# A Quantitative Comparison of Large Language Models and Commercial Services for the Translation of Chinese Legal Texts

Fei Qu

School of Foreign Languages, Southwest University of Political Science and Law, Chongqing, China

**Abstract**—The proliferation of Large Language Models (LLMs) presents transformative potential for professional domains, yet their application in the high-stakes field of legal translation requires rigorous empirical validation. This study conducts a quantitative comparison of the translation quality between two leading LLMs (Gemini 2.5 Pro, ChatGPT 4o) and two reputable commercial translation (CT) services (PKU Law, Wolters Kluwer). The evaluation uses the English translations of the General Provisions of the Criminal Law of the People's Republic of China, with quality assessed through the automated metrics of Bilingual Evaluation Understudy (BLEU) and Translation Edit Rate (TER). Statistical analysis of the four individual sources revealed significant performance differences, with Gemini demonstrating a superior output compared to ChatGPT and, on some measures, PKU Law. However, a subsequent comparison between the aggregated LLM and CT groups found no statistically significant difference in translation quality for either BLEU or TER scores. This study posits that this apparent parity is a methodological illusion that stems from the profound limitations of lexical-based metrics. These metrics reward the superficial fluency of LLMs but are incapable of assessing functional equivalence, thereby failing to penalize critical semantic and legal errors. The findings conclude that despite the impressive coherence of LLM outputs, the nuanced, jurisdiction-specific expertise of human professionals remains the indispensable arbiter of quality and validity in legal translation.

**Index Terms**—Large Language Models, ChatGPT, BLEU, TER, legal translation

## I. INTRODUCTION

The contemporary technological landscape is defined by the rapid development and pervasive integration of Large Language Models (LLMs) across a multitude of professional sectors [1]. These models, representing a significant evolution from earlier paradigms such as Neural Machine Translation (NMT), possess sophisticated capabilities for processing, generating, and interpreting human language. Their advanced architectures enable them to tackle complex informational tasks, driving innovation and transforming workflows in fields as diverse as medicine, finance, and, increasingly, law. The potential of LLMs to automate document drafting, assist in legal research, and provide translation services has generated considerable interest within the legal community.

Within the broader field of language services, legal translation constitutes a uniquely demanding and high-stakes domain. Unlike general-purpose translation, legal translation requires not only linguistic fluency but also profound domain-specific knowledge. The fidelity of a legal translation is paramount, as it must maintain absolute terminological precision, correctly interpret concepts specific to the source and target legal jurisdictions, and preserve the precise legal intent and nuances of the original text. The consequences of error are severe; a single mistranslated term or misinterpreted clause can lead to contractual disputes, regulatory non-compliance, and the invalidation of legal documents in court. Research has documented systematic errors that Al systems make in legal contexts, which underscore these risks. For instance, machine translation models have been observed to mistranslate the legal term "warrant" as the more generic "court order, " a substitution that significantly downplays the legal severity of the document. Similarly, contextual misunderstandings can lead to absurd yet dangerous outputs, such as translating "charged with a battery" as "loaded with a case of batteries". Such errors highlight a critical gap between the general linguistic competence of Al and the specialized precision required for legal practice.

While the fluency and general capabilities of modern LLMs are impressive, their efficacy and reliability for specialized legal translation remain insufficiently quantified. There is a pressing need for empirical evidence that compares the output of these new generative models against the established, human-curated translations provided by professional services, which have long been the standard in the legal industry. This study addresses this gap by conducting a rigorous quantitative analysis of translation quality. The primary objective of this paper is to compare the translation quality of two state-of-the-art LLMs (Gemini 2.5 Pro and ChatGPT 4o) with two reputable commercial legal translation databases (PKU Law and Wolters Kluwer). The evaluation is performed on their respective English translations of the General Provisions of the Criminal Law of the People's Republic of China, using the widely accepted automated metrics BLEU and TER. This objective is operationalized through the following research questions (RQs):

RQ1: Are there statistically significant differences in the translation quality, as measured by BLEU and TER scores, among the four translation sources (Gemini, ChatGPT, PKU Law, Wolters Kluwer) when translating the General Provisions of the PRC Criminal Law?

RQ2: Is there a statistically significant difference in the aggregate translation quality, as measured by BLEU and TER scores, between the Large Language Model (LLM) group and the Commercial Translation (CT) group?

## II. Literature Review

### *A. The Evolution of Automated Translation: From NMT to Large Language Models*

The field of automated translation has undergone a profound transformation over the past decade, moving from the paradigm of Neural Machine Translation (NMT) to the current era dominated by Large Language Models (LLMs). NMT, which utilizes deep neural networks to process entire sentences, marked a significant advance over previous statistical methods, substantially improving the fluency and accuracy of machine-generated text. These systems, however, were fundamentally designed as specialized engines optimized for the singular task of translation, trained on curated parallel corpora of source and target sentence pairs [1].

The emergence of LLMs represents not an incremental improvement but a fundamental paradigm shift in artificial intelligence and its application to language [2]. Unlike NMT systems, LLMs are general-purpose models trained on exceptionally vast and diverse text corpora, which endows them with a more comprehensive, context-aware understanding of language and a much broader range of capabilities. This generalist training allows LLMs to adapt to new tasks with minimal or no task-specific data, a technique known as zero-shot or few-shot learning. This versatility has enabled their rapid application across numerous specialized domains, including healthcare and, increasingly, law [3]. In the context of translation, this generalized capability manifests as superior fluency and an enhanced ability to handle context across longer documents, with some models designed to mimic complex human-like processes such as analyzing a source text before rendering a translation.

This technological evolution has introduced a critical tradeoff between fluency and accuracy. The training objective of an LLM, which involves predicting the next word in a sequence based on massive general-domain data, inherently optimizes for linguistic coherence and naturalness. This results in outputs that are exceptionally fluent and conversational. However, this same process makes them susceptible to generating plausible but factually incorrect information, a phenomenon known as hallucination, because their primary goal is linguistic plausibility rather than strict fidelity to a source text. Conversely, NMT systems are trained specifically on parallel texts to optimize for accurate source-to-target mapping, which often makes them more reliable for direct translation accuracy within their trained domains, though their output may be more rigid and less natural-sounding. This inherent tension between the generalist fluency of LLMs and the specialist accuracy of NMT creates a central challenge for evaluation, particularly in high-stakes fields where precision is paramount.

### *B. The Exigencies of Legal Translation and the Principle of Functional Equivalence*

Legal translation constitutes a uniquely demanding and high-stakes domain that magnifies the aforementioned challenges. The task requires not only bilingual fluency but also profound domain-specific knowledge, including an understanding of disparate legal systems and their distinct terminologies [4]. The consequences of error are severe; a mistranslated term or a misinterpreted clause can invalidate contracts, create regulatory liabilities, and subvert judicial outcomes. Therefore, the core of legal translation is the pursuit of absolute precision and consistency in the use of specialized terminology [5,6].

Given these exigencies, the theoretical benchmark for quality in legal translation has long since moved beyond simplistic notions of literalism. Functional equivalence posits that the primary objective of a translator is not to achieve formal, word-for-word correspondence but to produce a target text that has an equivalent effect on its audience. In the legal sphere, this translates to producing a text that has an equivalent legal effect within the target jurisdiction. Achieving this often requires the translator to employ sophisticated strategies such as adaptation, explanation, and the substitution of functionally analogous terms, particularly when a direct conceptual counterpart is absent in the target legal system. This process is an act of cross-jurisdictional communication that demands a deep understanding of the cultural and systemic underpinnings of law in both the source and target contexts. This principle is therefore fundamentally at odds with any evaluation methodology that relies on simple lexical or structural similarity [7,5,6].

### *C. Evaluating Machine Translation: A Critical Review of Automated Metrics*

The practical need for rapid, scalable, and inexpensive evaluation methods has led to the widespread adoption of automated metrics in machine translation research. Human evaluation, while considered the gold standard, is a slow, costly, and subjective process, making it unsuitable for the iterative development cycle of modern MT systems. Metrics such as the Bilingual Evaluation Understudy (BLEU) and Translation Edit Rate (TER) were developed to serve as automated proxies for human judgment [8,9,10].

The BLEU metric is founded on the principle that "the closer a machine translation is to a professional human translation, the better it is". It operates by calculating the precision of matching n-grams (contiguous word sequences) between a candidate translation and one or more human-authored reference translations, applying a brevity penalty to penalize outputs that are too short [9]. TER, conversely, is designed to approximate the post-editing effort required by a human. It calculates the

minimum number of edits, insertions, deletions, substitutions, and shifts, needed to transform a candidate translation to match a reference translation exactly [11].

Despite their widespread use, these lexical-based metrics have been the subject of extensive and long-standing scholarly criticism. The most fundamental critique is that they do not measure translation quality but rather surface-level string similarity to a given reference, a crucial distinction that is often overlooked. Their primary flaw is an inability to account for lexical variation; they cannot recognize synonyms or valid paraphrases, meaning a perfectly accurate translation that uses different wording from the reference is unfairly penalized. This can lead to paradoxical results where nonsensical sentences that happen to contain correct n-grams receive high scores. Consequently, numerous studies have demonstrated a poor correlation between these metrics and human judgments of quality, leading expert bodies to describe them as "artificial and irrelevant for production environments" for over a decade.

This critique becomes particularly salient when applying these metrics to legal translation, creating a profound mismatch between the theoretical requirements of the task and the operational mechanism of the evaluation tool. The theoretical goal of legal translation is to preserve the legal function of a text, which may necessitate altering its linguistic form. The methodological tools, BLEU and TER, operate by rewarding the preservation of linguistic form (lexical and structural matching) and are incapable of assessing legal function. This creates a situation where the more skillfully a system achieves functional equivalence through non-literal but legally correct phrasing, the more likely it is to be penalized. This fundamental conflict highlights why these metrics are theoretically inappropriate for this specific task. Recognizing these deficiencies, the research community has actively developed and transitioned toward semantically-aware metrics based on contextual embeddings, such as BERTScore and COMET, which demonstrate a significantly higher correlation with human judgments by capturing semantic similarity rather than mere lexical overlap [12,8].

### *D. Identifying the Research Gap: Quantifying LLM Performance in Legal Translation*

The existing body of literature reveals a clear and pressing research gap. While LLMs exhibit transformative potential for language tasks, their application to the high-stakes, nuanced domain of legal translation remains a nascent area of inquiry, fraught with documented challenges related to terminological consistency and accuracy. Scholarly consensus indicates a scarcity of empirical studies that rigorously evaluate the performance of state-of-the-art LLMs in this domain, particularly in direct comparison to the established commercial translation services that are the de facto standard in professional legal practice [2].

This study is positioned as a foundational contribution that directly addresses this gap. It provides a crucial, initial quantitative benchmark comparing the outputs of leading LLMs against reputable commercial services for a significant and complex legal text. The selection of BLEU and TER as evaluation metrics is a deliberate methodological choice. While the profound limitations of these metrics are well-established and acknowledged, their use serves a dual purpose in this context. First, it furnishes a baseline measurement using widely understood, albeit dated, metrics, making the results interpretable within the broader history of MT evaluation. Second, and more importantly, it offers a powerful empirical case study that illustrates the "evaluation crisis" and the methodological-theoretical mismatch inherent in applying lexical metrics to a functionally-driven task like legal translation. By quantifying the performance of these distinct system types, fluency-optimized LLMs and professionally curated commercial translations, this research provides the necessary groundwork to motivate and inform future investigations that must adopt more sophisticated, semantically-aware evaluation paradigms to assess generative models in specialized domains.

## III. METHODOLOGY

### *A. Corpus and Translation Sources*

The source text for this empirical investigation consists of the General Provisions (Articles 1 through 101) of the Criminal Law of the People's Republic of China. The General Provisions are highly representative for evaluating legal translation quality. They articulate the foundational doctrines of Chinese criminal law, such as definitions of crime, culpability, punishment, and legal defenses, which apply system-wide. Linguistically, this section exemplifies the core features of PRC legislative style, normative modality, complex parataxis, dense definitions, and exception structures, making it a rigorous testbed for assessing semantic precision and syntactic fidelity. Moreover, it foregrounds key translation challenges such as conceptual calibration, liability structure, and penalty conditions. As a result, it offers both doctrinal depth and linguistic generalizability, serving as a valid and methodologically sound corpus for comparative translation studies. Four distinct sources were used to generate English translations of this corpus. These sources were categorized into two groups for comparative analysis as follows:

Large Language Models (LLMs): This group includes two of the most advanced, Gemini 2.5 Pro developed by Google and ChatGPT 4o developed by OpenAI, publicly available generative AI models at the time of the study.

Commercial Translations (CT): This group includes English translations sourced from two highly reputable commercial databases known for providing professional translations of Chinese legal materials. PKU Law is a leading legal information database in China. Wolters Kluwer is a global

Table 1 Descriptive statistics of four translation sources under BLEU (unit: %)

| Translation sources | Min. | 25% Percentile | Median | 75% Percentile | Max. | Range | Mean | Std. Deviation |
|---|---|---|---|---|---|---|---|---|
| Gemini | 5.538 | 24.99 | 31.21 | 43.99 | 75.09 | 69.55 | 35.03 | 14.74 |
| ChatGPT | 6.058 | 18.64 | 27.62 | 37.25 | 60.32 | 54.26 | 28.37 | 13.00 |
| PKU Law | 4.309 | 20.75 | 27.79 | 39.43 | 88.53 | 84.22 | 31.07 | 15.93 |
| Wolters Kluwer | 5.581 | 21.49 | 30.17 | 43.66 | 66.60 | 61.02 | 32.24 | 14.33 |

Table 2 Normality test under BLEU

| Translation sources | Kolmogorov-Smirnov | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Test statistic | Degrees of freedom | P Value | Test statistic | Degrees of freedom | P Value |
| Gemini | .115 | 101 | .002 | .965 | 101 | .010 |
| ChatGPT | .068 | 101 | >.010 | .976 | 101 | .063 |
| PKU Law | .100 | 101 | .015 | .958 | 101 | .003 |
| Wolters Kluwer | .071 | 101 | >.010 | .980 | 101 | .120 |

Table 4 Multiple comparisons of Translation sources under BLEU

| Dunn's multiple comparisons test | Mean rank diff. | Significant? | Summary | Adjusted P Value |
|---|---|---|---|---|
| Gemini vs. ChatGPT | 50.47 | Yes | * | .013 |
| Gemini vs. PKU Law | 34.21 | No | ns | .224 |
| Gemini vs. Wolters Kluwer | 20.27 | No | ns | >.999 |
| ChatGPT vs. PKU Law | -16.26 | No | ns | >.999 |
| ChatGPT vs. Wolters Kluwer | -30.20 | No | ns | .396 |
| PKU Law vs. Wolters Kluwer | -13.95 | No | ns | >.999 |

provider of professional information, software solutions, and services for the legal and regulatory sectors.

*B. Evaluation protocol*

The evaluation protocol involved generating or collecting the English translations of all 101 articles from each of the four sources. These 404 translations (101 articles × 4 sources) were then systematically evaluated against the reference translation. Automated scripts were used to compute sentence-level BLEU and TER scores for each article from each source, resulting in a dataset of 101 data points per metric for each of the four translation systems.

*C. Procedure*

A two-stage statistical analysis was conducted using a significance level (alpha) of $p<.05$ for all inferential tests. All statistical analyses were performed on the provided dataset.

The first stage aimed to answer RQ1 by comparing the performance of the four individual translation sources. The assumption of normality for the score distributions of each source was assessed using the Shapiro-Wilk test, as it is generally more powerful for smaller sample sizes. The Shapiro-Wilk test indicated that the BLEU score data for the Gemini and PKU Law groups significantly deviated from a normal distribution. Consequently, the non-parametric Kruskal-Wallis H test was selected as the appropriate method to compare the median BLEU scores across the four groups. Dunn's test with Bonferroni correction for multiple comparisons was chosen as the post-hoc test to identify which specific pairs of groups differed significantly. The Shapiro-Wilk test indicated that the TER score data for all four groups did not significantly violate the assumption of normality ($p>.05$ for all). Therefore, a one-

way Analysis of Variance (ANOVA) was employed to compare the mean TER scores. Tukey's Honestly Significant Difference (HSD) test was selected for post-hoc pairwise comparisons to determine where the significant differences lay.

The second stage aimed to answer RQ2 by comparing the performance of the LLM group against the CT group. The data from the individual sources were aggregated into the

Table 5 Descriptive statistics of four translation sources under TER (unit: %)

| Translation sources | Min. | 25% Percentile | Median | 75% Percentile | Max. | Range | Mean | Std. Deviation |
|---|---|---|---|---|---|---|---|---|
| Gemini | 11.11 | 37.25 | 49.21 | 55.56 | 81.25 | 70.14 | 46.55 | 14.49 |
| ChatGPT | 14.29 | 47.50 | 55.81 | 62.79 | 85.00 | 70.71 | 54.28 | 13.81 |
| PKU Law | 4.309 | 20.75 | 27.79 | 39.43 | 88.53 | 84.22 | 31.07 | 15.93 |
| Wolters Kluwer | 6.667 | 40.37 | 52.86 | 60.19 | 78.31 | 67.20 | 50.81 | 13.85 |

Table 6 Normality test under TER

| Translation sources | Kolmogorov-Smirnov | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Test statistic | Degrees of freedom | P Value | Test statistic | Degrees of freedom | P Value |
| Gemini | .096 | 101 | .022 | .980 | 101 | .126 |
| ChatGPT | .077 | 101 | >.010 | .982 | 101 | .197 |
| PKU Law | .051 | 101 | >.010 | .987 | 101 | .431 |
| Wolters Kluwer | .072 | 101 | >.010 | .986 | 101 | .362 |

respective groups (LLM: Gemini + ChatGPT, n = 202; CT: PKU Law + Wolters Kluwer, n = 202). The Shapiro-Wilk test was again used to assess the normality of these aggregated distributions. For the BLEU score comparison, the Shapiro-Wilk test revealed that both the LLM group (p=.001) and the CT group violated the assumption of normality. For the TER score comparison, the LLM group also violated normality. Due to these violations, the non-parametric Mann-Whitney U test was selected as the appropriate method to compare the median scores between the LLM and CT groups for both the BLEU and TER metrics.

## IV. RESULTS

We conduct the full suite of statistical tests on the BLEU scores, after which the identical testing protocol is applied to the TER scores. Table 1 shows descriptive statistics of four translation tools under BLEU.r

A non-parametric approach was adopted for the subsequent analysis, as the BLEU score data did not meet the assumption of normality ($p<0.05$), as detailed in Table 2. Consequently, a Kruskal-Wallis H test was performed to evaluate for statistically significant differences in the BLEU scores among the four translation tools. The results, presented in Table 3, indicate a significant difference. The p-value was below the 0.05 significance level, leading to the conclusion that a statistically significant variance exists in the translation quality, as quantified by the BLEU metric, across the four translation sources.

To identify the specific sources of the observed variance, we conducted post-hoc pairwise comparisons. We applied Dunn's correction to the significance values to control for the increased risk of a Type I error from conducting multiple tests. The adjusted results, as shown in Table 4, reveal a statistically significant difference only in the comparison between Gemini and ChatGPT. The remaining pairwise comparisons did not achieve statistical significance.

Table 3 Kruskal-Wallis test

| P value | 0.017 |
|---|---|
| Exact or approximate P value? | Approximate |
| P value summary | * |
| Do the medians vary signif. | Yes |
| Kruskal-Wallis statistic | 10.18 |

Now that the statistical analysis of the BLEU scores is complete, we will perform the same set of tests on the TER scores. Table 5 provides the descriptive statistics for the four translation sources' TER scores.

We began the analysis of the TER scores with a normality test. The analysis of the TER scores commenced with an evaluation of data normality. As Table 6 shows, the Shapiro-Wilk test confirmed that the TER score data for all four sources adhered to the assumption of normality (p>.050 for all).

Because the data were normally distributed, a one-way Analysis of Variance (ANOVA) was the appropriate parametric test to compare the mean TER scores across the four sources.

Table 7 ANOVA results

| ANOVA summary | |
|---|---|
| F | 5.699 |
| P Value | .001 |
| P value summary | *** |

| | |
|---|---|
| Significant diff. among means (P < 0.05)? | Yes |
| R squared | .041 |

The ANOVA test yielded a statistically significant result (F=5.699, p=.001), which is presented in Table 7. This outcome demonstrates that a significant variance exists among the mean TER scores of the four translation sources. The ANOVA result confirms an overall difference, so a post-hoc test was necessary for the identification of specific pairwise differences between the sources. For this purpose, we utilized Tukey's Honestly Significant Difference (HSD) test.

Table 8 Multiple comparisons of Translation sources under TER

| Tukey's multiple comparisons test | Mean Diff. | Below threshold? | Summary | Adjusted P Value |
|---|---|---|---|---|
| Gemini vs. ChatGPT | -7.731 | Yes | *** | .001 |
| Gemini vs. PKU Law | -6.725 | Yes | ** | .006 |
| Gemini vs. Wolters Kluwer | -4.263 | No | ns | .158 |
| ChatGPT vs. PKU Law | 1.006 | No | ns | .961 |
| ChatGPT vs. Wolters Kluwer | 3.468 | No | ns | .325 |
| PKU Law vs. Wolters Kluwer | 2.462 | No | ns | .623 |

The results of the Tukey HSD post-hoc analysis are detailed in Table 8. The test identified statistically significant differences in two specific comparisons. The mean TER score for Gemini was significantly lower than that of ChatGPT

Table 9 Normality test between LLMs and CT under BLEU

| Translation sources | Kolmogorov-Smirnov | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Test statistic | Degrees of freedom | P Value | Test statistic | Degrees of freedom | P Value |
| LLMs | .071 | 202 | .015 | .976 | 202 | .001 |
| CT | 0.073 | 202 | .010 | .976 | 202 | .001 |

Table 10 Normality test between LLMs and CT under TER

| Translation sources | Kolmogorov-Smirnov | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Test statistic | Degrees of freedom | P Value | Test statistic | Degrees of freedom | P Value |
| LLMs | .076 | 202 | .006 | .984 | 202 | .019 |
| CT | 0.051 | 202 | >.01 | .990 | 202 | .195 |

Table 11 Mann-Whitney test result between LLMs and CT under BLEU

| P value | 0.8623 |
|---|---|
| Exact or approximate P value? | Approximate |
| P value summary | ns |
| Significantly different (P < 0.05)? | No |
| One- or two-tailed P value? | Two-tailed |
| Sum of ranks between LLMs and CT | 41109, 40701 |
| Mann-Whitney U | 20198 |

Table 12 Mann-Whitney test result between LLMs and CT under TER

| P value | 0.3033 |
|---|---|
| Exact or approximate P value? | Approximate |
| P value summary | ns |
| Significantly different (P < 0.05)? | No |
| One- or two-tailed P value? | Two-tailed |
| Sum of ranks between LLMs and CT | 39697,42114 |
| Mann-Whitney U | 19194 |

(p=.001) and also significantly lower than that of PKU Law (p=.006). The remaining pairwise comparisons between Gemini and Wolters Kluwer, ChatGPT and PKU Law, ChatGPT and Wolters Kluwer, and PKU Law and Wolters Kluwer did not produce statistically significant differences.

The analysis subsequently progressed to the second stage. This stage addressed the RQ2 and involved a comparison of the aggregated LLM group and the CT group. The evaluation commenced with the aggregated BLEU scores. A normality test was conducted on these two new groups. As Table 9 shows, the Shapiro-Wilk test indicated that both the LLM group (p=.001) and the CT group (p=.001) significantly deviated from a normal distribution. A parallel analytical procedure was then applied to the TER scores. The Shapiro-Wilk test for normality, presented in Table 10, revealed that the LLM group's data were not normally distributed (p=.019), while the CT group's data did not violate the assumption (p=.195).

Because the data violated the assumption of normality, the non-parametric Mann-Whitney U test was the correct statistical method for the comparison of the two independent groups. The results of this test are located in Table 11. The test yielded a p-value of 0.8623, a value that is substantially greater than the 0.05 significance threshold. Therefore, the analysis concludes that no statistically significant difference exists between the median BLEU scores of the LLM group and the CT group.

A non-parametric test is required when the assumption of normality is not met in at least one of the groups, so the Mann-Whitney U test was again employed. Table 12 displays the outcome of this comparison. The resulting p-value was 0.3033, which indicates the difference between the groups are not statistically significant. Consequently, there is no statistically significant difference between the median TER scores of the LLM and CT groups.

## V. DISCUSSION

This section provides a comprehensive interpretation of the statistical findings from the preceding analysis. It situates these results within the broader scholarly discourse on machine translation and legal language, critically appraises the methodological framework of the study, and delineates the implications of the findings for theory, practice, and pedagogy. Finally, it proposes a structured agenda for future research to address the limitations identified and to advance the field..

### *A. Qualitative Analysis Based on BLEU and TE*

A qualitative examination of translations that received low BLEU and high TER scores offers concrete evidence of the failures that lexical metrics can detect, even if they cannot capture the full nuance of legal meaning. An analysis confirms that these low scores are not arbitrary penalties for stylistic variation but are markers of substantive semantic and terminological failures.

ChatGPT produced a translation of Article 94 with a BLEU score of 6.05. The source text defines "司法工作人员"(sīfǎ gōngzuò rényuán), which translates to "judicial staff" or "judicial officers," by listing their specific functions: "侦查、检察、审判、监管" (zhēnchá, jiǎnchá, shěnpàn, jiānguǎn), meaning investigation, prosecution, adjudication, and supervision. ChatGPT's translation rendered this as "State functionary," which is a mistranslation. The term "State functionary" is a much broader category in Chinese law and fails to capture the specific functional roles that define a "judicial officer." Furthermore, the translation entirely omitted the enumerated duties, which are the core legal substance of the article. This omission constitutes a critical failure to achieve functional equivalence, as the definition is rendered legally meaningless without them.

PKU Law's translation of Article 3, which scored 4.3, demonstrates a failure in syntactic and semantic fidelity. The source text establishes the principle of legality (nullum crimen, nulla poena, sine lege) [13], a foundational doctrine in criminal law. The reference translation captures the parallel structure and deontic modality ("shall be subjected to... shall not be subjected to") which correctly conveys the mandatory nature of this legal principle. PKU Law's version, "is to be convicted... is not to be convicted," weakens this legal force by using a less definitive grammatical structure. Moreover, its phrasing is syntactically convoluted and less clear than the reference, which impairs the reader's ability to grasp the precise legal rule being articulated.

Similarly, Wolters Kluwer's translation of the same article, which scored 5.58, also fails to convey the correct legal meaning. The translation begins "For acts that are explicitly defined as criminal acts in law," which is a grammatically awkward and imprecise rendering of the original. More significantly, it introduces the term "offenders," which is not present in the source text and presupposes guilt. The Chinese text speaks of "acts"("行为"- xíngwéi), not the individuals committing them. This subtle shift alters the legal focus from the act itself to the actor, which represents a misinterpretation of the legal principle being established. This choice of terminology results in a translation that is not functionally equivalent to the source.

Gemini's translation of Article 74, which discusses the inapplicability of suspended sentences to specific categories of offenders, received a BLEU score of 5.53. The key legal terms in this article are "累犯" (lěifàn), meaning "recidivists," and "缓刑" (huǎnxíng), which translates to "probation" or "suspended sentence." While Gemini correctly translated "recidivists" and "ringleaders of criminal groups," its choice of "suspension of sentence" over "probation" created a lexical divergence from the reference text. Although "suspension of sentence" is a valid translation of "缓刑," the reference translation preferred "probation." This example highlights how even a seemingly minor lexical choice can contribute to a lower BLEU score, even when the translation may be legally acceptable. However, the accumulation of such minor deviations across a text can lead to a significant penalty under a strict lexical matching system.

A qualitative analysis of translations that received high TER scores provides further insight into the specific deficiencies of each system. The TER metric quantifies the post-editing effort

that a human would require, so a higher score indicates a greater number of necessary edits and, consequently, a lower-quality initial translation. An analysis of high-TER examples reveals that the required edits are not merely stylistic; they consistently involve substantive corrections to legal terminology and sentence structure that are essential for achieving functional equivalence.

ChatGPT's translation of Article 99 received a TER score of 85. The source text clarifies that numerical ranges in the law are inclusive. ChatGPT provided a literal translation of "以上" (yǐ shàng) and "以下" (yǐxià) as "above" and "below." These terms are not the standard phrasings that are used in English-language statutes to define inclusive numerical limits. The reference translation uses the correct legal functional equivalents, which are "not more than" and "not less than." The high TER score accurately reflects the significant post-editing effort that is needed. A human editor must replace the literal but functionally incorrect terms with the appropriate legal terminology to ensure the text has the correct legal effect.

The translation of Article 88 from PKU Law, which scored an exceptionally high 92.77, required extensive revision. The translation used the awkward and non-standard term "criminal element" instead of the more precise term "offender." Its sentence structure was convoluted and employed a weak passive voice, for example, "No limitation... is to be imposed." This phrasing fails to convey the direct and binding nature of the legal rule. The reference translation uses the active and definitive statement "the limitation period is not binding." The sheer number of edits that are necessary to correct the terminology, simplify the syntax, and restore the proper legal force justifies the extremely high TER score.

Wolters Kluwer's version of the same article, with a TER score of 78.31, was more competent but still flawed. It used the word "criminal" where "offender" or "suspect" would be more appropriate before a conviction. It also employed the term "dockets the case" as a translation for "立案" (lì'àn). While "docket" is a plausible choice, the reference translation's phrase "filed for investigation" is also a common and clear equivalent. The sentence structure, particularly in the second half of the article, was complex and less direct than the reference. The high TER score reflects the need for these terminological and syntactic adjustments so that the translation aligns with standard legal phraseology.

Gemini's translation of Article 74 received a TER score of 81.25, and this case highlights a specific characteristic of the metric. The translation used "a suspension of sentence" for the Chinese term "缓刑" (huǎnxíng). The reference translation selected "probation." Both terms are conceptually related and can be considered valid translations in different contexts. However, they are not perfect synonyms. The high TER score resulted almost entirely from this single major terminological substitution, along with minor grammatical shifts. This example demonstrates that TER heavily penalizes a translation when it deviates from the specific lexical choices of the single reference text, even if the chosen alternative is functionally similar.

### *B. Interpretation of Principal Findings: LLM Fluency and the Illusion of Statistical Parity*

The statistical analysis yielded a bifurcated set of results that demand careful interpretation. The RQ1 investigated the performance of the four individual translation sources. The findings revealed statistically significant differences among them, with Gemini demonstrating a notable performance advantage over ChatGPT, reflected in both a higher median BLEU score (p=.013) and a significantly lower mean TER score (p=.001). Furthermore, Gemini's output required significantly fewer edits than that of PKU Law, as indicated by its lower mean TER score (p=.006).This suggests that on a structural level, Gemini's translations were closer to the reference text than those from a reputable commercial database, a testament to the rapid, iterative advancements in the architecture of leading LLMs.

However, the analysis for RQ2, which compared the aggregated LLM group against the CT group, produced a starkly different outcome. The results of the Mann-Whitney U tests indicated no statistically significant difference between the two groups for either the BLEU metric (p=.8623) or the TER metric (p=.3033). A superficial reading of this finding might suggest that the translation quality of state-of-the-art LLMs has achieved parity with established, professional legal translation services. This discussion posits that such a conclusion is not only premature but is likely an artifact of the evaluation methodology itself. The observed statistical parity is, in effect, a methodological illusion that reveals more about the inherent limitations of the chosen metrics than it does about the true comparative quality of the translations.

This illusion of parity can be deconstructed by examining the interplay between the known strengths and weaknesses of both the LLMs and the evaluation metrics. LLMs are engineered to generate text that is exceptionally fluent and grammatically coherent, a capability derived from their training on vast text corpora. This inherent fluency naturally results in a high degree of n-gram overlap with any well-formed reference text, thereby inflating BLEU scores. Conversely, these same models are known to struggle with domain-specific knowledge and contextual nuance, leading to systematic errors in specialized fields such as law. These errors, which can be semantically and legally catastrophic, are often not adequately penalized by lexical metrics. For instance, an LLM might produce a highly fluent sentence that contains a critical terminological error, while a professional translation might use different phrasing, functionally correct but lexically divergent from the single reference, thereby receiving a comparatively lower BLEU score. Across a large corpus, the LLMs' high scores for fluency can effectively mask their low scores for accuracy, while the professional translations' perfect accuracy may be penalized for lexical divergence. The net effect is the potential cancellation of these differences, leading to the non-significant result observed in the group comparison. Therefore, the finding of no difference does not signify equivalent quality; rather, it highlights a fundamental misalignment between the evaluation tools and the complex nature of legal translation.

### *C. The Inadequacy of Lexical Metrics for Assessing Functional Equivalence in Legal Translation*

The central challenge in evaluating legal translation quality lies in defining an appropriate theoretical standard. The field has long moved past simplistic notions of literalism, converging instead on the principle of "functional equivalence" as the benchmark for high-quality translation. Advanced by theorists, this principle holds that the primary objective of a legal translator is not to achieve formal, word-for-word correspondence, but to produce a target text that has an equivalent legal effect within the target jurisdiction. This is a process of cross-jurisdictional communication that involves translating legal concepts and their intended consequences, an act that often requires adaptation, explanation, and the use of functionally analogous terms rather than literal equivalents.

This study's findings bring the profound conflict between this theoretical standard and the chosen evaluation metrics into sharp relief. The BLEU metric, by its very design, measures n-gram precision and is thus a proxy for formal, not functional, equivalence. It is fundamentally incapable of recognizing legitimate synonyms or valid paraphrasing, which are indispensable strategies for a translator attempting to achieve functional equivalence when a direct terminological counterpart is absent in the target legal system. Similarly, the TER metric measures the post-editing effort required to make a translation match a reference. However, it does so by counting edits without differentiating their severity; a single-word substitution that corrects a fundamental legal misinterpretation is weighted no more heavily than a trivial edit of punctuation or style. Consequently, TER quantifies the effort of correction but fails to capture the significance of the errors being corrected.

This misalignment creates what can be termed a fluency trap. Because modern LLMs excel at producing fluent text, they generate outputs that appear to be of high quality when assessed by metrics that prioritize fluency and lexical similarity. This masks deep-seated failures in achieving the functional equivalence that is the cornerstone of legal validity. This phenomenon poses a significant risk, as non-expert users or automated evaluation pipelines could erroneously conclude that raw LLM output is a viable substitute for professional legal translation, a conclusion that this study's results, when properly contextualized, strongly refute.

### *D. Implications for Legal Practice, AI Development, and Translation Pedagogy*

This study carries substantial implications for key stakeholder groups. The nuanced interpretation of statistical parity between LLMs and commercial services offers valuable insights for legal professionals, AI developers, and translation educators alike.

For legal professionals, the findings offer a critical caution: the appearance of parity in automated metrics can be misleading. LLMs such as Gemini, while fluent, remain prone to domain-specific semantic errors that pose risks in high-stakes contexts. This underscores the indispensability of human oversight. The most responsible application of LLMs in legal translation lies within a Machine Translation Post-Editing (MTPE) workflow, where AI-generated drafts are subject to expert review. Effective MTPE requires not only legal and linguistic competence, but also high-quality source texts, clear editorial standards, and well-maintained legal termbases to guide the post-editing process.

For AI developers and the MT research community, the study exposes the limitations of relying on traditional metrics like BLEU and TER, which privilege surface fluency over legal adequacy. Continued reliance on such metrics risks incentivizing models that perform well numerically but fail semantically. Progress in this field demands the development of domain-specific evaluation protocols, including high-quality multi-reference test sets for legal texts and the adoption of metrics better aligned with human judgments of legal accuracy. Moreover, the results suggest that substantial improvements in legal translation quality will likely come through domain-adaptive fine-tuning on curated legal corpora that equip models with the specialized knowledge they currently lack.

For translation pedagogy, the study signals a pressing need to update training programs. Legal translators of the future must possess critical AI literacy—understanding both the capabilities and the systemic limitations of LLMs. MTPE training should be a core component of the curriculum, emphasizing not only grammatical correction but the ability to detect and resolve subtle, high-stakes legal errors. Additionally, students must be equipped to critically assess automated evaluation metrics, recognize their limitations, and make informed decisions about tool adoption in professional contexts.

### *E. Methodological Limitation*

Beyond the theoretical shortcomings of the chosen metrics, this study faces several methodological limitations that constrain the interpretation and generalizability of its findings.

The most critical limitation lies in the exclusive reliance on automated evaluation metrics, particularly BLEU and TER, which have been widely criticized in recent scholarship. Studies have consistently shown that these metrics correlate poorly with human judgments, especially when comparing high-performing systems or evaluating legal translations at the sentence level. They fail to capture semantic adequacy, overemphasize surface-level lexical overlap, and rely heavily on a single reference translation. As a result, LLMs, which often produce lexically diverse yet semantically accurate output, may be systematically underrated. Furthermore, these metrics lack diagnostic power; they provide no insight into the nature of errors, whether lexical, syntactic, or terminological.

Another key constraint is the single-reference bottleneck. Evaluating translations against only one reference text penalizes legitimate variation, particularly in legal translation where multiple phrasings may convey equivalent legal meaning. This rigid comparison can obscure the true strengths of systems capable of producing functionally accurate but differently worded outputs.

The scope of the study also limits generalizability. The analysis was restricted to one document and to a single

language pair: Chinese-to-English. While this section offers important doctrinal and linguistic challenges, its terminological focus may not represent other legal domains such as contract, civil, or administrative law, which follow different conventions. Additionally, LLM performance varies across language pairs, particularly for low-resource languages, further limiting extrapolation.

Acknowledging these limitations is essential for interpreting the findings in context and for informing future research on evaluation strategies that better reflect the semantic and functional demands of legal translation.

### *F. Directions for Future Research*

The limitations identified in this study give rise to a clear and structured agenda for future research. To build upon these preliminary findings and to arrive at more robust and valid conclusions, subsequent investigations should proceed along four primary avenues.

First, the most critical next step is to conduct a parallel study that incorporates human-centric evaluation methodologies to triangulate the current findings. This would provide the "gold standard" assessment that is currently missing. Such a study could employ a multi-faceted approach, including Direct Assessment (DA), where bilingual legal experts rate the translations from all four sources on a continuous scale for adequacy and fluency. This should be complemented by a detailed error typology analysis, in which human annotators manually identify and categorize the errors made by each system (e.g., terminological, syntactic, semantic, omission). It is hypothesized that such an analysis would reveal significant differences in the types and severity of errors between the LLM and CT groups, thereby dismantling the illusion of parity produced by the lexical metrics.

Second, the existing corpus of translations should be re-evaluated using modern, semantically-aware automated metrics. Research has shown that metrics such as COMET and BERTScore, which are based on contextual embeddings from pretrained language models, correlate much more highly with human judgments of quality than BLEU or TER. A re-analysis using these advanced metrics would directly test the central hypothesis of this discussion: that superior evaluation tools will detect a statistically significant quality difference between the LLM and CT groups that the legacy metrics failed to capture. Such a finding would provide strong empirical support for the machine translation community's shift away from BLEU.

Third, future research should investigate the impact of domain-specific adaptation on LLM performance. The LLMs used in this study were general-purpose models. A powerful comparative study would involve fine-tuning a model like Gemini or ChatGPT on a large, high-quality corpus of parallel Chinese-English legal texts. A subsequent evaluation comparing the output of this fine-tuned model against the general-purpose model and the commercial services could quantify the performance gains attributable to domain specialization and determine whether such adaptation can genuinely close the quality gap with professional, human-led translation workflows.

Finally, the scope of the investigation must be broadened to enhance the generalizability of the findings. Future studies should include a wider variety of legal documents, such as contracts, judicial decisions, and patent applications, which present different linguistic and conceptual challenges. The analysis should also be extended to other language pairs, including low-resource languages, to assess the robustness of LLM performance across different linguistic contexts. An additional and highly practical avenue of research would be to measure the post-editing effort (in terms of time and cost) required to bring the output from each source to a publishable, commercially acceptable standard, which would provide a more pragmatic measure of each system's utility in a real-world professional setting.

To synthesize the core argument of this discussion and to guide future methodological choices, the following table compares the three primary classes of evaluation methodologies.

## VI. CONCLUSION

This study's quantitative comparison of LLM and commercial translation services for a specialized legal text yields a paradoxical result. The statistical findings, particularly the lack of a significant difference between the aggregated LLM and CT groups, do not demonstrate that generative AI has achieved parity with professional, human-centric translation. Instead, these results serve as a powerful illustration of the profound inadequacy of purely quantitative, lexical-based evaluation metrics for the high-stakes domain of legal translation. The investigation reveals that the tools commonly used to measure translation quality are fundamentally misaligned with the theoretical and practical requirements of the task, rewarding superficial fluency while remaining blind to critical errors in legal meaning and function.

The rapid advancement in the fluency of models like Gemini is undeniable and represents a significant technological achievement. However, this study concludes that the nuanced, context-aware, and jurisdiction-specific expertise that defines professional legal translation remains an exclusively human capability. For the foreseeable future, the human legal expert is not merely a participant in the quality assurance process but remains the indispensable and definitive arbiter of validity and quality in legal translation.

## REFERENCES


[1] S. M. Abdelhalim, A. A. Alsahil, and Z. A. Alsuhaibani, "Artificial intelligence tools and literary translation: A comparative investigation of ChatGPT and Google translate from novice and advanced EFL student translators' perspectives," Cogent Arts and Humanities, vol. 12, no. 1, pp. 1–20, 2025.

[2] M. Bajcic and D. Golenko, "Applying large language models in legal translation: The state-of-the-art," International Journal of LANGUAGE & LAW, pp. 171–196, 2024.

[3] X. Xuan et al., “TransLaw: Benchmarking large language models in multi-agent simulation of the collaborative translation,” arXiv, 2025. [Online]. Available: http://arxiv.org/abs/2507.00875v1.
[4] M. Bajcic and D. Golenko, “Applying large language models in legal translation: The state-of-the-art,” International Journal of LANGUAGE & LAW, pp. 171–196, 2024.
[5] C. Aliyev, “Functional equivalence in the translation of legal terms: Insights from Arabic and Azerbaijani,” vol. 1, no. 4, pp. 162–168, 2025.
[6] J. Huang and G. Wang, “A study on translation strategies of legal English texts from the perspective of functional equivalence theory,” Frontiers in Humanities and Social Sciences, vol. 4, no. 7, pp. 37–43, 2024.
[7] J. Jiang and Y. Zhi, “Translating euphemisms: Analyzing functional equivalence theory in context,” US-China Foreign Language, vol. 22, no. 5, pp. 270–275, 2024.
[8] S. Lee et al., “A survey on evaluation metrics for machine translation,” Mathematics, vol. 11, no. 4, pp. 1–22, 2023.
[9] K. Papineni et al., “Bleu: A method for automatic evaluation of machine translation,” in Proc. 40th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, 2001, pp. 1–9.
[10] C. Wu et al., “Statistical machine translation for biomedical text: Are we there yet?,” in AMIA ... Annual Symposium proceedings, 2011, pp. 1290–1299.
[11] M. Snover et al., “A study of translation edit rate with targeted human annotation,” in Proc. 7th Conference of the Association for Machine Translation in the Americas, 2006, pp. 223–231.
[12] T. Zhang et al., “BERTScore: Evaluating text generation with BERT,” in International Conference on Learning Representations, 2020, pp. 1–43. [Online]. Available: http://arxiv.org/abs/1904.09675v3.
[13] B. Xue, The compact English-Chinese dictionary of Anglo-American law. Beijing, China: Peking University Press, 2013.