Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Legal Aligner: Transforming Generic LLMs into Domain Experts for Enhanced Accuracy

Pengcheng Wen
The Hong Kong University of Science and Technology
Guoying LU
The Hong Kong University of Science and Technology
Sirui Han
siruihan@ust.hk
The Hong Kong University of Science and Technology
Yike GUO
The Hong Kong University of Science and Technology

Research Article

Keywords: Legal Aligner, Domain Adaptation, Legal AI, Factual Accuracy

Posted Date: March 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8868100/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Legal Aligner: Transforming Generic LLMs into Domain Experts for Enhanced Accuracy

First Author[1,2*], Second Author[2,3†] and Third Author[1,2†]

[1*]Department, Organization, Street, City, 100190, State, Country.
[2]Department, Organization, Street, City, 10587, State, Country.
[3]Department, Organization, Street, City, 610101, State, Country.

*Corresponding author(s). E-mail(s): iauthor@gmail.com;
Contributing authors: iiauthor@gmail.com; iiiauthor@gmail.com;
†These authors contributed equally to this work.

Abstract

Large Language Models (LLMs) demonstrate strong general-purpose capabilities but often underperform in specialized domains such as law, where jurisdiction-specific and long-tail knowledge is critical. Existing domain adaptation methods, particularly supervised fine-tuning, are computationally expensive, vulnerable to catastrophic forgetting, and infeasible for proprietary black-box models. We propose **Legal Aligner**, a lightweight, parameter-agnostic post-hoc refinement framework that enhances legal factuality without modifying upstream model weights. Acting as a plug-and-play correction layer, Legal Aligner identifies and rectifies unsupported doctrinal claims, jurisdictional inconsistencies, and citation errors in generated legal responses. We evaluate Legal Aligner on a curated Hong Kong legal consultation dataset across eight upstream LLMs, including both open-source and proprietary models. Results show consistent improvements in factual reliability, with gains of up to 14.9 percentage points in Factuality Coverage Score and a 58.9Our findings demonstrate that output-level alignment offers a scalable and maintainable alternative to parameter-level adaptation for high-stakes legal AI systems, particularly in jurisdictions underrepresented in large-scale pretraining corpora.

Keywords: Legal Aligner, Domain Adaptation, Legal AI, Factual Accuracy

1 Introduction

Large Language Models (LLMs) have rapidly emerged as a transformative technology for the legal domain, with proposed applications ranging from legal research and document analysis to automated legal consultation and decision support (Surden 2019). Unlike general-purpose natural language generation, however, legal tasks are fundamentally normative and authority-dependent: legal conclusions derive their validity not merely from linguistic plausibility, but from their correct grounding in statutes, regulations, and case law within a specific jurisdiction. In this setting, the cost of error is not simply reduced model accuracy, but the potential misrepresentation of legal rights, obligations, and remedies, which may mislead users and produce real-world legal harm (Gururangan et al. 2020).

Despite their impressive fluency, general-purpose LLMs frequently struggle when deployed in legal contexts that demand precise doctrinal distinctions and long-tail jurisdictional knowledge (Kandpal et al. 2023).Prior work has shown that such models may generate outputs that appear legally coherent while omitting decisive statutory constraints, conflating jurisdictions, or fabricating citations.

Foundational models like GPT-4 exhibit remarkable capabilities in broad contexts, they often fail to navigate the intricate long-tail knowledge characteristic of specialized legal systems (Liu et al. 2024). In legal practice, the core challenge is not merely linguistic competence, but the preservation of normative validity and accountability within legal advice systems. Such failures are particularly hazardous because authoritative tone and confident language can obscure epistemic uncertainty, creating an illusion of correctness that may distort professional judgment and compromise due process.

A primary impediment to the deployment of LLMs in legal consultation is the phenomenon of hallucination, which refers to the generation of legally plausible yet factually or doctrinally erroneous content (Huang et al. 2023). This lack of factual consistency is particularly exacerbated in knowledge-intensive scenarios where specific statutes, case precedents, and local regulations are paramount. In legal contexts, hallucinations are not merely factual inaccuracies, but normative failures that may misrepresent legal rights, obligations, or procedural options, thereby creating risks of harmful reliance. For legal practitioners, a model that prioritizes linguistic fluency over factual fidelity is not only unreliable but fundamentally hazardous (Lin et al. 2022). Bridging this expertise gap is therefore essential for the responsible and ethical adoption of Artificial Intelligence within the legal industry.

Despite these risks, existing approaches to improving LLM reliability largely optimize for technical accuracy or surface-level factual consistency, while falling short of addressing the normative, jurisdictional, and accountability requirements that are central to legal consultation. In practice, legal correctness is not solely a matter of retrieving the right facts, but of grounding conclusions in authoritative sources, respecting jurisdictional boundaries, and maintaining traceable responsibility for legal claims (Shui et al. 2023).

Current methodologies for domain adaptation, predominantly Supervised Fine-Tuning (SFT), face substantial hurdles in the legal sector. These resource-intensive approaches require vast quantities of high-quality, expert-annotated legal data and

prohibitive levels of computational power (Ji et al. 2024). Moreover, fine-tuning often risks catastrophic forgetting, where the model loses its general reasoning abilities, and remains inaccessible for closed-source proprietary models (Luo et al. 2023). Consequently, there is an urgent need for an efficient and parameter-free mechanism that can refine model outputs without the extensive costs and technical limitations of full-scale retraining.

To address these limitations, we propose *Legal Aligner*, a lightweight post-hoc alignment framework that enhances the factual reliability of large language models in legal consultation scenarios without requiring access to upstream model parameters.Unlike traditional adaptation approaches, Legal Aligner operates as a post-hoc refinement layer that aligns model outputs with verified legal knowledge bases (Ji et al. 2024). Its parameter-agnostic design supports plug-and-play deployment across heterogeneous upstream models, enabling systematic correction of domain-specific errors and substantial improvements in factual reliability (Gao et al. 2023). Beyond a technical enhancement, we conceptualize output-level alignment as a normative control layer embedded within socio-technical legal systems.In this study, we aim to systematically investigate the efficacy of the Legal Aligner framework by addressing the following two research questions:

RQ1:To what extent can the Legal Aligner framework enhance factual accuracy and mitigate hallucinations in legal tasks?

We evaluate the effectiveness of Legal Aligner through comprehensive benchmarks focused on factuality and helpfulness metrics. Our observations show that Legal Aligner consistently delivers substantial improvements in domain-specific precision across diverse upstream models, with the Factuality Coverage Score increasing by up to 14.9 percentage points. Qualitative analysis further confirms that the framework effectively mitigates hallucinations by correcting legal citations and statutory interpretations. In addition, human preference evaluations show an average win rate of 58.9%, demonstrating that our approach successfully bridges the gap between general linguistic fluency and expert-level legal accuracy.

RQ2: How do computational efficiency and architectural diversity affect the performance of the Aligner paradigm?

We evaluate the robustness of the Aligner paradigm through extensive experiments across 8 diverse upstream architectures, including both open-source and proprietary models. Our observations show that Legal Aligner functions as a highly resource-efficient, parameter-free module that provides stable performance gains without requiring access to core model weights. For generalizability, the results demonstrate that Legal Aligner maintains the models' intrinsic reasoning capabilities, as evidenced by stable scores on the Alpaca Eval 2 benchmark. In addition, this lightweight adaptation offers a scalable alternative to traditional fine-tuning by enabling a single training session to suffice for multiple models without the risk of catastrophic forgetting.

The core contributions of this work are summarized as follows:

1. We introduce a novel, resource-efficient alternative to traditional domain-specific fine-tuning. Operating as a post-hoc refinement layer, Legal Aligner enhances model performance without modifying core parameters, effectively mitigating the risk of catastrophic forgetting.

2. Through extensive evaluations on a specialized Hong Kong legal dataset, we show that Legal Aligner improves factual coverage by up to **14.9 percentage points** and achieves a **58.9% win rate** in human preference evaluations across multiple upstream architectures.
3. We present a scalable, lightweight solution for domain adaptation, compatible with both open-source and proprietary LLMs. This framework enables expert-level legal consultation with significantly lower computational overhead compared to traditional fine-tuning methods.

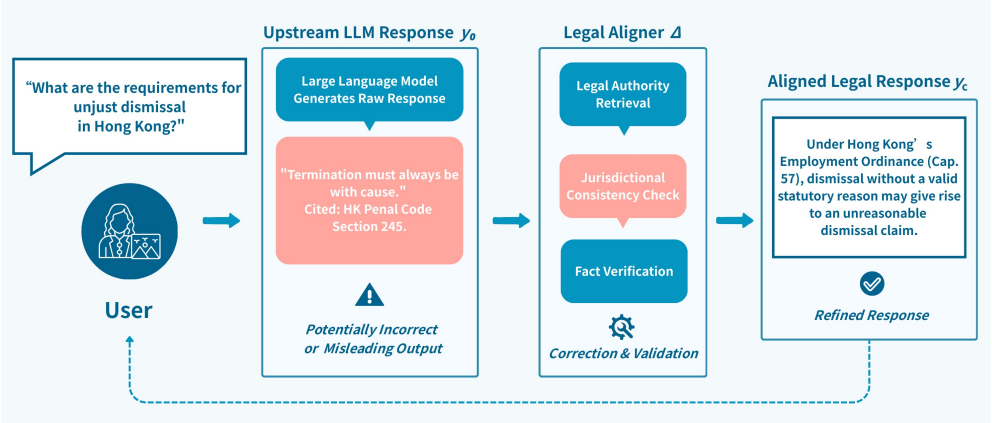


Fig. 1: **Overview of the Legal Aligner Framework**, illustrating the post-hoc refinement workflow of Legal Aligner. Given a legal query and an upstream model's initial response, Legal Aligner performs knowledge-base grounding and verification-based refinement to correct legal inaccuracies and reduce hallucinations, producing a refined response. Performance is evaluated in terms of factual coverage, legal coherence, and human judgments.

2 Related Works

The rapid integration of Large Language Models (LLMs) into the legal sphere has necessitated a multidisciplinary re-examination of how domain-specific expertise can be computationally encoded and verified (Guha et al. 2023). This section provides an exhaustive review of the literature across four pivotal dimensions: the evolution of legal AI evaluation benchmarks, the technical and economic impasses of current domain adaptation methodologies, the limitations of retrieval-based grounding, and the emerging theoretical framework of post-hoc factual alignment (Gururangan et al. 2020). This section reviews prior work along these axes and positions Legal Aligner as a lightweight, output-level refinement approach for high-stakes legal consultation.

2.1 Legal Reliability, Accountability, and Professional Risk

Unlike general-purpose information-seeking tasks, legal consultation operates within a professional and normative framework in which reliability, accountability, and risk mitigation are central concerns (Shui et al. 2023).In legal practice, the quality of an answer is traditionally evaluated through structured reasoning frameworks such as IRAC (Issue, Rule, Application, Conclusion), which require that every legal conclusion be explicitly derived from an applicable rule in response to a defined issue. Legal advice shapes individuals' procedural decisions, assessments of rights and liabilities, and compliance with regulatory obligations. Consequently, even minor doctrinal inaccuracies or misapplied authorities may generate harmful reliance, foreclose legal remedies, or compromise due process. As prior work in AI and Law has emphasized, legal correctness cannot be reduced to surface-level factual accuracy; it depends instead on authoritative grounding, jurisdictional validity, and traceable responsibility for legal claims .

Recent studies have documented that large language models frequently produce legally plausible yet unsound outputs, a phenomenon often described as *legal hallucination* (Dahl et al. 2024). Such failures include fabricated or misattributed citations, conflation of legal doctrines across jurisdictions, and overconfident statements in areas of unsettled law. These issues are particularly problematic in legal contexts because fluent and authoritative language can mask epistemic uncertainty, creating an illusion of correctness that may mislead both lay users and professionals. Empirical analyses of AI-assisted legal research and consultation tools demonstrate that hallucinated citations and unsupported doctrinal claims pose greater risks than generic factual errors, as they provide a false evidentiary basis for legal reasoning.

Beyond accuracy, accountability and auditability are foundational requirements for the responsible deployment of legal AI systems (National Institute of Standards and Technology 2023). Legal practice is governed by professional norms, regulatory oversight, and ethical obligations that demand transparency regarding the sources and limitations of legal advice. Prior work on responsible legal AI highlights the need for outputs to be verifiable, explainable, and attributable to authoritative legal sources, enabling both internal compliance checks and human oversight (Doshi-Velez and Kortz 2017). This shifts the evaluation focus from average-case performance to the mitigation of worst-case failures, emphasizing robustness, provenance, and controllability over purely statistical gains.

From a deployment perspective, these accountability requirements intersect with practical constraints. Many high-performing language models used in legal services are proprietary and accessible only through APIs, precluding parameter-level adaptation or continuous fine-tuning (Lai et al. 2024). At the same time, legal knowledge is inherently dynamic and jurisdiction-specific, making frequent retraining costly and operationally fragile (Atkinson et al. 2020). Consequently, recent work in legal AI increasingly advocates for lightweight, modular mechanisms that can improve output reliability, enforce citation faithfulness, and support post-hoc auditing without modifying upstream model parameters (Sadowski and Chudziak 2025).

Taken together, this body of literature underscores a critical gap between existing reliability-oriented techniques, which often optimize for general factual consistency,

and the normative, jurisdictional, and accountability requirements that define legal consultation. Addressing professional risk in legal AI therefore requires approaches that explicitly target output-level grounding, traceability, and correction, rather than relying solely on improvements in language modeling or aggregate accuracy metrics (Shui et al. 2023; Dahl et al. 2024).

2.2 Legal LLM Evaluation and Legal Reasoning Assessment

A fundamental challenge in deploying LLMs for legal applications is the lack of evaluation protocols that reflect the actual requirements of legal practice. Legal reasoning often involves multi-step argumentation, explicit reliance on authoritative sources, and careful handling of jurisdiction-specific rules, exceptions, and procedural constraints (Kumar and Dadhich 2024). Earlier benchmarks in legal NLP emphasized classification and prediction tasks, such as legal topic classification, judgment prediction, and outcome forecasting (Guha et al. 2023). These tasks enabled reproducible evaluation but did not capture the open-ended nature of consultation-style answers, where failures commonly occur in omissions of key elements, incorrect doctrinal statements, and fabricated or misapplied citations.

More recent benchmark suites attempt to bridge this gap by systematically evaluating LLMs on diverse legal tasks. LawBench (Fei et al. 2023) and LegalBench (Guha et al. 2023) cover statutory interpretation, case comparison, legal entailment, and drafting-related tasks, and they better reflect the need for coherent legal analysis. Complementary evaluations emphasize legal consistency, citation behavior, and robustness to adversarial or long-tail fact patterns.

Across these studies, general-purpose LLMs often show a characteristic failure mode: they generate fluent and plausible legal narratives while missing jurisdictional constraints, collapsing doctrinal distinctions, or producing confident but unsupported statements. Even when fine-tuning improves average scores, it may not eliminate these failures and can introduce sensitivity to prompt phrasing or benchmark artifacts (Cheng et al. 2023).

Another emerging issue is that legal evaluation must address not only accuracy but also normative and procedural dimensions. Responses that appear helpful can still be harmful if they encode bias, recommend improper procedural steps, or overstate certainty when the law is unsettled. As a result, evaluation increasingly combines factuality, legal coherence, and human or expert preference judgments, often emphasizing the need for verifiable grounding rather than mere linguistic plausibility.

2.3 Legal Knowledge Grounding, Retrieval, and Citation Fidelity

Legal consultation is structurally different from general Question & Answering because legal claims ideally link back to authority, including statutes, regulations, and case law, and because the hierarchy of sources matters (Fei et al. 2023). This has driven extensive work in legal information retrieval and legal question answering. Traditional legal information retrieval focuses on retrieving relevant cases and statutes, ranking by authority, and supporting downstream reasoning pipelines. With LLMs,

retrieval-augmented generation (RAG) has become the dominant approach to mitigate hallucinations by providing evidence passages for grounding (Lewis et al. 2020). In legal settings, RAG often requires additional components, such as authority-aware retrieval, temporal awareness for statutory amendments, and filtering for jurisdiction-specific sources.

However, grounding via retrieval does not automatically guarantee reliable outputs. Even with correct retrieval, LLMs may misinterpret holdings, conflate jurisdictions, overgeneralize from dicta, or synthesize an answer that is not entailed by the retrieved text (Magesh et al. 2024). A particularly costly failure mode in legal writing is citation hallucination, where a model fabricates case names, misstates citations, or attributes propositions to authorities that do not support them. Consequently, a growing body of work proposes citation-aware generation, constrained decoding, and post-generation verification to ensure that cited authority truly supports each claim (Lai et al. 2024). These approaches highlight an important distinction relevant to our work: retrieval primarily enhances access to relevant texts, while output-level mechanisms are needed to correct unsupported claims and to enforce citation faithfulness in the final answer.

2.4 Domain Adaptation for Legal LLMs Under Resource, Privacy, and Access Constraints

Given the performance gap revealed by legal benchmarks, domain adaptation has become a central strategy (Ling et al. 2023). Continued pre-training on legal corpora can improve coverage of legal terminology and document style (Gururangan et al. 2020). Domain-specific legal LLMs or instruction-tuned models often demonstrate gains on legal reasoning tasks, drafting, and consultation benchmarks. Despite these gains, large-scale pre-training and full fine-tuning are resource-intensive and difficult to replicate, requiring substantial compute and curated corpora with reliable provenance. Moreover, many high-performing upstream LLMs are proprietary and accessible only via APIs, making parameter updates impossible in practice.

The legal domain introduces additional barriers. High-quality supervision often requires expert annotation, and many realistic legal consultation datasets are constrained by confidentiality, licensing, and privacy regulations (Cui et al. 2023). Jurisdictional fragmentation further complicates adaptation, since legal rules vary across regions and change over time. A globally fine-tuned model may still fail on local doctrines or amended statutes, and maintaining a continually updated fine-tuned model can be operationally costly (Yue et al. 2023). Catastrophic forgetting is an additional concern, as adaptation can degrade general capabilities or safety alignment, which is particularly unacceptable in high-stakes legal contexts (Luo et al. 2023).

Parameter-efficient fine-tuning (PEFT), including LoRA (Hu et al. 2022) and prefix-tuning (Li and Liang 2021), reduces cost by updating a small subset of parameters. While PEFT improves feasibility for open-source models, it still requires parameter access, careful versioning across jurisdictions, and governance processes to manage updates and regressions . These constraints motivate alternatives that can be applied without changing upstream weights and without requiring repeated retraining when the target jurisdiction or policy context shifts.

2.5 Post-hoc Refinement, Verification, and the Aligner Paradigm

A complementary line of work seeks to improve reliability without modifying the base model by refining outputs after generation. Self-correction and critique-based prompting ask a model to review and revise its own responses (Gou et al. 2023). While inexpensive, such methods are not guaranteed to reduce hallucinations, especially when the model lacks access to authoritative evidence or when errors are systematic rather than incidental. Verification-based approaches introduce explicit checking, often at the claim level, to detect unsupported statements and to propose edits (Hong et al. 2024). These approaches align well with legal requirements, where each legal proposition ideally traces back to authority and unsupported claims are unacceptable.

The aligner paradigm (Ji et al. 2024) formalizes output-level refinement through a smaller auxiliary model that edits or rewrites the upstream response. This paradigm is motivated by the idea that correction can be easier than generation and that a dedicated editor can be more efficient and controllable than retraining a general-purpose generator (Yang et al. 2024). Tools such as GenAudit (Krishna et al. 2024) demonstrate the potential of post-hoc auditing and editing for improving factuality in generated text. Decoupled alignment further supports plug-and-play adaptation by separating alignment objectives from base model parameters, enabling modular updates and deployment across heterogeneous upstream models.

In legal tasks, refinement has also been explored at the level of reasoning chains and intermediate steps. LegalReasoner (Shi et al. 2025) introduces stepwise verification for legal judgment prediction, validating reasoning steps against evidence and constraints. This work highlights the value of explicit checks but is often tailored to prediction or structured tasks rather than consultation-style narrative answers. In addition, legal consultation responses pose distinct requirements, such as citation faithfulness, jurisdiction-specific doctrine, and the ability to present legally coherent advice with appropriate caveats. These requirements suggest that an output-level editor designed specifically for consultation may be a practical and impactful complement to retrieval and prompting (Jacovi and Goldberg 2020).

2.6 Risk, Accountability, and Practical Deployment in Legal AI

Unlike general-purpose generative applications, AI-assisted legal consultation operates within a professional architecture necessitating a high standard of *epistemic reliability* and provenance traceability. Legal advice informs critical procedural choices and the substantive assessment of rights and liabilities (Ariai et al. 2025). Consequently, minor doctrinal misinterpretations can precipitate harmful reliance or foreclose legal remedies, especially when fluent outputs create an illusory veneer of jurisprudential legitimacy. Given the strict hierarchy of legal authority, a fabricated citation is more deleterious than an ambiguous response because it provides a false evidentiary basis and undermines adjudicatory integrity (Magesh et al. 2024).

Beyond accuracy, legal AI must satisfy the principles of fairness, accountability, and transparency (FAccT) (Atkinson et al. 2020). Documented concerns regarding the

overconfidence trap, in which assertive language masks epistemic uncertainty, necessitate robust verification mechanisms and human-in-the-loop oversight. Professional regulatory norms mandate that generated content remain auditable to satisfy internal compliance standards (Zhou et al. 2023). This requirement shifts the strategic focus from average-case performance to the mitigation of *worst-case failures*, prioritizing the reduction of hallucinations to facilitate responsible professional stewardship (Dhuliawala et al. 2024).

Operational exigencies further constrain the feasibility of current adaptation methods (Li et al. 2025). Many legal organizations rely on proprietary, closed-source LLMs that preclude parameter-level access and render continuous fine-tuning technically impossible. Even in open-source environments, the fluid nature of legal precedents imposes significant maintenance burdens, as frequent retraining risks version drift and regression. Consequently, there is a pragmatic preference for lightweight, post-hoc refinement layers compatible with black-box architectures. Such modular approaches allow for iterative, domain-specific auditing and integrated safeguards without the prohibitive costs of full-scale model retraining.

Building on these strands, our work targets reliable legal consultation under realistic constraints. Legal Aligner is a lightweight post-hoc refinement module that can be applied to upstream LLM outputs without parameter access (Ji et al. 2024). Compared to RAG, which primarily grounds inputs by retrieving relevant texts, Legal Aligner targets errors that persist after generation, including unsupported doctrinal claims and citation inaccuracies. Compared to task-specific verifiers, Legal Aligner focuses on consultation-style responses, where factual coverage, legal coherence, and faithful reference behavior are central. By operationalizing the aligner paradigm in a jurisdiction-specific legal setting, our approach offers a practical pathway to improve factual accuracy and mitigate hallucinations while remaining compatible with both open-source and proprietary LLMs (Huang et al. 2023).

3 Methodology

The Legal Aligner is designed as a deployment-oriented framework for enhancing epistemic reliability in legal consultation systems. Rather than modifying the internal parameters of upstream language models, the approach adopts an output-level correction paradigm that responds directly to the normative and jurisdiction-specific requirements of legal reasoning.

At its core, the methodology separates linguistic generation from legal verification, thereby addressing the structural mismatch between probabilistic language modeling and authority-dependent legal analysis. The aligner operates as a post-hoc refinement module that reviews, corrects, and strengthens doctrinal grounding in generated responses.

The following subsections formalize the Aligner Paradigm, describe the auxiliary model architecture and dataset construction process, and define the evaluation metrics used to assess factual coverage, citation faithfulness, and practical utility in

high-stakes legal consultation settings. By decoupling domain-specific legal correction from upstream model parameters, this framework enables modular, scalable, and accountable legal AI deployment.

3.1 The Legal Aligner Paradigm

The Legal Aligner is developed under the Aligner Paradigm, which emphasizes correcting an upstream LLM's generated output rather than modifying its internal parameters (Ji et al. 2024). This design choice yields two practical benefits. First, it substantially reduces computational cost compared with full fine-tuning. Second, because the upstream weights remain untouched, it avoids the risk of catastrophic forgetting and preserves the upstream model's general reasoning and instruction-following capabilities (Ngweta et al. 2024). In this way, domain adaptation is decoupled from the base model, while legal precision is introduced through a dedicated refinement layer.

Formally, we denote the Legal Aligner as A_θ, implemented as a comparatively smaller language model (e.g., LLaMA 7B). It is deployed in a plug-and-play fashion and requires no access to the upstream model parameters. Given a user query x and the initial response y_0 produced by an upstream model M_{Upstream}, the aligner analyzes the draft, identifies legal inaccuracies or unsupported statements, and rewrites it into a corrected response that is more faithful to relevant legal facts and citations. The refinement process is summarized as:

$$y_C = A_\theta(x, y_0)$$

where y_C is the final output after post-hoc correction.

This division of labor allows the upstream LLM to focus on producing a fluent first-pass answer, while Legal Aligner specializes in domain-specific verification and correction, resulting in a resource-efficient and scalable approach for legal AI deployment.

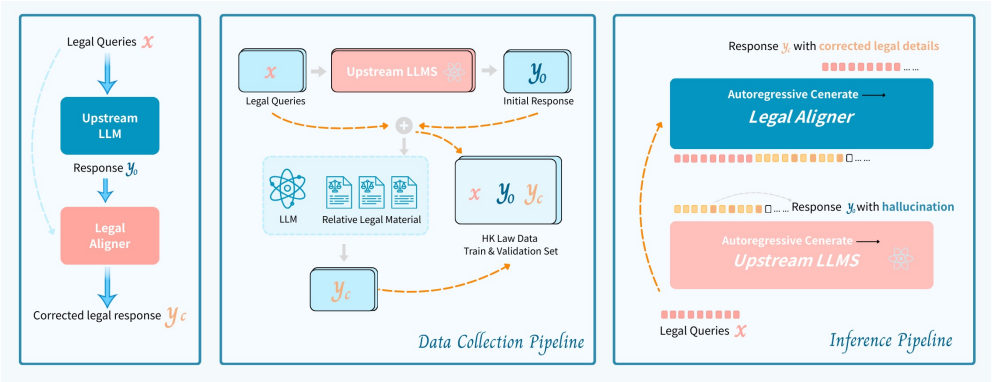


Fig. 2: **Data Collection Pipeline and Inference Pipeline of *Legal Aligner*.** Following Ji et al. (2024), we assembled a comprehensive dataset by collecting legal queries paired with authoritative legal materials, creating robust training and validation sets. The *Legal Aligner*, trained on these specialized resources, functions as a plug-and-play module that augments the inference pipeline without modifying upstream models. During operation, it processes the initial response y_0 from the upstream model, refining legal content and correcting inaccuracies to produce a final output y_c that precisely aligns with legal facts.

3.2 Evaluation Metrics for Legal Consulting Scenarios

To comprehensively assess the effectiveness of the Legal Aligner in knowledge-intensive legal tasks, we employ two critical metrics tailored to legal consultation scenarios: the Factuality Coverage Score (FCS) and the Helpfulness Score (HS). In addition, to examine whether integrating the Legal Aligner alters the upstream model's general instruction-following ability and alignment with human preferences beyond the legal domain, we conduct a general-purpose evaluation using the Alpaca Eval2 methodology (Dubois et al. 2024), reporting the win rate as the primary metric.

3.2.1 Factuality Coverage Score (FCS)

The FCS is designed to measure the extent to which the model's response accurately covers the key legal factual points. In the context of legal consultation, these points include legal knowledge, statutory statements, regulations, and case law precedents.

We first extract the set of key factual points K from the reference answer and the set of key factual points K_M from the model's response. The FCS is calculated as follows:

$$\text{FCS} = \frac{1}{N} \sum_{i=1}^{N} \frac{|K_{M,i} \cap K_i|}{|K_i|} \times 100\%$$

Where N is the total number of samples in the validation set, K_i is the set of key factual points extracted from the reference answer for the i-th query, and $K_{M,i}$ is the set of key factual points extracted from the model's response (Samarinas et al. 2025).

This metric quantifies the model's performance in terms of legal knowledge accuracy and completeness.

3.2.2 Helpfulness Score (HS)

The HS is designed to measure human preference for the model's response, reflecting the overall quality and utility of the output Dubois et al. (2024). In the legal domain, a helpful response must be not only factually accurate but also easy to understand, well-structured, and directive.

We adopt a human-preference-based evaluation method, utilizing a powerful LLM (e.g., GPT-4o) as an evaluator to simulate human preference between the Legal Aligner's corrected response (y_C) and the upstream LLM's original response (y_U).

We define N_W as the number of times y_C is preferred (wins) and N_L as the number of times y_U is preferred (loses). The Helpfulness Score is calculated as:

$$\text{HS} = \frac{N_W - N_L}{N_W + N_L} \times 100\%$$

The score ranges from $[-100\%, 100\%]$. A positive value indicates that the Legal Aligner's corrected response is generally preferred, while a negative value suggests the upstream LLM's original response is preferred.

3.2.3 General Instruction Following

To evaluate whether the integration of the Legal Aligner affects the upstream model's ability to follow general instructions and align with human preferences, we use the Alpaca Eval2 methodology (Dubois et al. 2024). This evaluation compares the performance of the model with and without the Legal Aligner across a variety of general tasks. The win rate is reported as the evaluation metric, where:

- 50% indicates no preference between the aligned and original responses,
- 100% indicates a complete preference for the aligned responses.

This metric ensures that the Legal Aligner does not compromise the model's ability to follow general instructions while enhancing domain-specific performance.

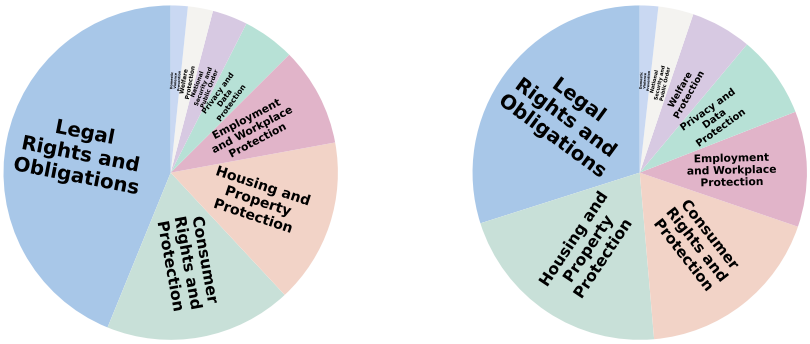
3.3 Dataset Construction and Preprocessing

This study utilizes a specially constructed Hong Kong legal domain dataset to ensure the domain specificity and validity of the evaluation.

1. **Content Collection and Segmentation**: We collected legal materials from The **Community Legal Information Centre** (CLIC,a public legal information platform in Hong Kong that provides accessible explanations of common legal issues) website and segmented them into coherent chunks based on the structure of the document's headings. Each chunk corresponded to a specific legal topic or subtopic.
2. **Training Set Construction**: Using GPT-4o, we generated diverse legal consultation questions (x) for each chunk of legal material. To obtain high-quality answers, we employed six upstream models, including the LLaMA 3.1 series (8B and 70B) and the Qwen 2.5 series (7B, 14B, 32B, and 72B), to generate candidate responses.

We then used Llama-3.3-70B to revise and refine the initial answers to enhance factual accuracy and ensure the responses were clearly grounded in the original legal materials. After quality filtering and deduplication, we constructed a dataset comprising 37,164 question-answer (QA) pairs. A breakdown of the number of QA pairs contributed by each model is provided in Appendix.

3. **Test Set Construction**: For evaluation, we curated the test set separately using the Frequently Asked Questions (FAQ) section from the CLIC website. These real-world legal inquiries reflect actual user concerns and provide a practical benchmark for assessing model performance. The final test set consists of 1,524 QA pairs covering a wide range of legal topics.



(a) Validation Set (b) Train Set

Fig. 3: **Legal Category Distributions in the Validation and Training Sets of our dataset.** Both datasets encompass a diverse range of legal categories, demonstrating varied coverage. While category proportions differ, the overall distributions remain relatively balanced, supporting balanced learning and evaluation across multiple legal domains.

During the training process, the original responses provided by CLIC serve as the Ground Truth to guide A_θ in learning how to refine the upstream LLM's response into a factually accurate and legally compliant answer.

4 Experiment Setup

4.1 Benchmark Models and Aligner Configuration

To assess the effect of model scale in knowledge-intensive settings, **we employ Qwen2.5 (3B, 14B)** (Yang et al. 2024) and **LLaMA 3.1 (8B)** (Grattafiori et al. 2024) as *Legal Aligner* models. These models were specifically selected for their lightweight nature and their ability to perform efficient output-level factual correction

in legal consultation tasks 3. The Legal Aligner was trained on the Hong Kong legal dataset described in Section 2.

For upstream models, we select a diverse range of open-source and closed-source large language models, including Llama 3.1 series (8B, 70B), Qwen2.5 series (7B, 14B, 32B, 72B), as well as closed-source commercial models GPT-4o (OpenAI 2024) and Claude-3.5-Sonnet (Anthropic 2024) This selection allows us to comprehensively assess the adaptability and effectiveness of the *Legal Aligner* across different model sizes and model families 2.

Table 1: Overview of upstream models evaluated in this study.

Model	Version	Model Size	Access	Creator
`GPT-4o`	20240806	undisclosed	API	OpenAI
`Claude-3.5-Sonnet`	20241022	undisclosed	API	Anthropic
`Qwen2.5-7B-Instruct`	–	7.61B	weights	Alibaba
`Qwen2.5-14B-Instruct`	–	14.7B	weights	Alibaba
`Qwen2.5-32B-Instruct`	–	32.5B	weights	Alibaba
`Qwen2.5-72B-Instruct`	–	72.7B	weights	Alibaba
`Llama-3.1-8B-Instruct`	–	8.03B	weights	Meta AI
`Llama-3.1-70B-Instruct`	–	70.6B	weights	Meta AI

Table 2: Training set composition by upstream model.

Model	Number of QA Pairs
Meta-Llama-3.1-8B-Instruct	4,586
Meta-Llama-3.1-70B-Instruct	6,503
Qwen2.5-7B-Instruct	6,910
Qwen2.5-14B-Instruct	6,834
Qwen2.5-32B-Instruct	6,562
Qwen2.5-72B-Instruct	5,769
Total	**37,164**

To ensure a fair comparison, all models, including the upstream LLMs and the Legal Aligner, were trained using the same hyperparameters, with a learning rate of 5×10^{-6} and for 3 epochs.

This configuration allows us to evaluate the performance of the Legal Aligner across a range of model sizes and architectures, ensuring that the improvements in factual accuracy and domain relevance are attributable to the alignment process rather than differences in model scale.

4.2 Evaluation Procedure

We evaluate all models on the validation set using the metrics defined in Section 3.2. For each legal query x, we first see an upstream LLM generate an initial response $y_0 = M_{\text{Upstream}}(x)$. This response is then passed, together with the original query, to the Legal Aligner to produce a refined response $y_C = A_\theta(x, y_0)$.

Evaluation is conducted by directly comparing the baseline response y_0 and the aligned response y_C under identical conditions:

- **Baseline**: The upstream LLM response y_0 is evaluated directly.
- **Aligned**: The response y_C, obtained by applying the Legal Aligner to y_0, is evaluated using the same metrics.

For legal-domain performance, we report the Factuality Coverage Score (FCS) and the Helpfulness Score (HS) for both baseline and aligned responses, and analyze the relative improvements introduced by the Legal Aligner.

In addition, to verify that output-level legal correction does not negatively affect general-purpose instruction following, we evaluate the same models using Alpaca Eval2. This evaluation compares the upstream model with and without the Legal Aligner on a set of general, non-legal tasks, reporting the win rate as an auxiliary metric.

5 Results

The results demonstrate how output-level alignment affects LLM-based legal consultation in the Hong Kong context. Across diverse upstream models, Legal Aligner improves domain-specific performance, particularly in terms of factual coverage and perceived helpfulness 4. To evaluate potential trade-offs, we also assess general instruction-following on a standard benchmark, examining whether post-hoc legal refinement introduces regressions in broader reasoning capabilities.

In addition to aggregate performance metrics, the analysis further investigates structural changes in model outputs, including citation behavior and jurisdiction-specific grounding, in order to clarify the mechanisms underlying the observed improvements.

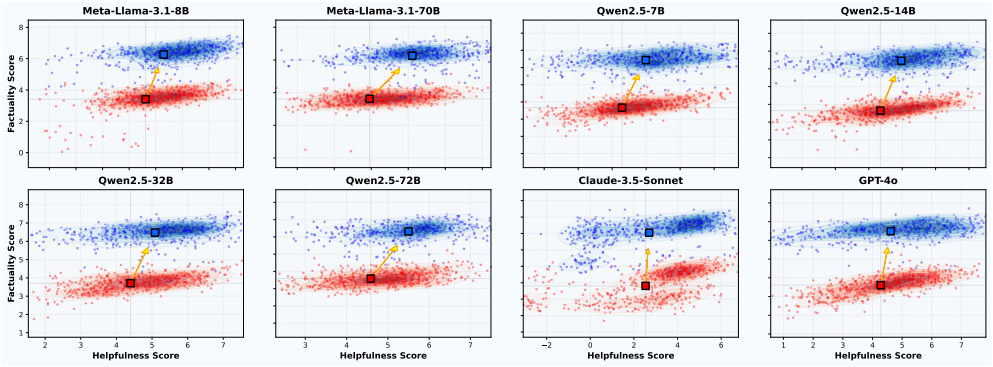


Fig. 4: Instance-level distribution shift in the joint Factuality–Helpfulness space after applying Legal Aligner. Each subplot corresponds to an upstream model evaluated on the HK Law validation set. Red points denote original upstream responses, while blue points denote aligned responses. Squares indicate distribution centroids, and arrows illustrate the mean shift introduced by the aligner. Across models, responses consistently move toward the upper-right region, indicating simultaneous improvements in factual grounding and perceived helpfulness rather than a trade-off between the two dimensions.

5.1 Performance on Legal Consultation

Figure 5 shows that integrating the Legal Aligner yields consistent improvements in both factual coverage and helpfulness across diverse upstream models, regardless of model family and openness. The gains in factual coverage remain substantial even for strong upstream models. In particular, the original study reports improvements ranging from 7.3 to 14.9 percentage points in factual coverage, including the strongest upstream model (Qwen2.5-72B-Instruct).

We also observe a scale-related pattern: as the upstream model size increases, the incremental gain in factual coverage tends to diminish, which suggests that larger upstream models may already contain more long-tail Hong Kong legal knowledge. Nevertheless, the Legal Aligner continues to provide non-trivial factuality gains even at the high end of model scale.

In addition to factuality, the LLM-as-a-judge helpfulness evaluation indicates that all aligned models produce more helpful responses in legal consulting scenarios, reflecting better structure, stronger actionability, and more domain-appropriate grounding of legal claims.

Legal Aligner	Upstream LLM	HK Law Validation set			
		Factuality Coverage Score ↑			Helpfulness ↑
		w/o *Aligner*	w/ *Aligner*	Gain	/
3B	GPT-4o	32.5%	41.2%	↑8.7%	62.8%
	Claude-3.5-Sonnet	37.0%	42.6%	↑5.6%	22.0%
	Qwen2.5-7B-Instruct	26.4%	37.6%	↑11.2%	53.0%
	Llama3.1-8B-Instruct	26.1%	36.4%	↑10.3%	29.7%
	Qwen2.5-14B-Instruct	28.9%	38.7%	↑9.8%	55.1%
	Qwen2.5-32B-Instruct	30.1%	40.0%	↑9.9%	52.5%
	Llama3.1-70B-Instruct	34.6%	41.5%	↑6.9%	23.8%
	Qwen2.5-72B-Instruct	36.8%	44.1%	↑7.3%	52.9%
	Average	31.6%	40.3%	↑8.7%	44.0%
8B	GPT-4o	32.5%	46.7%	↑14.2%	64.5%
	Claude-3.5-Sonnet	37.0%	46.2%	↑9.2%	48.1%
	Qwen2.5-7B-Instruct	26.4%	44.5%	↑18.1%	66.0%
	Llama3.1-8B-Instruct	26.1%	41.2%	↑15.1%	48.2%
	Qwen2.5-14B-Instruct	28.9%	45.9%	↑17.0%	65.7%
	Qwen2.5-32B-Instruct	30.1%	47.0%	↑16.9%	64.7%
	Llama3.1-70B-Instruct	34.6%	46.7%	↑12.1%	45.9%
	Qwen2.5-72B-Instruct	36.8%	50.7%	↑13.9%	55.2%
	Average	31.6%	46.1%	↑14.5%	57.3%
14B	GPT-4o	32.5%	48.3%	↑15.8%	77.8%
	Claude-3.5-Sonnet	37.0%	47.1%	↑10.1%	58.9%
	Qwen2.5-7B-Instruct	26.4%	46.1%	↑19.7%	63.6%
	Llama3.1-8B-Instruct	26.1%	43.4%	↑17.3%	48.6%
	Qwen2.5-14B-Instruct	28.9%	47.8%	↑18.9%	62.0%
	Qwen2.5-32B-Instruct	30.1%	48.6%	↑18.5%	63.7%
	Llama3.1-70B-Instruct	34.6%	47.7%	↑13.1%	42.5%
	Qwen2.5-72B-Instruct	36.8%	51.7%	↑14.9%	54.0%
	Average	31.6%	47.6%	↑16.0%	58.9%

Fig. 5: Performance of Legal Aligner on the HK Law validation set. We report Factuality Coverage Score (FCS) and Helpfulness Score (HS) for multiple upstream LLMs with and without Legal Aligner integration, and show the corresponding performance gains. Results are grouped by the size of the Legal Aligner model (3B, 8B, 14B), demonstrating consistent improvements across both open-source and closed-source upstream models.

5.2 General Instruction Following

To examine whether output-level legal correction compromises general instruction-following and broader human preference alignment, the original work evaluates models with AlpacaEval2 and reports win rate against the corresponding official instruct model as a reference baseline. Win rate is interpreted such that 50% indicates no preference between aligned and original responses.

Overall, the results show that integrating the Legal Aligner maintains or even improves win rate on general instruction-following, suggesting that the domain-specific correction does not degrade, and may sometimes enhance, general-purpose response quality.

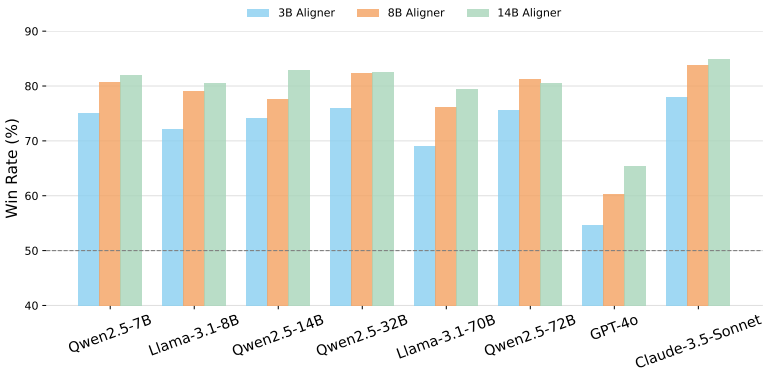


Fig. 6: **Win Rates on the Alpaca Eval2 Benchmark across Different Model Sizes and Aligners.** This bar chart compares the win rates of various language models evaluated with three aligner sizes (3B, 8B, and 14B). Models generally achieve higher win rates as aligner size increases, with Claude-3.5-Sonnet consistently outperforming others and GPT-4o showing lower performance across aligners. Win rates are measured with the official instruct model as the reference. The dashed line at 50% represents the baseline, indicating equal preference between the original and aligned responses.

Figure 6 further indicates a positive relationship between aligner size and win rate across multiple upstream models, with stronger gains generally appearing as the aligner capacity increases.

5.3 Comparison with Fine-tuning

To further validate the effectiveness of our Legal Aligner approach, we conduct an ablation study comparing three distinct strategies for legal domain adaptation: (1) using the upstream base model without any domain adaptation, (2) traditional supervised fine-tuning on legal consultation data, and (3) applying our Legal Aligner to the base model. This comparison directly addresses whether the modular aligner architecture offers advantages over the conventional approach of directly fine-tuning the target model.

Table 4 presents the results across three model scales (3B, 8B, and 14B parameters). The Legal Aligner consistently outperforms both baselines across all model sizes on both factuality coverage and helpfulness metrics. Notably, even when compared against models that have been directly fine-tuned on the same legal consultation

data, the Legal Aligner achieves superior performance. For instance, on the Qwen2.5-14B model, the Legal Aligner improves factual coverage by 9.1 percentage points over fine-tuning (47.8% vs. 38.7%) and helpfulness by 6.04 percentage points (76.51% vs. 70.47%).

These results demonstrate that the Legal Aligner's modular architecture—where a separate, specialized model refines the outputs of a frozen upstream model—not only provides greater flexibility and maintainability but also achieves better performance outcomes than traditional monolithic fine-tuning. We hypothesize that this advantage stems from the aligner's ability to specialize purely in legal correction and refinement without the competing objective of preserving general capabilities, which fine-tuning must balance. Furthermore, the modular design allows the aligner to be trained with a more focused objective, potentially leading to more effective domain-specific adaptation.

6 Analysis

To clarify the mechanisms behind the observed gains, we examine how Legal Aligner alters response characteristics and content structure relative to the upstream outputs. We complement these quantitative analyses with qualitative case studies that illustrate recurring correction patterns, such as fixing unsupported legal assertions, improving citation faithfulness, and strengthening jurisdiction-specific reasoning.

Table 3: Overview of upstream models evaluated in this study.

Model	Version	Model Size	Access	Creator
`GPT-4o`	20240806	undisclosed	API	OpenAI
`Claude-3.5-Sonnet`	20241022	undisclosed	API	Anthropic
`Qwen2.5-7B-Instruct`	–	7.61B	weights	Alibaba
`Qwen2.5-14B-Instruct`	–	14.7B	weights	Alibaba
`Qwen2.5-32B-Instruct`	–	32.5B	weights	Alibaba
`Qwen2.5-72B-Instruct`	–	72.7B	weights	Alibaba
`Llama-3.1-8B-Instruct`	–	8.03B	weights	Meta AI
`Llama-3.1-70B-Instruct`	–	70.6B	weights	Meta AI

6.1 Textual Feature Changes

Beyond aggregate metrics, we examine response-level textual features that are directly tied to the quality of legal consultation. Specifically, we analyze whether the gains in factuality and helpfulness coincide with systematic changes in how answers are structured and grounded, including shifts in content density, citation behavior, and the presence of jurisdiction-relevant legal entities. Consistent trends across models indicate that Legal Aligner does not merely improve scores, but also measurably reshapes outputs toward more legally anchored responses.

Response Length.

Aligned responses are longer on average. Across models, applying Legal Aligner yields an approximate 33.5% increase in response length, indicating that refinement generally adds substantively new legal content rather than merely rephrasing the upstream draft. The additional content is concentrated in legally salient material, including (i) jurisdiction-specific constraints, exceptions, and procedural requirements, (ii) clearer statutory framing and articulation of the governing legal standard, and (iii) more structured, stepwise guidance that enhances actionability in consultation settings.

Legal Citation Behavior.

Alignment also changes citation patterns in a model-dependent manner. For upstream models with weaker Hong Kong legal citation capabilities, the Legal Aligner substantially increases the number of legal citations, which suggests that part of the improvement stems from injecting jurisdiction-specific statutory references and anchoring conclusions to identifiable legal sources. In contrast, for upstream models that already produce frequent citations, the aligner primarily improves the *relevance* and *specificity* of citations by reducing generic or weakly grounded references and emphasizing provisions that directly support the answer.

Legal Entity Mentions.

Finally, aligned responses tend to mention more domain-relevant legal entities (e.g., ordinances, regulatory bodies, or protected groups), reflecting a shift toward more legally structured answers. This feature-level change is consistent with the intended role of the Legal Aligner as an output-level factual corrector that strengthens domain grounding while preserving the upstream model's general language ability.

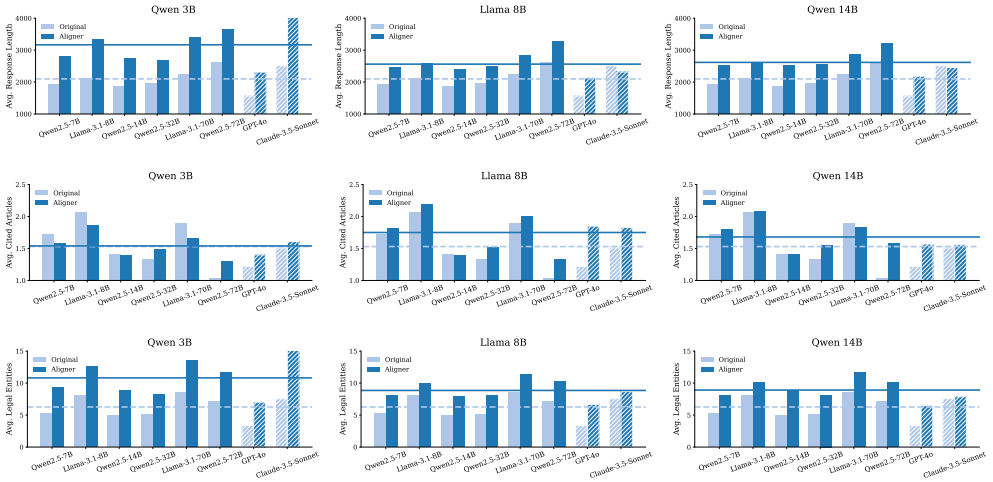


Fig. 7: **Comparison of Average Response Length, Legal Articles Cited, and Legal Entities Mentioned across Models and Aligner Sizes.** The figure presents three metrics (average response length in the top row, average number of legal articles cited in the middle row, and average number of legal entities mentioned in the bottom row) for various language models evaluated with three aligner sizes (3B, 8B, and 14B). Results are shown for Qwen, Llama, and Claude models under both original and aligner conditions, illustrating the impact of aligner size and type on model outputs.

In addition, we conduct an ablation study to compare three domain adaptation strategies: the upstream base model without adaptation, traditional supervised fine-tuning on legal consultation data, and the proposed Legal Aligner applied to the base model 4. Across all evaluated model scales (3B, 8B, and 14B), the Legal Aligner consistently achieves higher scores on both Factuality Coverage Score (FCS) and Helpfulness Score (HS) than either the unadapted base model or the fine-tuned variant. These findings suggest that output-level alignment provides a more effective mechanism for improving legal consultation performance than direct parameter updating under identical training conditions. In each experiment, the aligner size is matched to the corresponding base model, ensuring that performance differences are attributable to the adaptation paradigm rather than model capacity.

6.2 Qualitative Case Studies

To complement the quantitative evaluation, we conduct qualitative case studies to examine how the Legal Aligner modifies upstream model outputs in realistic legal consultation scenarios. Detailed before-and-after comparisons for all cases are provided in the appendix; here, we focus on summarizing the recurring correction patterns and their implications (Ribeiro et al. 2020).

Across representative examples, we observe three dominant correction behaviors: (i) transforming uncertainty or refusal into legally grounded conclusions, (ii) refining

Table 4: **Ablation Study: Comparing Fine-tuning vs. Legal Aligner for Domain Adaptation.** We compare three approaches to legal domain adaptation: (1) the upstream **Base** model without adaptation, (2) traditional **Fine-tuning** on legal data, and (3) our proposed ***Legal Aligner*** applied to the base model. Results show that the Legal Aligner consistently outperforms both baselines across different model sizes (3B, 8B, 14B) on both Factuality Coverage Score and Helpfulness Score. The Legal Aligner size is matched to the base model in each experiment.

Model	**Factuality Coverage Score ↑**			**Helpfulness Score ↑**		
	Base	Fine-tuned	w/ *Legal Aligner*	Base	Fine-tuned	w/ *Legal Aligner*
Qwen2.5-3B	21.3%	30.9%	**34.4%** (w/ Aligner 3B)	-	33.60%	**71.19%** (w/ Aligner 3B)
Llama3.1-8B	26.1%	39.1%	**41.2%** (w/ Aligner 8B)	-	53.54%	**74.80%** (w/ Aligner 8B)
Qwen2.5-14B	28.9%	38.7%	**47.8%** (w/ Aligner 14B)	-	70.47%	**76.51%** (w/ Aligner 14B)

broadly correct but underspecified responses with precise, jurisdiction-specific guidance, and (iii) correcting legally inaccurate statements through explicit statutory grounding.

In Case 1, which concerns the legality of "rabbitfish taxi," a practice involving taxi-sharing by unknown passengers at a fixed fare in Hong Kong, the upstream model failed to provide a definitive answer. It stated that it was unable to verify whether the practice was lawful. After applying the *Legal Aligner*, the response correctly identified the practice as illegal under Hong Kong law and cited the relevant provisions of the Road Traffic (Public Service Vehicles) Regulations (Cap. 374D) (Hong Kong e-Legislation 1997). The aligned response explained that Regulation 37(d) prohibits a taxi driver from permitting persons other than the hirer to enter the vehicle without consent, and Regulation 37(e) further restricts allowing anyone other than the driver and the hirer to enter and pay a fare. It also specified the applicable penalties for contravention, namely a fine of $5,000 and imprisonment for up to six months.

Case 2 illustrates how the *Legal Aligner* improves legal accuracy by refining responses that are partially correct and aligning them with authoritative local guidance. The initial response to the question of whether a prospective employer may collect a copy of an applicant's identity card during a job interview provided a general overview based on the Personal Data (Privacy) Ordinance (Cap. 486) (Hong Kong e-Legislation 1996). Although broadly accurate, it failed to incorporate situation-specific regulatory guidance. The aligned response addressed this gap by referencing the Code of Practice on the Identity Card Number and Other Personal Identifiers issued by the Office of the Privacy Commissioner for Personal Data and summarized in the Community Legal Information Centre. It clarified that employers should not collect a copy of an applicant's identity card unless and until a job offer has been made and accepted. The revised answer also included practical safeguards, such as marking the word "copy" on reproductions of identity cards to reduce the risk of misuse. By grounding the analysis in authoritative Hong Kong regulatory guidance, the *Legal Aligner* enhanced both the precision and reliability of the advice provided.

Case 3 demonstrates the correction of substantively inaccurate legal information. When asked about taxi drivers' obligations to assist passengers with physical disabilities, the upstream model incorrectly stated that there were no strict legal requirements

mandating such assistance in every instance. The *Legal Aligner* rectified this misstatement by citing the Disability Discrimination Ordinance and explaining that discrimination on the grounds of disability is unlawful. The aligned response clarified that drivers are required to provide assistance unless they have a reasonable excuse and must make reasonable accommodations to ensure accessibility. This correction transformed a potentially harmful answer that understated legal protections into an accurate explanation of the rights of disabled passengers under Hong Kong law.

These cases highlight how the *Legal Aligner* significantly improves legal responses in three distinct scenarios: (1) enabling models to provide substantive legal responses when the upstream model completely lacks the necessary domain knowledge, (2) enhancing already knowledgeable responses with more precise legal citations, specific regulatory details, and practical guidance, and (3) correcting substantively inaccurate legal information that could mislead users about their rights and protections under Hong Kong law. This demonstrates the aligner's versatility and effectiveness across different types of legal queries and varying levels of initial model performance(See Appendix A.3 for cases details.).

Taken together, these cases suggest that output-level alignment is particularly effective in legal consultation settings, where uncertainty, partial correctness, or informal guidance can lead to legally misleading outcomes. Rather than altering the reasoning process of the upstream model, the Legal Aligner systematically injects jurisdiction-specific legal constraints at the response level, thereby improving factual reliability while preserving the original response structure.

7 Discussion

7.1 From Factual Correction to Normative Grounding

The empirical results indicate that Legal Aligner improves factual coverage and perceived helpfulness in legal consultation scenarios while preserving general instruction-following capacity. However, in the context of AI and Law, the importance of these gains cannot be reduced to incremental improvements in factual accuracy. Legal correctness is fundamentally justificatory rather than merely descriptive. A legally adequate response must be anchored in authoritative sources, respect jurisdictional boundaries, and remain traceable within institutional structures of responsibility.

The literature on legal AI consistently emphasizes that legal outputs differ from general question answering because they derive validity from normative authority and institutional context rather than from linguistic plausibility alone (Surden 2019).Recent empirical studies demonstrate that large language models frequently generate legally plausible yet substantively incorrect claims, including fabricated or misattributed authorities. These so called legal hallucinations are especially problematic because fluent and confident language may conceal underlying errors and mislead users (Dahl et al. 2024).

Within this framework, the central contribution of the aligner paradigm lies in reframing domain adaptation as normative reinforcement at the output level. Instead of embedding jurisdiction-specific doctrine into upstream parameters, the aligner performs justificatory correction after an initial response is generated. This approach

resonates with the broader alignment literature that conceptualizes correction as a learnable residual editing task (Ji et al. 2024). In legal consultation, this separation is particularly valuable because it targets failure modes that are directly tied to professional risk, including missing statutory constraints, jurisdictional conflation, and unreliable citation behavior.

By separating generation from correction, the system effectively introduces a layered epistemic structure. The upstream model proposes candidate explanations, while the aligner enforces doctrinal coherence and authority-consistent reformulation. In this sense, improvements in factual coverage reflect stronger normative grounding rather than mere informational augmentation.

7.2 Alignment as Governance Architecture

Beyond epistemic considerations, output-level alignment has institutional implications. Legal services operate within regulatory and professional frameworks that require accountability, auditability, and transparency. The role of explanation and institutional responsibility in AI systems has been widely discussed in the accountability literature (Doshi-Velez and Kortz 2017). When AI systems influence legally consequential outcomes, it becomes essential to identify where responsibility is located and how corrective interventions can be implemented.

Traditional fine-tuning integrates domain-specific behavior into internal parameters, which can obscure the boundaries between general reasoning capabilities and jurisdiction-specific adjustments. In contrast, a modular aligner layer can be independently evaluated, updated, and audited. This structural separation creates a clearer locus of normative accountability. Updates to jurisdictional knowledge can be implemented within the aligner without modifying the upstream model. Such modularity aligns with governance expectations that legally sensitive AI systems should be controllable and reviewable.

This architecture also complements retrieval-based grounding methods. Retrieval augmented generation was introduced to reduce hallucinations by conditioning generation on retrieved sources (Lewis et al. 2020). However, empirical studies show that retrieval does not guarantee faithful interpretation of legal authority. Models may still overgeneralize, misinterpret holdings, or produce unsupported synthesis even when relevant materials are available. Output-level correction therefore functions as an additional layer of institutional control that enforces authority-consistent reformulation after generation.

7.3 Structural Limits of Output-Level Alignment

Despite its advantages, output-level alignment cannot resolve all features of legal reasoning. Legal interpretation frequently involves contestable doctrinal positions, conflicting authorities, and normative judgment. While alignment can improve statutory grounding and citation fidelity in cases where applicable rules are clear, it does not fully model interpretive pluralism.

Moreover, any aligner is bounded by the coverage of its domain corpus. Specialized or rapidly evolving areas of law may remain underrepresented. Legal reliability must

therefore be assessed not only through average-case performance but also through stress testing on difficult and ambiguous scenarios. The benchmark literature reinforces this concern. LegalBench and related efforts demonstrate that legal reasoning encompasses diverse task types that extend beyond factual recall or rule articulation (Guha et al. 2023). Evaluation frameworks for legal AI must therefore include jurisdiction selection, authority hierarchy analysis, and adversarial fact patterns. Output-level alignment contributes meaningfully to reliability but does not substitute for comprehensive reasoning frameworks or structured argumentation models.

7.4 Future Directions and Theoretical Extensions

The findings of this study delineate several critical pathways for future scholarly inquiry into the robustness and scalability of the Aligner Paradigm within specialized domains. One primary trajectory involves the transition toward *multi-jurisdictional alignment frameworks*. As legal practice increasingly encompasses cross-border complexities, future research should investigate the dynamic composition of multiple, jurisdiction-specific aligners. This would involve the development of a routing mechanism capable of selecting or synthesizing specialized correction modules based on the geographical and statutory context of the query, thereby facilitating comparative legal analysis and international regulatory compliance (Cheong et al. 2024).

Furthermore, the integration of *Human-in-the-Loop expertise* into the alignment loop remains a vital frontier. While automated rectification significantly enhances factual fidelity, the inherent ambiguity of judicial interpretation necessitates a more nuanced synergy between artificial intelligence and human legal reasoning (Crootof et al. 2023). Future work should explore reinforcement learning from legal expert feedback to calibrate the Aligner's performance in high-stakes or ethically sensitive cases. Such a collaborative framework would ensure that the model's outputs not only adhere to statutory facts but also align with the normative and teleological nuances of professional legal practice (Ouyang et al. 2022).

Another promising avenue lies in the conceptualization of *Adaptive Knowledge Architectures*. Given the fluid nature of case law and the frequent enactment of new statutes, future iterations of the Legal Aligner could adopt modular, incrementally updatable designs. This approach would minimize the maintenance overhead associated with temporal validity by allowing individual knowledge modules to be refreshed without necessitating a full retraining of the auxiliary model (Meng et al. 2022). Finally, the empirical success of the Aligner in the legal domain suggests its potential applicability to other high-stakes environments characterized by strict evidentiary standards, such as medical diagnostics, public policy formulation, and technical safety auditing (Farley and Lansang 2025). Extending this paradigm across diverse sectors would further validate its efficacy as a universal framework for securing the *epistemic reliability* of Large Language Models in the era of autonomous intelligence (Mitchell et al. 2022).

8 Conclusion

This paper introduced the aligner paradigm as a resource-efficient approach to domain adaptation that operates at the level of model outputs rather than internal parameters. Instantiated in the legal domain, Legal Aligner demonstrates that post-hoc corrective mechanisms can substantially enhance factual grounding and doctrinal fidelity in knowledge-intensive legal consultation tasks without requiring access to or modification of upstream model weights.

The central insight of this work is that improving legal reliability does not necessarily require embedding all jurisdiction-specific knowledge into increasingly large foundation models. Instead, reliability can be strengthened through structured output-level correction that reinforces statutory anchoring, jurisdictional consistency, and context-sensitive legal constraints. In high-stakes domains such as law, where professional accountability and normative validity are central, this separation between generative fluency and justificatory correction offers both practical and conceptual advantages.

Empirically, Legal Aligner produces consistent gains across heterogeneous upstream systems, including both open-source and proprietary models. These findings suggest that persistent failure modes in legal consultation are not solely attributable to insufficient model scale, but to systematic gaps in the integration and application of jurisdiction-specific constraints. By targeting these gaps directly, output-level alignment mitigates risks associated with generic framing, omission of decisive local rules, and citation inconsistency, while preserving the communicative strengths of upstream language models.

Beyond empirical performance, this study advances a deployment-oriented perspective on legal AI design. The aligner paradigm promotes modularity, enabling jurisdiction-specific knowledge to be updated independently of rapidly evolving foundation models. It supports portability, allowing domain correction layers to be paired with different upstream architectures. It also enhances governability, since an output-level correction component can be independently audited, stress-tested, and iteratively refined in accordance with institutional and regulatory requirements.

The successful application to Hong Kong law highlights a broader implication. Many jurisdictions and subdomains remain underrepresented in mainstream training corpora, and sustainable localization requires mechanisms that are adaptable and maintainable over time. Treating jurisdiction-specific expertise as a structured corrective layer, rather than as a permanently embedded feature of a base model, offers a pragmatic path toward scalable legal AI in diverse institutional contexts.

Future research should explore multi-jurisdictional alignment architectures capable of dynamically selecting and reconciling applicable legal regimes. Further work is also needed to integrate calibrated uncertainty communication, stronger safeguards against overconfident outputs, and structured human oversight into alignment pipelines. Combining output-level correction with complementary techniques such as retrieval augmentation and structured verification may further enhance transparency and traceability.

In sum, this study demonstrates that output-level alignment can function as a scalable architectural strategy for strengthening normative grounding and factual reliability in legal AI systems. By decoupling generative capacity from jurisdiction-specific correction, the aligner paradigm contributes to a layered model of epistemic governance that is particularly suited to the institutional realities of legal practice.

Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Atkinson, K., Bench-Capon, T., Bollegala, D.: Explanation in ai and law: Past, present and future. Artificial Intelligence **289**, 103387 (2020) https://doi.org/10.1016/j.artint.2020.103387

Ariai, F., Mackenzie, J., Demartini, G.: Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. ACM Computing Surveys **58**(6), 1–37 (2025) https://doi.org/10.1145/3777009

Anthropic: Claude 3.5 Sonnet Model Card Addendum. Available at https://www.anthropic.com/news/claude-3-5-sonnet (2024)

Cheng, D., Huang, S., Wei, F.: Adapting Large Language Models to Domains via Reading Comprehension (2023). https://arxiv.org/abs/2309.09530

Crootof, R., Kaminski, M.E., Price II, W.N.: Humans in the loop. Vanderbilt Law Review **76**(2), 429–510 (2023)

Cui, J., Ning, M., Li, Z., Chen, B., Yan, Y., Li, H., Ling, B., Tian, Y., Yuan, L.: ChatLaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. Preprint at https://arxiv.org/abs/2306.16092 (arXiv:2306.16092) (2023)

Cheong, I., Xia, K., Feng, K.J.K., Chen, Q.Z., Zhang, A.X.: (a)i am not a lawyer, but...: Engaging legal experts towards responsible LLM policies for legal advice. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), pp. 1–22 (2024). https://doi.org/10.1145/3630106.3659048

Dubois, Y., Galambosi, B., Liang, P., Hashimoto, T.B.: Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators (2024). https://arxiv.org/abs/2404.04475

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., Weston, J.: Chain-of-verification reduces hallucination in large language models. In: Findings of the Association for Computational Linguistics: ACL 2024, pp. 212–226 (2024). https://doi.org/10.18653/v1/2024.findings-acl.212

Dahl, M., Magesh, V., Suzgun, M., Ho, D.E.: Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive. Stanford Law School and Stanford HAI (2024). https://law.stanford.edu/2024/01/11/hallucinating-law-legal-mistakes-with-large-language-models-are-pervasive/

Doshi-Velez, F., Kortz, M.: Accountability of AI Under the Law: The Role of Explanation. SSRN Working Paper (2017). https://doi.org/10.2139/ssrn.3064761

Farley, E.A., Lansang, C.R.: Ai auditing: First steps towards the effective regulation of artificial intelligence systems. Harvard Journal of Law & Technology Digest (2025)

Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Zhang, S., Chen, K., Shen, Z., et al.: LawBench: Benchmarking Legal Knowledge of Large Language Models (2023). https://arxiv.org/abs/2309.16289

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The Llama 3 Herd of Models. Technical report, arXiv (2024). https://arxiv.org/abs/2407.21783

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: Adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8342–8360. Association for Computational Linguistics, ??? (2020). https://doi.org/10.18653/v1/2020.acl-main.740

Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., *et al.*: Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. Advances in Neural Information Processing Systems **36**, 44123–44279 (2023)

Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., Chen, W.: CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing (2023). https://arxiv.org/abs/2305.11738

Gao, L., Zhao, V., Lao, N., Lee, H., Juan, D.-C., Guu, K.: Rarr: Researching and revising what language models say, using language models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pp. 16477–16508 (2023). https://doi.org/10.18653/v1/2023.acl-long.910

Hong Kong e-Legislation: Personal Data (Privacy) Ordinance, Chapter 486. Hong Kong e-Legislation. Ordinance governing collection, use and handling of personal data in Hong Kong (1996). https://www.elegislation.gov.hk/hk/cap486

Hong Kong e-Legislation: Road Traffic (Public Service Vehicles) Regulations, Chapter 374D. Hong Kong e-Legislation. Subsidiary legislation under the Road Traffic Ordinance (Cap. 374), relating to regulation of public service vehicles including penalty provisions (1997). https://www.elegislation.gov.hk/hk/cap374D

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations (ICLR) (2022)

Hong, R., Xie, Y., Zhou, M., Wang, X.: A closer look at the self-verification abilities of large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 900–925 (2024). https://doi.org/10.18653/v1/2024.naacl-long.52

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions (2023). https://arxiv.org/abs/2311.05232

Ji, J., Chen, B., Lou, H., Hong, D., Zhang, B., Pan, X., Qiu, T.A., Dai, J., Yang, Y.: Aligner: Efficient Alignment by Learning to Correct (2024). https://arxiv.org/abs/2402.02416

Jacovi, A., Goldberg, Y.: Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4198–4205 (2020). https://doi.org/10.18653/v1/2020.acl-main.386

Kumar, A., Dadhich, H.: Regulatory frameworks for artificial intelligence in law: Ensuring accountability and fairness. NUJS Journal of Regulatory Studies **9**(2), 13–25 (2024)

Kandpal, N., Deng, H., Roberts, A., Wallace, E., Raffel, C.: Large language models struggle to learn long-tail knowledge. In: International Conference on Machine Learning, pp. 15696–15707 (2023). PMLR

Krishna, K., Ramprasad, S., Gupta, P., Wallace, B.C., Lipton, Z.C., Bigham, J.P.: GenAudit: Fixing Factual Errors in Language Model Outputs with Evidence (2024). https://arxiv.org/abs/2402.12566

Li, H., Ai, Q., Chen, J., Dong, Q., Wu, Z., Liu, Y.: Blade: Enhancing black-box large language models with small domain-specific models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 24422–24430 (2025). https://doi.org/10.1609/aaai.v39i23.34620

Lai, J., Gan, W., Wu, J., Qi, Z., Yu, P.S.: Large language models in law: A survey. AI Open **5**, 181–196 (2024) https://doi.org/10.1016/j.aiopen.2024.09.002

Lin, S., Hilton, J., Evans, O.: Truthfulqa: Measuring how models mimic human falsehoods. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pp. 3214–3252 (2022). https://doi.org/10.18653/v1/2022.acl-long.229

Liu, C., Kang, Y., Wang, S., Qing, L., Zhao, F., Sun, C., Kuang, K., Wu, F.: More than Catastrophic Forgetting: Integrating General Capabilities for Domain-Specific LLMs (2024). https://arxiv.org/abs/2405.17830

Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, pp. 4582–4597 (2021). https://doi.org/10.18653/v1/2021.acl-long.353

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474 (2020)

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., Zhang, Y.: An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-Tuning (2023). https://arxiv.org/abs/2308.08747

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X.: Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey (2023). https://arxiv.org/abs/2305.18703

Mitchell, E., Lin, C., Bosselut, A., Finn, C., Manning, C.D.: Fast model editing at scale. In: International Conference on Learning Representations (2022)

Meng, K., Sharma, A., Andonian, A., Belinkov, Y., Bau, D.: Locating and editing factual associations in gpt. In: Advances in Neural Information Processing Systems, vol. 35, pp. 17359–17372 (2022)

Magesh, V., Suzgun, M., Dahl, M., Ho, D.E.: Hallucination-Free? Assessing the Reliability of Leading AI-Driven Legal Research Tools (2024). https://arxiv.org/abs/2405.20362

Ngweta, L., Agarwal, M., Maity, S., Gittens, A., Sun, Y., Yurochkin, M.: Aligners: Decoupling LLMs and Alignment (2024). https://arxiv.org/abs/2403.04224

National Institute of Standards and Technology: Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST Special Publication 1270 (2023). https://doi.org/10.6028/NIST.AI.100-1

OpenAI: GPT-4o System Card. Preprint at urlhttps://arxiv.org/abs/2410.21276 (2024). https://doi.org/10.48550/arXiv.2410.21276

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems, vol. 35, pp. 27730–27744 (2022)

Ribeiro, M.T., Singh, S., Guestrin, C.: Beyond accuracy: Behavioral testing of nlp models with checklist. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4902–4912 (2020). https://doi.org/10.18653/v1/2020.acl-main.442

Sadowski, A., Chudziak, J.A.: On verifiable legal reasoning: A multi-agent framework with formalized knowledge representations. In: Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM 2025), pp. 2535–2545. ACM, ??? (2025). https://doi.org/10.1145/3746252.3761057

Shui, R., Cao, Y., Wang, X., Chua, T.-S.: A comprehensive evaluation of large language models on legal judgment prediction. In: Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 7337–7348. Association for Computational Linguistics, ??? (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.490

Samarinas, C., Krubner, A., Salemi, A., Kim, Y., Zamani, H.: Beyond Factual Accuracy: Evaluating Coverage of Diverse Factual Information in Long-Form Text Generation (2025). https://arxiv.org/abs/2501.03545

Surden, H.: Artificial intelligence and law: An overview. Georgia State University Law Review **35**(4), 1305–1337 (2019)

Shi, W., Zhu, H., Ji, J., Li, M., Zhang, J., Zhang, R., Zhu, J., Xu, J., Han, S., Guo, Y.: Legalreasoner: Step-wised verification-correction for legal judgment reasoning. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, pp. 7297–7313. Association for Computational Linguistics, ??? (2025). https://doi.org/10.18653/v1/2025.acl-long.361

Yue, S., Chen, W., Wang, S., Li, B., Shen, C., Liu, S., Zhou, Y., Xiao, Y., Yun, S., Huang, X.: DISC-LawLLM: Fine-Tuning Large Language Models for Intelligent Legal Services (2023). https://arxiv.org/abs/2309.11325

Yang, K., Liu, Z., Xie, Q., Huang, J., Zhang, T., Ananiadou, S.: MetaAligner: Towards Generalizable Multi-Objective Alignment of Language Models (2024). https://arxiv.org/abs/2403.17141

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al.: Qwen2.5 Technical Report. Technical report, arXiv (2024). https://arxiv.org/abs/2412.15115

Zhou, K., Jurafsky, D., Hashimoto, T.B.: Navigating the grey area: How expressions of

uncertainty and overconfidence affect language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 5506–5524 (2023). https://doi.org/10.18653/v1/2023.emnlp-main.335

A Appendix

A.1 Evaluation with Alternative Judge Model

To address potential concerns regarding judge bias when using GPT-4o as the sole evaluator, we conducted additional experiments employing o4-mini as an alternative judge model. Table 5 presents a comprehensive comparison of preference scores across different upstream models evaluated by both GPT-4o and o4-mini.

Our findings demonstrate that using o4-mini as a judge also supports our main conclusions and findings. While the absolute preference scores obtained from o4-mini are generally lower than those from GPT-4o, the overall trend remains remarkably consistent. Specifically, most preference scores are consistently positive across both judges, indicating that responses corrected through Legal Aligner are indeed superior to the original responses generated by upstream models.

This cross-validation with an alternative judge model strengthens the robustness of our evaluation methodology and confirms that the improvements brought by Legal Aligner are not artifacts of a particular judge's bias.

A.2 Prompt Design

We designed several prompt templates tailored for different subtasks in our evaluation and data generation pipeline, summarized as follows:

- **Choice Question Helpfulness Evaluation:** This prompt instructs the model to focus solely on evaluating the helpfulness of responses, explicitly disregarding factual correctness. Given two responses to the same prompt, the model is asked to compare them and output a standardized judgment format indicating which response is better, with detailed reasoning guiding the comparison. The prompt encourages the model to avoid judging responses as equal unless clearly justified.
- **Law Article Extraction:** This prompt requires the model to extract only those law articles explicitly mentioned or referenced in the given text. The output must be a valid JSON array of strings, containing exactly the mentioned laws without any extra explanation or formatting. Strict rules ensure the output is machine-readable and unambiguous.
- **Answer Coverage Evaluation:** The model is tasked to assess the extent to which each key sentence from a reference legal text is covered by the model's answer. For each key sentence, the model assigns a coverage score between 0 and 1, provides the corresponding content from the answer if present, and a brief analysis. The prompt concludes with an overall average coverage score and a check for irrelevant or inaccurate information.

Legal Aligner	Upstream Model	Preference Score (Win Rate) ↑	
		Judge: GPT-4o	**Judge: o4-mini**
3B	GPT-4o	62.8%	15.0%
	Claude-3.5-Sonnet	22.0%	4.9%
	Qwen2.5-7B-Instruct	53.0%	18.64%
	Llama3.1-8B-Instruct	29.7%	5.9%
	Qwen2.5-14B-Instruct	55.1%	19.8%
	Qwen2.5-32B-Instruct	52.5%	12.7%
	Llama3.1-70B-Instruct	23.8%	-9.5%
	Qwen2.5-72B-Instruct	52.9%	13.2%
8B	GPT-4o	64.5%	33.1%
	Claude-3.5-Sonnet	48.1%	30.0%
	Qwen2.5-7B-Instruct	66.0%	35.9%
	Llama3.1-8B-Instruct	48.2%	19.6%
	Qwen2.5-14B-Instruct	65.7%	32.4%
	Qwen2.5-32B-Instruct	64.7%	32.8%
	Llama3.1-70B-Instruct	45.9%	5.8%
	Qwen2.5-72B-Instruct	55.2%	25.3%
14B	GPT-4o	77.8%	36.9%
	Claude-3.5-Sonnet	58.9%	37.4%
	Qwen2.5-7B-Instruct	63.6%	34.9%
	Llama3.1-8B-Instruct	48.6%	23.5%
	Qwen2.5-14B-Instruct	62.0%	36.7%
	Qwen2.5-32B-Instruct	63.7%	33.5%
	Llama3.1-70B-Instruct	42.5%	6.9%
	Qwen2.5-72B-Instruct	54.0%	25.7%

Table 5: **Evaluation of Human Preference with Different Judges.** Addressing the concern of potential bias in using GPT-4o as the sole judge, we introduced **o4-mini** to cross-validate the results.

- **Key Sentence Extraction:** The model is asked to extract the most important and representative legal sentences from a standard answer. Instructions emphasize completeness, conciseness, accuracy, and avoidance of duplicate or irrelevant information. The extracted key sentences serve as ground truth for coverage evaluation.

Prompt template for generating QA pairs from legal materials

```
OPEN_QUESTION_SYSTEM_PROMPT: str = '''You are a legal expert.
Your task is to extract 1-5 question-answer (QA) pairs from
the given legal material. Each QA pair should reflect the key
concepts, facts, or principles discussed in the content.

Task Objectives:
- Extract question-answer pairs that reflect the core ideas.
- Questions can focus on definitions, legal principles,
  implications, or factual details.
- Answers should be concise, accurate, and directly
  supported by the text.

Input Structure:
[[Title]]: l1 title, l2 title
[[Content]]: Specific legal content

Output Format:
[
    {
        "id": 1,
        "l1": "l1 title",
        "l2": "l2 title",
        "question": "What is contract fraud?",
        "answer": "Contract fraud involves intentional
                  misrepresentation of material facts..."
    }
]
'''

OPEN_QUESTION_USER_PROMPT: str = '''[[Title]]: {l1_title},
{l2_title}

[[Content]]: {content}
'''
```

Prompt template for dataset generation

```
OPEN_QUESTION_SYSTEM_PROMPT: str = '''Based on the legal
materials below, please revise the original answer,
maintaining the original style and framework as much as
possible, only adding, removing, or modifying the legal
facts in the original answer according to the known
information in the legal materials.'''

OPEN_QUESTION_USER_PROMPT: str = '''Legal materials:
{materials}

Question: In Hong Kong, {question}

Original answer: {answer}'''
```

Prompt template for helpfulness evaluation

```
CHOICE_QUESTION_SYSTEM_PROMPT = '''You are a specialist in
the domain of large language models and are presently engaged
in enhancing the helpfulness of these models. We view
helpfulness and factuality as two separate, unrelated
aspects, so we kindly ask you to disregard any
factuality-related elements in this context.'''

CHOICE_QUESTION_USER_PROMPT = '''Next, I will show two
responses from a large language model to the same prompt.
Please help me decide which response is better, but try not
to consider factors related to usefulness. Provide detailed
reasoning first, and then the conclusion at the end. Be
careful about whether the responses are long and redundant.
If response A is better, return - Better: [[responseA]].
If response B is better, return - Better: [[responseB]].
If they are equally useful, return - Better: [[Equal]].
Do not output your reasoning. Instead, if the two are
unequal, output the differences between the responses that
led you to your decision. Judge the two judiciously as
equals, and as often as possible as to determine that they
are unequal. And directly output the answer Better:
[[responseA]], Better: [[responseB]], or Better: [[Equal]].

<Prompt>: {question}
<response 1>: {original_answer}
<response 2>: {correct_answer}'''
```

Prompt template for law article extraction

```
OPEN_QUESTION_SYSTEM_PROMPT = '''You are a legal expert
assistant specialized in Hong Kong law. Your only task is to
extract all law articles explicitly mentioned or referenced
in the given text.

IMPORTANT: You must follow these rules strictly:
1. Only extract law articles that are explicitly mentioned or
   referenced in the text.
2. You must return a valid JSON array containing ONLY the law
   articles as strings.
3. The output must be a valid JSON array, nothing else. No
   explanations, no descriptions, no extra text.
4. If no law articles are mentioned, return an empty array: []
5. The output must start with [ and end with ].
6. Each law article must be enclosed in double quotes.
7. Include the law name and article/section number if
   explicitly mentioned in the text.
8. If only the law name is mentioned without specific
   article/section, just include the law name.

Example of correct output:
[
  "Crimes Ordinance",
  "Theft Ordinance, Section 9",
  "Basic Law, Article 45"
]

Example of incorrect output (DO NOT DO THIS):
- Crimes Ordinance
- Theft Ordinance

Or:
The laws mentioned are:
1. Crimes Ordinance
2. Theft Ordinance

Or:
{"laws": ["Crimes Ordinance", "Theft Ordinance"]}

Remember: Only return a valid JSON array of strings,
containing exactly what is mentioned in the text - no more,
no less.'''

LAW_ARTICLE_EXTRACTION_PROMPT = '''Extract all law articles
explicitly mentioned or referenced in the following text.
Return ONLY a valid JSON array of strings, containing exactly
what is mentioned in the text:

Text: {answer}'''
```

Prompt template for evaluating answer coverage

```
OPEN_QUESTION_SYSTEM_PROMPT: str = '''You are a professional
legal text analysis assistant. Your task is to evaluate
whether the model's answer adequately covers the key
sentences. Evaluation criteria:

1. Assess each key sentence and determine if the model's
   answer includes this key information
2. Provide a coverage score of 0-1:
   - 1: Fully covers the key information
   - 0.5: Partially covers the key information
   - 0: Does not cover the key information
3. Note if the model's answer contains any additional
   inaccurate or irrelevant information

Please output the evaluation in the following format:
Key Sentence 1: [key sentence content]
- Coverage Score: [0-1]
- Corresponding Content: [matching content from model's
  answer, if any]
- Analysis: [brief explanation of the score]

Key Sentence 2: [key sentence content]
...

Overall Assessment:
- Average Coverage: [mean coverage score of all key sentences]
- Additional Information: [Yes/No, explain if yes]
- Analysis: [overall evaluation]'''

OPEN_QUESTION_USER_PROMPT: str = '''Please evaluate if the
model's answer adequately covers the key sentences:

Original Question:
{question}

Key Sentences:
{keys}

Model's Answer:
{answer}'''
```

Prompt template for extracting key sentences from standard answers

```
OPEN_QUESTION_SYSTEM_PROMPT: str = '''You are a professional
legal text analysis assistant. Your task is to extract key
sentences from given legal texts. These key sentences should:
1. Contain the most important legal concepts and provisions
   from the text
2. Each sentence should be complete and self-contained
3. Avoid duplicate information
4. Use clear and concise language
5. Maintain accuracy of the original text

Please output these key sentences in a list format.'''

OPEN_QUESTION_USER_PROMPT: str = '''Please extract key
sentences from the following legal text:

{answer}'''
```

A.3 Cases Details

Table 12: Original Answers and Corresponding Full Revisions by Legal Aligner. This table presents the complete responses corresponding to the three representative examples shown earlier. Each example illustrates the original output from the upstream language model and the improved version generated by Legal Aligner. The original responses are marked as before correction, and the revised responses are labeled as after correction.

Case 1:
Question: Is the practice of taxi-sharing by unknown passengers at a fixed fare (known as "rabbitfish taxi") legal?

Original answer: I am unable to verify if the practice of taxi-sharing by unknown passengers at a fixed fare, known as "rabbitfish taxi," is legal in Hong Kong.

Corrected answer: In Hong Kong, the practice of taxi-sharing by unknown passengers at a fixed fare, known as "rabbitfish taxi," is not legal. According to the Road Traffic (Public Service Vehicles) Regulations (Cap. 374D), taxi drivers are required to follow specific rules. Regulation 37(d) states that a taxi driver shall not, without reasonable excuse, when his taxi is hired, permit any person other than the hirer to enter the taxi without the consent of the hirer. Additionally, Regulation 37(e) states that a taxi driver shall not, without reasonable excuse, when his taxi is hired, allow any person other than the driver and the hirer to enter the taxi and pay the fare, unless the person is a child under the age of 3 years. If a taxi driver contravenes these regulations without reasonable excuse, they commit an offence and are liable on conviction to a fine of $5,000 and to imprisonment for 6 months. Therefore, in Hong Kong, "rabbitfish taxi" is not a permissible practice as it does not comply with the regulations governing taxi operations.

Case 2:
Question: Can a prospective employer record my ID card number or collect a copy of my ID card when I attend a job interview?

Original answer: In Hong Kong, the Employment Ordinance (Cap. 57) regulates the collection and use of personal data by employers. According to the Personal Data (Privacy) Ordinance (Cap. 486), employers are required to comply with the principles of good data practices when collecting and using personal data. In general, it is not necessary for an employer to collect a copy of your ID card or record your ID card number during a job interview. However, there are some exceptions: 1. Pre-employment checks: If the employer is required to conduct pre-employment checks, such as verifying your identity or work experience, they may request to see your ID card or record your ID card number. However, they should only collect the information necessary for the purpose of the check and should not retain it for any other purpose. 2. Statutory requirements: In some cases, employers may be required by law to collect and retain certain personal data, such as your ID card number, for tax or social security purposes. 3. Industry-specific requirements: Certain industries, such as finance or healthcare, may have specific regulations or guidelines that require employers to collect and verify personal data, including ID card numbers. If an employer asks to collect your ID card or record your ID card number during a job interview, you can politely decline and ask why they need this information. If they provide a valid reason, you can consider providing the information. However, if you are unsure or uncomfortable, you can ask if there are alternative ways to verify your identity or work experience. It's also worth noting that you have the right to request access to and correction of your personal data held by the employer under the Personal Data (Privacy) Ordinance.

Corrected answer: In Hong Kong, the Code on the topic of PERSONAL DATA PRIVACY from CLIC regulates the collection and use of personal data by employers. According to the Code, employers are required to comply with the principles of good data practices when collecting and using personal data. In general, it is not necessary for an employer to collect a copy of your ID card or record your ID card number during a job interview. In fact, the Code specifically states that a copy of an ID card should not be collected unless and until the individual has accepted an offer of employment. However, there are some exceptions: 1. Pre-employment checks: If the employer is required to conduct pre-employment checks, such as verifying your identity or work experience, they may request to see your ID card or record your ID card number. But they should only collect the information necessary for the purpose of the check and should not retain it for any other purpose. 2. Statutory requirements: In some cases, employers may be required by law to collect and retain certain personal data, such as your ID card number, for tax or social security purposes. 3. Industry-specific requirements: Certain industries, such as finance or healthcare, may have specific regulations or guidelines that require employers to collect and verify personal data, including ID card numbers. If an employer asks to collect your ID card or record your ID card number during a job interview, you can politely decline and ask why they need this information. If they provide a valid reason, you can consider providing the information. However, if you are unsure or uncomfortable, you can ask if there are alternative ways to verify your identity or work experience. It's also worth noting that you have the right to request access to and correction of your personal data held by the employer under the Personal Data (Privacy) Ordinance. Additionally, employers should mark the word "copy" across the image of copies of ID cards to reduce the chance for misuse and abuse, as recommended by the Code.

Case 3:
Question: I am a physically disabled person and I always have difficulty in taking a taxi. Should the taxi driver help me on every occasion? What if the driver refuses to offer taxi services to me?

Original answer: In Hong Kong, taxi drivers are generally expected to assist passengers with disabilities, but there are no strict legal requirements mandating that they provide such assistance in every instance. However, there are guidelines and regulations that promote accessibility and support for disabled passengers. What You Can Expect: 1. Assistance: Taxi drivers are encouraged to provide reasonable assistance to passengers with disabilities, such as helping them get in and out of the taxi. 2. Accessible Taxis: There are designated accessible taxis in Hong Kong that are equipped with ramps and other facilities to accommodate passengers with mobility issues. These taxis are often marked with special signage. What to Do if a Driver Refuses: 1. Report the Incident: If a taxi driver refuses to provide services to you, you can report the incident to the Transport Department or the relevant taxi company. They can investigate and take appropriate action. 2. Use Designated Services: Consider using designated accessible taxis or other forms of accessible transportation, such as the Accessible Transport Service (ATS) provided by the Transport Department. 3. Contact Support Services: You can also contact organizations that support people with disabilities, such as the Hong Kong Society for the Rehabilitation of the Disabled, for assistance and advice. Additional Resources: - Transport Department: You can contact the Transport Department for more information on accessible transportation services and to report any issues. - Accessible Transport Service (ATS): This service provides door-to-door transportation for people with disabilities who have difficulty using regular public transport. Conclusion: While there is no legal obligation for taxi drivers to assist you in every instance, there are resources and support systems in place to help you navigate these challenges. If you face difficulties, it's important to report the incident and seek alternative transportation options or support.

Corrected answer: In Hong Kong, as a physically disabled person, you have the right to access taxi services, and taxi drivers are expected to assist you. According to the Disability Discrimination Ordinance (DDO), it is unlawful for taxi drivers to discriminate against or harass you on the grounds of your disability. What You Can Expect: 1. Assistance: Taxi drivers are required to offer assistance to passengers with disabilities, such as helping them get in and out of the taxi, unless they have a reasonable excuse. For example, if a driver has difficulty providing such assistance due to their own disability, they should not willfully refuse to do so. 2. Reasonable Accommodation: Drivers should make reasonable accommodation to ensure that you can board and alight from the taxi independently. This may include placing the taxi in a position where you can easily access it, such as near the kerb, or providing assistance if you are unable to open the door yourself. What to Do if a Driver Refuses: 1. Report the Incident: If a taxi driver refuses to provide services to you, you can report the incident to the Equal Opportunities Commission or the Transport Department. They can investigate and take appropriate action. 2. Use Designated Services: Consider using designated accessible taxis or other forms of accessible transportation, such as the Accessible Transport Service (ATS) provided by the Transport Department. 3. Contact Support Services: You can also contact organizations that support people with disabilities, such as the Hong Kong Society for the Rehabilitation of the Disabled, for assistance and advice. Additional Resources: - Equal Opportunities Commission: You can contact the Equal Opportunities Commission for more information on anti-discrimination laws and to report any issues. - Transport Department: You can contact the Transport Department for more information on accessible transportation services and to report any issues. - Accessible Transport Service (ATS): This service provides door-to-door transportation for people with disabilities who have difficulty using regular public transport. Conclusion: As a physically disabled person in Hong Kong, you have the right to access taxi services, and taxi drivers are expected to assist you. If you face difficulties, it's essential to report the incident and seek alternative transportation options or support. Remember that refusing to provide taxi services to you on the grounds of your disability is unlawful, and you have the right to equal access to transportation services.