Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 73-85 recibido 14-12-2023 revisado 10-01-2024 aceptado 28-01-2024

Gradable ChatGPT Translation Evaluation

Evaluación de traducción graduada de ChatGPT

Hui Jiao,[1] **Bei Peng,**[2] **Lu Zong,**[1] **Xiaojun Zhang,**[1,4] **Xinwei Li**[3]
[1]Xi'an Jiaotong-Liverpool University
[2]University of Liverpool
[3]Southeast University
[4]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
Xiaojun.Zhang01@xjtlu.edu.cn

Abstract: ChatGPT, as a language model based on large-scale pre-training, has exerted a profound influence on the domain of machine translation. In ChatGPT, a "Prompt" refers to a segment of text or instruction employed to steer the model towards generating a specific category of response. The design of the translation prompt emerges as a key aspect that can wield influence over factors such as the style, precision and accuracy of the translation to a certain extent. However, there is a lack of a common standard and methodology on how to design and select a translation prompt. Accordingly, this paper proposes a generic taxonomy, which defines gradable translation prompts in terms of expression type, translation style, Part-of-Speech information and explicit statement, thus facilitating the construction of prompts endowed with distinct attributes tailored for various translation tasks. Specific experiments and cases are selected to validate and illustrate the effectiveness of the method.
Keywords: Translation prompt, ChatGPT, T3S taxonomy, Evaluation.

Resumen: ChatGPT, un modelo de lenguaje basado en un pre-entrenamiento a gran escala que ha tenido un profundo impacto en la traducción automática. En este contexto, un "prompt" se refiere a un segmento de texto o instrucción utilizada para dirigir el modelo hacia la generación de una respuesta específica. El diseño del prompt de traducción es crucial y puede influir en aspectos como el estilo, la precisión y la exactitud de la traducción. Sin embargo, actualmente carecemos de un estándar y metodología común para diseñar y seleccionar prompts de traducción. Por lo tanto, este artículo propone una taxonomía genérica que define prompts de traducción evaluables en términos de expresión, estilo, Part-Of-Speech (POS) y declaración explícita, facilitando la construcción de prompts con distintos atributos adaptados a diversas tareas de traducción. Se han seleccionado experimentos y casos específicos para validar e ilustrar la eficacia del método.
Palabras clave: Prompt de traducción, ChatGPT, Taxonomía T3S, Evaluación.

1 Introduction

Machine translation (MT), one of the oldest branches of research in the field of natural language processing, involves techniques for transforming one natural language into another. As a key research area within the field of Artificial Intelligence, MT has been widely used in a wide range of fields and has attracted extensive attention from both academia and industry (Yang, Wang, and Chu, 2020). In recent years, there has been a growing trend towards the use of large-scale pre-trained language models for natural language processing (NLP) (Yang, Wang, and Chu, 2020; Brown et al., 2020; Amplayo, Yoo, and Lee, 2022). Large Language Models (LLMs), such as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), and LLaMA (Tou-

Corresponding author: Xiaojun Zhang

ISSN 1135-5948 DOI 10.26342/2024-72-5

abundant representations of the input text, and have greatly contributed to the development of MT technology (Lewis et al., 2019).

NLP tasks have recently been significantly influenced by the emergence of ChatGPT, a powerful pre-trained LLM developed by OpenAI. The model has been trained to perform a large number of human-like tasks (e.g., question answering, code debugging, generating evaluations, etc.) with human feedback. However, the overall performance of ChatGPT on translation tasks using simple prompts and basic settings is not as good as the commercial translation products such as Google Translate and Microsoft Translate (He et al., 2022; Zhou et al., 2022; Jiao et al., 2023; Hendy et al., 2023)but is promising to surpass them with complex and explicit prompts. This is due to the fact that for a given complex translation task, significant differences in the performance of LLMs arise when different types/styles/levels of detail of prompts are introduced (Karmaker and Feng, 2023).

The effective use of LLMs requires elaborate prompt engineering, which refers to the process of designing and refining the prompts or instructions provided to a large language model (Zhou et al., 2022; Liu et al., 2023). For translation tasks of different styles, the relevant prompts that the user can try while performing the task are different. In addition, the number of details included in the prompts also largely affects the performance of LLMs when dealing with target complex translation tasks (Karmaker and Feng, 2023). By providing clear and well-structured prompts, users can help guide LLMs in the right direction and reduce potential biases or errors. To achieve this, we propose a unified gradable prompting taxonomy for ChatGPT translation called T3S, which employs standardized criteria to categorize various types of translation prompts, thereby further enhancing ChatGPT's translation capabilities.

On the one hand, ChatGPT is utilized to conduct very broad and various types of translation tasks from general text translation to highly specialised domains (e.g., medical, legal, technical, etc.), and categorising prompts will help ChatGPT distinguish between different prompts and implement different translation strategies to meet translation needs in different domains and contexts. On the other hand, the proposed T3S taxonomy can help researchers explore in depth the working principles of LLMs and their performance differences in different domains, thus promoting further research and innovation in LLMs. Specifically, the widespread adoption of this taxonomy can potentially promote a more accurate performance assessment of ChatGPT in different translation tasks, thereby identifying specific problems that may exist in the model, providing targeted feedback for improvement and guiding the direction of model optimisation, and achieving continuous progress of the model.

2 Related Work

A prompt is a set of instructions provided to an LLM that enhances the functionality of the LLM by customising it (Liu et al., 2023). In the field of large language modelling, complex tasks refer to those involving multiple steps or subtasks that require a higher level of semantic understanding, planning, reasoning, and natural language generation capabilities, which makes prompt engineering particularly critical and challenging (Tan et al., 2022).

In recent years, many researchers have proposed different approaches to engineering prompts. For example, one of the best-known (and easiest to implement) prompt engineering techniques is to add "Think step by step" to the end of the prompt. Adding this phrase improves the accuracy of the GPT-3 (text-DaVinci-002 model) across multiple tasks (Wei et al., 2022). Moreover, Brown et al. (2020) presented a standard question-answer pair prompting technique which produces a few-shot effect. By providing suitable output instances, LLMs are more likely to produce the desired output (Zhao et al., 2021). Similar to "Think step by step", the Chain of Thought (CoT) prompting method guides LLMs to break down a complex task into multiple intermediate steps (Wang et al., 2023). Fu et al. (2022)showed that separating each step with a new line in exemplar reasoning is much more effective than separating each step with a full stop. Researchers have also explored other prompt design techniques such as Reasoning and Acting (ReAct) (Yao et al., 2022), which overcomes the illusions and error propagation problems prevalent in CoT reasoning by interacting with a simple Wikipedia API. Other techniques such as Zero-shot-CoT (Kojima et al., 2022) and Self-Ask (Press et al., 2022), improve

LLMs' reasoning and action in solving questions and answering tasks. Meanwhile, Kim, Baldi, and McAleer (2023) suggested that recursively criticising and improving its output (RCI) is superior to CoT prompts in terms of its effectiveness in reasoning ability in a range of natural language reasoning tasks.

More specifically, conversational LLMs, such as ChatGPT, have generated considerable research interest in a range of domains, with tasks ranging from answering questions for medical licensing exams to generating code snippets (Gilson et al., 2023). Correct prompt engineering has become a key skill for users wishing to utilise the full potential of ChatGPT and obtain optimal results in a variety of applications. Accordingly, there has been an influx of prompt engineering research in many different areas. Thirunavukarasu et al. (2023) explored how prompts can be used to enhance the efficiency and effectiveness of ChatGPT in medical clinics, education, and research, while Trautmann, Petrova, and Schilder (2022) proposed zero-sample legal prompts engineering (LPE) to guide and enhance LLMs in natural legal language processing (NLLP) capabilities. These studies focused on NLP tasks rather than MT. In the field of machine translation, however, previous studies (Liu et al., 2019; Guo et al., 2020) have shown that while LLMs can enhance a translation system's understanding of the source text, improving its generative capabilities is more difficult. A well-developed translation system requires strong language comprehension and generation capabilities to achieve accurate and fluent translation results. Although some studies (Brown et al., 2020; Chowdhery et al., 2022) have explored the effects of different prompts on translation results, there is still a lack of systematic research on how to improve MT using prompts. With the popularity of LLM-based prompting approaches, researchers are starting to recognise the importance of introducing prompts into neural machine translation (NMT) (Li et al., 2022; Tan et al., 2022; Wei et al., 2022). Nevertheless, these approaches still rely on pre-training or fine-tuning the models, rather than directly applying them to "frozen" LLMs. Therefore, it is critical to study how to make the most effective use of these prompts in order to balance language comprehension and generation capabilities and achieve better results in MT.

To summarise, prompt engineering is a crucial step for the effective utilisation of LLMs. However, existing works have mainly emphasised the use of diversified prompts to improve the ability of LLMs to perform general-purpose NLP tasks, while the specification of prompts for machine translation remains under-explored. Therefore, this paper aims to fill the research gap by providing translators with a systematic approach to selecting and designing prompts, which can improve the consistency, reliability and quality of ChatGPT translations, as well as promote the development and innovation in the field of MT.

3 ChatGPT Translation Prompting Taxonomy Design

Through training, ChatGPT is capable of generating appropriate responses based on the given prompts. This attribute determines its high sensitivity to the information provided by the prompts. Differences in several key factors of the prompts can have a significant impact on the accuracy and performance of large language models in translation tasks (Karmaker and Feng, 2023). These key factors are described below, where the prompt is defined by the combination of the instruction (intended target) and the source text provided to ChatGPT for performing the translation task. The prompts present differences in the instruction section under the assumption that the source text has been provided.

Ambiguous or contextually inadequate prompts can easily confuse LLMs, leading to inaccurate or irrelevant responses from these models (Jiang et al., 2022). Therefore, providing clear and specific instructions in the prompts can help guide ChatGPT to generate the desired translation results. In the MT domain, detailed prompts usually consist of elaborating on various aspects of the translation task specification. In the following, we will delve into two basic aspects of translation task specification.

1) Explicit Descriptions. In translation task prompts, a clear task description is essential for obtaining accurate and relevant translation results. Specifying the translation objectives clearly and asking ChatGPT to proofread before generating the translation results helps to guide ChatGPT to a full understanding of the translation task and maintain a consistent focus throughout the trans-

lation process. This ensures a higher level of accuracy, relevance and fluency in the translation results, thereby increasing the probability of obtaining the desired translation outcome.

2) Contextual Information. Contextual information plays a key role in enhancing ChatGPT's understanding of words, phrases, and sentences in the source text, thus helping to reduce ambiguity and misunderstanding. Models can optimise translation decisions based on context, so as to avoid translation errors (Popescu-Belis, 2019). For example, for polysemous words, context can help ChatGPT determine the correct word meaning. This is crucial for producing accurate translations, as the same word may be translated differently in different contexts. Additionally, contextual background helps to maintain consistency in the model's expression throughout the translation process. When the source text has certain specific usages or terms in the context, the model can retain them in the translation based on the context, ensuring a coherent translation.

Translation prompts consist of instructions and source text. Assuming that the source text is invariant in a given translation task, we believe that the difference between prompts lies in the explicit descriptions and contextual information they contain. To address the key factors of prompt selection and design, we will classify the prompts for ChatGPT translation tasks according to the following four aspects.

3.1 Expression Types

There are two main types of expressing translation prompts: the single-turn prompt and the multiple-turn prompt. Both types of expression can be effective in different contexts. Single-turn prompts involve presenting the model with a solitary input, typically in the form of a single sentence or a brief textual segment ("*Please translate the following text...*"), for the model to translate into the target language. Whereas multiple-turn prompts incorporate conversational interactions, usually consisting of multiple dialogue rounds, so that the model can better understand and perform the translation task. For example, a user can ask ChatGPT to check and revise the translation after it responds to the first round of prompts ("*Please translate it again / Please revise the translation*"). Such prompts can be used for more complex translation tasks, where context, clarification, follow-up questions, etc., may need to be taken into account. Compared to single-turn prompts, the use of multiple-turn prompts can significantly enhance the comprehension of ChatGPT, and reduce its tendency to generate irrelevant or inaccurate responses (Pan et al., 2023). Depending on the nature and needs of the translation task, choosing the appropriate prompt expression form can help ChatGPT generate the required translation output more accurately.

3.2 Translation Style

Depending on the translation task in different translation domains to deal with different genres such as literary, medicine, legal and commercial texts, the desired prompt needs to be defined and selected according to the translation style. Defining translation style involves determining the content and expression to be conveyed in a translation task. It relates to the specific translation target and audience. Encompassing facets such as affective undertones, tonal modulations, and the amplitude of linguistic exposition, the election of a fitting translation style exerts a discernible impact on the ultimate rendition. Thus, the inclusion of relevant contextual information such as target audience and level of expression in the prompt can provide ChatGPT with additional information to produce a more accurate translation result.

3.3 POS Information

Part of Speech (POS) is a grammatical category that includes nouns, verbs, adjectives, adverbs, etc. (Hlaing et al., 2022). Many NLP tasks benefit from the use of POS tags. In translation tasks, POS tags help ChatGPT to capture the grammatical structure of a sentence in the source language, and accurately locate the grammatical roles of each word in the sentence, so as to better translate it into the grammatical structure of the target language, eliminating word ambiguities, and further enhancing the natural fluency of the translation result. For instance, Feng et al. (2020) demonstrated that incorporating POS tagging information into the target side can significantly improve the translation performance of the NMT system in both Chinese-to-English and German-to-English translation pairs. Further, Hlaing et al. (2022) con-

ducted an NMT study using POS tagging information on low-resource language pairs, explicitly pointing out the necessity of integrating POS tags when using NMT models that include linguistic features.

3.4 Few-shot Prompts

LLMs can benefit from example-based learning, which involves providing specific input-output pair examples (a small number of examples). This can help models to better understand task requirements and generate appropriate output (Brown et al., 2020). We believe that including a few input-output examples in the prompt will improve the performance of the LLMs without any adjustments to the parameters or architecture. For example, by including specific terminology and styles in the few-shot prompts, ChatGPT is able to adapt quickly across different translation domains, generating domain-appropriate translations.

We propose a gradable prompting taxonomy for ChatGPT translation which is categorised into five different levels based on the above four key elements in prompt design including expression type, translation style, POS information, and few-shot prompts. We named it T3S standing for expression type, translation style, POS information and few-shot 5-level prompting taxonomy.

More precisely, level "0" represents the lowest level of detail, where only basic translation is required with the general prompt such as "*Please translate the following text. . .*"; Level "1" distinguishes the expression type of the single-turn form with the multiple-turn form; Level "2" adds translation style instruction providing contextual information; Level "3" integrates POS information into translation instruction; and Level "4" represents the highest level of detail, where the multiple-turn prompts include clear instructions, explicit contextual information of translation with few-shot examples, and an explicit statement asking ChatGPT to check and revise the results.

4 Experimental Validation

In order to verify the rationality and validity of the T3S taxonomy, we set up an evaluation experiment. In the following, we show the details of the experimental setup, including the adopted dataset and evaluation metrics. The results and analyses of the experiment are also presented.

4.1 Dataset

We evaluated the translation quality of ChatGPT at different levels of prompt on the Flores-101 (Goyal et al., 2022)dataset. The dataset consists of 1012 sentences extracted from the English Wikipedia covering a wide variety of topics and domains. In real translation applications, ChatGPT needs to process texts from a variety of topics and domains. A dataset covering different topic domains can help us evaluate ChatGPT's generalisation ability and gain a more comprehensive understanding of ChatGPT's translation performance under different contexts. Moreover, these sentences have been translated into 101 languages by professional translators through a rigorously controlled process with automated and manual quality checks. Furthermore, all translations are multilingual aligned. Such a high-quality and high-coverage dataset ensures the accuracy and consistency of reference translations, and better helps us understand and evaluate the quality and performance of ChatGPT's translations. However, a graded translation quality assessment for multiple languages may add complexity and resource requirements. Therefore, for a clearer direct comparison of translation quality and to save time and cost, we only used the Chinese-English bilingual corpus for the assessment.

4.2 Metrics

To assess translation quality at all levels, we employed the most commonly used BLEU score (Papineni et al., 2002). Additionally, we utilised CHrF (Popović, 2015), TER (Snover et al., 2006) and ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004), and calculated the F1 average of the scores of the ROUGE series to provide a more comprehensive assessment of translation quality at all levels. This is because one indicator may be more sensitive to certain aspects of translation quality, while another may capture different aspects of quality.

4.3 Prompt Construction

Prompts for each level were meticulously crafted to align with the taxonomy's gradable elements. Concretely, regarding the translation styles in Level 2, we added the labels (domain & topic) in the dataset as translation styles to the prompt, such as wikinews(business),

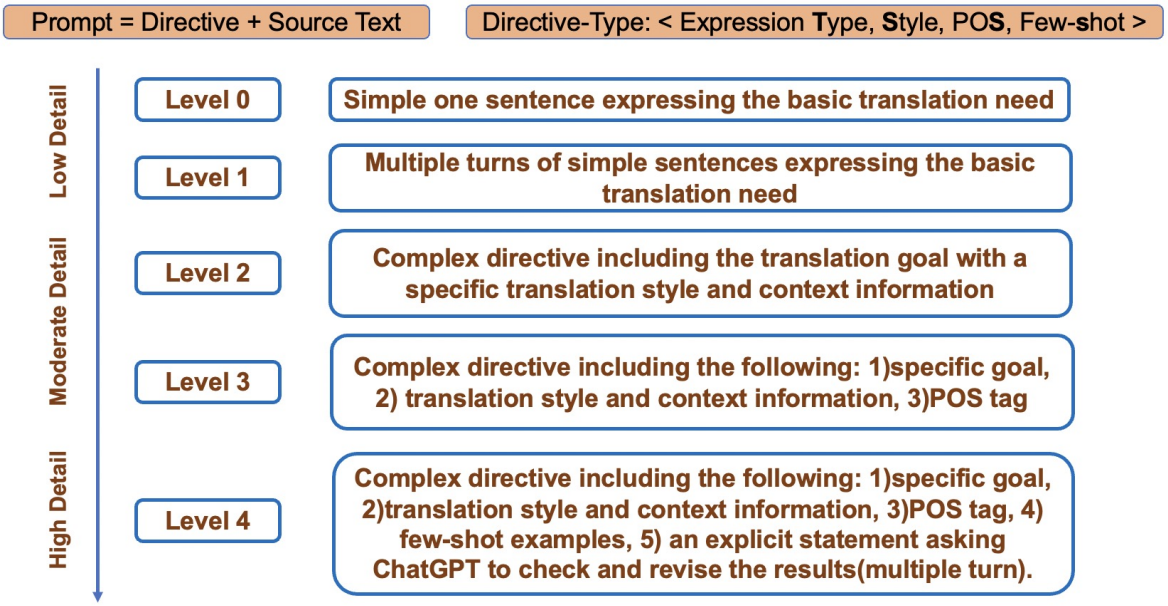

Figure 1: T3S Taxonomy.

Translation	BLEU	CHrF	ROUGE F1(avg)	TER
Level 0	38.42	30.77	0.6132	160.19
Level 1	38.93	31.07	0.6165	146.90
Level 2	40.25	32.41	0.6256	117.12
Level 3	41.25	33.57	0.6303	122.78
Level 4	42.88	36.24	0.6523	112.78

Table 1: ChatGPT's Translation Performance at All Levels.

wikivoyage(travel), wikivoyage(sports), wikibooks(sociology/culture), etc. As for the POS information, we used the open-source natural language processing tool spaCy as our lexical annotation tool to preprocess the source text. With respect to the few-shot examples in Level 4, we randomly selected two sets of source text and target text pairs under the same domain and topic as examples to guide ChatGPT for translation.

4.4 Results

The experimental results, as shown in Table 1, indicate that the translation quality of ChatGPT improves accordingly as the prompt level increases. Specifically, ChatGPT obtains a BLEU score of 38.42 for the basic translation prompt at level 0. This level of prompt is only the most basic translation requirement and does not contain any additional contextual or guidance information. When the prompt upgrades to Level 1, which distinguishes between single-turn and multiple-turn expression types, the BLEU score improves slightly (0.51). However, a more significant increase occurred at Level 2, when the translation style of contextualisation was added to the prompt, which increased the BLEU score to 40.25. This suggests that the inclusion of the translation style has a significant positive impact on the quality of the translation compared to the base translation requirement. At Level 3, the integration of lexical information into the translation prompt further increased the BLEU score to 41.2571. This result suggests that the introduction of lexical information can provide the model with more precise linguistic information, which helps to generate more accurate translations. Finally, the Level 4 prompt achieved the highest BLEU score of 42.88 by including clear instructions, explicit information about the translation context, few-shot examples, and asking the model to check and revise the translation results. Compared to Level 3, the increase was 1.63 points. This largest increase fully demonstrates the remarkable effectiveness of the T3S taxonomy in guiding high-quality translation. Importantly, the Level 4 translation quality is higher than the zero-shot translation quality (BLEU score of 42.50) of GPT-4 under the

same dataset and evaluation metrics (Jiao et al., 2023). Moreover, the results for the ChrF, ROUGE series (F1 average), and TER mainly support the above findings.

Overall, the experimental results consistently show that the translation performance of ChatGPT improves significantly as the level of the prompts increases. This suggests that the T3S taxonomy is not only reasonable but also effective in providing targeted guidance for ChatGPT's translation tasks. Future work can explore the application of the T3S taxonomy to more language pairs, as well as the applicability and optimisation space in different domains and scenarios.

5 Case Analysis

This section presents how ChatGPT's performance under the same translation task can be assessed by our T3S taxonomy through a specific translation case. We demonstrate the translation example with different levels of prompts below.

Example Text: MacBook Air is all you — pick your size, pick your color, then go. Whichever model you choose, it's built with the planet in mind, with a durable 100 percent recycled aluminum enclosure. And a fanless design means it stays silent even under intense workloads.

Expert Translation: MacBook Air尽显你的风格，挑个合你的款，选个配你的色，直接搞定。两款机型都采用100％再生铝金属外壳，不管选哪款，都能体现对地球的心意。而无风扇设计意味着即使工作强度高，也能保持静谧无声。

Level 0 Prompt: *Please translate the following text into Chinese:* MacBook Air is all you — pick your size, pick your color, then go. Whichever model you choose, it's built with the planet in mind, with a durable 100 percent recycled aluminum enclosure. And a fanless design means it stays silent even under intense workloads.

Level 0 Translation: MacBook Air 是完全为您打造的— 选择您的尺寸，选择您的颜色，然后开始使用。无论您选择哪个型号，它都是考虑地球环保的，采用了耐用的100％回收铝外壳。而且无风扇设计意味着即使在高强度工作下，它也能保持静音。

Level 1 Prompt (Multiple-Turn form): *First turn: Please translate the following text into Chinese:* MacBook Air is all you — pick your size, pick your color, then go. Whichever model you choose, it's built with the planet in mind, with a durable 100 percent recycled aluminum enclosure. And a fanless design means it stays silent even under intense workloads. *Second turn: Please check and revise the translation results.*

Level 1 Translation: MacBook Air就是为你而设计的— 选择你的尺寸，选择你的颜色，然后开始使用。无论你选择哪个型号，它都是为了地球着想，采用了耐用的100％回收铝外壳。而且，它的无风扇设计意味着即使在高负荷工作下也能保持宁静。

Level 2 Prompt (Advertising style): *Please translate the following text into Chinese in a concise, impressive and advertising style:* MacBook Air is all you — pick your size, pick your color, then go. Whichever model you choose, it's built with the planet in mind, with a durable 100 percent recycled aluminum enclosure. And a fanless design means it stays silent even under intense workloads.

Level 2 Translation: MacBook Air全由您决定— 选择您的尺寸、挑选您的颜色，即刻出发。无论您选择哪个型号，它都以地球为念，采用坚固的100％再生铝外壳打造。无风扇设计确保在高强度工作中也能保持绝对宁静。

Level 3 Prompt (with POS information): *Given the context of an advertisement for an electronic product and the POS tags, please translate this specific sentence into Chinese in a concise, impressive and advertising style:* MacBook Air (Noun) is (Verb) all you (Pronoun) — pick (Verb) your (Pronoun) size (Noun), pick (Verb) your (Pronoun) color (Noun), then (Adverb) go (Verb). Whichever (Determiner) model (Noun) you (Pronoun) choose (Verb), it's (Verb) built (Verb) with (Preposition) the (Determiner) planet (Noun) in (Preposition) mind (Noun), with (Preposition) a (Determiner) durable (Adjective) 100 percent (Percentage) recycled (Verb) aluminum (Noun) enclosure (Noun). And (Conjunction) a (Determiner) fanless (Adjective) design (Noun) means (Verb) it (Pronoun) stays (Verb) silent (Adjective) even (Adverb) under (Preposition) intense (Adjective) workloads (Noun).

Level 3 Translation: MacBook Air就是您的综合选择— 选择您的理想尺寸，选取心仪的颜色，然后出发。无论您选择哪个型号，它都以地球为念，采用坚固的百分之百再生铝质外壳。而无风扇设计确保即使在高强度工作负荷下，也能保持绝对宁静。

Level 4 Prompt (with Few-shot exam-

ples): *First turn: Context Information: It is extracted from an advertisement for an electronic product. Few-shot Examples: 1. Translate "Two perfect sizes. Whether you pick the 13- or 15-inch laptop, both models are superlight and measure just under half an inch thin, so you can take yours anywhere." into* "两个尺寸两相宜。*13*英寸和*15*英寸两款笔电都十分轻巧纤薄，厚度仅有*1*厘米多，选哪款都称心称手，去哪里都自由自在。" *2. Translate "Four stellar colors. Your options are out of this world — and each one comes with a matching MagSafe charging cable." into* "四款配色都惹眼。每种选择都靓出天际，还配有同色系的*MagSafe*磁吸充电线。" *Considering the context information, few-shot examples and POS tags, please translate the following text into Chinese in a concise, impressive and advertising style:* MacBook Air (Noun) is (Verb) all you (Pronoun) — pick (Verb) your (Pronoun) size (Noun), pick (Verb) your (Pronoun) color (Noun), then (Adverb) go (Verb). Whichever (Determiner) model (Noun) you (Pronoun) choose (Verb), it's (Verb) built (Verb) with (Preposition) the (Determiner) planet (Noun) in (Preposition) mind (Noun), with (Preposition) a (Determiner) durable (Adjective) 100 percent (Percentage) recycled (Verb) aluminum (Noun) enclosure (Noun). And (Conjunction) a (Determiner) fanless (Adjective) design (Noun) means (Verb) it (Pronoun) stays (Verb) silent (Adjective) even (Adverb) under (Preposition) intense (Adjective) workloads (Noun). *Second turn: Please check and proofread the translation to ensure that no errors have been made.*

Level 4 Translation: MacBook Air 轻薄本，是你的不二之选——选尺寸、选颜色，然后出发吧！无论您选择哪种型号，它都专为地球环保而设计，采用100％可回收的铝合金外壳，坚固耐用。而且，无风扇设计确保即使它在高强度工作下也能保持宁静。

As evident from the foregoing, Level 4 provides the closest results to the expert translation. To further validate the efficacy of this taxonomy, we set up both human and machine evaluation methods.

1) We invite three proficient translators to evaluate these five levels of translations according to the quantitative assessment criteria for translation quality proposed by Li (1991), employing the expert translation as the reference standard. This assessment method was set up with four scoring metrics, namely accuracy, fluency, style and textual coherence, with the following weights: W_A=3.5,W_F=2.5,W_S=2,W_C=2. The source text is an advertisement text, which aims at conveying specific messages, emotions and appeals to attract the attention and resonance of the target audience, thus making accuracy and fluency the primary principles of advertisement translation (Xiao, 2010). Advertising's main objective is to promote a product or service, rather than a literature. While style and textual coherence can enhance the appeal and taste of an advert, it should not normally come at the expense of accuracy and fluency. Hence, for this case, the weighting cited as such can provide reasonable constraints on the role played by the scoring indicators in terms of importance and priority.

More specifically, accuracy involves the key purpose of translation, which is to ensure that the translation accurately conveys the message and meaning expressed in the original text. Fluency, on the other hand, emphasises the quality of the written expression of the translation, including regularity, clarity and linguistic fluency. Style stresses the importance of appropriately conveying the stylistic, social and local characteristics of the original text, as well as the extent of the use of rhetorical devices, while maintaining the accuracy of the actual meaning. At last, textual coherence considers whether the arrangement of utterances in the translation adequately takes into account the primary and secondary relationships of the information in the original text, whether contextual co-ordination is achieved, and whether coherence of tone is maintained (Li, 1991).

Each criterion had a maximum attainable score of ten. The ensuing scores represent the respective assessments of the aforementioned trio of professional translators. Moreover, with a view to ensuring the reliability of the results and presenting the evaluation results more explicitly, we calculated the final weighted scores using the average scores of the three translators for all levels of translations under different factors.

As shown in Equation **??** and Table 2, we show the calculation method for the final scores, as well as the different scores and final scores for each level of translation in terms of accuracy, fluency, style, and textual coherence. In the Equation **??**, n stands for the number of professional translators, n=3; W_A, W_F,

W_S, W_C represent the different weights of the four indicators; A_i stands for the specific translator's score for each level of translation in terms of accuracy, F_i stands for the specific translator's score for each level of translation in terms of fluency, S_i refers to the specific translator's score for each level of translation in terms of Style, and Ci denotes the specific translator's score for each level of translation in terms of textual Coherence, i=1, 2, 3.

Based on the score data presented in Table 2, it can be clearly observed that the progression from Level 0 to Level 4 is marked by a discernible trend toward translations that exhibit a greater proximity to expert translation. Specifically, the lowest rated Level 0 (6.8) and Level 1 (7.3) translations score approximately the same in terms of accuracy, and differs significantly in terms of fluency, style and textual coherence. This is due to the fact that their prompts are only different in expression types. Moreover, the improvement (1.0; 0.7; 0.4) in fluency, style and textual coherence from Level 0 to Level 1 proves, to some extent, the effectiveness of multiple-turn prompts in improving the quality of ChatGPT translation results. Compared to Level 2, Level 3 has only a slight improvement (0.2), which is due to the fact that the lexical nature of the source text does not confuse ChatGPT. However, specific texts like legal documents, medical literature, and technical documents often contain specialized terminology, intricate grammatical structures, and polysemous words, and demand a high level of precision and professionalism. In such cases, the inclusion of POS tags becomes crucial as they furnish essential grammatical and semantic information necessary for effectively processing these texts. It should be noted that the addition of POS tags may also increase the preprocessing workload, thus requiring a comprehensive consideration of task requirements and efficiency. Finally, the highest-scoring Level 4 (8.8) translation results exemplify the importance of few-shot examples, which enable ChatGPT to understand the task requirements as well as possible and generate the most brand-specific translations.

2) LLMs not only show excellent capabilities in several NLP tasks such as machine translation, text summarisation, etc., but they are also state-of-the-art translation quality evaluators (Kocmi and Federmann, 2023). Kocmi and Federmann (2023) proposed a GPT-based metric for translation quality assessment, namely GPT Estimation Metric Based Assessment (GEMBA). By conducting experiments on nine versions of GPT models, including ChatGPT and GPT-4, they demonstrated the usefulness and accuracy of pre-trained generative LLMs for translation quality assessment at the system level by using a zero-shot standard prompt. However, this prompt performed poorly at the segment level. Subsequently, Lu et al. (2023b) further validated the capability of LLMs in assessing machine-translated translations. They combined the Chain-of-Thought (CoT) prompting strategy (Wei et al., 2022) and the Error Analysis (EA) paradigm (Lu et al., 2023a) to propose a novel prompting strategy, Error Analysis Prompting (EAPrompt). EAPrompt divides the scoring process into two stages: first, the LLM is prompted to identify the major and minor errors in the translation. The LLM is then asked to count the number of errors in both categories and calculate the final score. Unlike standard prompts, EAPrompt can produce human-like evaluations of machine translations at both the system and segment levels.

Based on such a pioneering discovery, we perform evaluations of all levels of translations in this case with EAPrompt. Below we show the specific prompt template and the scores for each level of translation. Due to space constraints, the Q&A sessions for each level of translation are not presented.

First-turn:
(Source Text)
(Reference)
(Translation)
Based on the given source and reference, identify the major and minor errors in this translation. Note that Major errors refer to actual translation or grammatical errors, and Minor errors refer to smaller imperfections, and purely subjective opinions about the translation.

Second-turn:
Count the number of major and minor errors identified in your last response and compute the final score for this translation. Deduct 5 points for each major error. Deduct 1 point for each minor error. If the translation has no errors, its score will be 0.

	Accuracy			Fluency			Style			Coherence			Final Score
	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	
Level 0	7	6	7	7	7	7	6	7	6	7	8	7	6.8
Level 1	7	7	7	8	8	7	7	7	7	8	8	7	7.3
Level 2	9	8	8	8	9	8	8	8	8	8	9	9	8.0
Level 3	8	9	8	9	7	9	8	7	7	9	9	9	8.2
Level 4	9	9	8	9	8	9	9	8	9	9	9	9	8.8

Table 2: Assessment of Accuracy, Fluency, Style and Coherence of Translations at All Levels.

$$FinalScore = \frac{\sum_{i=1}^{n} (W_A * A_i + W_F * F_i + W_S * S_i + W_C * C_i)}{n} \quad (1)$$

	Level 0	Level 1	Level 2	Level 3	Level 4
Results	-27	-23	-22	-18	-12

Table 3: Results of ChatGPT's Quality Assessment of Five Levels of Translations under EAPrompt.

Based on Table 3, it can be concluded that the translation quality exhibits an upward trend as the prompt level increases, which is roughly in line with the results of the human-based translation quality assessment. This result further confirms the effectiveness of T3S Taxonomy and the potential of LLMs in translation quality assessment. However, it is also noted that even higher quality translations (e.g. Level 4 Translation) still received negative scores. This suggests that ChatGPT using EAPrompt may have some rigour in the assessment process or be highly sensitive to subtle differences in translations. This could be due to the fact that the LLMs can capture subtle semantic differences and expressive inconsistencies that may seem acceptable to a human evaluator.

6 Conclusions and Future Directions

This paper highlighted the significance of a taxonomy of prompts for translation tasks, identifying critical design elements such as expression type, style, POS tagging, and few-shot examples. Furthermore, we explored in detail the key roles of gradable translation prompting taxonomy with explicit descriptions and contextual information to enhance the prompts' quality. The synergistic effect of these factors helps to improve translation quality and avoid misunderstanding and ambiguity, which in turn provides more precise guidance for the ChatGPT translation task. Based on the above, we conducted the T3S taxonomy of prompts for ChatGPT translation tasks.

In our study, we evaluated the effectiveness of our translation taxonomy by conducting an experiment using open-source datasets and standard evaluation metrics to rate translation quality across five levels. We also showcased the taxonomy's usefulness through a case study with ChatGPT, highlighting how prompt design impacts translation performance. Our findings offer valuable insights for enhancing ChatGPT's translation applications, prompt optimization, and overall translation quality and efficiency. Building on the insights gathered from our current investigation, future research could focus on comparing ChatGPT, prompted with our refined taxonomy, with dedicated translation services like Google Translate. This would assess whether our taxonomy-based approach can improve the translation capabilities of LLMs to outperform established translation services.

7 Acknowledgements

This work was supported by Gongdong Provincial Key Laboratory of Novel Security Intelligence Technologies (Grant No.2022B1212010005) and XJTLU Research Development Funding (Grant No.RDF-22-01-053).

References

Amplayo, R. K., K. M. Yoo, and S.-W. Lee. 2022. Attribute injection for pretrained language models: A new benchmark and an efficient method. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1051–1064, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chowdhery, A., S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Feng, X., Z. Feng, W. Zhao, B. Qin, and T. Liu. 2020. Enhanced neural machine translation by joint decoding with word and pos-tagging sequences. *Mobile Networks and Applications*, 25(5):1722–1728, oct.

Fu, Y., H. Peng, A. Sabharwal, P. Clark, and T. Khot. 2022. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Gilson, A., C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, D. Chartash, et al. 2023. How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9(1):e45312.

Goyal, N., C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Guo, J., Z. Zhang, L. Xu, H.-R. Wei, B. Chen, and E. Chen. 2020. Incorporating bert into parallel sequence decoding with adapters. *Advances in Neural Information Processing Systems*, 33:10843–10854.

He, Z., X. Wang, Z. Tu, S. Shi, and R. Wang. 2022. Tencent ai lab-shanghai jiao tong university low-resource translation system for the wmt22 translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 260–267, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hendy, A., M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Hlaing, Z. Z., Y. K. Thu, T. Supnithi, and P. Netisopakul. 2022. Improving neural machine translation with pos-tag features for low-resource language pairs. *Heliyon*, 8(8).

Jiang, E., E. Toh, A. Molina, K. Olson, C. Kayacik, A. Donsbach, C. J. Cai, and M. Terry. 2022. Discovering the syntax and strategies of natural language programming with generative language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Jiao, W., W. Wang, J.-t. Huang, X. Wang, and Z. Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Karmaker, S. K. and D. Feng. 2023. Teler: A general taxonomy of llm prompts for benchmarking complex tasks. *arXiv preprint arXiv:2305.11430*.

Kim, G., P. Baldi, and S. McAleer. 2023. Language models can solve computer tasks. *arXiv preprint arXiv:2303.17491*.

Kocmi, T. and C. Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton, and H. Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland, June. Eu-

ropean Association for Machine Translation.

Kojima, T., S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, X. 1991. Integral optimisation of translations. *Foreign Languages Research*, 04:54–59.

Li, Y., Y. Yin, J. Li, and Y. Zhang. 2022. Prompt-driven neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2579–2590, Dublin, Ireland, May. Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Liu, P., W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Liu, Z., Y. Xu, G. I. Winata, and P. Fung. 2019. Incorporating word and subword units in unsupervised machine translation using language model rescoring. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 275–282, Florence, Italy, August. Association for Computational Linguistics.

Lu, Q., L. Ding, L. Xie, K. Zhang, D. F. Wong, and D. Tao. 2023a. Toward human-like evaluation for natural language generation with error analysis. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5892–5907, Toronto, Canada, July. Association for Computational Linguistics.

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

Pan, W., Q. Chen, X. Xu, W. Che, and L. Qin. 2023. A preliminary evaluation of chatgpt for zero-shot dialogue understanding. *arXiv preprint arXiv:2304.04256*.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Popescu-Belis, A. 2019. Context in neural machine translation: A review of models and evaluations. *arXiv preprint arXiv:1901.09115*.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, and P. Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Press, O., M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12. Association for Machine Translation in the Americas.

Tan, Z., X. Zhang, S. Wang, and Y. Liu. 2022. MSP: Multi-stage prompting for making pre-trained language models better translators. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6131–6142, Dublin,

Ireland, May. Association for Computational Linguistics.

Thirunavukarasu, A. J., D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Trautmann, D., A. Petrova, and F. Schilder. 2022. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*.

Wang, B., S. Min, X. Deng, J. Shen, Y. Wu, L. Zettlemoyer, and H. Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada, July. Association for Computational Linguistics.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Xiao, X. 2010. Towards the three principles for translation of english advertising texts. *Journal of Jiangxi University of Finance and Economics*, pages 81–85.

Yang, S., Y. Wang, and X. Chu. 2020. A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*.

Yao, S., J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Zhao, Z., E. Wallace, S. Feng, D. Klein, and S. Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Zhou, Y., A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.