, amounts to 2.73 errors per sentence which is very high. It should be noted that, according to MQM main dimensions, the errors are distributed as follows.

- Fluency errors come in front with 302 errors (64.39% of the annotated errors).

ISSN : 2410-1818

- Accuracy errors are next with 77 errors (16.42% of the annotated errors).
- Terminology errors follow with 62 errors (13.22% of the annotated errors).
- Last come the style errors with 28 errors (5.97% of the annotated errors).

The distribution of errors according to MQM main dimensions is represented in Figure 1 below.

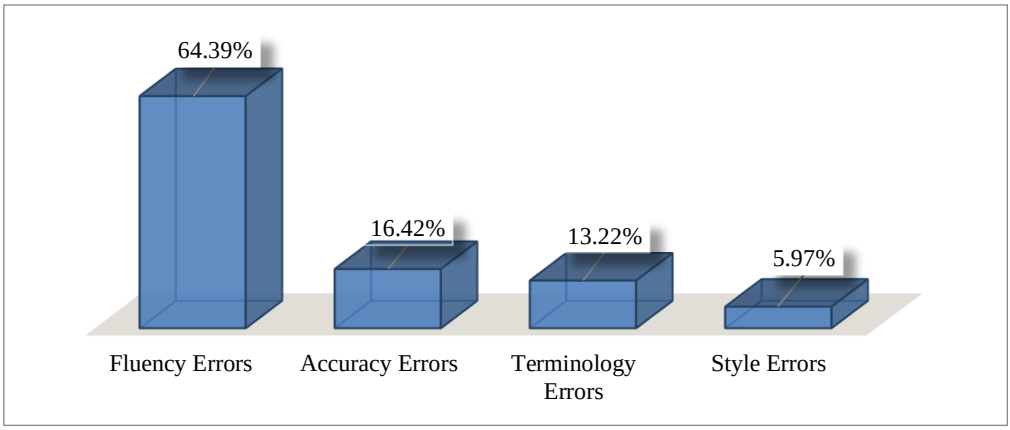


**Figure 1: Error Distribution per MQM Main Dimensions**

### 3.4 Prompting

As discussed earlier, it is established that the use of informed prompts plays a crucial role in improving the quality of ChatGPT translation outputs. This section introduces the dataset and the phases of the prompting operation taken to utilize informed prompts effectively in this study. It also presents the results of the evaluation of the original translation and the prompted translation of the prompting dataset sentences.

#### 3.4.1 Prompting Dataset

After performing the error taxonomy, the research team uses the sentences with the most errors as the prompting dataset. Thirty-six sentences with the most errors are extracted from the main dataset to be used in the various prompting phases.

The prompting operation is divided into four main phases: the instruction phase, the initial training phase, the testing phase, and the application phase.

ISSN : 2410-1818

### 3.4.2 Initial Training Phase

- **Instruction phase (pre-prompting):**

The model was initially trained with specific instructions to identify common translation errors. The model is presented with the errors annotated in the main dataset classified based on MQM dimensions. Explanation of errors was provided to the model to aid in understanding the types of errors.

- **Training Set:**

Out of the prompting dataset, 10 sentences were used as a training set to train the model. The research team ensured that these sentences covered all the types of annotated errors to guarantee that the training set served as a foundation for the model to learn the nuances of English to Arabic translation and the specific areas where errors frequently occurred. Using prompting, the model was provided with a source sentence, its original translation, and the errors in the original translation, and it was required to retranslate the given sentence.

### 3.4.3 Testing Phase

After the training phase, a set of six test sentences was provided to the model. These sentences were selected based on the presence of errors as established in the initial training set. These sentences were retranslated to evaluate the performance of the model in improving the translation. In this phase, the model was provided with the source sentence and the original translation.

The new translation outputs were compared to the original translation outputs. It was observed that the model's performance improved significantly, with a noticeable reduction of errors and more accurate translations. This confirmed the effectiveness of the training and the used prompts. Accordingly, the research team moved to the next phase.

### 3.4.4 Application Phase

Following the successful testing phase, the remaining sentences from the dataset were provided to the model for retranslation. In this phase, the sentences were given without specifying the errors as well. The translation outputs of this phase were evaluated along with the original translation to evaluate the effectiveness of the used prompting operation in improving ChatGPT’s translation performance.

ISSN : 2410-1818

### 3.4.5 Evaluation:

The application of informed prompts proved to be effective and continued to yield positive results. The retranslations showed a significant reduction in errors compared to the initial translations. Manual evaluation by is performed by the three professional annotators using as a secular quality metric (Freitag et al, 2021).The 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.

The BLEU metric is also used to perform the automatic evaluation. Both evaluations indicated that the translation significantly improved. The following table shows the evaluation of the original translation as well as the retranslation outputs.

**Table 2: Manual and BLEU Evaluations of Original Translation and Retranslation**

| Evaluation | Original Translation | Retranslation |
|---|---|---|
| **Manual Evaluation** | 62.76% | 85.42% |
| **BLEU** | 32.21% | 83.49% |

ISSN : 2410-1818

## 4. Discussion and Conclusion

Based on the nature and number of the annotated errors, the following points can be stated.

- ChatGPT translation outputs show a high number of errors when translating English complex sentences into Arabic. The frequency of errors indicates that ChatGPT still falls short when translating complex sentences from English into Arabic.
- Fluency errors show the highest percentage among errors annotated in ChatGPT translation outputs, which indicates that the model struggles to grasp the structural variations between English and Arabic. This indicates that ChatGPT faces a big challenge when translating from English as a language with poor morphology to Arabic as a language with rich morphology.
- There are a lot of terminology errors, which indicates that the model needs more training on Arabic texts and that there is an urgent need to build annotated Arabic corpora.

The results of the evaluation of original translation outputs and the retranslations using the adopted prompts indicate that the proposed strategy to improve ChatGPT translation outputs is very effective. Despite the difference in the evaluation of the original translation between manual evaluation and automatic evaluation, both evaluations show that there is a great improvement in the quality of translation.

Accordingly, it can be concluded here that complex sentences form a great challenge for MT and that informed prompts are effective in improving the translation output. It is recommended that further research in the area of informed prompts is conducted to reach a very high level in terms of automatic translation.

## Acknowledgement

This research received grant no. (76/2023) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

ISSN : 2410-1818

## 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 preprint arXiv:2306.03856*. https://doi.org/10.48550/arXiv.2306.03856

Diessel, H. (2004). *The acquisition of complex sentences* (Vol. 105). Cambridge University Press.

Farghal, M. (2017). Textual issues relating to cohesion and coherence in Arabic/English translation. *Jordan Journal of Modern Languages and Literature*, *9*(1), 29-50. https://journals.yu.edu.jo/jjmll/Issues/vol9no12017/Nom3.pdf

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 preprint 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 preprint 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 preprint arXiv:2303.15587*. https://doi.org/10.48550/arXiv.2303.15587

Hasler, E., de Gispert, A., Stahlberg, F., Waite, A., & Byrne, B. (2017). Source sentence simplification for statistical machine translation. *Computer Speech & Language*, *45*, 221-235. https://doi.org/10.1016/j.csl.2016.12.001

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., ... & Zhou, M. (2018). Achieving human parity on automatic Chinese to English news

ISSN : 2410-1818

translation. *arXiv preprint 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 preprint arXiv:2302.09210*. https://doi.org/10.48550/arXiv.2302.09210

Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. arXiv preprint arXiv:2301.08745, 1(10). https://doi.org/10.48550/arXiv.2301.08745

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 (WMT) (pp. 1-45). https://aclanthology.org/2022.wmt-1.1

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). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1512

Li, Y. (2023). The study of evolution and application related to the Chat-GPT. *Highlights in Science, Engineering and Technology*, *57*, 185-188. https://doi.org/10.54097/hset.v57i.9999

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 natural language processing. *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), 0455-463. https://doi.org/10.5565/rev/tradumatica.77

Lu, X., Qiang, J., Li, Y., Yuan, Y., & Zhu, Y. (2021). An unsupervised method for building sentence simplification corpora in multiple languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 227-237).https://aclanthology.org/2021.findings-emnlp.22

Miller, J. (2002). *Introduction to English syntax*. Edinburgh University Press.

Nagi, K. A. (2023). Arabic and English relative clauses and machine translation challenges. *Journal of Social Studies*, *29*(3), 145-165. https://doi.org/10.20428/jss.v29i3.2180

ISSN : 2410-1818

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

Qasmi, N. H., Zia, H. B., Athar, A., & Raza, A. A. (2020). SimplifyUR: unsupervised lexical text simplification for Urdu. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3484-3489). https://aclanthology.org/2020.lrec-1.428

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 large language models in translation. *Available at SSRN 4448091*. http://dx.doi.org/10.2139/ssrn.4448091

Štajner, S., & Popović, M. (2018). Improving machine translation of english relative clauses with automatic text simplification. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)* (pp. 39-48). https://doi.org/10.18653/v1/W18-7006

Sulem, E., Abend, O., & Rappoport, A. (2018). Simple and effective text simplification using semantic and neural methods. In *56th Annual Meeting of the Association for Computational Linguistics, ACL 2018* (pp. 162-173). Association for Computational Linguistics (ACL).https://aclanthology.org/P18-1016

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

Turganbayeva, A., Rakhimova, D., Karyukin, V., Karibayeva, A., & Turarbek, A. (2022). Semantic connections in the complex sentences for post-editing machine translation in the Kazakh language. *Information*, *13*(9), 411. https://doi.org/10.3390/info13090411

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 preprint arXiv:2302.11382.https://doi.org/10.48550/arXiv.2302.11382