English Language Teaching and Linguistics Studies
ISSN 2640-9836 (Print) ISSN 2640-9844 (Online)
Vol. 8, No. 2, 2026
www.scholink.org/ojs/index.php/eltls

Original Paper

Prompting Strategies Enhance GPT-5's Chinese-English Legal Translation Quality: Versus DeepL

Weiyue Feng[1]

[1] Shanxi University of Finance and Economics, Taiyuan, Shanxi, China

Received: January 29, 2026 Accepted: April 19, 2026 Online Published: April 30, 2026

doi:10.22158/eltls.v8n2p257 URL: http://dx.doi.org/10.22158/eltls.v8n2p257

Abstract

Generative artificial intelligence does perform very well in machine translation, but in the legal scene of English-Chinese translation, controllability and professional reliability are not enough. Legal translation itself is characterized by intensive terminology, complex sentence patterns and strict logical requirements, so the requirements for accuracy and efficiency are extremely high. Because of this, the actual effect of this kind of AI in the legal field has not been fully verified, especially in consistency and coherence at the discourse level. In addition, there is a lack of in-depth research on the extent to which the project can improve the quality of translation. This study compares the performance of ChatGPT-5 and DeepL in English-Chinese legal translation, and examines the differences in the effects of different prompt strategies. Five prompt strategies are designed, from simple to complex, to deal with legal texts. There are three evaluation dimensions : lexical richness, semantic accuracy and discourse coherence, and then the statistical analysis method is used to find out the significant differences. The results show that the prompt design has a great influence on the translation quality. The more structured the prompt is, the more consistent and accurate the output is. In terms of semantic accuracy and coherence, GPT-5 is similar to DeepL. However, under the structured prompt, the lexical richness of GPT-5 is better. On the whole, this study promotes the development of legal machine translation, reveals the different effects of prompt engineering, and shows that structured prompts can improve lexical richness, while the two systems have their own advantages. Based on this, an evidence-based prompt optimization framework is proposed.

Keywords

Generative artificial intelligence translation, Legal translation, Prompt engineering, Translation evaluation

1. Introduction

In the practice of legal translation, there is a recognized direction with high difficulty, that is, the translation of legal texts (Cao, 2023). This work should not only ensure the accuracy of semantics and grammar, but also make the connection between sentences natural and logical, and keep consistent with the thinking framework of the source legal system. Recent studies have found that the specific design of prompt sentences will significantly affect the translation results of ChatGPT, such as whether the words are unified, whether the sentence structure is reasonable, and whether the style is in line with the norms (He, 2024; Wu & Hu, 2023; Yamada, 2023; Zhang et al., 2023). For example, if we clearly tell the model what role it should play, what kind of readers it should target, and what task constraints it has, it can often improve the accuracy of vocabulary, make the expression between sentences more consistent, and guide the system to produce more formal and contextual translation. Based on these understandings, this study puts the prompt project into the scene of legal sentence translation, because accuracy and contextual coherence are particularly important in legal translation. As Moneus and Sahari (2024) noted, AI-based translation systems often fail to fully capture the meaning of the original text and the complex legal background, so that the legal effectiveness of the translated content may be compromised.

In this study, we treat translation based on large language model (LLM) as a form of machine translation (MT). This approach continues the basic division of manual translation and machine-assisted translation in previous studies (Kenny, 2022). Although LLM-based systems are different from traditional neural machine translation (NMT) models in terms of structural design and user interaction, they are often classified as machine translation in practical applications, because such systems can automatically complete the translation process without direct human intervention. However, this paradigm is now undergoing a big change: from the traditional NMT which is biased towards certainty and pattern matching, to the generative AI translation which pays more attention to context and is more biased towards generative features. In such a highly professional field as legal translation, this technical turn has brought about a very practical problem, that is, the quality of translation is not reliable and accurate. In the past, most of the quality comparative studies within the scope of machine translation mainly focused on literary works (Gao et al., 2024) and news texts (Lin, 2024). In contrast, there are relatively few systematic studies on the performance of machine translation in the legal field, whether it is generative AI translation or traditional NMT system. There are some comprehensive assessments (Kamaluddin et al., 2024; Matviienko et al., 2024) provide useful baseline information, but they tend to ignore something unique to the legal language, such as cohesive devices. Therefore, a direct comparison between the representative of the current generative direction GPT-5 (OpenAI, 2025) and the representative of the optimized neural machine translation DeepL (2024) is critical for judging whether the generative AI translation or the NMT can better meet the standards of legal translation.

There is a distinctive feature in this study, that is, the project is applied to the quality control of legal

translation. The so-called prompt engineering is to design structured input instructions to optimize the output effect of the model (Han, 2025; Kong et al., 2024). In a translation task, the prompt can be as simple as “translate the text below”, or as complex as having a small number of examples that explicitly specify the role, purpose, tone, and example. Some previous studies have shown that prompt design affects the model 's understanding of tasks and the organization of contexts, thus affecting the final output results (Wei et al., 2022; Zhao et al., 2021). In the scene of legal translation, a reasonable prompt design allows the model to “play the role of a professional legal translator”, striving to achieve consistent terminology, formal tone, and accurate expression of terms. However, although this design has its potential, few studies have systematically explored the effect of prompt strategies designed specifically for the characteristics of legal texts (such as interdependence between clauses, coherence at the discourse level).

How to evaluate the quality of translation is also an issue that cannot be left. Today’s methods, whether automatic scoring (such as BLEU, TER) or artificial scoring, are not enough to measure the good or bad of legal translation (Peng et al., 2023; Stahlberg, 2020). Zhang (2025) mentioned that when studying ChatGPT-5 translation performance, people often think of the prompt information as a neglected variable, but in fact, it has a great impact on the model output. Although automatic scoring indicators are convenient to quantify contrast, they cannot grasp the nuances in context and the law. In contrast, manual evaluation， mainly whether translation is faithful and readily unsmooth (Fitria, 2023), remains the recognized “gold standard,” but it is necessary to refine it. As Kang and Zong (2022) demonstrated, it can be noticeably improved if the structure of the text level is taken into account in neural machine translation. This also shows that it is not enough to look at the sentence level alone, and it is necessary to evaluate the quality of translation in a larger context.

In the past, when doing legal translation research, whether it is manual translation or machine translation, we mostly stare at the terminology is not accurate, the common method is error analysis, corpus comparison, or traditional machine translation evaluation indicators. (Hu & Cheng, 2016; Clay, 2022; Alkatheery, 2023). While these methods provide some useful references, they often overlook higher-order text features such as coherence, interdependence between clauses, and consistency of style, which are the key to the construction of legal meaning. In addition, existing machine translation studies rarely consider the tip design as an analytical variable, especially in the context of legal text. In response to these two blanks – one is ignoring the features of higher-order text and the other is the lack of design in machine translation research – this study incorporates the hint engineering into the analytical model, hoping to introduce a new perspective on human-machine interaction: the quality of translation depends not only on the system itself, but also on how the instructions are designed. From this perspective, the prompt is like an external boot that connects user intentions and model outputs, affecting how the model understands tasks, how to choose language forms, and how to keep consistent between sentences. That is, translation results rely on both the ability of the model and the writing of the user input, which further illustrates that manual guidance is becoming more and more important in

AI-assisted translation.

Based on the above understanding, the study has built a multidimensional assessment framework covering terminology accuracy, syntax accuracy, stylistic suitability, language coherence and consistency. By comparing ChatGPT-5 and DeepL together, the study not only examined the impact of prompt design on translation results, but also gave a practical tips framework for future legal translation applications. In the process, it is highlighted that different tip strategies will not only affect the accuracy of sentences, but also affect broader text features such as consistency, structural organization and other. This allows for a more comprehensive assessment of the quality of translation, rather than the sight of isolated language units in the past.

Overall, the study aims to respond to the shortcomings identified in previous studies: the lack of systematic comparison between generative AI and neural machine translation in legal translation; the lack of operable tips for legal text characteristics; and the insufficient emphasis on consistency in translation assessments.

By addressing these issues, the study aims to theoretically contribute to understanding the role of generative AI translation in professional translation and, in practice, provide guidance to translators and language service providers to use AI systems responsibly and effectively in legal environments (Fu & Liu, 2024; Telaumbanua et al., 2024). Specifically, the study analyzes how prompts design affects translation quality in different language dimensions, thus revealing the interaction between artificial input and machine output. The study also proposes a structured tip framework that can be used for professional scenarios, which can help achieve more consistent and reliable results in legal translation.

2. Methods

2.1 Material

The corpus of this study are derived from three official legislative texts: *the Family Violence Reforms Act 2022*, *the Plain Language Act 2022* and *the Statute of the International Court of Justice*. There are several reasons to choose these texts. First, they do not have an official Chinese translation that minimizes the impact of the model in training in a bilingual control version. Second, these texts are publicly available, the legal provisions are complete and standardized, and the data is reliable and uniform. Third, the selected documents are closer to the legal and legislative characteristics, suitable for controlled contrast under different translation conditions.

It is necessary to comprehensively grasp the characteristics of the source text in terms of language complexity and consistency, and the analysis mainly uses mean and standard deviation (SD) to establish statistical benchmarks. These values come from language indicators such as lexical richness, coherence and semantic fluency, which are extracted from texts processed by the Coh-Metrix tool (Graesser et al., 2004). Coh-Metrix is chosen because it is recognized as a reliable and comprehensive text evaluation tool in the field of translation and readability research. It can take into account multi-dimensional coherence and lexical complexity, and is especially suitable for analyzing formal

legal texts with high language complexity and strict structural requirements.

Each statement is limited to 30 to 40 words in order to reduce the influence of text length on translation results. Lexical richness reflects the difference in language complexity by counting the diversity of words used. Coherence is measured by Coh-Metrix, where CNCAll reflects overall cohesion, and CNCLogic represents a clear logical relationship between sentences. In addition, Fleischer 's reading simplicity formula is used to evaluate readability and provide a standardized measure of text readability. The mean and standard deviation of each index are calculated to confirm that the linguistic features of the corpus are consistent (see Table 1). Finally, 27 legal English texts with highly similar linguistic features were selected to ensure the consistency and reliability of the data. The purpose of this selection is to minimize data pollution and improve the effectiveness of experimental results.

Table 1. Means and Standard Deviations of Linguistic Measures in the Legal Texts

	Mean (Standard Deviation)
Word Count	33.63 (3.51)
CNCAll	111.11 (46.14)
Lexical Richness	46.57 (19.58)
CNCLogic	56.49 (35.28)
Readability	28.49 (13.69)

2.2 Translation Procedure

This study uses factorial design to systematically examine the impact of different tools and prompt engineering on the quality of legal translation. There are two independent variables : translation system and prompt strategy. The former includes GPT-5 and DeepL, which represent generative AI translation and traditional neural machine translation respectively. DeepL does not support prompt input and can only be operated through a fixed interface. The latter covers five cue conditions, from simple to highly customized: GPT-Simple, GPT-Task-Specified, GPT-Personna, GPT-Constrained and GPT-Composite. The dependent variable is the overall legal translation performance, which is measured by four indicators : lexical richness, noun accuracy, verb accuracy and coherence. This two-factor design not only allows the study to look at the main effects of translation system and prompt type respectively, but also to analyze their interaction on the output of professional legal translation.

In order to find out how these language results are affected, this study uses a hierarchical explicit guidance framework to apply the prompt engineering strategy. Five prompt strategies are designed for GPT-5 to observe how different prompt designs affect the quality of machine translation. These hints are gradually constructed : starting with the simplest zero-sample hint (requiring direct translation only), they are gradually upgraded to a more detailed version, which clarifies the translator’s role, task objectives, structural constraints, and also emphasizes terminological consistency and legislative style

(Kong et al., 2024; Wang et al., 2024; White et al., 2023). The design should make the difference between different prompt strategies clear and comparable. All tips are written concisely and directly, avoiding ambiguity and ensuring consistent response under multiple tests. In addition, the privacy mode of GPT-5 is opened to prevent data retention and ensure that the subsequent experimental links are independent of each other. Figure 1 shows an evolution path from zero samples to a small number of samples. The detailed tips listed in Table 2 are intended to ensure transparency and repeatability. All the translations of this study are completed with GPT-5.

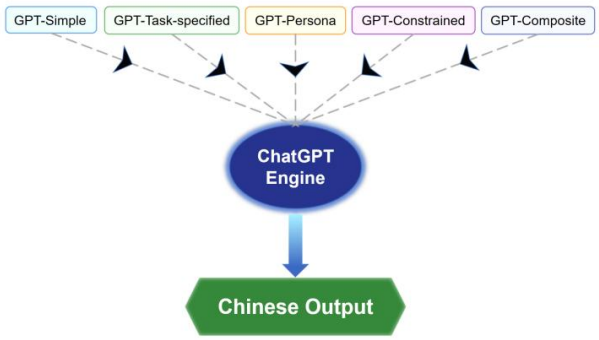

Figure 1. Diagram of GPT-5 based Translation Process Using Five prompt Strategies

Table 2. Design and Strategies of Prompt Engineering for Legal Translation

No.	Prompt	Design Focus
1	Before translating, please state the model name and version you are currently using. Translate the following English legal provisions into Chinese. [Insert text here]	Evaluates the model's raw translation capability without persona or stylistic guidance. Acts as the primary AI control group.
2	Before translating, please state the model name and version you are currently using. You are asked to translate the following legal provisions from English into formal written Chinese. Maintain legal meaning, preserve structure, and avoid summarization. [Insert text here]	Focuses on enforcing a "Formal Written" register and strict structural fidelity by explicitly forbidding summarization or paraphrasing.
3	You are an experienced legal translator. Translate the following legal text from the [Name of Law] into Chinese.	Employs a specific professional persona (Experienced Translator) and provides the specific legal act

	Your task is to translate the following provisions faithfully and precisely, using standard Chinese legislative terminology and formal legal syntax. [Insert text here]	name to trigger domain-specific knowledge.
4	Before translating, please state the model name and version you are currently using. You are a professional legal translator. Translate the following provisions from English into Chinese for publication in an official bilingual law database. Follow these constraints: 1.Keep the numbering and clause structure unchanged. 2.Ensure accuracy and terminological consistency throughout. [Insert text here]	Sets a high-stakes application scenario (Official Database) and introduces numbered constraints for structural and terminological consistency.
5	Before translating, please state the model name and version you are currently using. You are an expert legal translator. Translate the following section into formal written Chinese, ensuring: 1.Clause alignment and consistent terminology. 2.Logical coherence across sections (e.g., references to “declaration” or “court or judge” remain semantically linked). 3.Legislative style and tone remain consistent with official Chinese statutes. [Insert text here]	The most advanced configuration; targets document-level logic, semantic linking of recurring terms, and stylistic alignment with official PRC statutory tone.

2.3 Apparatus

This study uses several digital tools to complete translation generation, text evaluation and statistical analysis. The choice of these tools is based on whether they are relevant, reliable, and can be applied to the analysis of legal translation performance under different prompt conditions. By integrating multiple platforms, the research not only ensures the standardization of methods, but also has the ability of in-depth analysis.

2.3.1 Translation Tools

This study uses two translation systems : ChatGPT-5 and DeepL. The former generates translations

under five different prompt strategies to facilitate the systematic evaluation of the impact of prompt design on output. ChatGPT, as a large language model, can flexibly assign tasks by prompting engineering, which is suitable for studying the controllable changes of translation conditions. All prompts are executed on the same platform to ensure that the process is consistent.

DeepL is used as a benchmark system for comparison. As a widely recognized neural machine translation system, DeepL is famous for its excellent performance in general translation tasks and relatively stable output quality. Incorporating it into this study provides a benchmark for evaluating the performance of prompt-based generative translation.

2.3.2 Evaluation Tools

In order to evaluate the linguistic features of the source text and the target text, we use a variety of tools. Among them, AlphaReadabilityChinese (Lei, Wei, & Liu, 2024) is used to evaluate the linguistic features of Chinese translations. The tool can give quantifiable readability and language quality indicators, such as semantics, lexical richness, accuracy, complexity, etc. It is selected because it is specifically designed for Chinese text, with large processing capacity, and can output multiple indicators that reflect different aspects of the text. It is suitable for Chinese language analysis, which is in line with the research.

In addition, Coh-Metrix is also used in the original English text. This tool has a good reputation in linguistics and translation studies because it can capture multiple dimensions of the text, such as coherence, lexical diversity and logical structure. With it, the comparability between source texts can be systematically evaluated before translation.

In addition, an online readability platform was used to evaluate readability using Fleisch 's reading simplicity index (Flesch, 1948). This metric estimates text difficulty based on sentence length and word complexity, providing an additional standardized reference for readability.

2.4 Experimental Workflow

2.4.1 Automated Text Analysis

This experimental procedure is designed to ensure the consistency, comparability and repeatability of the results under different translation conditions. The whole process is divided into three steps : translation generation, data collation, and evaluation preparation.

A total of 27 selected legal provisions are translated under six different conditions: five prompt strategies under GPT-5 (GPT-Simple, GPT-Task-Specified, GPT-Persona, GPT-Constrained, GPT-Composite), plus a benchmark condition of DeepL. All conditions use the same source text, so that the difference in output can be attributed to the difference in the prompt strategy or translation system, rather than the input material itself. Each prompt is applied only once, without iterative optimization or post-editing, and the original system output is retained for analysis. For the consistency of the process, all prompts are executed in a fixed order, and the prompt structure is the same under each condition, and the text input order is also unchanged. This ensures that each source text follows the same experimental parameters under each prompt strategy. In addition, all translations are

completed in a short time to avoid the impact of system updates or environmental fluctuations.

After the translation is generated, all the outputs are systematically recorded and organized to form a structured data set. This dataset consists of three parts: (1) The original text and its pre-evaluated linguistic features; (2) The translation results corresponding to each experimental condition; (3) Translation language features extracted by ARC tool. Each translation is marked with its own conditions, which facilitates a clear horizontal comparison between different prompt strategies and benchmark systems.

In order to unify the data format of subsequent evaluations, all translation outputs are normalized before analysis. Non-essential formatting elements such as clause numbers or sub-paragraph references in parentheses (such as "(6)(b)" in legal texts) have been removed for compatibility with evaluation tools. Because this study only wants to see the original performance of each system under the set conditions, there is no manual modification of the translation content.

This part ensures that all translations are generated, recorded and prepared in a systematic and consistent manner, thus supporting reliable comparative analysis under different conditions and minimizing the influence of external variables.

2.4.2 Human Evaluation of Discourse Coherence

In addition to automatic evaluation indicators, textual coherence is also judged manually. In legal texts, logical structure and clarity are not optional issues, but the key to translation function, so this dimension is worth looking at separately. Three graduate students majoring in translation serve as raters. Before the assessment, they received training based on the existing assessment methods. The purpose is to enable the three people to have a relatively unified understanding of what each standard means in practice.

The score uses the 5-point Like Scale, from 1 point (very poor) to 5 points (excellent), to see three dimensions: logical coherence, connectivity, information fluidity. This framework is chosen because more and more studies point out that the characteristics of the language level deserve more systematic attention in translation evaluation (Li, & Gao, 2025; Ma et al., 2026). Logical coherence is the first dimension. It asks whether the translation faithfully reproduces the discourse relationship coded in the source text, such as causal reasoning, conditional structure and concession framework. Cohesion is more internal: it involves whether the reference chain is coherent, whether conjunctions are used appropriately, and whether legal terms are consistent in each clause. The flow of information is the third dimension, which may be the most extensive of these three dimensions. The focus is on whether the overall arrangement of the text is clear and logically orderly from the reader's perspective.

In order to reduce the fatigue and subjective bias of evaluators, we have taken several specific measures. All translations are randomly sorted before distribution to prevent evaluators from seeing different versions of the same original text continuously. The evaluation is carried out in several stages, the duration of each stage is controllable, and the evaluator is encouraged to rest between stages. At the same time, the blind method is used: the evaluator does not know which system the translation comes from or which prompts, and ensures that the results reflect the text itself, not any expectations of the system. These practices together ensure the reliability and effectiveness of manual evaluation.

2.5 Data Analysis

We use SPSS (version 27) for statistical analysis, including repeated measurement variance analysis, two-to-two comparison, and descriptive statistics such as mean and standard deviation. SPSS is chosen because it is reliable, friendly interface, widely used in quantitative linguistics and social sciences, and can ensure that the analysis process is standardized and repeatable.

In addition, R language is also used for data visualization. The graphics of the generated results are used to show the differences between different prompt strategies and also help explain the statistical results. Using R language can make the patterns and changes in the data clearer, thus improving the clarity and transparency of the results.

In order to strictly examine the impact of different prompt engineering strategies on the effect of machine translation, this study adopts an intra-group experimental design, and the main statistical method is Re-ANOVA. The reason why this model is chosen is that the same batch of legal texts (N = 27) should be measured under multiple experimental conditions, which can control the differences in the text itself and improve the statistical effectiveness of the analysis.

In statistical analysis, prompt conditions are divided into five levels as independent variables in the group: GPT-simple, GPT-Task-Specified, GPT-Persona, GPT-Constrained, GPT-Composite. The result variable selects four quantitative indicators to measure the quality of translation: vocabulary richness, noun accuracy, verb accuracy (these two reflect semantic accuracy) and coherence. Use repeated measurement variance analysis to test the effect of independent variables in the group under different prompt conditions, to see whether there is a significant overall difference between these five prompt strategies in each language indicator, and lay a statistical basis for subsequent pair comparison.

In addition, the partial eta square (η_p^2) is used to represent the amount of effect, which shows how big the observed difference is. The amount of reporting effects allows later researchers to judge the practical significance of the prompt project in the translation task, and also facilitates the subsequent meta-analysis.

While using automatic indicators, it is also combined with manual scoring based on the Likter scale as a supplementary perspective to measure the quality of coherence. The three scorers score the coherence of the translation respectively, and the results are summarized before analysis. The scoring is divided into two steps: for each translation, the scores of the three scorers in each dimension are first averaged to obtain a single score of each dimension; then the average of the three dimensions is added to obtain

the overall coherence score, and the weight of the three dimensions (logical coherence, coherence, information flow) is the same. It is intentional to separate and then summarize, so that logical relationships, text connectivity and structural organization can be analyzed from the sub-item and overall levels respectively, instead of combining them into a number at the beginning.

After the test, I made two comparisons according to the plan to see the specific differences between the five experimental prompt conditions and the benchmark system (DeepL). These comparisons are theoretically based on the purpose of testing whether structured hints can make large language models reach or even exceed the level of professional machine translation engines in the professional field of legal translation.

In order to avoid the risk of the first type of error (false positive) caused by multiple comparisons (five times for each indicator), we strictly adopted Beauferni correction. The adjusted significance level of each comparison is obtained by the initial value ($\alpha = 0.05$) divided by the number of comparisons. This conservative approach can ensure that any reported hint strategy advantage is statistically significant, not caused by cumulative probability.

3. Results

3.1 Descriptive Statistics of Translation Quality Measures

Table 3 lists the average score (M) and standard deviation (SD) of each translation quality indicator under six experimental conditions. Descriptive analysis mainly looks at the changes in relative performance between different prompt strategies. In 27 legal text samples, the average score of the five GPT-5 variants is very close, especially in terms of semantic accuracy. These descriptive results provide an empirical basis for the subsequent analysis of repeated measurement variance, which is used to judge whether these differences are statistically significant.

Table 3. Descriptive Statistics of Translation Quality across Experimental Conditions (N = 27)

Condition	Lexical Richness (M ± SD)	Noun Accuracy (M ± SD)	Verb Accuracy (M ± SD)	Cohesion (M ± SD)
GPT-Simple	3.21 ± .15	4.48 ± 1.88	7.56 ± 1.83	4.32 ± .42
GPT-Task-specified	3.23 ± .14	4.13 ± 1.78	7.14 ± 2.00	4.05 ± .50
GPT-Persona	3.20 ± .15	4.42 ± 1.74	7.24 ± 1.61	4.25 ± .40
GPT-Constrained	3.17 ± .14	4.40 ± 1.98	7.19 ± 1.88	4.11 ± .51
GPT-Composite	3.23 ± .22	4.24 ± 1.69	7.27 ± 1.90	4.31 ± .53
DeepL (Baseline)	3.13 ± .12	4.32 ± 1.48	7.50 ± 2.22	4.18 ± .42

Notes. 1. Lexical Richness, Noun Accuracy, and Verb Accuracy were calculated using automated computational metrics.

2. Cohesion was evaluated through human assessment using a 5-point Likert scale (1 = Very Poor, 5 =

Excellent).

3.2 General Analysis of Translation Quality Conditions

The repeated-measures ANOVA indicated that the effect of prompting strategies on lexical richness was not statistically significant among the five GPT-5 outputs, F (4,104) = 1.56, p = .19, = .06. Descriptive statistics showed that lexical richness values across GPT-Simple to GPT-Composite were closely distributed, with overlapping ranges and similar standard deviations across conditions. However, a distinct distributional shift was observed when these GPT-5 based conditions were collectively contrasted against the DeepL baseline. Planned pairwise comparisons revealed that while the five GPT-5 strategies maintained internal parity, they consistently diverged from baseline NMT performance. GPT-Task-specified (p < .001) and GPT-Persona (p = .002) exhibited significantly higher lexical richness scores than DeepL, with effect sizes (η_p^2) of .48 and .32, respectively. The horizontal distance between the GPT-5 group means and the DeepL mean in the density plots confirmed that even the minimal-prompting condition (GPT-Simple) sustained a statistically significant margin over the baseline (p = .006).

Regarding semantic accuracy, both noun and verb dimensions exhibited robust stability across the experimental conditions. The omnibus test for noun accuracy approached marginal significance, F (4, 104) = 2.21, p = .07, = .08. As illustrated in the violin plots, the density curved for all six translation systems, including the five GPT-5 variants and the DeepL baseline, which displayed a highly symmetrical and concentrated distribution. The kernels of these distributions were predominantly aligned between the 4.0 and 5.0 score range, with their respective medians showing minimal vertical displacement. Planned pairwise comparisons against DeepL confirmed this trend, yielding no significant deviations (all p > .05). Even the strategy with the highest observed mean, GPT-Constrained, maintained a performance profile statistically equivalent to the professional MT engine (p = .71, = .005), as evidenced by the extensive overlap of the 95% confidence intervals across the horizontal axis.

Similarly, the analysis of verb accuracy revealed a nearly flat performance profile across all prompting conditions, F (4, 104) = .96, p = .44, = .04. The internal variability within each condition remained consistent, with the dispersion patterns of the 27 legal text samples showing a high degree of homogeneity. The minimal variance observed between GPT-Constrained and GPT-Composite suggests that for predicate logic in legal discourse, the model's performance adheres to a rigid accuracy ceiling.

Regarding cohesion, the repeated-measures ANOVA revealed a significant main effect of prompting strategy among the five GPT outputs, F (4, 104) = 2.97, p = .02, = .10. Descriptive statistics indicated that the mean cohesion scores for GPT-generated texts fluctuated between 4.05 and 4.32 on the 5-point Likert scale. The distributional patterns across the experimental conditions showed varying degrees of density; specifically, GPT-Simple and GPT-Composite maintained higher numerical concentrations compared to other variants. Despite the significant variance observed within the GPT-5 experimental group, planned pairwise comparisons between each GPT-5 strategy and the DeepL baseline (M = 4.18)

yielded no statistically significant differences. For instance, the comparison between the highest-scoring GPT-Simple and DeepL resulted in p = .268, = .047, while the lowest-scoring GPT-Task-specified showed p = .321, = .038. The results show that the output of GPT-5 is statistically comparable to that of the DeepL benchmark in terms of coherence. The score distribution of the two coincides significantly, and the amount of effect in all comparisons is relatively small.

3.3 Distributional Analysis of Translation Outputs

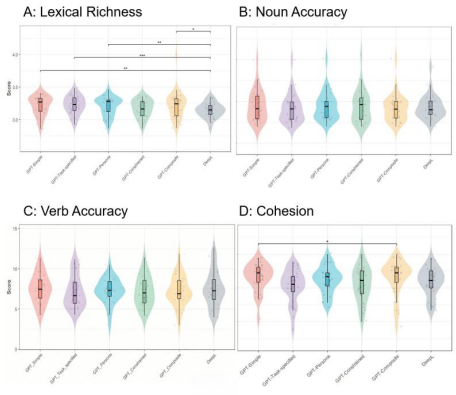

Figure 2. Violin Plot Comparison of Translation Systems on Lexical Richness, Noun Accuracy, Verb Accuracy, and Cohesion Metrics (*p < .05, **p < .01, ***p < .001)

Figure 2 gives four violin diagrams, which visually show the distribution of translation quality scores under different prompt conditions. The horizontal axis is the translation conditions (GPT-Simple, GPT-Task-Specified, GPT-Persona, GPT-Contrained, GPT-Composite and DeepL), and the vertical axis is the specifics of each quality index. Values, including vocabulary richness, noun accuracy, verb accuracy and coherence scores. The width of each violin diagram represents the density of the data point, and the wider the place, the more concentrated the data.

From the perspective of lexical richness (Figure-A), the distribution density of GPT versions is on the rise compared with the DeepL baseline as a whole. Specifically, in both cases, the interquartile range (IQR) was completely higher than the median of the DeepL group. In addition, the distribution of DeepL is more concentrated, while the density distribution of GPT-composite strategy is wider, and the long tail feature of upturn is obvious. In the violin plots specified by the GPT-composite and GPT-Task-Specified, there are individual data points exceeding 4.0, indicating that these strategies may achieve the highest vocabulary performance that the baseline system cannot do.

In terms of noun accuracy (see Figure-B), the distribution in all six cases is very symmetrical and is mainly concentrated in the range of 4.0 to 5.0. The obvious vertical overlap and similar median positions in all line charts visually confirm the statistically significant differences not found in the two-to-two comparison.

In terms of verb accuracy (see Figure-C), this indicator is most evenly distributed under all hint

conditions. The density curves of GPT-Constrained and GPT-Composite are particularly compact, indicating that the performance consistency is very high, which is comparable to the professional benchmark.

Coherence (see Figure-D): Although the coherence scores evaluated manually are more scattered than automatic indicators, the "excellent" ratings of GPT-Simple and GPT-Composite are more concentrated. Specifically, in both cases, the density core is obviously wider at the top of the score (5.0 points), while the curve of DeepL is sharper. The midline of GPT-Simple and GPT-Composite falls on the upper boundary of the quartile range of the GPT-Task-Specified group, indicating that the intermediate data has obviously moved up. In contrast, the distribution of DeepL is more like a uniform cylinder, with a more even density between 3.0 and 5.0. Although the total score range of all experimental groups (about -2.0 to 6.0 on the visual score table) is the same, the internal box diagrams of each variant of GPT-5 have a higher degree of vertical compression in the 4.0-5.0 range, indicating that more observations have been gathered near the highest mass threshold.

4. Discussion

This study found that the impact of prompting engineering on the quality of legal translation is not different in different language dimensions. Specifically, the role of prompts is not unified, but changes significantly with different dimensions: it has a significant impact on vocabulary richness, has a limited impact on connection and depends on the specific context, and has a very small impact on the semantic accuracy of core legal terms. The comparison with the benchm-5ark neural machine translation system (DeepL) further shows that although the style of GPT output is somewhat different, there is no statistically significant difference in the semantic accuracy of the core legal terms between the two systems.

This study did not make a wide range of comparisons, but conducted a controlled empirical test of the performance of the prompt-based generation model and the neural machine translation system on the same batch of legal texts under clearly specified conditions. In this way, we can see more intensively where the differences appear and where the performance converges. In this way, the study fills a gap in the literature: in the legal field, there are not many studies that directly and conditionally compare generative AI translation and neural machine translation. This practice provides concrete evidence of the comparative advantages of the two types of systems, especially explaining why prompt strategies can bring obvious differences in vocabulary richness and coherence. The next discussion will link the above findings with the core issues in the research and put them in the broader context of translation research and natural language processing research.

4.1 Lexical Richness

In terms of vocabulary richness, GPT-5 is obviously better than DeepL under different prompt conditions, especially under structured prompts such as GPT-Task-Specified and GPT-Persona. This discovery fills the research gap of the lack of actionable prompt strategies in legal translation. This

study clearly defines the prompt conditions and changes systematically at five different levels, from the simplest instructions (GPT-Simple) to more structural and binding settings (GPT-Composite). By presenting prompt strategies in a structured and comparable form, this study transcends abstract prompt engineering discussions and provides a specific framework for testing and adjustment, which can be used for future legal translation research. Previous studies have shown that prompt design can affect translation quality by shaping model behavior (Yamada, 2023; Brown et al., 2020)[35], but few people have studied the application of such strategies in professional fields. Current research results show that structured hints can provide clearer context guidance, so that the model can adopt a more formal and professional expression. This is consistent with the research results of Wang et al. (2024), who pointed out that role-based prompts can improve the style suitability and domain adaptability of language model output.

A more detailed explanation can be obtained by using teleology and model theory. On the basis of teleology (Vermeer, 1978; Nord, 1997), translation is guided by its expected function in the target context. In this study, the vocabulary performance of the two types of prompts, GPT-Task-Specified and GPT-Persona setting, is obviously better than the benchmark DeepL system and the basic GPT-simple conditions. Unlike the simple translation instructions of zero samples, these two tips clarify the task as a formal and official legislative statement and give the status of a professional legal translator. This goal-oriented constraint has prompted GPT-5 to shift from colloquial to precise legal expression, which directly enhances the diversity of vocabulary. In contrast, GPT-Constraint and GPT-Composite value more structural consistency and text logical coherence, and do not make additional optimization in vocabulary refinement, so they do not bring obvious vocabulary improvement. From the perspective of pattern theory, role and task-oriented prompts effectively activate the formal legal language mode of model storage and guide it to choose domain terms. In general, this targeted prompt adjustment changes the language style output, but does not change the core semantics. This also theoretically confirms a conclusion that structured prompts can selectively improve the vocabulary richness in AI-assisted legal translation.

4.2 Semantic Accuracy

From the perspective of semantic accuracy, the performance under various cue conditions is highly consistent, and the difference between the output of GPT-5 and the DeepL baseline is almost negligible. At the noun level (usually corresponding to the core legal terms and entities), all systems can give stable and accurate expressions, which reflects the normative nature of the legal terms themselves (Sun & Chen, 2024). The verb level (representing predicate structure and relational action) is also the same. No matter how the prompt is designed, the translation output is very stable. These results indicate that, unlike lexical richness, the accurate transmission of legal meaning is less sensitive to the structural or functional constraints brought by cues.

The possible reason for this phenomenon is that the legal language is very normative, and there is not much room for semantic change (Prieto, 2021). In this case, the accuracy of translation depends more

on the training data and internal representation of the model itself, rather than how the external prompts are written. This view complements the previous studies that emphasize that prompts can shape the behavior of generative models (Zhang et al., 2023), indicating that the role of prompts may be limited in those highly regulated fields. In addition, the similar performance of GPT-5 and DeepL also reminds us that a strong neural machine translation baseline is still important, and the progress of the generated model does not mean that the technical semantic accuracy will naturally improve. From the perspective of stylistic theory and computational linguistics, it can be said in detail:

According to corpus theory (Hatim & Mason, 1990), legal language is a highly restricted genre, in which the semantic mapping of core words is constrained by established norms and institutional usage. Key legal expressions such as modal verbs expressing obligations or verbs expressing responsibilities cannot be replaced at will, but should follow long-established legal interpretations and usage practices. Therefore, even if the prompt strategies are different, the acceptable vocabulary selection range is still limited, and the model will naturally tend to output stable and compliant translations.

From the perspective of computational linguistics, this linguistic limitation can also be reflected in the stability of the potential semantic representation of the model (Mikolov et al., 2013; Devlin et al., 2019). In the pre-training stage, the large language model learns through a large amount of text. In the legal text, many terms will form a stable symbiosis relationship, which are encoded into dense and fixed representations in the semantic space, so the core legal terms are not affected by external instructions. Although the role setting or task description in the prompt project can adjust attention and affect the style, the effect is more on the surface of language. When the model encounters legal words, the pre-trained semantic association intensity often dominates, and almost the same words will be selected under different prompt conditions. Therefore, the binding characteristics of the legal style are further enhanced by the stability expressed within the model, which explains why there is no obvious statistical difference in semantic accuracy between different GPT-5 configurations and DeepL baselines.

4.3 Cohesion

As for the results of coherence, a third research gap is also pointed out : insufficient attention has been paid to the discourse-level features in translation assessment. Although there are statistical differences in GPT-5 under different cue conditions, there is no significant difference compared with the DeepL baseline, indicating that the overall coherence level of the two systems is similar. However, among the four indicators, coherence is most affected by cue conditions and has the strongest statistical effect, indicating that it is more sensitive to changes in cue design. This result further shows that it is necessary to incorporate coherence indicators into machine translation evaluation, because features such as coherence can capture differences that cannot be reflected by sentence-level indicators (Kang & Zong, 2022). This is also consistent with previous research findings : many studies adopt the univariate method, only focusing on isolated linguistic features, ignoring the attributes at the discourse level (Li & Gao, 2025).

This concept can be explained more clearly by using connection theory and discourse theory.

According to Halliday and Hasan's (2014) believe that the connection relies on the repetition of these language means of references, conjunctions and words, which together make the meanings between sentences coherent. When translating, these connections must be retained or reconstructed to maintain the coherence of the text. From the perspective of discourse processing, the connection is not only a structural attribute, but also depends on how the reader understands the connection between sentences (Kintsch, 1998). It involves the cognitive relationship between propositions and views, making readers feel that the text is coherent and meaningful. In the occasion where the logical relationship of legal translation is very strict, changes in the prompt design will affect how the model maintains the connection between sentences. These results show that although large language models can reliably retain semantic content, the ability to adjust coherence is more sensitive to guidance information, which reminds us that we should not only look at isolated words or grammar to evaluate translation quality. This also explains why coherence is more sensitive to different prompt strategies: those prompts that combine multiple constraints such as role description, style guidance, structural instructions, etc. will affect the way the model organizes information between different subordinate clauses. In addition, the reason why there is no obvious difference from the benchmark system is that whether it is GPT or DeepL, it can maintain the basic subordinate sentence coherence and logical flow in the legal text. Although the output of GPT-5 has a certain sensitivity to prompt changes, the standardized structure and standardized wording of the legal language itself provide strong support for all systems. The inherent regularity of the source text limits the range of change, and also limits the measurable impact of different prompt strategies on the overall coherence. Overall, the coherence in translation is determined by language restrictions, cognitive interpretation and model-level processing, which explains why diversity is observed under different prompt conditions, although there is no significant difference compared with the benchmark system.

4.4 Limitation

There are also some things that need to be paid attention to in these results, and it is necessary to mention them. The prompt strategy is artificially designed. Although we have made adjustments to different structures and constraints, it does not guarantee that these configurations cover all possible situations. Different prompt designs may bring different results, so pay attention to this when interpreting. The data set is also a problem that needs to be treated with caution. The legal clauses used were deliberately controlled for consistency, but a small and carefully selected sample is not the same as a representative one, and legal discourse is far more varied in practice than any single study can adequately capture. There is also the question of evaluation. Quantitative metrics formed the backbone of the analysis, with human assessment playing a more limited supporting role. This is a defensible choice given the scale of the study, but it does mean that some of the more interpretive, context-dependent dimensions of translation quality remain underexplored. Future work could productively pick up on any of these threads: automated prompt generation would help move beyond the constraints of manual design; a wider range of legal text types would test how far the current

findings actually travel; and more integrated evaluation approaches, ones that bring human judgment into closer and more systematic dialogue with quantitative measurement, would give a fuller picture of what translation quality in this domain really involves.

4.5 Summary

In a word, this study finds that prompting engineering has different effects on different linguistic dimensions of English-Chinese legal translation. Structured prompts significantly improve the richness of vocabulary, so that GPT-5 can generate more formal and domain-characteristic expressions than the DeepL baseline. In contrast, the core semantic accuracy remains stable under different prompts, because fixed legal terms and established professional practices limit changes and ensure the overall consistency of generative and neural machine translation systems. The coherence of the discourse is more complicated. The prompt adjustment has a significant impact on the order and logical connection of subordinate clauses in the GPT-5 output, but the overall coherence level is comparable to that of DeepL, which is related to the inherent strict norms of the legal discourse itself. These results show that the prompted-enhanced large language model and the traditional neural machine translation system have their own advantages in legal translation. When following carefully designed prompts, the language model is especially good at optimizing style characteristics such as vocabulary selection and text fluency, while the neural machine translation system is stable and reliable in accurately conveying the core legal meaning. This brings a practical revelation: which tool to choose is very important, especially when the translation priority is clear. The hint project is not omnipotent, but it does work.

5. Conclusion

This study compares the generative AI translation system with the professional neural machine translation benchmark (DeepL) to examine the effect of structured prompt engineering in English-Chinese legal translation. Through five prompt strategies (from the simplest instructions to multi-constrained composite specifications), the study has delimited the functional boundaries of artificial prompts in shaping the translation quality of different language dimensions. The result does not present a single and unified pattern, but shows that the hints play a role in different ways and to different degrees in different languages.

The most obvious influence of the project is the richness of vocabulary. In terms of vocabulary diversity and context matching, structured tips (especially those with character settings or clear task frameworks) make the model significantly exceed the DeepL baseline. The reason is not complicated: if the functional requirements of the task are clear, the model will go in the direction of formal legal terms and will not rely on general or simplified expressions. In other words, roles and mission instructions replace the domain knowledge of human translators to a certain extent. This shows that prompt design can effectively shape the style characteristics of translation, especially when there are multiple acceptable language options.

In contrast, there is an obvious upper limit on semantic accuracy. Whether it is a core legal noun or a

verb, the translation accuracy of all prompt strategies and DeepL baselines is statistically similar, and there is no significant difference. This shows that the expression of the key legal meaning is not much affected by the prompt conditions, but is more determined by the existing knowledge of the model and the practice of the legal language itself. The structure of legal terms and predicates is in a relatively fixed semantic space, and the acceptable equivalent expression is mainly constrained by established usage and domain norms. Therefore, whether it is a generative model or a traditional neural machine translation, when dealing with legally significant units, even if the prompt conditions are different, the output results are often very close.

In addition, the results of coherence are more complicated. Coherence itself is a tricky topic. It is indeed more sensitive to prompt design than semantic accuracy, and there are also meaningful differences between different configurations of GPT-5, but these differences are not big to surpass DeepL statistically. This shows that although prompts can affect the information connection between sentences to a certain extent, this influence has an upper limit. Legal texts are not created out of thin air: they follow a set of evolved standardized organizational models, which are formed to achieve specific communication functions, and also limit the connection between sentences, no matter what instructions the model receives. The actual effect is that all systems tend to be close to a similar level in terms of overall coherence, even if their internal strategies are different.

Taken together, these results give us a clearer understanding of how cueing techniques work in specific translation scenarios. The cueing strategy does not bring all-round improvement, but selectively affects some aspects of the language. While flexibly adjusting, it has little effect on those components that are more constrained by domain norms. This also shows that the evaluation of translation quality requires a multi-dimensional perspective, because different linguistic features respond differently to experimental operations.

References

Cao, D. (2023). On the challenges of legal translation. *Comparative Legilinguistics*, *55*, 109-117.

He, S. (2024, June). Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation* (Vol. 1) (pp. 316-326).

Wu, Y., & Hu, G. (2023, December). Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 166-169).

Yamada, M. (2023). *Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability*.

Zhang, B., Haddow, B., & Birch, A. (2023, July). Prompting large language model for machine translation: A case study. In *International conference on machine learning* (pp. 41092-41110). PMLR.

Moneus, A. M., & Sahari, Y. (2024). Artificial intelligence and human translation: A contrastive study based on legal texts. *Heliyon*, *10*(6), e28613.

Kenny, D. (2022). Machine translation for everyone: Empowering users in the age of artificial intelligence. *Language Science Press*.

Gao, R., Lin, Y., Zhao, N., & Cai, Z. G. (2024). Machine translation of Chinese classical poetry: A comparison among ChatGPT, Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications*, *11*(1), 1-10.

Lin, J. (2024). Artificial intelligence applications in English news translation: Strategies and research. *International Conference on Educational Information Technology* (Vol. 6, No. 2).

Kamaluddin, M. I., Rasyid, M. W. K., Abqoriyyah, F. H., & Saehu, A. (2024, June). Accuracy analysis of DeepL: Breakthroughs in machine translation technology. *Journal of English Education Forum (JEEF)*, *4*(2), 122-126.

Matviienko, L., Khomenko, L., Denysovets, I., Horodenska, K., Nikolashyna, T., & Pavlova, I. (2024). Comparative analysis of online translators in the machine translation system. *Revista Romaneasca pentru Educatie Multidimensionala*, *16*(3), 101-118.

OpenAI. (2025). *ChatGPT (GPT-5.2)* [Large language model].

DeepL SE. (2024). *DeepL Translator* [Machine translation software].

Han, J. (2025). *Conveying Imagistic Thinking in TCM Translation: A Prompt Engineering and LLM-Based Evaluation Framework*. arXiv e-prints, arXiv-2511.

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., & Zhou, X. (2023). *Better zero-shot reasoning with role-play prompting*. arXiv preprint arXiv:2308.07702.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

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

Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., … Tao, D. (2023). *Towards making the most of ChatGPT for machine translation*.

Stahlberg, F. (2020). Neural machine translation: A review. *Journal of Artificial Intelligence Research*, *69*, 343-418.

Zhang, Y. (2025). Consideration of prompts as a neglected factor in research on evaluating ChatGPT's translation performance. *Journal of China Computer-Assisted Language Learning*.

Fitria, T. N. (2023). Performance of Google Translate, Microsoft Translator, and DeepL Translator: Error analysis of translation result. *Al-Lisan: Jurnal Bahasa (e-Journal)*, *8*(2), 115-138.

Kang, X., & Zong, C. (2022). Enhancing neural machine translation with discourse-level context modeling: A survey and new perspectives. *Journal of Chinese Information Processing*, *36*(4),

1-15.

Hu, P.-C., & Cheng, L. (2016). A study of legal translation from the perspective of error analysis. *International Journal of Language and Law*, *5*, 1-15.

Clay, E. (2022). A corpus-based approach to examining terminological variation in EU law. *JLL*, *11*, 142.

Alkatheery, E. R. (2023). Google translate errors in legal texts: Machine translation quality assessment. *AWEJ for Translation & Literary Studies*, *7*(1), 208-219.

Fu, L., & Liu, L. (2024). What are the differences? A comparative study of generative artificial intelligence translation and human translation of scientific texts. *Humanities and Social Sciences Communications*, *11*(1), 1-12.

Telaumbanua, Y. A., Marpaung, A., Gulo, C. P. D., Waruwu, D. K. W., Zalukhu, E., & Zai, N. P. (2024). Analysis of two translation applications: Why is DeepL Translate more accurate than Google Translate? *Journal of Artificial Intelligence and Engineering Applications*, *4*(1), 82-86.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004) Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193-202.

Wang, R., Mi, F., Chen, Y., Xue, B., Wang, H., Zhu, Q., ... & Xu, R. (2024, June). Role prompting guided domain adaptation with general capability preserve for large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2243-2255).

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT*. arXiv preprint arXiv:2302.11382.

Lei, L., Wei, Y., & Liu, K. (2024). AlphaReadabilityChinese: A tool for the measurement of readability in Chinese texts and its applications. *Foreign Language Teaching*, *46*(1), 83-93.

Rudolf Flesch (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221-233.

Li, J., & Gao, Y. (2025). Variability of cohesion and coherence in Chinese-to-English translation: measuring the effect of translation variety and register divergence. *Humanities and Social Sciences Communications*, *12*(1), 1-18.

Ma, Y. R., Ye, Y., & Xie, H. (2026). *CLASE: A hybrid method for Chinese legalese stylistic evaluation*. arXiv preprint arXiv:2602.12639.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Vermeer, H. J. (1978). *Ein Rahmen für eine allgemeine Translationstheorie.*

Nord, C. (1997). *Translating as a purposeful activity: Functionalist approaches explained. Routledge.*

Sun, X., & Chen, X. (2024). Review of legal terminology translation research. *International Journal of Education and Humanities*, *17*, 343-348.

Prieto Ramos, F. (2021). Translating legal terminology and phraseology: Between inter-systemic

incongruity and multilingual harmonization. *Perspectives*, *29*(2), 175-183.

Hatim, B., & Mason, I. (2014). *Discourse and the Translator*. Routledge.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 4171-4186). Association for Computational Linguistics.

Halliday, M. A. K., & Hasan, R. (2014). *Cohesion in English*. Routledge.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.