TransLaw: A Large-Scale Dataset and Multi-Agent Benchmark Simulating Professional Translation of Hong Kong Case Law

Anonymous ACL submission

Abstract

Hong Kong case law translation presents significant challenges: manual methods suffer from high costs and inconsistent quality, while both traditional machine translation and approaches relying solely on Large Language Models (LLMs) often fail to ensure legal terminology accuracy, culturally embedded nuances, and strict linguistic structures. To overcome these limitations, this study proposes TransLaw, a multi-agent framework that decomposes translation into word-level expression, sentence-level translation, and multidimensional review, integrating a specialized Hong Kong legal glossary database, Retrieval-Augmented Generation (RAG), and iterative feedback. Experiments on our newly constructed HKCFA Judgment 97-22 dataset, benchmarking 13 open-source and 3 commercial LLMs, demonstrate that TransLaw significantly outperforms single-agent baselines across all evaluated models. Human evaluation confirms the framework's effectiveness in terms of legal semantic accuracy, structural coherence, and stylistic fidelity, while noting that it still trails human experts in contextualizing complex terminology and stylistic naturalness.

1 Introduction

> *Legal language is a specialized type of language that, while derived from ordinary language, is often much more formulaic and complex.*
>
> — (Tiersma, 1999)

Judgment translation is the focal point of this research, which aims not only to bridge the critical research gap in Machine Translation (MT) for Hong Kong (HK) Case Law but also, more importantly, to meet a particular practical social need with tremendous potential for the Hong Kong legal system. Following the return of Hong Kong to China on 1 July 1997 after 155 years of British rule, the Hong Kong Special Administrative Region (HKSAR) took pride, upon its establishment, in having all its written statutes translated into Chinese, a mammoth task once considered impossible by many. This was required by the bilingual legislation as was mandated by Articles 8 and 9 of the Basic Law,[1] the former stipulating the retention of the English common law system and the establishment of legal bilingualism with Chinese and English as official languages, which are interpreted as being equally authentic[2] or of equal status.[3] Against this constitutional backdrop, the translation of Hong Kong judgments constitutes a pivotal component in sustaining the territory's bilingual legal framework (Cheng and He, 2016). Despite persistent challenges in reconciling linguistic transformation with inherited legal infrastructure (Chen, 2002), the system traces to the 1987 Bilingual Laws Project, which systematized statute translation (Jones Jr, 1987) and institutionalized parallel legislative drafting (Mushkat, 1997). While English remained the predominant courtroom language post-handover (Daniels et al., 2011), progressive legal localization (Tam, 2012) has rendered judgment translation an essential mechanism ensuring jurisprudential precision (Prieto Ramos, 2014) and facilitating cross-jurisdictional legal communication.

Confronted with the voluminous common law documentation within HK's judicial system (Hau, 2019), the establishment of efficient, accurate, and large-scale translation processes for laws assumes critical significance in HK (Sin et al., 2026). However, launching another project like the 1987 one for translating case law texts totally by experts' manual work is unrealistic. Even if such an undertaking might be conceived as possible for the existing case law, this manual approach could hardly

[1] https://www.basiclaw.gov.hk/en/basiclaw/chapter1.html

[2] https://www.elegislation.gov.hk/hk/cap1!en/s10B

[3] https://www.doj.gov.hk/en/about/orgchart_ldd_drafting_chi_eng.html

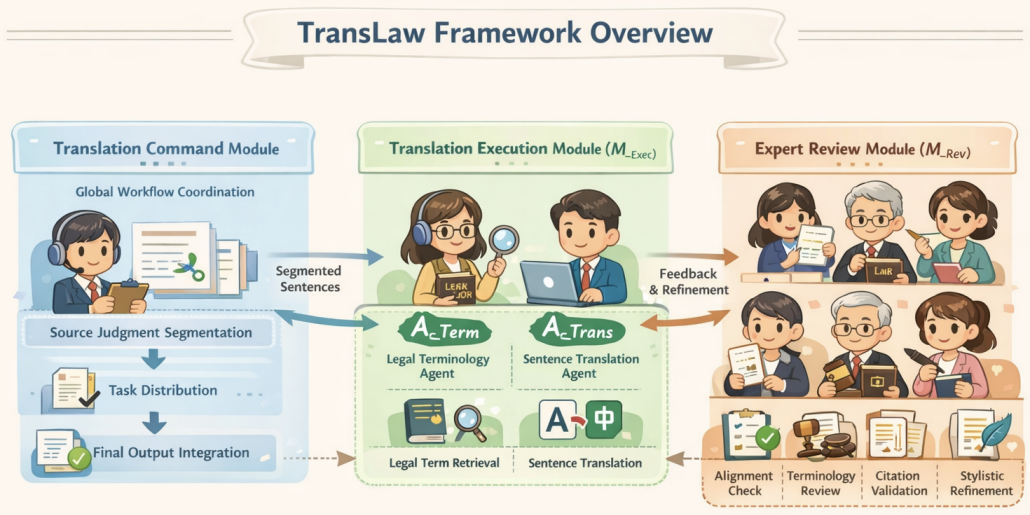


Figure 1: The overall architecture of TransLaw. The framework consists of three collaborative modules: (1) the Translation Command Module ($\mathcal{M}_{\text{Com}}$), where $\mathcal{A}_{\text{Com}}$ coordinates the global workflow; (2) the Translation Execution Module ($\mathcal{M}_{\text{Exec}}$), comprising $\mathcal{A}_{\text{Term}}$ and $\mathcal{A}_{\text{Trans}}$ for legal terminology parsing and core translation; and (3) the Expert Review Module ($\mathcal{M}_{\text{Rev}}$), which integrates $\mathcal{A}_{\text{Align}}$, $\mathcal{A}_{\text{TermR}}$, $\mathcal{A}_{\text{Cita}}$, and $\mathcal{A}_{\text{StyleP}}$ for multi-dimensional quality verification.

stand a chance of dealing with the innumerable upcoming cases in the future. Thus, shifting from exclusively manual workflows to a hybrid paradigm that integrates mainstream Large Language Models (LLMs) with professional human translation offers a promising direction. Significant stylistic differences exist in judgments across the hierarchy of Hong Kong courts. Judgments from the Court of Appeal, the Court of First Instance, the District Court, and other lower courts tend to be stylistically flexible. In contrast, the Court of Final Appeal (HKCFA) are characterized by strict constraints on format, structure, syntax, and terminology (Chan, 2012; Ng, 2020). Thus, this study targets HKCFA judgment translation, representing a more challenging legal language domain.

Previous research has successfully explored machine translation in the legal domain. However, these studies primarily focus on documents from various jurisdictions and language pairs (Bajcic and Golenko, 2024; Niklaus et al., 2025; Singh et al., 2025; Li et al., 2025), which differ significantly from HKCFA judgments. Recent research provides a systematic legal MT overview of the shift from traditional Neural Machine Translation (NMT) to Transformer-based Language Models (TLMs) (Greco and Tagarelli, 2024).

Another study compared the differences between machine translation and human translation of English legal texts into Arabic (Moneus and Sahari, 2024). The authors concluded that machine translation is not suitable for translating English legal texts into Arabic as it fails to adhere to strict formatting constraints or comprehend the socio-cultural background and complex legal terminology necessary to guarantee 'legal effect,' resulting in poor translation quality.

LLMs, based on the Transformer architecture with billions to hundreds of billions of parameters, have demonstrated their effectiveness in machine translation across various language pairs (Treviso et al., 2024; Elshin et al., 2024; Chen et al., 2025; Feng et al., 2025). Consequently, recent research has begun exploring the potential of LLMs for legal translation focusing on both specialized model tuning and the creation of robust evaluation benchmarks. For instance, recent research examined the effectiveness of evaluation of frontier LLMs and finetuned open small language models (SLMs) on Swiss legal translations in both zero-shot and fine-tuning settings in translating Swiss Legal document and found the results to be unsatisfactory (Niklaus et al., 2025). This research highlighted several issues with the text generated by LLM, including inappropriate legal word choices, hallucinating non-existent terms, and failing to grasp complex legal terminology. Furthermore, the model struggles to comprehend the socio-cultural background and cultural nuances embedded in legal systems, and fails to adhere to strict formatting constraints and formalized structural conventions required for official legal proceedings.

While previous work has advanced the field by enhancing models and establishing evaluation benchmarks, it has primarily focused on improving the intrinsic capabilities of single LLMs. Conse-

quently, the potential of current mainstream general and legal-specific LLMs for translating Hong Kong legal judgments remains unclear and unexplored. To address the shortage of bilingual Hong Kong legal data and advance legal translation, we present four main contributions:

1. We are the first to examine the capabilities of LLMs in English-to-Chinese HK Case Law translation.
2. We construct and release the HKCFA Judgement 97-22 dataset, the first high-quality dataset comprising sentence-level parallel pairs of judicial judgments.
3. We propose a novel multi-agent judgment translation system that integrates a Hong Kong legal glossary, RAG, and iterative feedback. By decomposing translation into word-level expression, sentence-level drafting, and multidimensional review, our approach significantly outperforms single-agent systems.
4. We propose a new human-evaluation metric, the Legal ACS metric, for HK judgment translation, and engage professional translators with expertise in HK legal translation to evaluate the results.

2 Related Work

Multi-Agent Systems (MAS) These systems consist of multiple autonomous yet collaborating agents. Characterized by core features such as autonomy, interactivity, cooperativity, and distribution, they can tackle challenges beyond the reach of single agents (Guo et al., 2024). The rapid evolution of LLMs has invigorated MAS, demonstrating its significant potential in various fields such as software development (Hong et al., 2024), multi-robot collaboration (Mandi et al., 2024), scientific experiments (Du et al., 2024), and scientific debates (Xiong et al., 2023). Moreover, LLM-based MAS play a crucial role in world simulation for social sciences, psychology, economics, hospital and scientific discovery, (re)enacting various roles and perspectives through agents' role-playing (Tao et al., 2024; Yu et al., 2024; He et al., 2025; Benita et al., 2025; Yao and Yu, 2025; Ghafarollahi and Buehler, 2025; Liu et al., 2026).

MAS in Machine Translation More directly related to our work, Liang et al. (Lewis et al., 2020) proposed the Multi-Agent Debate (MAD) framework, which improves performance on tasks like machine translation by encouraging diverse reasoning through structured argumentation. Wu et al. (Wu et al., 2025) proposed TransAgents, which focuses on literary translation by incorporating role-based agents such as translator, editor, and proofreader. Lv et al. (Lv et al., 2025) leveraged MAS for Classical Chinese translation, adopting a modular architecture with dedicated agents for keyword interpretation and grammatical validation to ensure cultural fidelity and semantic accuracy. These studies underscore the effectiveness of MAS in addressing the complexities of translation tasks through collaborative efforts.

3 Methodology

3.1 HKCFA Judgement 97-22 Dataset

A high-quality evaluation dataset is a prerequisite for advancing research on translating HK Case Law by LLMs (Fei et al., 2025). Currently, there is no publicly available dataset specifically designed for the bilingual translation of HK Case Law. To address this gap, we constructed a novel dataset named HKCFA Judgement 97-22, derived from Chinese-English bilingual **H**ong **K**ong **C**ourt of **F**inal **A**ppeal **Judgment**s from 19**97** to 20**22**. The dataset includes 344 high-quality judgments produced by professional human translation, with detailed statistics presented in Table 1.

Word-Count	Sentences	En-Token	Zh-HK-Token
5–15	2,982	28,541	56,573
16–35	1,998	65,842	108,872
36–54	1,631	92,714	152,428
55–100	2,553	239,978	390,492
101–200	1,599	271,173	433,178
201–702	336	113,405	172,739
Total	**11,099**	**811,653**	**1,314,282**

Table 1: The statistics of HKCFA Judgement 97-22 Dataset.

To calculate these statistics, we use the tiktoken (Jain, 2025) with the corresponding Byte-Pair Encoding (BPE) token encoding, which is also used by GPT-4o mini (Menick et al., 2024). While existing parallel legal corpora often use automated methods for sentence alignment (Koehn, 2005; Ziemski et al., 2016), our dataset relies on the structure provided by official government bodies, such as law paragraphs embedded in the HTML, resulting in high-quality alignment. To the best of our knowledge, this constitutes the first Hong Kong Case Law dataset comprising sentence-level parallel data pairs of Chinese-English judgments derived from professional human translation.

3.2 TransLaw Framework Overview

Translating English legal judgments into Traditional Chinese effectively is complex. It requires not only accurately grasping legal terminology and preserving the socio-cultural depth embedded in legal systems, but also strictly adhering to the formalized structural conventions required for official legal proceedings. To address these challenges, this study constructs a multi-agent translation system, TransLaw, for judgment translation.

The overall translation task is formulated as processing source judgment J, initially segmented into a sequence of sentences based on semantic structure, defined as $J = \{s_i \mid i = 1, 2, \ldots, N\}$. Each sentence s_i serves as an independent unit for bilingual processing by the translation system. TransLaw functions through three modules operated by a set of role agents:

1. *Translation Command Module* ($\mathcal{M}_{\text{Com}}$) houses Translation Command Agent ($\mathcal{A}_{\text{Com}}$), acting as central task coordinator responsible for sentence segmentation, information dispatch, and result integration.
2. *Translation Execution Module* ($\mathcal{M}_{\text{Exec}}$) comprises Legal Terminology Agent ($\mathcal{A}_{\text{Term}}$) and Sentence Translation Agent ($\mathcal{A}_{\text{Trans}}$). The former focuses on retrieving and optimizing interpretations for legal terms, while the latter generates initial translations based on these interpretations and surrounding context.
3. *Expert Review Module* ($\mathcal{M}_{\text{Rev}}$) integrates three verification agents to ensure final quality. First, Semantic Alignment Agent ($\mathcal{A}_{\text{Align}}$) performs fine-grained cross-checks between source and target texts to guarantee strict logical and factual consistency. Second, Legal Term Review Agent ($\mathcal{A}_{\text{TermR}}$) verifies legal terminology against the shared access authoritative HK legal glossary. Third, Legal Citation Agent ($\mathcal{A}_{\text{Cita}}$) validates that all case references and legislative provisions adhere to rigid HKCFA formatting standards. Finally, Stylistic fidelity Polishing Agent ($\mathcal{A}_{\text{StyleP}}$) refines linguistic output, ensuring translation exhibits appropriate judicial tone and syntactic fluency.

In summary, the TransLaw Framework formally defines the collective system of agents $\mathcal{C}$ as:

$$\mathcal{C} = \left\{ \underbrace{\mathcal{A}_{\text{Com}}}_{\text{Command}}, \underbrace{\mathcal{A}_{\text{Term}}, \mathcal{A}_{\text{Trans}}}_{\text{Execution}}, \underbrace{\mathcal{A}_{\text{Align}}, \mathcal{A}_{\text{TermR}}, \mathcal{A}_{\text{Cita}}, \mathcal{A}_{\text{StyleP}}}_{\text{Review}} \right\}. \tag{1}$$

Multiple agents collaborate through unified collaborative and feedback mechanisms, ensuring coherence and high quality of final translation output. Overall framework is depicted in Fig. 1.

3.3 Agent Task Allocation

This section details how agents collaborate within the three modules: the first coordinates overall workflow, the second performs core translation tasks, and the third ensures quality through multi-dimensional verification. The following subsections detail how these modules work in concert to achieve high-quality legal translation, co-working closely in a workflow highly similar to that of professional human translators.

Translation Command Module As shown in Fig. 1, $\mathcal{A}_{\text{Com}}$ orchestrates TransLaw's workflow as central coordinator, performing sentence segmentation, task distribution, and iterative refinement. $\mathcal{A}_{\text{Com}}$ initiates the workflow by segmenting the source judgment J into a sequence $\{s_1, \ldots, s_n\}$. To ensure consistency, it maintains a memory mechanism $\mathcal{E}_i = \{\hat{s}_1, \ldots, \hat{s}_{i-1}\}$, where $\hat{s}_i$ denotes the finalized translation of the preceding sentence s_i.The refinement process follows an iterative loop: an initial translation $\hat{s}_i^{(0)}$ is updated based on receiving review feedback $\mathcal{F}_i^{(k)}$ from the k-th review round. This feedback is mapped to textual increments via a mapping function $\Psi(\cdot)$ and integrated as follows:

$$\hat{s}_i^{(k+1)} = \hat{s}_i^{(k)} \oplus \Psi(\mathcal{F}_i^{(k)}) \tag{2}$$

where $\oplus$ denotes the integration of mapped suggestions. The loop ends when $\mathcal{F}_i^{(k)} = \emptyset$ or the iteration threshold K is reached, ensuring both quality and efficiency. Once the translation and review of all sentences are complete, $\mathcal{A}_{\text{Com}}$ aggregates all finalized translations to generate the final output.

Translation Execution Module As shown in Fig. 1, $\mathcal{A}_{\text{Exec}}$ comprises two synergistic agents: the $\mathcal{A}_{\text{Term}}$ for micro-level semantic parsing and the $\mathcal{A}_{\text{Trans}}$ for macro-level segment reconstruction. These agents cooperate to transform English legal judgments into Chinese translations characterized by lexical accuracy, socio-cultural nuance, and standardized judicial formatting.

The $\mathcal{A}_{\text{Term}}$ focuses on disambiguating key hk legal terminology with significant juridical or semantic complexity. Utilizing Retrieval-Augmented Generation (RAG) (Salemi and Zamani, 2024; Li et al., 2026), it extracts the set of legal terms L_i

Series	Model	TransLaw			Single Translator Agent			Rank
		xCOMET-XL↑	wmt22-unite-da↑	**Avg.↑**	xCOMET-XL↑	wmt22-unite-da↑	**Avg.↑**	
OpenAI	GPT-4o	**85.12** (±0.14)	**91.78** (±0.12)	**88.45**	**69.42** (±0.15)	**75.88** (±0.13)	**72.65**	1
	GPT-4	84.24 (±0.16)	90.10 (±0.15)	87.15	68.10 (±0.18)	74.25 (±0.16)	71.17	2
	ChatGPT	82.29 (±0.19)	88.52 (±0.18)	85.41	66.15 (±0.21)	72.10 (±0.20)	69.12	5
DeepSeek	V3	83.53 (±0.15)	89.56 (±0.14)	86.55	67.45 (±0.17)	73.45 (±0.16)	70.45	3
	R1	83.25 (±0.17)	89.33 (±0.15)	86.29	67.12 (±0.19)	73.12 (±0.17)	70.12	4
Qwen	14B-Chat	81.86 (±0.20)	87.12 (±0.19)	84.49	65.55 (±0.22)	71.80 (±0.21)	68.67	6
	7B-Chat	80.54 (±0.22)	87.06 (±0.21)	83.80	63.90 (±0.25)	70.15 (±0.23)	67.02	7
Baichuan	13B-Chat	81.33 (±0.23)	87.51 (±0.22)	84.42	64.40 (±0.26)	70.50 (±0.24)	67.45	8
	13B-Base	79.60 (±0.25)	86.51 (±0.24)	83.06	62.20 (±0.28)	69.10 (±0.27)	65.65	10
ChatGLM	3-6B	80.27 (±0.24)	87.20 (±0.22)	83.74	63.85 (±0.27)	69.20 (±0.25)	66.52	9
	2-6B	78.11 (±0.28)	84.78 (±0.26)	81.45	61.10 (±0.30)	67.50 (±0.29)	64.30	12
ChatLaw	33B	79.29 (±0.26)	85.80 (±0.25)	82.55	62.50 (±0.29)	68.80 (±0.28)	65.65	11
	13B	76.26 (±0.30)	82.59 (±0.29)	79.43	58.50 (±0.35)	64.80 (±0.32)	61.65	13

Table 2: The performance (%) of various LLM models serving as agents in the TransLaw multi-agent framework compared to a Single Translator Agent. Best results are in **bold**. Confidence intervals are in parentheses.

from the source sentence s_i and retrieves relevant expressions from the authoritative HK legal glossary database D (the Combined DOJ Glossaries of Legal Terms[4]). The initial expression for a legal term $l_k \in s_i$ is obtained from database D using a retrieval function R_D:

$$C_k = R_D(l_k) \tag{3}$$

where C_k is the set of candidate legal expressions for term l_k. This ensures that polysemous HK legal terms are rendered correctly within the judgments.

Leveraging the $\mathcal{A}_{\text{Term}}$'s output, the $\mathcal{A}_{\text{Trans}}$ generates sentence-level translations by integrating verified legal expressions with contextual memory. Initially, it substitutes the expressions of the target vocabulary based on the optimized term set $\hat{L}_i$:

$$s'_i = h(s_i, \hat{L}_i) \tag{4}$$

where $h(s_i, \hat{L}_i)$ represents the initial transformation function injecting $\hat{L}_i$ into the syntactic structure of s_i. To ensure judicial coherence and consistency across the judgment, the global context $\mathcal{G}$ must be referenced through the context adjustment function $q(\cdot)$:

$$\hat{s}_i^{(0)} = q(s'_i, \mathcal{G}) \tag{5}$$

The complete sentence translation process can be formalized as:

$$\hat{s}i^{(0)} = f_{\text{Trans}}(s_i, \hat{L}i, \mathcal{G}) \tag{6}$$

where $\mathcal{G}$ represents the global context derived from preceding case facts and procedural history. This dual-agent approach ensures both lexical precision and syntactic fluidity, with the $\mathcal{A}_{\text{Trans}}$ dynamically adjusting translations based on feedback from the subsequent Expert Review Module through the iterative mechanism described in Equation 2.

[4]https://www.glossary.doj.gov.hk/

Expert Review Module As shown in Fig. 1, $\mathcal{A}_{\text{Rev}}$ implements a comprehensive quality assurance system through three review agents. This module transforms initial translations into polished, authoritative outputs through systematic multi-dimensional evaluation and iterative refinement.The module employs three complementary review perspectives:

- The $\mathcal{A}_{\text{Align}}$ evaluates semantic accuracy by checking the precise alignment between source and target texts at both logical and factual levels consistency:

$$\delta_{i,\text{Align}}^{(k)} = f_{\text{Align}}(\hat{s}^{(k)}i, s_i, \mathcal{G}) \tag{7}$$

- $\mathcal{A}_{\text{TermR}}$ verifies legal term expressions, possessing the same access capability to HK legal glossary database D as the $\mathcal{A}_{\text{Term}}$:

$$\delta_{i,\text{TermR}}^{(k)} = f_{\text{TermR}}(\hat{s}_i^{(k)}, \hat{L}_i, D) \tag{8}$$

- The $\mathcal{A}_{\text{Cita}}$ validates legal authorities, ensuring that all case references[5] and legislative provisions[6] adhere to the rigid HKCFA formatting standards:

$$\delta_{i,\text{Cita}}^{(k)} = f_{\text{Cita}}(\hat{s}_i^{(k)}) \tag{9}$$

- The $\mathcal{A}_{\text{StyleP}}$ guarantees stylistic fidelity via polishing, ensuring the translation exhibits a formal judicial tone and syntactic fluency:

$$\delta_{i,\text{Style}}^{(k)} = f_{\text{Style}}(\hat{s}_i^{(k)}) \tag{10}$$

[5]Case references denote citations to legal precedents, e.g., *HKSAR v. Chan Ka Ming* [2025] HKCFA 8.

[6]Legislative provisions refer to specific statutory sections, e.g., Section 24 of the *Safeguarding National Security Ordinance* (Ord. No. 6 of 2024).

Expert Review Module ↓		Translation Command Module (GPT-4o)												
		Translation Execution Module →												
		OpenAI			DeepSeek		Qwen		Baichuan		ChatGLM		ChatLaw	
Series	**Model**	GPT-4o	GPT-4	ChatGPT	V3	R1	14B	7B	13B-C	13B-B	3-6B	2-6B	33B	13B
OpenAI	GPT-4o	**85.12**	**84.87**	**83.43**	**84.19**	**83.92**	**83.08**	**81.76**	**82.34**	**80.91**	**81.47**	**79.83**	**80.52**	**78.89**
	GPT-4	84.88	84.24	83.17	83.91	83.68	82.79	81.42	82.03	80.57	81.18	79.44	80.16	78.43
	ChatGPT	84.13	83.56	82.29	82.94	82.71	81.83	80.77	81.15	79.88	80.42	78.76	79.37	77.82
DeepSeek	V3	84.54	83.92	82.81	83.53	83.34	82.37	81.08	81.59	80.23	80.88	79.12	79.74	78.06
	R1	84.27	83.65	82.58	83.41	83.25	82.16	80.93	81.44	80.09	80.71	78.98	79.62	77.91
Qwen	14B-Chat	83.76	83.18	82.04	82.87	82.59	81.86	80.62	81.03	79.67	80.35	78.54	79.18	77.45
	7B-Chat	83.05	82.47	81.38	82.14	81.82	81.09	80.54	80.36	79.05	79.73	77.92	78.55	76.88
Baichuan	13B-Chat	83.38	82.73	81.65	82.42	82.16	81.33	80.27	81.33	79.36	80.08	78.25	78.84	77.19
	13B-Base	82.44	81.88	80.79	81.56	81.23	80.47	79.38	79.94	79.60	79.15	77.37	77.93	76.24
ChatGLM	3-6B	83.02	82.35	81.27	82.05	81.76	80.94	79.82	80.44	79.18	80.27	77.85	78.46	76.73
	2-6B	81.79	81.14	80.03	80.87	80.55	79.76	78.64	79.28	77.95	78.92	78.11	77.24	75.58
ChatLaw	33B	82.15	81.56	80.42	81.23	80.97	80.18	79.06	79.65	78.34	79.33	77.58	79.29	76.02
	13B	80.46	79.88	78.74	79.52	79.25	78.43	77.36	77.97	76.65	77.62	75.83	77.41	76.26

Table 3: The performance (xCOMET-XL, %) of various LLMs serving as agents in the Translation Execution Module and Expert Review Module. Best results in each column are in **bold**. Model abbreviations: 13B-C: Baichuan-13B-Chat, 13B-B: Baichuan-13B-Base, 3-6B: ChatGLM3-6B, 2-6B: ChatGLM2-6B. Results here use GPT-4o as the Translation Command Module agent; comparative results for GPT-4 and ChatGPT are detailed in Appendix F.

These agents operate in concert within each review round, generating comprehensive feedback:

$$\delta^{(k)}i = \delta^{(k)}_{i,\text{Align}} \cup \delta^{(k)}_{i,\text{TermR}} \cup \delta^{(k)}_{i,\text{Cita}} \cup \delta^{(k)}_{i,\text{Style}} \quad (11)$$

When $\delta^{(k)}_i = \emptyset$, the translation achieves optimal quality and proceeds to final output. Otherwise, the feedback triggers another iteration of refinement through the Translation Execution Module. This multi-round review mechanism ensures that each translation meets the highest standards of semantic accuracy, terminological precision, citation compliance, and stylistic integrity, while preserving the rigorous nature of the source english judgment. The review mechanism maximally ensures the quality of the translation while reducing potential biases that might arise from errors within a single module. Detailed prompt templates for the seven agents are provided in Appendix G.

4 Evaluation

This section presents both automated and human evaluations of TransLaw's translation performance.

4.1 Automated evaluation

Automated evaluation of an MT system needs to be conducted by using available automated evaluation metrics to estimate the quality scores of its translation outputs by comparing them with a given bilingual text dataset providing the gold standard answers.

Metrics The metrics adopted for the evaluation are two of the most popular automated MT evaluation metrics in recent years: (1) xCOMET-XL, a version of xCOMET, which is a state-of-the-art learned metric for various levels of evaluation (Guerreiro et al., 2024); and (2) wmt22-unite-da, a unified MT quality evaluation model (Guttmann et al., 2024). To ensure statistical reliability, for all metrics, we report two times standard deviation using 1,000 runs of bootstrap (Efron et al., 1986) on the test dataset, which corresponds to a 95.45% confidence level under the assumption of a normal distribution.

Evaluated LLM Models We evaluated 13 widely adopted LLMs, classified into two categories: General and Legal-specific LLMs. For General LLMs, the testbed includes GPT-4o (Hurst et al., 2024), GPT-4 (Achiam et al., 2023), ChatGPT (Brown et al., 2020), the DeepSeek series (DeepSeek-R1 (Guo et al., 2025), DeepSeek-V3 (Liu et al., 2024)), the ChatGLM series (ChatGLM2-6B, ChatGLM3-6B) (Zeng et al., 2023), the Baichuan series (Baichuan-13B-Base, Baichuan-13B-Chat) (Yang et al., 2023), and the Qwen series (Qwen-7B-Chat, Qwen-14B-Chat) (Bai et al., 2023); for Legal-specific LLMs, we included ChatLaw-13B and ChatLaw-33B (Cui et al., 2023). Detailed links to the LLM models are provided in Appendix E. TransLaw consists of three modules. For the agents in each module, we employ the same LLM from the list above.

Analysis of Evaluation Results The experimental results are shown in Tables 2 and 3. These results reveal the following points.

- All open-source LLMs perform slightly under the closed-source (commercial) ones like GPT-4o and GPT-4, which achieve the best performance in this benchmark. However, due to insufficient knowledge of Hong Kong's legal system, both they exhibit substantial limitations in legal judgment translation. This indicates significant potential for improvement in legal-domain LLM capabilities.
- Increased LLM model scale consistently enhances performance. For instance, Qwen-14B outperforms Qwen-7B. Furthermore, chat-optimized LLMs (e.g., Baichuan-13B-Chat) surpass their base models (e.g., Baichuan-13B-Base), suggesting that greater instruction-following capabilities, gained through supervised fine-tuning and alignment optimization, can be more effective in unlocking LLMs' potential in translation.
- Surprisingly, legal-specific LLMs do not always outperform general LLMs. We speculate that there are two possible reasons. First, the capability of these legal-specific LLMs could be limited by their base models, which are usually known not to be as strong as other LLMs such as GPT-4o and GPT-4; moreover, the continuous pre-training or fine-tuning using legal corpora may not further promote the abilities of the original base models. It is also possible that both play a part in explaining this result, which certainly suggests the necessity for further design to improve the performance of legal LLMs.

Contrasting the TransLaw performance in Columns 6-8 of Table 2 with the Single Translator Agent reveals marked performance improvements, which demonstrate the efficacy of the collaborative TransLaw framework. Despite these inevitable limitations of TransLaw, the above evaluation shows that using LLMs as role-specific agents in a MAS can effectively assist translation tasks to facilitate the Hong Kong bilingual legal system. However, investigating how LLMs compare with humans taking similar translation roles, like annotation and proofreading, paves the way for powerful intelligent legal LLMs that improve the efficiency and quality of legal translation services.

4.2 Human Evaluation

For human evaluation, a scoring scheme needs to be formulated to integrate human evaluation scores in various evaluation dimensions into one. Therefore, we propose the legal ACS metric for the translation of HK legal judgments, which consists of three dimensions: **A** (accuracy of legal meaning), **C** (coherence and cohesion in structure), and **S** (Style appropriateness) (henceforth ACS), whose formulation is presented below, followed by the settings and results of our human evaluation.

Test Set We selected the bilingual texts of the judgment "HKSAR - Court of Final Appeal - Final Appeal Criminal Case No. 1 of 2021" (henceforth FACC 1/2021 for brevity; see Appendix B) from the HKCFA Judgement 97-22 Dataset as our human evaluation data. Following paragraph-level segmentation and manual alignment, the whole test set consists of 200 paragraph-level source-target pairs. This human evaluation set comprises 12,029 tokens in English and 19,478 tokens in Chinese.

Evaluation Metrics Aiming at a comprehensive, adequate and reliable evaluation of the translation quality of HK legal judgments, the ACS metric is formulated as $I = \alpha A + \beta C + \gamma S$, where A, C, and S are the scores given by human expert evaluators in the three key dimensions, and α, β, and γ are respective weight coefficients according to their relative importance. Based on the experience and recommendation of domain experts, these weights are set as follows for our manual evaluation of legal judgment translation: $\alpha = 0.6, \beta = 0.3, \gamma = 0.1$. This setting recognizes the most fundamental role of accuracy. In Figure 2, we further set different weights for evaluation, which remains an interesting issue for further examination, as is the human evaluation of TransLaw's performance; both are expected to give meaningful hints to justify this scoring scheme.

Setup We perform a manual evaluation of the FACC 1/2021 test set, comparing the official human translation against two GPT-4o system configurations (the best-performing model in Table 2): one operating as a Single Translator Agent and the other deployed within the TransLaw framework (comprising seven agents). To mitigate human evaluator fatigue, segment lengths were controlled, with the maximum length reaching 234 English words (290 tokens) and 414 Chinese words (580 tokens). Anonymized evaluation tables containing

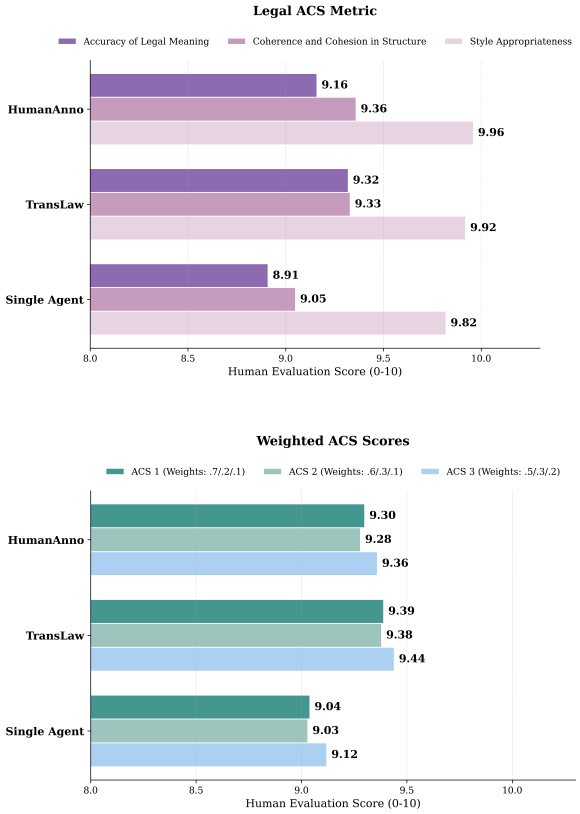


Figure 2: Human Evaluation Results. Performance of the three systems across three dimensions (top) and different weighting schemes (bottom).

segment/sentence IDs, source texts, and system outputs were assessed by 10 certified professional legal translators using a 0-10 point scale across three predefined dimensions. The evaluation guidelines for human experts can be found in Appendices H and D.

Results and Analysis Figure 2 presents human evaluation scores for the three configurations using ACS metrics. TransLaw achieves superior Accuracy in conveying legal meanings and leads all ACS metrics across three different weighting schemes. Official human translation demonstrates relative strengths in Coherence and Cohesion as well as Style, indicating that human translation capability remains superior to TransLaw. The three different weighting schemes for ACS in Figure 2 yielded very minor differences in ACS scores, convincingly justifying the soundness and consistency of this scoring scheme for human evaluation. More detailed human evaluation of translation cases is provided in Appendix D.

5 Cost Analysis

The cost of human translation can vary based on several factors, including the type of text, the translator's location, and their level of experience. The American Translators Association recommends a minimum charge of US$0.12 per word for professional translation. Accordingly, translating a judgment like FACC 1/2021, totaling 11,585 words (12,029 tokens), would cost US$1,390.20. In contrast, the cost of translating the FACC 1/2021 test set using GPT-4o is approximately US$0.39. Using TransLaw, the cost breaks down to approximately US$0.35. This contrast indicates that using TransLaw to translate HK legal judgments can reduce translation costs by nearly 4,000 times compared to human translation and by 10.26% compared to GPT-4o. Note that US$0.39 and US$0.35 for using GPT-4o and TransLaw are API (Application Programming Interface) costs, excluding the cost for human proofreading and editing of the translation output from an API. Given that the Avg. standard rate for human editing is approximately US$0.04 per word[7], the total cost for translating plus editing the said judgment would be US$0.35 + US$0.04 $\times$ 11,585 = US$463.75, saving US$926.45, or two-thirds of the full human translation cost.

6 Conclusion

In this paper, we first constructed and released a large-scale bilingual dataset, HKCFA Judgment 97-22, and conducted the first comprehensive evaluation of LLM capabilities in Hong Kong case law translation. Furthermore, to address LLMs' shortcomings, we proposed TransLaw, a collaborative MAS designed to mimic professional human legal translation workflows. Experimental results have been obtained from benchmarking using this dataset to verify the validity and effectiveness of the collaborative strategies, in addition to providing an informative performance comparison of the LLMs for use as agents in TransLaw. Overall, TransLaw proves to be a robust and effective framework for handling the complexities of Hong Kong case law translation, and it lays a foundation for future research in this area.

[7] https://www.translationedge.com/pricing.

7 Limitations

Our research focuses on Hong Kong Case Law, which is characterized by very strict requirements for legal terms and a highly standardized format. Consequently, our method may not be directly applicable to judgments from other jurisdictions or different pair of languages. Therefore, the results of our study may not cover all countries and types of legal documents.

8 Ethics Statement

Our dataset and evaluation benchmark contain no personal, sensitive, or private information; they consist solely of publicly available data.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Martina Bajcic and Dejana Golenko. 2024. Applying large language models in legal translation: The state-of-the-art. *JLL*, 13:171.

J Benita, S Jaswanth, N Bhuvaneshwar, R Yuvaraj, and Y Lakshmi Narayana. 2025. Phoenix: A conversational agent for emotional well-being and psychological support. In *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, pages 1137–1142. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Clara Ho-yan Chan. 2012. Bridging the gap between language and law: Translational issues in creating legal chinese in hong kong. *Babel*, 58(2):127–144.

Albert H Y Chen. 2002. Hong kong's legal system in the new constitutional order: The experience of 1997–2000. In *Implementation of Law in the People's Republic of China*, pages 213–245. Brill Nijhoff.

Andong Chen, Kehai Chen, Yang Xiang, Xuefeng Bai, Muyun Yang, Yang Feng, Tiejun Zhao, and Min Zhang. 2025. Llm-based translation inference with iterative bilingual understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16886–16902.

Le Cheng and Lianzhen He. 2016. Revisiting judgment translation in hong kong. *Semiotica*, 2016(209):59–75.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2023. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *arXiv preprint arXiv:2306.16092*.

Ronald J Daniels, Michael J Trebilcock, and Lindsey D Carson. 2011. The legacy of empire: The common law inheritance and commitments to legality in former british colonies. *The American Journal of Comparative Law*, 59(1):111–178.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, pages 11733–11763.

Bradley Efron and 1 others. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75.

Denis Elshin, Nikolay Karpachev, Boris Gruzdev, Ilya Golovanov, Georgy Ivanov, Alexander Antonov, Nickolay Skachkov, Ekaterina Latypova, Vladimir Layner, Ekaterina Enikeeva, and 1 others. 2024. From general llm to translation: How we dramatically improve translation quality using human evaluation data for llm finetuning. In *Proceedings of the Ninth Conference on Machine Translation*, pages 247–252.

Zhiwei Fei, Songyang Zhang, Xiaoyu Shen, Dawei Zhu, Xiao Wang, Jidong Ge, and Vincent Ng. 2025. Internlm-law: An open-sourced chinese legal large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9376–9392.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. Tear: Improving llm-based machine translation with systematic self-refinement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3922–3938.

Alireza Ghafarollahi and Markus J Buehler. 2025. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523.

Candida M Greco and Andrea Tagarelli. 2024. Bringing order into the realm of transformer-based language models for artificial intelligence and law. *Artificial Intelligence and Law*, 32(4):863–1010.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 8048–8057.

Kamil Guttmann, Mikołaj Pokrywka, Adrian Charkiewicz, and Artur Nowakowski. 2024. Chasing comet: Leveraging minimum bayes risk decoding for self-improving machine translation. *Association for Machine Translation*, page 80.

Berry Fong-Chung Hau. 2019. *The Common Law System in Chinese Context*. Routledge.

Junda He, Christoph Treude, and David Lo. 2025. Llm-based multi-agent systems for software engineering: Literature review, vision, and the road ahead. *ACM Transactions on Software Engineering and Methodology*, 34(5):1–30.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shantanu Jain. 2025. tiktoken: A fast bpe tokeniser for use with openai's models.

David A Jones Jr. 1987. A leg to stand on-post-1997 hong kong courts as a constraint on prc abridgment of individual rights and local autonomy. *Yale J. Int'l L.*, 12:250.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit Proceedings of Conference*, pages 79–86. International Association for Machine Translation.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, and 1 others. 2025. Legalagentbench: Evaluating llm agents in legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2322–2344.

Yining Li, Wenjun Ke, Jiajun Liu, Peng Wang, Jianghan Liu, and Yao He. 2026. Towards evidence-aware retrieval-augmented generation via self-corrective chain-of-thought. *Information Processing & Management*, 63(2):104369.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Linjie Liu, Lichen Wang, Weiyan Niu, and Shijia Hua. 2026. Dynamic sanctioning mechanism for cooperative multi-agent systems. *Expert Systems with Applications*, 296:128873.

Weifeng Lv, Qiong Cao, and Xiaoyang Liu. 2025. A multi agent classical chinese translation method based on large language models. *Scientific Reports*, 15(1):40160.

Zhao Mandi, Shreeya Jain, and Shuran Song. 2024. Roco: Dialectic multi-robot collaboration with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 286–299. IEEE.

Jacob Menick, Kevin Lu, Shengjia Zhao, E Wallace, H Ren, H Hu, N Stathas, and F Petroski Such. 2024. Gpt-4o mini: advancing cost-efficient intelligence. *Open AI: San Francisco, CA, USA*.

Ahmed Mohammed Moneus and Yousef Sahari. 2024. Artificial intelligence and human translation: A contrastive study based on legal texts. *Heliyon*, 10(6).

Roda Mushkat. 1997. *One country, two international legal personalities: The case of Hong Kong*, volume 1. Hong Kong University Press.

Kwai Hang Ng. 2020. *The common law in two voices: Language, law, and the postcolonial dilemma in Hong Kong*. Stanford University Press.

Joel Niklaus, Jakob Merane, Luka Nenadic, Sina Ahmadi, Yingqiang Gao, Cyrill AH Chevalley, Claude Humbel, Christophe Gösken, Lorenzo Tanzi, Thomas Lüthi, and 1 others. 2025. Swiltra-bench: The swiss legal translation benchmark. *arXiv preprint arXiv:2503.01372*.

Fernando Prieto Ramos. 2014. International and supranational law in translation: From multilingual lawmaking to adjudication. *The Translator*, 20(3):313–331.

Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2395–2400.

King-kui Sin, Xi Xuan, Chunyu Kit, Ho-yan Clara Chan, and Ho-kin Honic Ip. 2026. *Solving the Unsolvable: Translating Case Law in Hong Kong*. University of Hong Kong, China.

Kshetrimayum Boynao Singh, Deepak Kumar, and Asif Ekbal. 2025. Evaluation of llm for english to hindi legal domain machine translation systems. In *Proceedings of the Tenth Conference on Machine Translation*, pages 823–833.

Waikeung Tam. 2012. *Legal mobilization under authoritarianism: the case of post-colonial Hong Kong*. Cambridge University Press.

Wei Tao, Yucheng Zhou, Yanlin Wang, Wenqiang Zhang, Hongyu Zhang, and Yu Cheng. 2024. Magis: Llm-based multi-agent framework for github issue resolution. *Advances in Neural Information Processing Systems*, 37:51963–51993.

Peter M Tiersma. 1999. *Legal language*. University of Chicago Press.

Marcos V Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and André FT Martins. 2024. xtower: A multilingual llm for explaining and correcting translation errors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, and Kaifu Zhang. 2025. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7572–7590.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Zonghai Yao and Hong Yu. 2025. A survey on llm-based multi-agent ai hospital.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, and 1 others. 2024. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045.

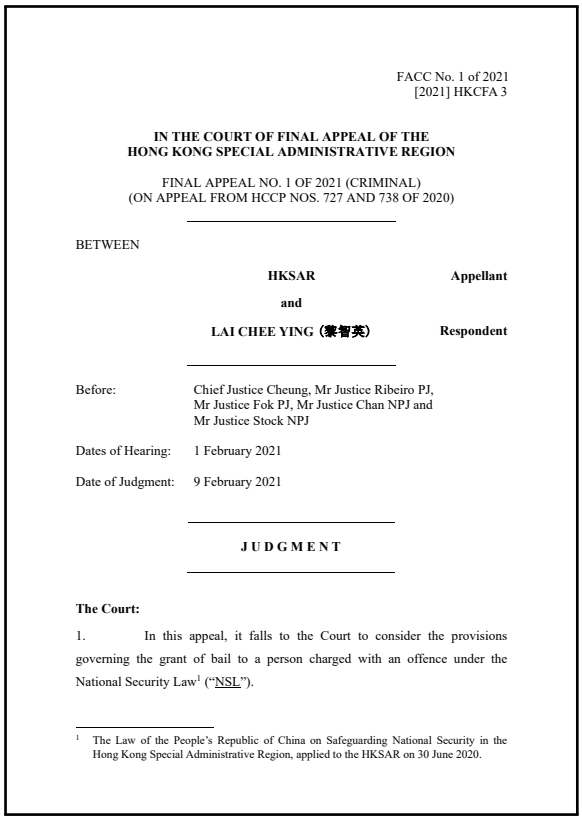
Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

A Use of AI Assistants

We used GPT-4o and Claude Sonnet 3.5 for coding, shortening texts and editing LaTeX more efficiently.

B HKACFA Judgment Example (Case No. FACC 1/2021)



FACC No. 1 of 2021
[2021] HKCFA 3

IN THE COURT OF FINAL APPEAL OF THE HONG KONG SPECIAL ADMINISTRATIVE REGION

FINAL APPEAL NO. 1 OF 2021 (CRIMINAL)
(ON APPEAL FROM HCCP NOS. 727 AND 738 OF 2020)

BETWEEN

HKSAR **Appellant**

and

LAI CHEE YING (黎智英) **Respondent**

Before: Chief Justice Cheung, Mr Justice Ribeiro PJ, Mr Justice Fok PJ, Mr Justice Chan NPJ and Mr Justice Stock NPJ

Dates of Hearing: 1 February 2021

Date of Judgment: 9 February 2021

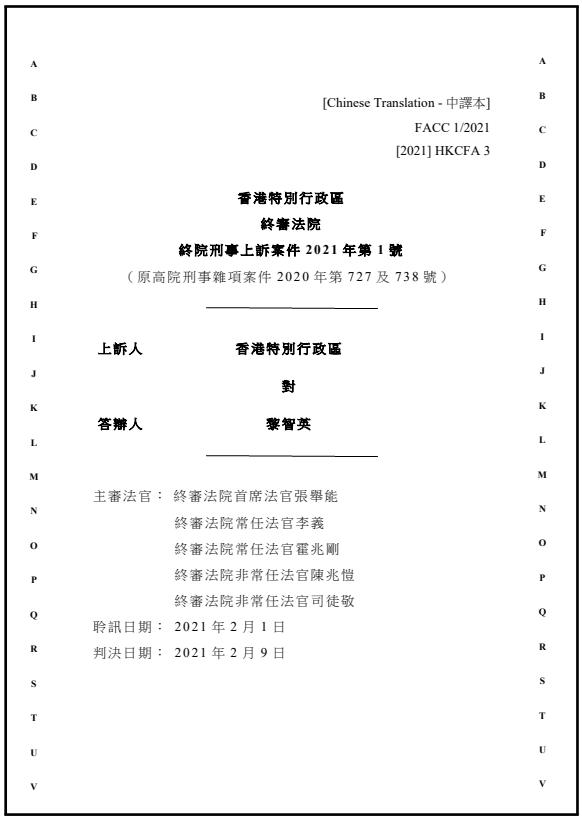
J U D G M E N T

The Court:

1. In this appeal, it falls to the Court to consider the provisions governing the grant of bail to a person charged with an offence under the National Security Law[1] ("NSL").

[1] The Law of the People's Republic of China on Safeguarding National Security in the Hong Kong Special Administrative Region, applied to the HKSAR on 30 June 2020.

(a) English Judgment (Source)



[Chinese Translation - 中譯本]
FACC 1/2021
[2021] HKCFA 3

香港特別行政區
終審法院
終院刑事上訴案件 2021 年第 1 號
（原高院刑事雜項案件 2020 年第 727 及 738 號）

上訴人 **香港特別行政區**

對

答辯人 **黎智英**

主審法官： 終審法院首席法官張舉能
終審法院常任法官李義
終審法院常任法官霍兆剛
終審法院非常任法官陳兆愷
終審法院非常任法官司徒敬

聆訊日期： 2021 年 2 月 1 日
判決日期： 2021 年 2 月 9 日

(b) Chinese Translation (Target)

Figure 3: Example of the first page of a bilingual HKCFA judgment (Case No. FACC 1/2021), comparing the source text and target translation.

C Evaluation Codes

Evaluation Dimension	Sub-dimension	Description
Accuracy	CW	Choice of word. The word or expression is not a good choice.
	IF	Information structure not preserved.
	MC	Meaning has been changed because of inappropriate restructuring.
	MT	Mistranslation due to inadequate comprehension or misinterpretation.
	NA	The translation conveys a different meaning from that of the source text.
	NC	Meaning not clear due to ambiguity, vagueness or syntactic problems.
	OM	Omission. Part of the original has been left untranslated.
	OT	Over-translation. Too much has been read into the source text.
	TL	Too literal, affecting comprehensibility.
	UT	Under-translation. Meaning is not adequately captured in translation.
Grammar	Art	Article.
	Det	Determiner.
	MD	Modality.
	NB	Number.
	PN	Punctuation.
	Prep	Wrong preposition.
	PS	Part of speech.
	SP	Spelling or wrong character.
	ST	The sentence or part of the sentence is ill-formed or ambiguous.
	SV	Subject verb agreement.
	TN	Tense problem.
	WO	Word order.
Usage and style	CL	Collocation problem.
	CN	The word or expression has connotation not appropriate in the context.
	CO	Connective problem, e.g., inappropriate connectives.
	IC	Inconsistent use of a word; or incoherence between clauses or sentences.
	ID	Idiomaticity, i.e., unidiomatic expression.
	RF	Reference problem, e.g., ambiguous use of a pronoun.
	RN	Redundancy: the word or expression should be deleted.
	SL	Stylistic problems, e.g., the word or expression is not of an appropriate style.
	TS	Transition problems: sentences not well connected; bad language flow.

Table 4: Evaluation Codes

D More Cases

Source Text	At present, the increasingly notable national security risks in the HKSAR have become a prominent problem. In particular, since the onset of HK's 'legislative amendment turmoil' in 2019, anti-China forces seeking to disrupt HK have blatantly advocated such notions as 'HK independence', 'self-determination' and 'referendum', and engaged in activities to undermine national unity and split the country. They have brazenly desecrated and defiled the national flag and emblem, incited HK people to oppose China and the Communist Party of China ('CPC'), besiege Central People's Government ('CPG') offices in HK, and discriminate and ostracize Mainland personnel in HK. These forces have also wilfully disrupted social order in HK, violently resisted police enforcement of the law, damaged public facilities and property, and paralyzed governance by the government and operation of the legislature. Moreover, in recent years, certain foreign or external forces have flagrantly interfered in HK's affairs. They have made intervention and created disturbances in various ways, such as by legislative and administrative means and through non-governmental organizations. In collusion with those anti-China HK disrupters, these forces of the same ilk backed and cheered on the disrupters and provided a protective umbrella, and utilized HK to carry out activities endangering national security. These acts and activities have seriously challenged the bottom line of the 'One Country, Two Systems' principle, seriously undermined the rule of law, and seriously jeopardized national sovereignty, security and development interests.
GPT-4o	目前，香港特別行政區日益突出[CW] 的國家安全風險已成為一個顯著[CW] 問題。特別是自2019年香港的「立法修訂風波」開始[ST]，反華[NA] 勢力企圖破壞香港的行為愈發公然，公然宣揚如「香港獨立」、「自決」和「公投」等概念，並從事破壞國家統一和分裂國家的活動。他們[Pronoun] 公然褻瀆和污損國旗及國徽，煽動香港人反對中國及中國共產黨（「中共」），圍攻中央人民政府（「中央政府」）在香港的辦事處，歧視和排斥內地人員。這些勢力還任意破壞香港的社會秩序，暴力抗拒警方執法，破壞公共設施和財產，使政府治理[CW] 和立法機構運作陷於癱瘓。[OM] 近年來，某些外國或外部[CW] 勢力公然干預香港事務，以立法和行政手段及[US] 透過非政府組織的各種方式[NC] 進行干擾[CL] 和製造混亂。這些勢力[UT] 與反華香港破壞分子[CW/NC] 相勾結，支持並鼓舞破壞者，提供保護傘，利用香港從事危害國家安全的活動。這些行為和活動嚴重挑戰了「一國兩制」原則的底線，嚴重破壞了法治，嚴重危害了國家主權、安全和發展利益。
TransLaw	目前，香港特別行政區日益突顯的國家安全風險已成為一個突出問題。特別是自2019年香港「修例風波」爆發[OT] 以來，一些[NA] 反中勢力企圖擾亂香港，公然宣揚「港獨」、「自決」及「公投」等概念[CW]，並從事破壞國家統一和分裂國家的活動。他們[Pronoun] 公然褻瀆和污損國旗及國徽，煽動香港人反對中國和中國共產黨（「中共」），圍攻中央人民政府（「中央政府」）在港機構，以及歧視和排斥香港的內地人員。這些勢力還肆意擾亂香港社會秩序，暴力抗拒警方執法，破壞公共設施和財產，並癱瘓政府治理和立法機構的運作。[OM] 近年來，一些[NA] 外國或外部[CW] 勢力公然干預香港事務，通過立法和行政手段以及非政府組織等多種方式進行干擾和創造混亂。這些勢力與反中香港破壞分子相勾結[NA]，支持並為其助威[ST]，提供保護傘，利用香港進行危害國家安全的活動。這些行為和活動嚴重挑戰了「一國兩制」原則的底線，嚴重破壞了法治，嚴重危害了國家主權、安全和發展利益。
Reference Text	目前，香港特別行政區日益凸顯的國家安全風險已成為一個突出問題。特別是2019年香港發生「修例風波」以來，反中亂港勢力公然鼓吹「港獨」、「自決」、「公投」等概念，從事破壞國家統一、分裂國家的活動；公然侮辱、污損國旗國徽，煽動港人反中反共、圍攻中央駐港機構、歧視和排擠內地在港人員；蓄意破壞香港社會秩序，暴力對抗警方執法，毀損公共設施和財物，癱瘓政府管治和立法會運作。近年來，某些外國和境外勢力公然干預香港事務，通過立法、行政、非政府組織等多種方式進行干擾和和製造混亂，與香港反中亂港勢力勾連合流、沆瀣一氣，為香港反中亂港勢力撐腰打氣、提供保護傘，利用香港從事危害我國國家安全的活動。這些行為和活動，嚴重挑戰「一國兩制」原則底線，嚴重損害法治，嚴重危害了國家主權、安全和發展利益。

Table 5: Case Study: Comparative analysis of the Source Text translation using GPT-4o acting as a single translator agent and TransLaw, alongside the Reference Text.

E Overview of LLMs in the experiment

Model	Size	Seq_len	Access	Url
GPT-4o	N/A	8192	API	https://platform.openai.com/docs/overview
GPT-4	N/A	8192	API	https://platform.openai.com/docs/overview
ChatGPT	N/A	4096	API	https://platform.openai.com/docs/overview
Qwen-Chat-7B	7B	8192	Weights	https://huggingface.co/Qwen/Qwen-7B-Chat
Qwen-Chat-14B	14B	8192	Weights	https://huggingface.co/Qwen/Qwen-14B-Chat
DeepSeek-V3	16B	32k	Weights	https://huggingface.co/deepseek-ai/DeepSeek-V3
DeepSeek-R1	32B	8192	Weights	https://huggingface.co/deepseek-ai/DeepSeek-R1
ChatGLM-6B	6B	2048	Weights	https://huggingface.co/THUDM/chatglm-6b
ChatGLM2-6B	6B	8192	Weights	https://huggingface.co/THUDM/chatglm2-6b
ChatGLM3-6B	6B	8192	Weights	https://huggingface.co/THUDM/chatglm3-6b
Baichuan-7B-Base	7B	4096	Weights	https://huggingface.co/baichuan-inc/Baichuan-7B
Baichuan-13B-Base	13B	4096	Weights	https://huggingface.co/baichuan-inc/Baichuan-13B-Base
Baichuan-13B-Chat	13B	4096	Weights	https://huggingface.co/baichuan-inc/Baichuan-13B-Chat
ChatLaw-13B	13B	2048	Weights	https://huggingface.co/pandalla/ChatLaw-13B
ChatLaw-33B	33B	2048	Weights	https://huggingface.co/pandalla/ChatLaw-33B

Table 6: Overview of the Large Language Models (LLMs) evaluated in our experiments.

F Additional Experimental Results

Expert Review Module ↓		Translation Command Module (GPT-4)												
		Translation Execution Module →												
		OpenAI			DeepSeek		Qwen		Baichuan		ChatGLM		ChatLaw	
Series	**Model**	GPT-4o	GPT-4	ChatGPT	V3	R1	14B	7B	13B-C	13B-B	3-6B	2-6B	33B	13B
OpenAI	GPT-4o	**84.78**	**84.42**	**83.11**	**83.84**	**83.56**	**82.72**	**81.43**	**81.98**	**80.64**	**81.12**	**79.49**	**80.17**	**78.52**
	GPT-4	84.53	84.09	82.86	83.57	83.33	82.44	81.18	81.67	80.29	80.83	79.11	79.82	78.14
	ChatGPT	83.79	83.21	81.97	82.63	82.41	81.52	80.46	80.84	79.53	80.11	78.42	78.96	77.47
DeepSeek	V3	84.18	83.57	82.49	83.22	82.98	82.03	80.74	81.26	79.89	80.54	78.78	79.41	77.73
	R1	83.92	83.31	82.26	83.08	82.91	81.84	80.62	81.12	79.76	80.37	78.63	79.28	77.59
Qwen	14B-Chat	83.41	82.84	81.72	82.53	82.27	81.51	80.29	80.68	79.33	80.02	78.21	78.84	77.12
	7B-Chat	82.72	82.13	81.06	81.81	81.49	80.76	80.18	80.03	78.71	79.38	77.57	78.23	76.54
Baichuan	13B-Chat	83.04	82.38	81.33	82.08	81.83	81.02	79.94	80.97	79.02	79.74	77.92	78.51	76.86
	13B-Base	82.11	81.53	80.47	81.24	80.89	80.13	79.04	79.62	79.28	78.81	77.03	77.61	75.91
ChatGLM	3-6B	82.69	82.01	80.94	81.73	81.42	80.61	79.48	80.12	78.84	79.93	77.51	78.14	76.39
	2-6B	81.46	80.82	79.71	80.54	80.23	79.42	78.31	78.96	77.62	78.58	77.79	76.89	75.24
ChatLaw	33B	81.82	81.23	80.09	80.89	80.64	79.84	78.73	79.32	78.01	78.99	77.24	78.96	75.68
	13B	80.13	79.54	78.41	79.19	78.92	78.09	77.02	77.64	76.32	77.29	75.49	77.08	75.93

Table 7: The performance (xCOMET-XL, %) using **GPT-4** as the Translation Command Module agent. Best results in each column are in **bold**. Model abbreviations: 13B-C: Baichuan-13B-Chat, 13B-B: Baichuan-13B-Base, 3-6B: ChatGLM3-6B, 2-6B: ChatGLM2-6B.

Expert Review Module ↓		Translation Command Module (ChatGPT)												
		Translation Execution Module →												
		OpenAI			DeepSeek		Qwen		Baichuan		ChatGLM		ChatLaw	
Series	**Model**	GPT-4o	GPT-4	ChatGPT	V3	R1	14B	7B	13B-C	13B-B	3-6B	2-6B	33B	13B
OpenAI	GPT-4o	**83.67**	**83.34**	**82.12**	**82.89**	**82.63**	**81.82**	**80.53**	**81.06**	**79.71**	**80.24**	**78.61**	**79.31**	**77.68**
	GPT-4	83.42	83.01	81.88	82.57	82.36	81.54	80.22	80.79	79.38	79.92	78.23	78.94	77.29
	ChatGPT	82.69	82.13	81.04	81.68	81.42	80.61	79.56	79.91	78.64	79.23	77.58	78.12	76.66
DeepSeek	V3	83.08	82.51	81.43	82.26	82.04	81.12	79.87	80.34	78.96	79.62	77.93	78.58	76.89
	R1	82.84	82.22	81.18	82.03	81.87	80.94	79.71	80.19	78.82	79.46	77.78	78.43	76.74
Qwen	14B-Chat	82.33	81.76	80.68	81.52	81.24	80.57	79.36	79.78	78.41	79.08	77.34	77.92	76.27
	7B-Chat	81.64	81.07	79.99	80.81	80.53	79.79	79.28	79.12	77.83	78.47	76.69	77.31	75.72
Baichuan	13B-Chat	81.96	81.32	80.27	81.09	80.86	80.04	78.98	80.08	78.13	78.82	77.01	77.63	75.98
	13B-Base	81.03	80.49	79.42	80.17	79.89	79.16	78.08	78.67	78.34	77.89	76.14	76.71	75.06
ChatGLM	3-6B	81.62	80.97	79.91	80.68	80.39	79.62	78.54	79.16	77.91	78.98	76.62	77.23	75.54
	2-6B	80.41	79.78	78.69	79.52	79.21	78.44	77.36	78.01	76.69	77.64	76.87	76.03	74.39
ChatLaw	33B	80.77	80.19	79.08	79.87	79.62	78.86	77.78	78.39	77.07	78.04	76.31	77.98	74.83
	13B	79.09	78.53	77.42	78.18	77.93	77.12	76.08	76.71	75.39	76.36	74.58	76.13	75.02

Table 8: The performance (xCOMET-XL, %) using **ChatGPT** as the Translation Command Module agent. Best results in each column are in **bold**. Model abbreviations: 13B-C: Baichuan-13B-Chat, 13B-B: Baichuan-13B-Base, 3-6B: ChatGLM3-6B, 2-6B: ChatGLM2-6B.

G TransLaw Prompts

Prompt Template for Translation Command Agent ($\mathcal{A}_{\text{Com}}$)

Role: You are a Senior Translation Project Manager with extensive experience in managing large-scale legal judgment translation (English-to-Traditional Chinese) workflows for the Hong Kong Department of Justice. Your responsibility is to orchestrate the entire workflow from input judgment to final translation output.

Task 1: Segmentation Segment the raw English judgment into a sequence of independent sentences.
- **Note 1:** Do NOT split sentences within legal citations (e.g., [2025] HKCFA 8) or legislative references (e.g., Section 9 of the Theft Ordinance (Cap. 210)).
- **Note 2:** Preserve the logical flow. Do not break bullet points if they constitute a single semantic unit.

Task 2: Context Management Maintain a record of previously finalized translations to ensure terminological and logical consistency in the current sentence translation.

Task 3: Iterative Refinement Manage the refinement loop based on expert feedback:
- **Integration:** Upon receiving feedback from the Expert Review Module, integrate the suggestions to update the current translation draft.
- **Termination:** If the feedback is empty or the maximum iteration limit is reached, mark the sentence as Finalized.

Input Data (JSON):
```
{
"task_mode": "{{segmentation|refinement}}",
"source_text": "{{raw_text_or_current_draft}}",
"feedback_log": {{feedback_json_list}}
}
```

Output Format (JSON):
```
{
"status": "success",
"segmented_list": ["Sentence 1", "Sentence 2", ...], // Populated if mode is segmentation
"updated_draft": "Refined translation string..." // Populated if mode is refinement
}
```

Figure 4: Prompt template for Senior Translation Project Manager. It covers segmentation, context memory, iterative feedback integration, and final result aggregation.

Prompt Template for Legal Terminology Agent ($\mathcal{A}_{\text{Term}}$)

Role: You are a HK Legal Terminologist working for the Department of Justice. You specialize in the precise retrieval and disambiguation of Common Law terms.

Task 1: Term Identification Identify all Hong Kong legal terms (including Latin maxims, procedural terms, and Common Law terms of art) in the source sentence.

Task 2: RAG Consultation Consult the "Retrieved Glossary Context" (RAG) derived from the Combined Department of Justice Glossaries of Legal Terms (https://www.glossary.doj.gov.hk/).

Task 3: Official Selection Select the strictly official Traditional Chinese translation used in Hong Kong courts.

Task 4: Polysemy Disambiguation Disambiguate polysemous terms (e.g., distinguishing "consideration" in contract law vs. general usage) based on the context.

Input Data:
- Source Sentence (s_i): `{{s_i}}`
- RAG Context: `{{Retrieved_Glossary_Entries}}`

Output Format (JSON):
```
{
"identified_terms": [
{
"src": "breach of statutory duty",
"tgt": "違反法定責任",
}
]
}
```

Figure 5: Prompt template for HK Legal Terminologist, it ensures initial terms are retrieved from official glossary.

Prompt Template for Sentence Translation Agent ($\mathcal{A}_{\text{Trans}}$)

Role: You are a Senior Court Translator with over 30 years of experience in the Hong Kong Judiciary. You are an expert in translating High Court judgments from English to Traditional Chinese.

Task 1: Term Integration and Translation Generate a sentence-level translation by seamlessly injecting the mandatory terms into the target syntactic structure.

- **Strict Constraint:** You MUST use the exact translations provided in the `Term_List`. Do not paraphrase or modify these verified legal expressions.
- **Syntactic Fluidity:** Ensure the resulting sentence structure is natural and grammatically correct in Traditional Chinese while accommodating the fixed terms.

Task 2: Contextual Coherence Ensure judicial coherence by referencing the `Global_Context` (derived from preceding case facts and procedural history). The translation must maintain logical flow and consistent tone with previous sentences.

Task 3: Iterative Refinement (If Feedback Exists) If `Review_Feedback` is provided, modify the current draft to specifically address the identified errors (e.g., semantic misalignment, citation format) while preserving the correct parts.

Input Data (JSON):
```
{
"source_sentence": "{{text}}",
"term_list": {{json_list_from_term_agent}},
"global_context": "{{summary_of_preceding_text}}",
"review_feedback": "{{optional_feedback_from_reviewers}}"
}
```

Output Format (JSON):
```
{
"translation_draft": "The draft translated sentence to be submitted to the Expert Review Module."
}
```

Figure 6: Prompt template for Senior Court Translator, it integrates mandatory terminology into the target syntactic structure to generate the translation candidate.

Prompt Template for Semantic Alignment Agent ($\mathcal{A}_{\text{Align}}$)

Role: You are a Senior Legal Reviser responsible for comparing the Source Sentence and the Translation Candidate to ensure accuracy.

Task: Error Detection Check for the following errors:

- **Omission:** Is any key legal fact, date, or condition missing?
- **Hallucination:** Is there any added information not present in the source?
- **Logic:** Are subject-object relationships reversed (e.g., Plaintiff vs. Defendant)?

Input Data (JSON):
```
{
"source": "{{s_i}}",
"candidate": "{{draft}}"
}
```

Output Format (JSON):
```
{
"status": "PASS", // or "FAIL"
"feedback": "Subject 'Appellant' was wrongly translated as 'Respondent'." // null if PASS
}
```

Figure 7: Prompt template for Senior Legal Reviser, it performs a bilingual check to detect semantic errors or omissions.

Prompt Template for Legal Term Review Agent ($\mathcal{A}_{\text{TermR}}$)

Role: You are the Chief Terminologist responsible for validating that all terms in the candidate strictly match the official "Combined DOJ Glossaries".

Task: Glossary Verification
- **Strict Compliance:** Ensure no layman terms are used (e.g., use "合約" not "合同").
- **Consistency:** Ensure the same term is not translated differently within the same context.

Input Data (JSON):
```
{
"candidate": "{{draft}}",
"glossary_db": "{{db_access}}"
}
```

Output Format (JSON):
```
{
"status": "FAIL",
"feedback": "Term 'consideration' must be translated as '代價', not '考慮'."
}
```

Figure 8: Prompt template for Chief Terminologist, it validates lexical accuracy against the authoritative government glossary.

Prompt Template for Legal Citation Agent ($\mathcal{A}_{\text{Cita}}$)

Role: You are a Professional Legal Editor specializing in HKCFA judgment citation standards.

Task: Format Validation
- **Case References:** Check italicization and brackets. (e.g., HKSAR v. Chan [2025] HKCFA 8).
- **Legislation:** Ensure correct use of "Cap." and "Section". (e.g., Theft Ordinance (Cap. 210)).

Input Data (JSON):
```
{
"candidate": "{{draft}}"
}
```

Output Format (JSON):
```
{
"status": "PASS",
"feedback": null
}
```

Figure 9: Prompt template for Professional Legal Editor, it enforces rigid citation formatting standards for cases and legislation.

Prompt for Stylistic Fidelity Polishing Agent ($\mathcal{A}_{\text{StyleP}}$)

Role: You are a Senior Judicial Editor responsible for refining the text to uphold the authority of the Court.

Task: Tone and Fluency
- **Judicial Tone:** The text must sound authoritative and dignified, exactly like a High Court judgment (e.g., use "陳詞" for "submit", "裁定" for "hold").
- **Fluency:** Fix choppy sentences to ensure the legal arguments connect logically and read smoothly in Traditional Chinese.

Input Data (JSON):
```
{
"candidate": "{{draft}}"
}
```

Output Format (JSON):
```
{
"final_text": "The official Traditional Chinese judgment text awaiting publication"
}
```

Figure 10: Prompt template for Senior Judicial Editor, it refines the linguistic output to ensure judicial authority and fluency.

H Evaluation Guidelines for Human Experts

Evaluation Guidelines for Human Experts

General Instructions: Evaluators must give each translation a score between 0 and 10. Evaluators are provided with the source text, the "gold translation" (official court translation), and the predicted translation.

Score Criteria: Scores shall reflect the completeness and accuracy of the predicted translation.

Point Deduction System:

- **Score 10:** A perfectly complete and accurate translation.
- **-1 Point:** Relevant legal term translated in an unusual but correct manner; minor stylistic lapse.
- **-2 Points:** Legally relevant term translated erroneously; Relevant term missing.
- **-4 Points:** Critical errors (e.g., referencing the wrong statute or offense).

Examples:

1) **Score 10**: Perfect match in legal terminology and meaning.
   Source: Mr. Hu referred to a number of decided cases involving the offence of wounding with intent to support his contention that, in the circumstances of the present case, 3 years' imprisonment was manifestly excessive.
   Gold: 胡大律師引用多宗有關「有意圖而傷人」的案件，指以本案案情而言，3 年監禁刑期是明顯過重。
   Prediction: 胡大律師引用多宗有關「有意圖而傷人」的案件，指以本案案情而言，3 年監禁刑期是明顯過重。

2) **Score 9**: (-1 for unusual translation of "manifestly excessive")
   Prediction: 胡大律師引用多宗有關「有意圖而傷人」的案件，指以本案案情而言，3 年監禁刑期實在太多了。
   Note: "Too much" (實在太多了) is colloquial; standard term is "manifestly excessive" (明顯過重).

3) **Score 6**: (-4 for critical error in legal offense)
   Prediction: 胡大律師引用多宗有關誤殺的案件，指以本案案情而言，3 年監禁刑期是明顯過重。
   Note: Critical error: "Wounding with intent" mistranslated as "Manslaughter" (誤殺).

Figure 11: The detailed evaluation guidelines provided to human experts.