Bridging Legal Language Gaps in Hong Kong: A Multi-Agent Framework for Context-Aware Translation of Judgments

Anonymous ACL submission

Abstract

Multi-agent systems empowered by large language models (LLMs) have demonstrated remarkable capabilities in a wide range of downstream applications, including machine translation. However, translating Hong Kong legal judgments remains an exceptionally challenging task due to its intricate legal lexicon, culturally embedded nuances, and complex linguistic structures. In this work, we introduce TAPAGENTS, a novel multi-agent translation system inspired by real-world case law translation workflow. TAPAGENTS employs specialized agents — Translator, Annotator, and Proofreader — to collaboratively produce translations that are Accuracy in Legal Meaning, Appropriateness in Style, and Coherence and Cohesion in Structure. Our system supports customizable LLM configurations and achieves 3,972× cost reduction compared to professional human services. Evaluations show TAPAGENTS surpasses ChatGPT-4o in legal semantic accuracy, structural coherence, and stylistic fidelity, yet trails human experts in contextualizing complex terminology and stylistic naturalness.Our live demo website is available at [1]. Our demonstration video is available at [2].

1 Introduction

The translation of Hong Kong judicial judgments constitutes a pivotal component in sustaining the territory's bilingual legal framework operating in both Chinese and English (Cheng and He, 2016). Since the 1997 handover, Hong Kong has confronted persistent challenges in reconciling linguistic transformation within its judicial system while preserving its inherited legal infrastructure (Chen, 2002). The foundation of this bilingual legal architecture traces back to the 1987 Bilingual Laws Project – a landmark initiative that not only systematized the translation of existing statutes into Chinese (Jones Jr, 1987) but also institutionalized parallel legislative drafting in both official languages (Mushkat, 1997). Nevertheless, judicial practice reveals that English remained the predominant courtroom language in higher courts throughout the postcolonial transitional period (Daniels et al., 2011). Through the progressive localization of legal institutions (Tam, 2012), judgment translation has evolved into an essential mechanism ensuring jurisprudential precision (Prieto Ramos, 2014) and facilitating cross-jurisdictional legal communication (Lin et al., 2023). Confronted with the voluminous corpus of common law documentation within Hong Kong's judicial system (Hau, 2019), the establishment of efficient, accurate, and large-scale translation processes assumes critical significance (Sin et al., 2025).

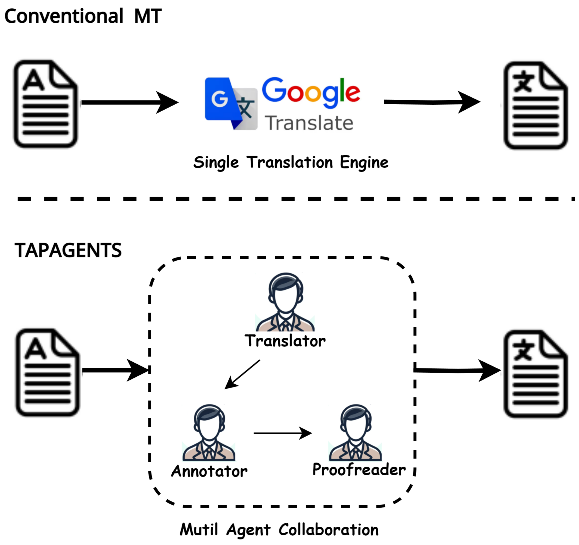


Figure 1: Compared to conventional machine translation (MT) systems that utilize a single MT engine, TAPAGENTSleverages the collaboration among multiple language agents, each powered by large language models (LLMs), for translation.

Consequently, to address the aforementioned

[1]

[2]

challenges and inspired by multi-agent systems (Durante et al., 2024; Tao et al., 2025; Yu et al., 2025; He et al., 2025) and real-world case law translation workflows, we propose TAPAGENTS (as shown in Figure 1). Similar to human translation studios, TAPAGENTS functions as a virtual multi-agent translation system. It mitigates challenges in generating high-quality translations through process decomposition and collaborative specialization. Specifically, each agent in TAPAGENTS manages discrete translation phases, aimed at producing translations comparable to human translators in accuracy and naturalness. Each of our agents plays a specialized role, including Translator, Annotator, and Proofreader. Together, these agents replicate the traditional human translation judgment process, delivering translations that are accurate in Legal Meaning, Appropriateness in Style, and of Coherence and Cohesion in Structure. Finally, we evaluate TAPAGENTS alongside other state-of-the-art translation systems using our proposed judicial judgment test dataset [3]. Our experimental results show that, despite TAPAGENTS higher XCOMET-XL scores and surpasses ChatGPT-4o in legal semantic accuracy, stylistic fidelity, and structural coherence, yet trails human legal experts in contextualizing complex terminology and stylistic naturalness.

2 Related Work

Large Language Models Large Language Models(LLMs) have revolutionized not only the field of natural language processing (NLP) but the entire Artificial Intelligence (AI). They are typically pre-trained on massive text data, so as to learn to predict the next word in a sentence (Brown et al., 2020; Chowdhery et al., 2022; Fan, 2023; Team et al., 2023; Touvron et al., 2023; Bai et al., 2023; Anil et al., 2023). After pre-training, they are fine-tuned with instructions, through a process known as Supervised Fine Tuning (SFT) or Instruction Tuning (IT), so as to turn their capacity of language understanding into capability of following and executing human instructions (Sanh et al., 2021; Wei et al., 2021; Tay, 2023; Longpre et al., 2023; Shen et al., 2023; Chung et al., 2024; Wang et al., 2024). Additionally, the performance of these models can be further enhanced by Reinforcement Learning from Human Feedback (RLHF), an approach to fine-tuning using feedback from humans or other large language models for rating the quality of model outputs (Ouyang et al., 2023; Hejna et al., 2023; Rafailov et al., 2024; Ethayarajh et al., 2024; Hong, 2024).

Multi Agent Systems Multi Agent Systems (MAS) emphasize effective communication and interaction among agents with unique characteristics and the process of their collective decision-making. Multiple autonomous agents handle more dynamic and complex tasks through communication and collaboration with one another while maintaining their own unique strategies and behaviors (Guo et al., 2024). Recent research has shown promising results of this approach in various fields such as software development (Hong et al., 2023), multi-robot collaboration (Mandi et al., 2024), scientific experiments (Du et al., 2023), and scientific debates (Xiong et al., 2023). Additionally, LLM-based multi-agent systems (LLM-MAS) play a crucial role in world simulation for social sciences, gaming, psychology, economics, and policymaking, (re)enacting various roles and perspectives through agents' role-playing (Park et al., 2022, 2023; Xu et al., 2023; Li et al., 2023; Mukobi et al., 2023; Liang et al., 2023)

Judicial Judgments Machine Translation Prior research on machine translation (MT) applications for judicial judgments has achieved only partial success, constrained by persistent challenges in managing domain-specific complexities—particularly the nuanced handling of legal terminology. Statistical machine translation (SMT) frameworks, for instance, have proven inadequate for translating specialized lexicons, as evidenced by systematic errors in Spanish Supreme Court summary translations (Farzindar and Lapalme, 2009). Neural machine translation (NMT) architectures, while advancing general-domain performance, exhibit critical shortcomings when processing the intricate logical scaffolding of judicial reasoning and syntactic structures unique to legal discourse (Killman, 2014). These limitations reveal fundamental gaps in conventional MT paradigms' capacity to address the semantic precision and rhetorical conventions required for authoritative legal texts.

Emerging studies highlight the transformative potential of large language models (LLMs) in legal domain, with ChatGPT demonstrating cross-task adaptability including multilingual translation (Elshin et al., 2024; Eschbach-Dymanus et al., 2024; Ji et al., 2024; Lee et al., 2025). While

[3]

foundational work has mapped its general translation capabilities (Hendy et al., 2023; Kudo et al., 2024; Feng et al., 2024), scholarly attention has increasingly focused on legal translation scenarios characterized by terminological density, jurisdictional logic variations, and cross-cultural conceptual asymmetries. Preliminary evaluations by (Briva-Iglesias et al., 2024) indicate that CHATGPT-4 achieves measurable improvements in contextual disambiguation for multilingual legal instruments—including contractual provisions and transnational treaties—through enhanced semantic parsing architectures. Nevertheless, empirical analyses reveal persistent deficiencies in its treatment of hyper-specialized legal nomenclature and inconsistencies in reconstructing the intricate logical progression of judicial ratio decidendi.

Ours In this work, we introduce TAPAGENTS, a novel multi-agent framework that harnesses collaborative efforts among agents for Hong Kong judicial judgments translation. These language agents are powered by the latest state-of-the-art LLMs.

3 TAPAGENTS

We have established a virtual professional studio of MAS for Hong Kong legal judgment translation and proofreading. Its overall architecture is given in Figure 1. The roles of its three agents are Translator, Annotator, and Proofreader. Following these typical roles in translation, we call it TAPAGENTS, or simply TAP.

This MAS simulates the entire translation process of a judgment (or any text), with these agents in different roles co-working together to ensure the quality and consistency of the final product throughout the whole translation process. In the following subsections, we will present the roles (Section 3.1) and core collaboration strategies of its agents (Section 3.2), and its workflow (Section 3.3) to carry out translation tasks.

3.1 Roles of Agents

To simulate the entire translation process of a judgment, the three agents in TAP take various roles as follows, according to each one's responsibilities.

1. **Translator:** Responsible for accurate translation of the judgment from English to traditional Chinese, ensuring the preservation of its legal meanings, terminology, and the tone of the judgment, ensuring the accuracy and completeness of the translation according to the context and background, and also ensure consistency in legal terminology throughout the translation process so as to avoid confusion or misunderstanding.

2. **Annotator:** Responsible for marking errors in the Translator's translation according to the multi-level translation evaluation annotation standard (Proofread Codes, see Appendix Table 1). The errors to be annotated include but are not limited to the following types: (1) Accuracy errors; (2) Grammatical errors; (3) Usage and style errors. The Annotator's role is to provide detailed error annotations and modification suggestions to the Proofreader.

3. **Proofreader:** Responsible for correcting and revising the initial translation from the Translator according to the Annotator's error annotations, conducting the final review, and finalizing the translation.

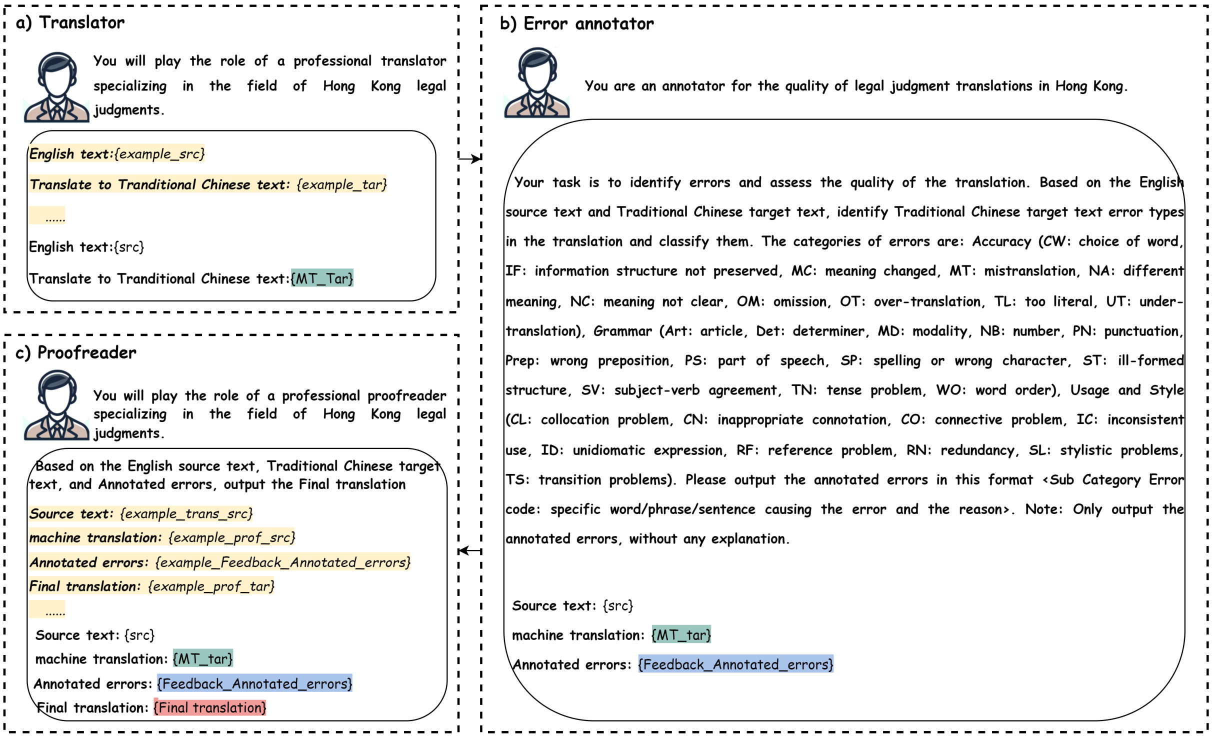
Through the collaborative work of the three agents in these roles, TAP seeks to maximize the accuracy, completeness, and professionalism of judgment translation up to a quality level to meet the rigorous requirements of the legal field. To examine the realism of TAP's translation process simulation, we use GPT-3.5 Turbo as the agent LLM for all three roles.

To ensure that the LLM fully understands the task content, avoids hallucinations, and produces precise and concise outputs, we have carefully formulated respective role prompts for these roles, as presented in Figure 2. We have detailed 30 subcategories of translation error in the prompts for the Annotator and Proofreader, corresponding to the multi-level translation evaluation annotation standard (Proofread Codes) developed by Hong Kong judgment translation experts.

The experiments we carried out to test TAP verified that this approach to error annotation feedback can guide the LLM effectively in correcting mistakes in translation, supporting and inspiring future research in this field.

3.2 Core Strategies of Agent Collaboration

Agent capability acquisition is a key process in LLM-MAS that enables agents to learn and evolve incrementally in a dynamic manner. In TAP, this acquisition process is crucial, ensuring the agents continuously enhance their ability and performance.



a) Translator

You will play the role of a professional translator specializing in the field of Hong Kong legal judgments.

English text:{example_src}

Translate to Tranditional Chinese text: {example_tar}

......

English text:{src}

Translate to Tranditional Chinese text:{MT_Tar}

b) Error annotator

You are an annotator for the quality of legal judgment translations in Hong Kong.

Your task is to identify errors and assess the quality of the translation. Based on the English source text and Traditional Chinese target text, identify Traditional Chinese target text error types in the translation and classify them. The categories of errors are: Accuracy (CW: choice of word, IF: information structure not preserved, MC: meaning changed, MT: mistranslation, NA: different meaning, NC: meaning not clear, OM: omission, OT: over-translation, TL: too literal, UT: under-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: ill-formed structure, SV: subject-verb agreement, TN: tense problem, WO: word order), Usage and Style (CL: collocation problem, CN: inappropriate connotation, CO: connective problem, IC: inconsistent use, ID: unidiomatic expression, RF: reference problem, RN: redundancy, SL: stylistic problems, TS: transition problems). Please output the annotated errors in this format <Sub Category Error code: specific word/phrase/sentence causing the error and the reason>. Note: Only output the annotated errors, without any explanation.

Source text: {src}

machine translation: {MT_tar}

Annotated errors: {Feedback_Annotated_errors}

c) Proofreader

You will play the role of a professional proofreader specializing in the field of Hong Kong legal judgments.

Based on the English source text, Traditional Chinese target text, and Annotated errors, output the Final translation

Source text: {example_trans_src}

machine translation: {example_prof_src}

Annotated errors: {example_Feedback_Annotated_errors}

Final translation: {example_prof_tar}

......

Source text: {src}

machine translation: {MT_tar}

Annotated errors: {Feedback_Annotated_errors}

Final translation: {Final translation}

Figure 2: Illustration of few-shot prompts used in TAPAGENTS (Green/blue/red highlights indicate the outputs of the T/A/P Agent, respectively).

Two fundamental issues need to be handled in this process: one is how the agents receive feedback of various types, and the other is how they adjust themselves accordingly in order to carry out their roles to address complex problems.

Feedback Types There are two basic types of feedback in TAP, as follows. (1) Feedback between agents for collaborative interaction: An agent receives feedback from another as response to or as judgment about its output through communication between agents. This form of feedback promotes cooperation and information sharing among agents, for the purpose of optimizing the overall performance of the whole MAS. (2) Human feedback: This kind of feedback from humans serves the purpose of ensuring what the LLM-MAS in question does or produces aligns well with human knowledge (such as the expertise of experts) and/or preference (such as translation style). This feedback mechanism aims at helping the system understand and meet user needs properly, in hopes of enhancing the accuracy and naturalness of translation.

Self-Adaptation To further enhance translation and proofreading performance, TAP incorporates two self-adaptation strategies. One is a memory module that allows the agents to store and retrieve their interaction records in the past, including translation and proofreading memory, and feedback information. It enables a continuous learning mechanism that allows the agents to improve their performance by utilizing available historical data.

The other is self-evolution, which allows the agents to adjust how they perform their roles via learning from their interactions with humans using feedback or communication logs. This strategy may lead to continuous changes in working methods and subtasks to fulfill the roles of the agents, aiming at further improvement of the overall intelligence and efficiency of the MAS.

3.3 TAPAGENTS Workflow

By virtue of the above strategies, TAP aims at efficient and accurate translation of complex legal documents via a highly collaborative process. This strategy relies on the close cooperation of its three agents, which play specific roles as specified in respective prompts. This multi-layered collaborative approach ensures meticulous handling at each stage and is thereby expected to have a high potential for enhancing the overall translation quality at the system level. The workflow for judgment translation and proofreading in the TAPAGENTS 's System Walkthrough to be detailed below. The user follows these steps to operate the system:

- **Step 1:** Enter API key.
- **Step 2:** Select agents for each of the three roles (Translator, Annotator, Proofreader) from the available options (NiuTrans, GPT-3.5-turbo, GPT-4-turbo, GPT-4).
- **Step 3:** Choose translation direction (default: English to Traditional Chinese).
- **Step 4:** Select terminology database (default: Combined DoJ Glossaries; custom option available).
- **Step 5:** The system performs the execution process in three phases:
  - **Phase 1 (Context-Aware Translation):** The Translator Agent uses GPT-3.5 Turbo with Physical Neighbor Sampling (PNS) to retrieve contextually relevant paragraph pairs for localized translation.
  - **Phase 2 (Error Annotation):** The Annotator Agent tags errors in translations, creating structured <src, ref, err> triplets, which are stored in the Proofreading Memory (PM) database.
  - **Phase 3 (Iterative Refinement):** The Proofreader Agent refines translations by retrieving similar error triplets from the PM and generating revisions based on Proofread Codes.
- **Step 6:** Once the translation is complete, the user can download the translated document, and the final translations and corrections are stored in both Translation Memory and PM databases for continuous improvement.

4 Evaluation

In this section, we report both automated and human evaluation of our TAPAGENTS.

4.1 Automated evaluation

Automated evaluation of an MT system is conducted by applying available authentic automated metrics to compute quality scores for its translation output by contrasting the output with the gold standard answers in a given bilingual text dataset.

Metrics The metrics that we adopted for our evaluation are the following three that have been the most popular in recent years for automated MT evaluation: (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); (2) Unified MT quality evaluation model wmt22-unite-da, a unified MT quality evaluation model (Guttmann et al., 2024).

Experiments for Evaluation **(1) Test Set** We selected the bilingual texts of the judgment "HKSAR - Court of Final Appeal - Final Appeal Criminal Case No. 1 of 2021" from the CFA Judgement Corpus 97-22 dataset as our test data. We may refer to this case as FACC 1/2021 henceforth for brevity. The main reasons for choosing it include its availability and our expert in legal translation ' familiarity with it. Including paragraph-level segmentation and manual alignment, the whole test set consists of 200 paragraph-level source-target pairs. According to the Tokenizer[4], the source text consists of 12,029 tokens (57,926 characters) in English.

(2) Models The LLM we used for all three agents in TAP MAS is GPT-3.5 Turbo. Our HMIT platform integrates two types of MT engines: NMT- and LLM-based ones. The former includes NiuTrans, Google Translate, and DeepL, and the latter GPT-3.5 Turbo(OpenAI, 2023a), GPT-4.0 Turbo,(OpenAI, 2023b) and GPT-4o Turbo (OpenAI, 2024). In the future, we will continue to update and integrate state-of-the-art LLMs for users' choices in our HMIT Platform.

(3) LLM Response Parameter Settings Specifically, TAP is composed of the three agents whose LLM response parameter setting is the same as follows: `Temperature = 0`, `max_tokens = 4,096`, `frequency_penalty = 0`, and `presence_penalty = 0`.

Evaluation Results and Analysis The performance of TAPAGENTS with different role configurations, in terms of K-shot example prompts (if applicable), for its three agents is reported in Table 1. Additionally, the comparative experiments of different NMT- and LLM-based (one-shot) models as Translator Agents are presented in Table 2. These results are based on our evaluation using the bilingual texts of FACC 1/2021 as the test set and XCOMET-XL and Wmt22-unite-da as evalua-

[4] https://platform.openai.com/tokenizer

Table 1: Performance of TAPAGENTS with different configurations for the three agents (T: Translator, A: Annotator, P: Proofreader; X: not used)

MAS	Agent: 0- vs 5-shot			Metric	
	T	A	P	XCOMET-XL	wmt22-unite-da
1	0	X	X	0.2192	0.6172
2	0	X	0	0.7635 (+0.5443)	0.8574 (+0.2402)
3	0	X	5	0.8028 (+0.5836)	0.8662 (+0.2490)
4	0	LLM	0	0.8466 (+0.6274)	0.8664 (+0.2492)
5	0	LLM	5	0.8633 (+0.6441)	0.8726 (+0.2554)
6	5	X	X	0.8381	0.8745
7	5	X	0	0.8330 (-0.0051)	0.8709 (-0.0036)
8	5	X	5	0.8486 (+0.0105)	**0.8749 (+0.0004)**
9	5	LLM	0	0.8435 (+0.0054)	0.8637 (-0.0108)
10	5	LLM	5	**0.8669 (+0.0288)**	0.8732 (-0.0013)
11	5	Manual	Manual	0.8290	0.8662

Table 2: Comparative experiments of different NMT- and LLM-based (one-shot) models as Translator Agents.

	System	**XCOMET-XL**	**wmt22-unite-da**
NMT	NiuTrans	0.7529	0.8450
	GoogleTranslate	0.7162	0.8523
	DeepL	0.8015	0.8573
LLM	GPT-3.5-turbo	0.8077	0.8697
	GPT-4-turbo	0.8176	0.8713
	GPT-4o	0.8410	**0.8775**
	Ours	**0.8467**	0.8688

tion metrics. The configurations can be grouped into two categories for the purpose of comparison, i.e., MAS 1-5 as one and MAS 6-10 as another. In each group, there is a baseline (i.e., the one with the smallest number) and other variations on top of it for possible enhancement. In addition, manual error annotation and Proofread by an expert in legal translation [5] is also brought in to replace the Annotator and Proofreader agent for comparison.

5 Human Evaluation

For human evaluation, we first need to formulate a scoring scheme for use to integrate a human evaluator's scores in various evaluation dimensions into one. The one we have developed specifically for the translation of Hong Kong legal judgments is the legal ACS metric (or simply ACS for brevity), whose formulation will presented in the next subsection, followed by the settings and results of our human evaluation.

5.1 Evaluation Metrics

Aimed at a comprehensive, adequate and reliable evaluation of the translation quality of Hong Kong legal judgments, the ACS metric is formulated as follows,

$$I = \alpha A + \beta C + \gamma S \tag{1}$$

where A, C, and S are the scores in the three key dimensions of evaluation by a human expert evaluator, namely, accuracy of legal meaning, coherence and cohesion in structure, and appropriateness in style, and α, β, and γ are their respective weight coefficients according to the relative importance of these dimensions. 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 the accuracy of legal meaning as the key criterion in determining the quality of legal translation. In Table 2, we further set different weights for evaluation.

5.2 Setup

Due to resource constraints, we randomly selected 10 segments from the FACC 1/2021 test set to evaluate three systems: GPT-4o (baseline), MAS 10 (highest configuration: 5-shot T & P + Annotator), and MAS 11 (manual A & P). The longest segment comprised 234 words (290 tokens/1,432 characters) in English and 414 words (580 tokens/460 characters) in Traditional Chinese. To mitigate evaluator fatigue, we manually split translations into 25

[5] The Second author.

System	A	C	S	ACS 1	ACS 2	ACS 3
GPT-4o	8.91	9.05	9.82	9.04	9.03	9.12
MAS 10	**9.32 (+4.60%)**	9.33 (+3.09%)	9.92 (+1.02%)	**9.39 (+4.85%)**	**9.38 (+4.64%)**	**9.44 (+4.82%)**
MAS 11	9.16 (+2.73%)	**9.36 (+3.43%)**	**9.96 (+1.43%)**	9.30 (+2.27%)	9.28 (+2.78%)	9.36 (+2.73%)

Table 3: Results of human evaluation for the three representative MT systems, with various ACS calculations based on different weightings for A, C, and S. The relative improvement in ACS, A, C, and S is shown in parentheses with a plus sign. ACS 1: .7/.2/.1; ACS 2: .6/.3/.1; ACS 3: .5/.3/.2.

sentence-level pairs (max: 91 EN words/486 characters; 92 words/135 characters) using the `OpenAI Tokenizer`[4]. These were anonymized in evaluation tables (with segment/sentence IDs, source text, and system labels) and assessed by legal translation experts using a 0–10 scale across three dimensions.

5.3 Results and Analysis

Both MAS 10 and MAS 11 using GPT-3.5 Turbo surpass GPT-4o across all three quality dimensions (A: legal accuracy, C: coherence, S: style) and their unified ACS scores. Key findings show that in terms of Legal Accuracy (A), MAS 10 achieves 9.32 (+4.60% vs GPT-4o), outperforming even human-annotated MAS 11 (9.16). For Structural Coherence (C), MAS 11 scores highest at 9.36 (+3.43% vs GPT-4o), with MAS 10 close behind at 9.33. When it comes to ACS Scores, MAS 10 consistently attains the highest values (9.39, 9.38, 9.44) across all weighting schemes (ACS 1–3), demonstrating robustness to metric design, while MAS 11 ranks second, and GPT-4o trails significantly with scores between 9.04 and 9.12. The prioritization of legal accuracy (A weighted 50–70%) amplifies MAS 10's advantage. In terms of Annotation Efficacy, automated annotation (MAS 10) yields superior ACS performance compared to human annotation (MAS 11), with a 0.85–2.58% gap across metrics. These results confirm that MAS 10's architecture optimizes translation quality for Hong Kong legal judgments, even when using a less advanced base LLM (GPT-3.5 Turbo vs GPT-4o).

6 Cost Analysis

The cost of human translation services 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 services. Therefore, translating *FACC 1/2021 [2021] HKCFA3* – a Final Criminal Appeal Case decided by the Court of Final Appeal, which contains 11,585 English words, would cost US$1,390.20.

In contrast, the cost of translating the entire test set using GPT-4o is approximately US$0.39. Using the TapAgents, the cost for translating the entire test set breaks down to approximately US$0.08 (Translator) + US$0.05 (Annotator) + US$0.22 (Proofreader) = US$0.35. Thus, using the TapAgents to translate Hong Kong legal judgments can reduce translation costs by 3,972 times compared to human translation and by 10.26% compared to GPT-4o.[6]

7 Case Study

In this section, we present two case studies from *FACC 1/2021 [2021] HKCFA3* – a Final Criminal Appeal Case test set to demonstrate the superiority of TAPAGENTS.

Accuracy in Legal Meaning As shown in Table 4, this case study examines the translation of the term "subversion of state power" under Article 23 of the National Security Law. The original English text uses "subversion of state power," a critical legal term. The reference translation correctly renders this as "被告人被控顛覆國家政權." However, GPT-4's translation, "被告人被指控顛覆國家權力," introduces a slight deviation by using "權力" (power) instead of "政權" (state power), which may cause ambiguity in legal interpretation. In contrast, the TapAgentstranslation maintains the correct legal meaning with "政權,"

[6] Note that US$0.39 for using GPT-4o is an API cost, and US$0.35 for using our multi-agent translator is also an API cost. As the name suggests, "API cost" refers to the monetary expense associated with using an Application Programming Interface. Such cost does not include the cost for using a human editor to proofread and edit the output translation of an API. The average standard rate for human editing is approximately US$0.04 per word (see e.g., https://www.translationedge.com/pricing). The editing cost for the said judgment would then be US$0.04 x 11,585 words = US$463.30. So the total cost for translating plus editing the judgment would be US$0.35 + US$463 = US$463.35, saving US$926.85, or 3 times the full human translation cost.

Original Text	The defendant is charged with subversion of state power, a crime under Article 23 of the National Security Law.
REFERENCE	被告人被控顛覆國家政權，根據《國家安全法》第23條的規定構成犯罪。
GPT-4o	被告人被指控顛覆國家權力，根據《國家安全法》第23條的犯罪。
TAPAGENTS	被告人被控顛覆國家政權，根據《國家安全法》第23條的規定構成犯罪。

Table 4: Case study for Accuracy in Legal Meaning. The text highlighted in red indicates incorrect translations across different chapters. The text highlighted in blue indicates correct translations.

Original Text	The defendant's actions have severely violated national security, endangering the country's stability and social order.
REFERENCE	被告人的行為已經嚴重違反國家安全，危及國家穩定及社會秩序。
GPT-4o	被告人的行為已經嚴重違背國家安全，威脅國家的安全和公共秩序。
TAPAGENTS	被告人的行為已經嚴重違反國家安全，危及國家穩定及社會秩序。

Table 5: Case study for Appropriateness in Style. The text highlighted in red indicates incorrect translations across different chapters. The text highlighted in blue indicates correct translations.

ensuring accuracy in both legal context and terminology.

Appropriateness in Style As shown in Table 5, the stylistic divergence manifests in register selection and formulaic patterns. GPT-4 adopts "威脅" ("threatening"), a term connoting interpersonal confrontation, which injects subjective urgency ill-suited to legal documentation. Its substitution of "公共秩序" ("public order") further deviates from the canonical "社會秩序" ("social order") enshrined in statutory phrasing. Conversely, TAPAGENTSreplicates the REFERENCE translation's detached bureaucratic syntax ("危及...社會秩序"), employing the clinically precise "危及" ("endangering") to reflect institutional objectivity.

8 Limitations and Future work

The number of evaluated judgments is limited Due to time constraints, we have used only one judgment for this paper as the evaluation set for the system proposed in this paper. If conditions allow in the future, we will use a large-scale set of judgments for further evaluation.

LLM's multi-turn dialogues exhibit hallucination When setting multiple rounds (3 rounds, 5 rounds) of dialogue between the Annotator LLM and the Proofreader LLM for repeated revisions, we found that the meaning of the translation often deviates from the original text after multiple revisions (hallucination phenomenon). The preliminary solution we propose, referencing (Wu et al., 2024), is to add an extra hallucination arbitrator LLM. This part of the work will be addressed in a subsequent paper.

9 Summary

This study proposes a cost-effective and efficient solution to address language disparities within Hong Kong's legal framework, introducing the TAPAGENTS system. The system's seamless coordination of three principal roles—Translator, Annotator, and Proofreader—addresses the intricacies and subtleties of legal texts. The system's exceptional efficacy is substantiated through advanced evaluation metrics such as `XCOMET-XL` and `Wmt22-unite-da`, as well as subjective assessments from domain experts with over three decades of experience. These evaluations underscore the system's superior translation quality relative to human-written references, particularly in legal precision, stylistic relevance, and structural integrity. Additionally, cost analyses reveal that TAPAGENTS delivers a 3,972-× reduction in translation expenses compared to GPT-4o. In sum, TAPAGENTS marks a substantial leap forward in the field of Hong Kong legal judgment translation and proofreading, with significant potential for broader implementation. Future research directions will prioritize the systematic integration of advanced LLMs and the refinement of agent coordination mechanisms to continuously advance the technical frontiers of legal translation.

Acknowledgments

References

Rohan Anil, Andrew M. Dai, Orhan Firat, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, and J. Qwen. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Vicent Briva-Iglesias, João Lucas Cavalheiro Camargo, and Gökhan Dogru. 2024. Large language models "ad referendum": How good are they at machine translation in the legal domain? *arXiv preprint*, arXiv:2402.07681.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

A. 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.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, et al. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Patrick Chung, Calvin T. Fong, Adam M. Walters, et al. 2024. Large language model capabilities in perioperative risk prediction and prognostication. *JAMA Surgery*.

R. J. Daniels, M. J. Trebilcock, and L. 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.

Y. Du, S. Li, A. Torralba, et al. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint*, cs.CL/2305.14325.

Z. Durante, Q. Huang, N. Wake, et al. 2024. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*.

D. Elshin, N. Karpachev, B. Gruzdev, et al. 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.

J. Eschbach-Dymanus, F. Essenberger, B. Buschbeck, et al. 2024. Exploring the effectiveness of llm domain adaptation for business it machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 610–622.

Kawin Ethayarajh, Wei Xu, Niklas Muennighoff, et al. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Angela Fan. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint at*.

Abolfazl Farzindar and Guy Lapalme. 2009. Machine translation of legal information and its evaluation. In *Advances in Artificial Intelligence: 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009, Kelowna, Canada, May 25-27, 2009, Proceedings*, volume 5549 of *Lecture Notes in Computer Science*, pages 64–73. Springer Berlin Heidelberg.

Z. Feng, R. Chen, Y. Zhang, et al. 2024. Ladder: A model-agnostic framework boosting llm-based machine translation to the next level. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15377–15393.

N. M. Guerreiro, R. Rei, D. van Stigt, et al. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Tianyu Guo, Xiaofei Chen, Yi Wang, et al. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

K. Guttmann, M. Pokrywka, A. Charkiewicz, et al. 2024. Chasing comet: Leveraging minimum bayes risk decoding for self-improving machine translation. *arXiv preprint*, arXiv:2405.11937.

B. F. C. Hau. 2019. *The Common Law System in Chinese Context*. Routledge.

J. He, C. Treude, and D. Lo. 2025. Llm-based multi-agent systems for software engineering: Literature review, vision and the road ahead. *ACM Transactions on Software Engineering and Methodology*.

Jakub Hejna, Raphael Rafailov, Harshit Sikchi, et al. 2023. Contrastive preference learning: Learning from human feedback without rl. *arXiv preprint arXiv:2310.13639*.

Ahmed Hendy, Mohamed Abdelrehim, Amr Sharaf, et al. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint*, arXiv:2302.09210.

Dawei Hong. 2024. How much is a "feedback" worth? user engagement and interaction for computer-supported adaptive quizzing. *Interactive Learning Environments*, 32(7):3398–3413.

S. Hong, M. Zhuge, J. Chen, et al. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. *arXiv preprint*, arXiv:2308.00352.

B. Ji, X. Duan, Y. Zhang, et al. 2024. Zero-shot prompting for llm-based machine translation using in-domain target sentences. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

D. 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.

J. Killman. 2014. Vocabulary accuracy of statistical machine translation in the legal context. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 85–98.

Kento Kudo, Haruka Deguchi, Makoto Morishita, et al. 2024. Document-level translation with llm reranking: Team-j at wmt 2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 210–226.

M. Lee, Y. Noh, and S. J. Lee. 2025. A testset for context-aware llm translation in korean-to-english discourse level translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1632–1646.

S. Li, J. Yang, and K. Zhao. 2023. Are you in a masquerade? exploring the behavior and impact of large language model driven social bots in online social networks. *arXiv preprint*, arXiv:2307.10337.

Z. Liang, W. Yu, T. Rajpurohit, et al. 2023. Let gpt be a math tutor: Teaching math word problem solvers with customized exercise generation. *arXiv preprint*, arXiv:2305.14386.

F. Lin, D. Holloway, L. C. Li, et al. 2023. Hong kong as a belt and road initiative dispute resolution hub. In *Hong Kong Professional Services and the Belt and Road Initiative*, pages 105–126. Routledge.

Shayne Longpre, Le Hou, Tu Vu, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Z. Mandi, S. Jain, and S. 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.

G. Mukobi, H. Erlebach, N. Lauffer, et al. 2023. Welfare diplomacy: Benchmarking language model cooperation. *arXiv preprint*, arXiv:2310.08901.

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

OpenAI. 2023a. Gpt-3.5 turbo documentation. Accessed: 2025-01-12.

OpenAI. 2023b. Gpt-4 turbo and gpt-4 documentation. Accessed: 2025-01-12.

OpenAI. 2024. Gpt-4o documentation. Accessed: 2025-01-12.

Shiqing Ouyang, Jie M. Zhang, Mark Harman, et al. 2023. Llm is like a box of chocolates: the non-determinism of chatgpt in code generation. *arXiv preprint arXiv:2308.02828*.

J. S. Park, J. O'Brien, C. J. Cai, et al. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

J. S. Park, L. Popowski, C. Cai, et al. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

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

Raphael Rafailov, Abhishek Sharma, Eric Mitchell, et al. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Victor Sanh, Albert Webson, Colin Raffel, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Tianxing Shen, Ruixiang Jin, Yongwei Huang, et al. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

K. Sin, X. Xuan, C. Kit, et al. 2025. Solving the unsolvable: Translating case law in hong kong. *arXiv preprint arXiv:2501.09444*.

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

W. Tao, Y. Zhou, Y. Wang, et al. 2025. Magis: Llm-based multi-agent framework for github issue resolution. *Advances in Neural Information Processing Systems*, 37:51963–51993.

Jonathan Q. Tay. 2023. Chatgpt and the future of plastic surgery research: evolutionary tool or revolutionary force in academic publishing? *European Journal of Plastic Surgery*, 46(4):643–644.

Google DeepMind Team, Rohan Anil, Sebastian Borgeaud, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yue Wang, Li Wang, Qian Zhou, et al. 2024. Multimodal llm enhanced cross-lingual cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8296–8305.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, et al. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

M. Wu, Y. Yuan, G. Haffari, et al. 2024. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *arXiv preprint arXiv:2405.11804*.

K. Xiong, X. Ding, Y. Cao, et al. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. *arXiv preprint*, arXiv:2305.11595.

Z. Xu, C. Yu, F. Fang, et al. 2023. Language agents with reinforcement learning for strategic play in the werewolf game. *arXiv preprint*, arXiv:2310.18940.

Y. Yu, Z. Yao, H. Li, et al. 2025. 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.

A Appendix

Table 1: Proofread Codes

Error Category	Subcategory	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, e.g., changing the passive to active or vice versa.
	MT	Mistranslation due to inadequate comprehension or misinterpretation of the source text.
	NA	The translation conveys a different meaning from that of the source text.
	NC	Meaning not clear, e.g., because of 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.