Semantic Specialization of Legal LLMs via Large-Scale Fine-Tuning on Jurisdictional MCQs

Hsuan-Lei Shao[1,*]

[1]*Graduate Institute of Health and Biotechnology Law,Taipei Medical University*

Abstract

This paper examines how instruction-based fine-tuning enhances the semantic competence of legal Large Language Models (LLMs) in answering multiple-choice questions (MCQs) drawn from jurisdiction-specific legal examinations. Using Breeze-7B, a Mandarin LLM adapted from Mistral-7B, we fine-tune the model on two corpora containing 5,000 and 70,000 legal MCQs to evaluate how training volume affects accuracy, norm identification, and alignment with local legal reasoning. We evaluate the resulting models across three datasets—the Taiwanese Bar Examination (TBE), the Taiwan Multimodal Legal Understanding benchmark (TMMLU), and MMLU—using both probability-based and zero-shot prompt-based protocols. Our findings show that large-scale fine-tuning substantially improves the model's ability to recognize legal obligations, interpret statutory structures, and reproduce jurisdiction-specific doctrinal reasoning. Comparisons with general-purpose models such as GPT-3.5 and GPT-4o highlight the trade-offs between broad linguistic generality and jurisdictional semantic specialization. We argue that MCQs function as structured, scalable signals for assessing and cultivating legal semantic competence in LLMs. Our results demonstrate that task-specific instruction tuning, combined with jurisdiction-grounded datasets, offers a cost-efficient pathway for developing explainable, localized legal AI systems. This work contributes to ongoing efforts in legal informatics to advance semantic analysis, model adaptation, and digital sovereignty within diverse legal environments.

Keywords

Legal Large Language Models (LLMs), Multiple-Choice Questions (MCQs), Instruction Fine-Tuning, Jurisdiction-Specific Datasets, Semantic Specialization

1. Introduction

1.1. The Evolution of Legal Informatics: Large Language Models in Legal Contexts

Legal informatics—an interdisciplinary field situated at the intersection of law and information technology—has undergone substantial transformation since its inception. Early efforts focused on automating legal documentation and constructing searchable databases of statutes and case law to meet the legal profession's growing need for efficient information management [1, 2]. These developments laid the foundation for more advanced systems capable of supporting legal analysis, document automation, and predictive tools for litigation forecasting [3, 4]. Visualization techniques later emerged as a methodological bridge, enabling legal texts to be translated into more formal, machine-readable structures [5].

Recent advances in artificial intelligence have accelerated innovation in legal informatics. Collaborative efforts among legal scholars, computer scientists, and data scientists now aim to build computational models that meaningfully improve legal processes and institutional functioning [6]. Surveys such as those by Katz and Dolin [7] highlight the field's evolution toward practical applications including automated document review, online dispute resolution, and intelligent legal analytics.

Proceedings of the Seventh International Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2025), 16 June, 2025, Chicago, USA.

*Corresponding author.

† A Chinese-language version of this work has appeared previously (DOI: 10.53106/1025593136204). The present paper is a substantially revised and expanded English-language edition and constitutes the first peer-reviewed version of this research.

hlshao@tmu.edu.tw (H. Shao)

0000-0002-7101-5272 (H. Shao)

1.2. Challenges in Implementing LLMs Across Diverse Legal Systems

Large Language Models (LLMs) have shown substantial promise in tasks such as legal drafting, summarization, and research assistance. However, significant challenges arise when deploying general-purpose LLMs within specialized legal systems, where statutory schemes, judicial doctrines, and regulatory structures vary widely across jurisdictions. Such tasks require models to internalize and adapt to heterogeneous legal frameworks—capabilities that current general LLMs lack [8, 9].

A key concern is the potential erosion of legal diversity. Smaller jurisdictions or culturally distinctive legal systems are often underrepresented in global training corpora, leading to the marginalization of their doctrinal nuances. This issue is increasingly discussed under the rubric of "sovereignty AI," which emphasizes the necessity of preserving jurisdiction-specific legal reasoning and institutional identity within AI systems [10, 11]. As a result, there is growing consensus on the need for customized, locally trained LLMs to ensure that generated legal analysis remains relevant, accurate, and compliant with jurisdictional requirements [12, 13].

1.3. Specializing Local Knowledge in Legal LLMs

In this study, we focus on building an LLM capable of incorporating jurisdiction-specific legal knowledge. Each jurisdiction possesses its own legal system, professional practices, examination mechanisms, and pedagogical traditions. Accordingly, we conceptualize a legal LLM not merely as a general-purpose model applied to legal tasks, but as a form of localized AI infrastructure—and potentially a foundation for "sovereign AI" development.

We employ Breeze-7B, a Mandarin derivative of the Mistral-7B architecture, and enhance it through instruction fine-tuning using curated legal MCQs. Two fine-tuned variants were constructed: one trained on 5,000 MCQs (bz5k) and another on 70,000 MCQs (bz70k). The comparative performance of these models reveals the intricate relationship between training volume and legal specialization. Fine-tuning with insufficient data can degrade overall performance by overfitting limited doctrinal patterns, thereby weakening the model's capacity for knowledge-intensive tasks. By contrast, large-scale fine-tuning (as with bz70k) substantially strengthens legal semantic competence—particularly in statutory interpretation and obligation recognition—while producing modest trade-offs in general-domain performance.

This balance underscores the inherent tension between specialization and generality. It also demonstrates that global LLMs frequently overlook smaller jurisdictions, such as Taiwan, whose legal corpora are rarely incorporated into mainstream training pipelines. By explicitly introducing localized materials, our approach shows that jurisdictions and expert communities can build domain-specific LLMs that preserve digital sovereignty and capture professional legal reasoning otherwise absent from globally trained models.

2. Literature Review

2.1. Overview of LLM Capabilities in Legal Domains

As we mentioned, LLMs have increasingly become integral to various applications within legal domains, demonstrating capabilities that span from basic legal information retrieval to complex reasoning and document generation. Studies have shown that LLMs, like the GPT series and its successors, can interpret, generate, and summarize legal texts with a high degree of accuracy. These models have been employed for contract analysis, litigation prediction, and even legal education assistance by generating hypothetical legal scenarios for study. This section reviews the extent of LLM integration in legal practices and evaluates their effectiveness in handling diverse legal tasks [8, 15, 19].

2.2. Current Methodologies in Prompt Engineering

Prompt engineering involves the strategic crafting of input prompts to LLMs to elicit the most accurate and relevant outputs. In legal applications, effective prompt engineering is critical as it significantly impacts the model's ability to provide legally sound advice and information. Recent advancements have focused on developing methodologies that enhance the specificity and context-awareness of prompts, thereby improving the precision of the model's responses. This includes techniques such as prompt chaining, in which a series of logically connected prompts are used to guide the model through complex reasoning tasks, but may negatively impact the model's architecture that is designed for knowledge-intensive tasks [15].

2.3. Current Methodologies in Instruction Fine-Tuning

Instruction fine-tuning is a recent development aimed at refining the training process of LLMs to better follow user instructions. Unlike traditional model training, instruction fine-tuning focuses on aligning the model's outputs with specific user expectations and requirements. In the legal field, this is particularly advantageous for ensuring that models adhere to legal reasoning patterns and comply with jurisdiction-specific regulations. This segment will cover the latest methodologies in instruction fine-tuning, including the application of specialized data sets (such as legal judgments or statutory provisions) that train models to recognize and replicate the nuanced decision-making processes typical in legal analyses [23,24].

3. Research Design

3.1. Multiple Choice Questions in Legal Evaluation

On the one hand, LLMs are important in our daily lives or research fields; on the other hand, LLMs are like a black-box that we hardly add new knowledge in. So, we use one of the basic indicates in legal education: multiple choice questions (MCQs). The MCQ plays a critical role in legal education and professional assessments. They are widely used in exams like the bar examination to efficiently test a broad range of legal knowledge. The format of MCQs allows for assessing students' understanding of key concepts and their ability to distinguish between closely related legal issues, a crucial skill in legal practice.

Moreover, the structured nature of MCQs makes them particularly suitable for automation using AI technologies like LLMs. By incorporating LLMs in creating and grading MCQs, educational institutions can enhance the objectivity and efficiency of assessments. LLMs can also be used to generate diverse question sets that cover a wide array of topics, providing a robust tool for comprehensive legal training [14, 15]. In Taiwan, this logic is particularly pronounced because the national judicial officer and bar examinations are not only pedagogical tools but also part of a state-sanctioned "knowledge industry." These exams are high-stakes gateways to the legal profession, and their official questions and answers embody codified statutory interpretation, case law reasoning, and doctrinal expectations. The fact that cram schools, publishers, and exam candidates invest enormous resources into these MCQs underscores their commercial value and national significance. Thus, by using Taiwanese bar exam MCQs as fine-tuning data, we are not only leveraging a structured educational tool but also converting a resource of immense social and economic value into semantically precise AI training material.

However, the effectiveness of LLMs in this area depends heavily on their training and the quality of data used. It is essential that the data reflects the specific legal principles and practices relevant to the jurisdiction where the education or assessment is taking place. This ensures that the questions are accurate and contextually appropriate, fostering a more effective and meaningful learning environment [16, 17].

To clarify the provenance and representativeness of our training data, we briefly summarise the composition and sampling strategy for the MCQ corpus. The 70k-sample pool is drawn from publicly

available multiple-choice questions used in Taiwanese legal education and professional examinations, including past judicial officer and bar exams as well as high–stakes preparatory materials. We then construct two training sets: a smaller subset of 5,000 MCQs (for bz5k) and a larger subset of 70,000 MCQs (for bz70k). Both sets were sampled to preserve diversity across core legal domains, including constitutional law, criminal law, civil law, administrative law, and procedural law.

To avoid data leakage with the 2023 Taiwanese Bar Examination (TBE) evaluation set, we rely on two institutional characteristics of the Taiwanese examination system. First, by regulation, the Ministry of Examination does not reuse questions across years. Each annual examination is produced by a newly convened committee that deliberates and drafts items specifically for that year, ensuring that questions are not duplicated across exam cycles. Second, because the TBE is typically administered in August, constructing our training corpus exclusively from materials dated prior to 2023 guarantees that no TBE-2023 items could have appeared in the fine-tuning data. This separation ensures that improvements observed on TBE 2023, TMMLU, and MMLU genuinely reflect generalisation from the training corpus rather than memorisation of evaluation items.

Moreover, even assuming arguendo that data leakage had occurred, one would reasonably expect a markedly elevated accuracy on the TBE 2023 set, as rote exposure to identical items typically produces an clear and disproportionate performance spike. The empirical pattern observed here is inconsistent with such an effect. Indeed, the only model with any plausible access to contemporaneous online material—GPT-4o accessed through an API—may be more susceptible to overestimation due to its capacity, in principle, to retrieve web-based information that indirectly references examination content. Such asymmetry further supports the inference that the improvements achieved by our fine-tuned models derive from genuine generalisation rather than contamination.

With respect to item difficulty, while we do not provide a full psychometric decomposition, it is well recognised within Taiwan's legal examination regime that yearly fluctuations in difficulty are routine and are often noted by examination specialists. Importantly, the system does not rely on absolute cut-scores but instead employs a relative-ranking mechanism: approximately the lowest 33.3% of candidates are eliminated based on cohort-wide performance. This institutional design mitigates the impact of year-to-year variance in difficulty and preserves comparability across examination cycles. Notably, in the years comprising our training corpus, an accuracy level near 66.6% closely corresponds to the empirical cumulative percentile distributions reported by the Ministry of Examination. Taken together, these structural features reinforce our position that MCQs produced within this regime constitute a stable, scalable, and jurisdiction-grounded signal for legal semantic specialisation.

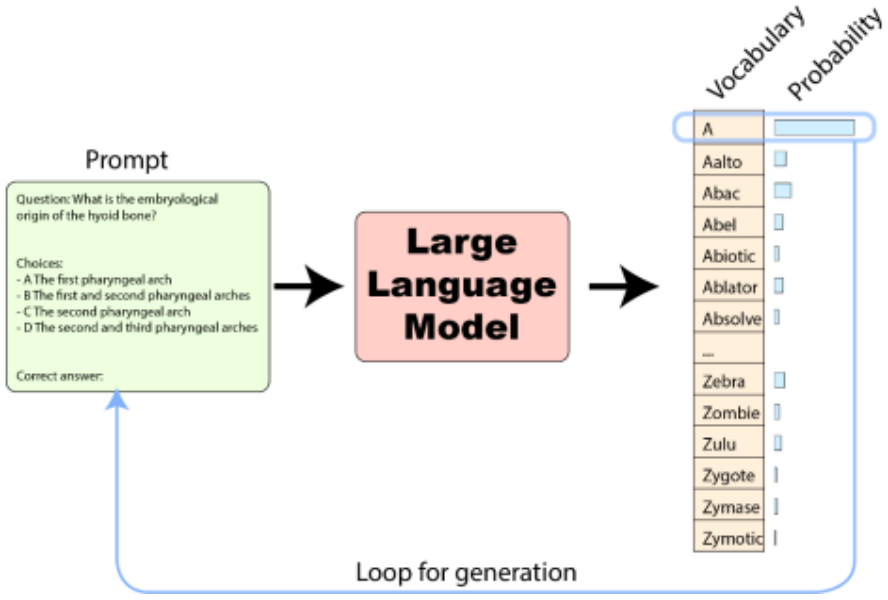

Figure 1: Probability-Based Evaluation[25]

3.2. Model Instruction Fine-tuning: Breeze-7B

In this study, we based the capabilities of the Breeze-7B-base model [27], which is built upon the foundations of the Mistral-7B architecture by incorporating an extensive set of MCQs into its training regimen. The original Breeze model, without any specific fine-tuning towards these datasets, serves as a control to understand the baseline capabilities of the LLM. The bz5k model, finetuned with 5,000 samples, represents a modest increase in dataset-specific training. The bz70k model, representing a substantial fine-tuning effort with 70,000 samples, aims to tailor the model towards the dataset characteristics significantly.

Our approach to instruction fine-tuning involves directly using the multiple-choice options as inputs. The output consists of the correct option (A, B, C, or D) along with the content of the option, which provides additional information. This method ensures that the model not only selects the correct answer but also captures the context and details associated with each option, enhancing its ability to handle similar questions effectively. For example:

This method ensures that the model not only selects the correct answer but also captures the context and details associated with each option, enhancing its ability to handle similar questions effectively.

For both bz5k and bz70k, we fine-tune Breeze-7B using the AdamW optimiser with a learning rate of 2e-5, and a linear warm-up over the first 10% of the training steps. We trained for one or three epochs, finally we choose the one epoch version because of better perfermance. It with a global batch size of 32 and a maximum sequence length of 1,024 tokens. All models are trained on one NVIDIA A6000 (80GB) GPU using bfloat16 mixed precision to reduce memory consumption. The random seed is fixed to 42 to ensure reproducibility. For LoRA, we adopt standard low-rank adaptation parameters with a rank (r) of 8, an [a] of 16, and a dropout rate of 0.05. Except for the size of the fine-tuning corpus (5k vs. 70k MCQs), all other hyperparameters remain identical so that observed performance differences can be attributed to differences in training volume rather than in the optimisation scheme or model architecture.

3.3. Evaluation Design

These models were benchmarked against a general baseline model, GPT-3.5, the more advanced GPT-4o. The datasets employed for evaluation included the Multimodal Legal Understanding (MMLU) [28], the Taiwanese Multimodal Legal Understanding (TMMLU) [29], and the 2023 Taiwanese Bar Examination questions [18]. We employ two complementary evaluation protocols to assess the semantic and jurisdictional alignment of the models: (i) probability-based scoring, and (ii) zero-shot prompt-based evaluation. Using both methods allows us to measure not only the model's underlying token-level preferences but also its behaviour in a more realistic prompting setting.

(i) Probability-Based Evaluation (Figure 1)

For probability-based scoring, we follow the standard MMLU evaluation protocol. Each MCQ is converted into four separate model queries, each corresponding to a candidate option ("A", "B", "C", or "D"). For each query, we extract the logits assigned to the option token by the final decoder layer with sampling disabled (temperature = 0, top-p = 0), and normalise them using a softmax over the four options. The option with the highest probability is taken as the model's prediction. This procedure yields fully deterministic results and enables systematic comparison across models. Because probability extraction does not require autoregressive generation, no new tokens are produced beyond the option token itself, ensuring consistency across runs.

(ii) Prompt-Based Zero-Shot Evaluation (Figure 2)

In addition to probability scoring, we evaluate the models in a zero-shot prompting setting that mirrors real-world usage. Each MCQ is presented as a single prompt containing the stem and the four candidate options, and the model is asked to answer directly. If the model explicitly outputs one of "A", "B", "C", or "D," that token is taken as the predicted answer. In the rare cases where no option token appears in the output (fewer than 10% of all prompts), we follow prior work and assign a neutral default answer ("C"). This convention avoids privileging extreme positions (e.g., always selecting "A" or "D")

and allows fair comparison across models with different verbosity levels.

(iii) Rationale for Two Protocols Probability-based scoring isolates the model's preference structure at the token level and is therefore sensitive to fine-grained semantic distinctions. In contrast, Prompt-based evaluation captures the model's behavior under naturalistic conditions where instruction following, formatting, and verbosity also influence performance. Using both provides a more comprehensive assessment of semantic specialization, jurisdictional alignment, and robustness. The evaluation of these models was conducted using two different methodologies to assess their performance in legal multiple-choice question scenarios.

```
{
    "input": "Question: Which of the following is NOT considered in assessing capacity for
liability? (A) The ability to recognize that an action is illegal (B) The ability to control one's
actions (C) The mental state at the time of the action (D) The ability to choose between legal and
illegal actions",
    "output": "(B) The ability to control one's actions "
},
{
    "input": "Question: After A grants B a permit for hillside development and B transfers the land
to C, does the original permit still apply to C? (A) Yes (B) No (C) Depends on the situation (D)
None of the above",
    "output": "(A) Yes "
},
```

Figure 2: Prompt-Based Zero-Shot Evaluation

4. Research Result and Discussion I: Fine-tuning Quantity Effect

4.1. Probability Selection Evaluation

Table 1
Comparing Different Fine-Tuning Quantity Effects (Probability-Based Evaluation on MMLU and TMMLU; Accuracy over N = 15908 and N = 472 questions respectively)

Dataset\Model	Breeze	bz5k	bz70k
TMMLU(Law)	0.407	0.401	**0.486**
TMMLU(Engineering)	**0.498**	0.493	0.458
MMLU	0.560	**0.562**	0.515
TBE	0.486	0.457	**0.514**

note: TBE = "the 2023 Taiwanese Bar Examination"

The table "Comparing Different Fine-tuning Quantity Effects" showcases the impact of varying quantities of data used in fine-tuning on the performance of the Breeze model across different datasets. These datasets encompass the Taiwanese Multi-Modal Legal Understanding (TMMLU) in Law and Engineering domains, the broader Multi-Modal Legal Understanding (MMLU), and the 2023 Taiwanese Bar Examination (TBE).

1. Dataset-Specific Performance: TMMLU (Law) and TMMLU (Engineering): For the Law subset of TMMLU, increasing the fine-tuning data quantity results in improved performance, as evident from the bz70k model's score of 0.486 compared to the bz5k's 0.401 and the baseline's 0.407. This suggests that a larger dataset helps the model better understand and adapt to legal nuances.
   Conversely, in the Engineering subset, the performance decreases as the quantity of fine-tuning increases (0.458 in bz70k down from 0.498 in the baseline). This could indicate overfitting or perhaps the introduction of noise or less relevant information through the additional data.
2. Different Language Performance: MMLU
   Here, we see a slight improvement in bz5k over the baseline (0.562 vs. 0.560), but a reduction with bz70k (0.515). This pattern suggests that while some targeted fine-tuning can be beneficial,

excessive fine-tuning may lead to diminishing returns or negative transfer, where too much specificity detracts from the model's general applicability. This demonstrates that if an LLM performs better in one language, it often performs worse in another. This may be related to the parameters of individual tokens, where fine-tuning can detrimentally affect the original linguistic structure of the LLM.

The observation that LLMs may exhibit improved performance in one language at the expense of another underscores a critical aspect of language model training known as the trade-off between specialization and generalization. In the context of LLMs, which are often trained on diverse multilingual corpora, each token's parameters carry the weight of representing linguistic features pertinent to multiple languages. During the fine-tuning process, these parameters are adjusted to optimize performance for specific tasks or languages. This optimization, while beneficial in the targeted context, can inadvertently lead to a degradation of the model's capabilities in other languages not emphasized during the fine-tuning phase.

3. The Latest Local Knowledge: TBE
   Performance on the Taiwanese Bar Examination dataset improves significantly with the highest data volume (bz70k), increasing from 0.486 to 0.514. This improvement indicates that comprehensive legal training data can enhance model performance on specialized legal tasks such as bar exams, which likely benefit from a deeper understanding of localized legal principles and practices.

4.2. Discussion

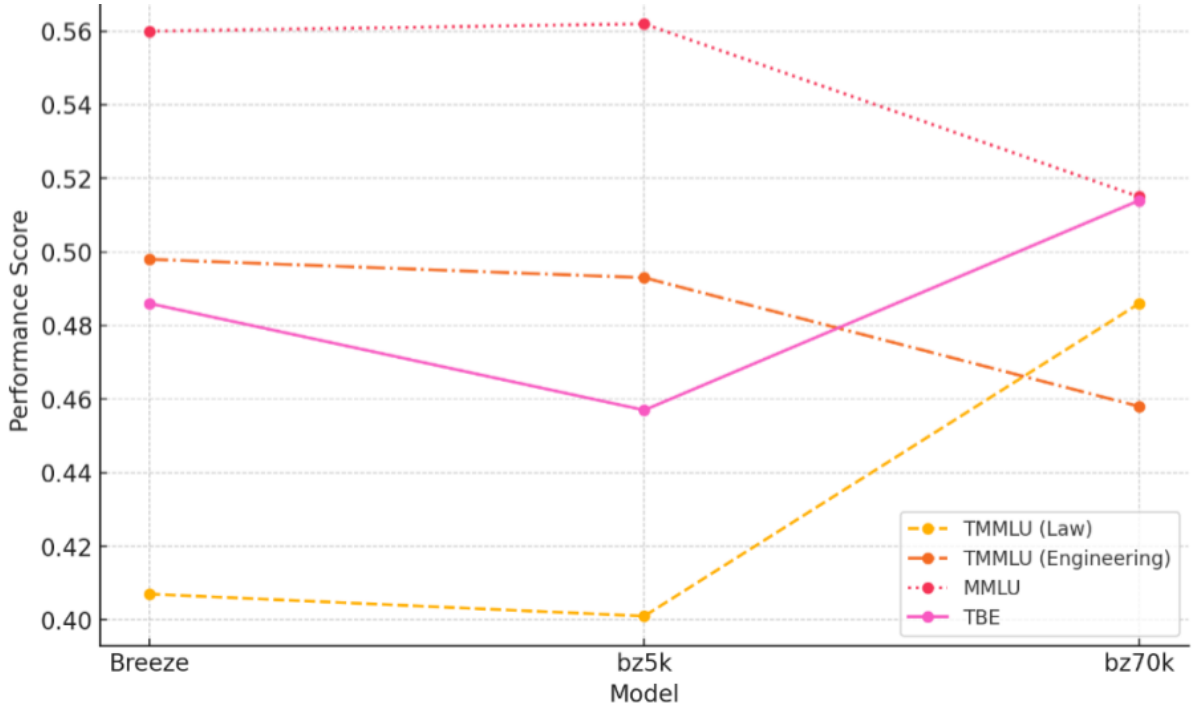

Figure 3: Impact of Fine-Tuning Data Volume on Model Accuracy Across Datasets (Probability-Based Evaluation; TMMLU-Law N = 472, TMMLU-Engineering N = 119, MMLU N = 15908, TBE N = 300)

The graph represents the performance comparison of three models (Breeze, bz5k, bz70k) across four different datasets (TMMLU-Law, TMMLU-Engineering, MMLU, TBE), with each model serving as a point on the x-axis and performance scores on the y-axis. Different line styles distinguish each dataset.

1. TMMLU (Law) (dotted line): Shows a trend of improvement as the fine-tuning data volume increases, peaking with the bz70k model.

2. TMMLU (Engineering) (dash-dot line): This line trends downward, indicating a decrease in performance with more extensive fine-tuning, potentially due to overfitting or less relevant fine-tuning data for engineering-specific content.
3. MMLU (dashed line): Performance slightly increases with moderate fine-tuning (bz5k) but decreases with extensive fine-tuning (bz70k), suggesting that a balance needs to be found to avoid diminishing returns.
4. TBE (solid line): Shows a recovery in performance with the most extensive fine-tuning (bz70k), indicating that larger, more focused datasets may be beneficial for specialized legal examinations like the bar exam.

This graph visually illustrates how varying the amount of fine-tuning data impacts model performance across different domains. It highlights the need for careful consideration of how much and what type of data to use for fine-tuning to optimize performance without compromising on the model's generalization capabilities. This insight is crucial for applying LLMs to specialized fields where accuracy and specificity are paramount.

Impact of Dataset Size in Fine-tuning: The varying effects of increased fine-tuning data volumes across different datasets underscore the need for a nuanced approach to model training, where the quantity and quality of data must be carefully managed to avoid overfitting while still achieving meaningful improvements.

5. Research Result and Discussion II: Prompt Evaluation

5.1. Prompt Evaluation

In the second phase, we input the MCQs string directly through the API (refer to 3.3. Evaluation Design), which allows us to compare our results with other LLMs.

Table 2
Comparing Different Fine-tuning Quantity Effects by Prompting Input.

Dataset\Model	Breeze	bz5k	bz70k	GPT-3.5	GPT-4o
TMMLU(admin_law)	0.250	0.380	**0.580**	0.336	0.650
MMLU	0.320	0.540	0.590	**0.660**	0.860
TBE	0.106	0.423	**0.640**	0.423	0.680

The table provides comparative performance data for different language models on MCQs across three datasets: TMMLU (administrative law, sub-category of the Law category), MMLU, and TBE. It shows how each model fares in accurately responding to prompts within these specific domains.

First of all, we wish to skip the discussion on GPT-4o since it is an outlier in terms of performance. Moreover, because it is much larger, its performance is not comparable to our approximately 7B parameter model. We can see the trends on the Breeze-series and other LLMs:

1. Impact of Fine-tuning Method: Because our instruct fine-tuning itself has enough MCQs diversity, the bz70k model can achieve high performance when we ask it directly. The bz70k in some TMMLU fields simply because it was not "familiar" with the instructions.
2. Quantity Affects Quality: This table clearly illustrates that fine-tuning with a larger volume of data specifically tailored to the task at hand can significantly enhance a model's performance. The bz70k's success across datasets indicates that the additional specific training it received is highly effective than Breeze-base and bz5k, even the GPT-3.5.
3. General vs. Specialized Models: The comparison between bz70k variants and GPT-4.0 highlights an essential aspect of language model application: general models can perform well across broad tasks, but under domain-specific fine-tuning processing, smaller LLM (7B) can reach the larger performance.

5.2. Discussion

The graph illustrates the performance comparison across different language models on three datasets: TMMLU (administrative law), MMLU, and TBE (Taiwanese Bar Examination). Each model is represented on the X-axis, and the performance score, likely accuracy or a similar metric, is represented on the Y-axis. Different line styles and colors distinguish the performance of each dataset.

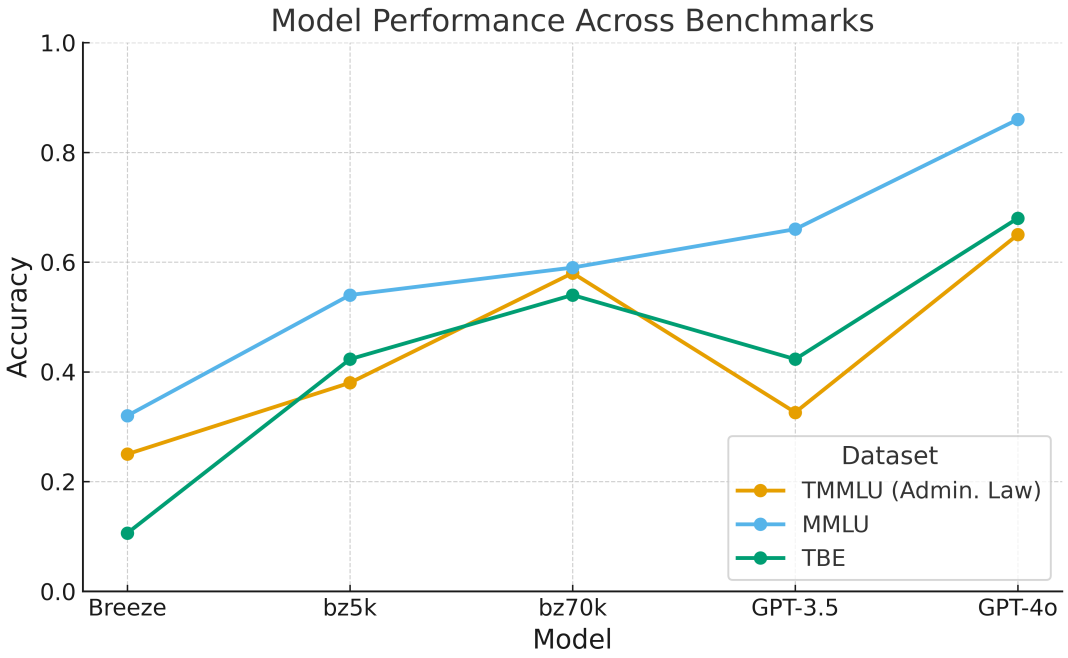

Figure 4: Prompt-Based Evaluation Across Specialized and General Models (Zero-Shot Prompting; TMMLU-Administrative Law N = 472, MMLU N = 10492, TBE N = 300

Comparison of Breeze, bz5k, bz70k, GPT-3.5, and GPT-4o on TMMLU (administrative law), MMLU, and TBE under direct prompting conditions. The bz70k model shows marked improvement in Taiwanese bar exam accuracy, rivaling larger general-purpose LLMs. These results highlight the potential of fine-tuned small LLMs to serve as cost-efficient, explainable, and localized AI tools for legal applications. Trend Analysis of the graph:

1. Incremental Improvements: The graph illustrates a clear trend of incremental performance improvements as we move from the baseline Breeze model to the bz5k and then to the bz70k. This trend is evident across all datasets but varies in magnitude.
2. TMMLU (Admin Law): For the TMMLU (Admin Law) dataset, the performance improvement from Breeze (0.25) to bz5k (0.38) and then to bz70k (0.58) is quite pronounced. This significant uptick suggests that the additional training samples used in fine-tuning the bz70k model are highly effective at enhancing the model's capabilities in handling complex administrative law scenarios.
3. MMLU: The trend in the MMLU dataset follows a similar pattern. Starting from a performance score of 0.32 with Breeze, there is a noticeable increase to 0.54 with bz5k, and further improvement to 0.59 with bz70k. This consistent increase across fine-tuning stages underscores the effectiveness of using larger, more targeted training sets for enhancing model performance in general legal contexts.
4. TBE: In the TBE dataset, the performance jumps considerably from Breeze (0.106) to bz5k (0.423), and sees a significant peak at bz70k (0.64). This demonstrates that extensive fine-tuning with a large volume of specialized data is particularly beneficial for models that navigate the complexities of bar examination questions, which likely involve nuanced legal reasoning and specific legal knowledge.

5.3. Qualitative Error Analysis

To complement the aggregate accuracy scores, we conducted a small-scale qualitative error analysis on a sample questions drawn from TBE 2023 and TMMLU-Law. representative error pattern arises in questions involving causal attribution and the foreseeability of aggravated results. Consider the following item from TBE 2023:

A and B bore longstanding animosity toward each other. One day, upon seeing B loitering on the street, A proceeded to strike B. Startled after being attacked, B attempted to flee but accidentally fell into a roadside construction pit. As a result, in addition to the contusions inflicted by A, B suffered a comminuted fracture of the right lower leg, causing destruction of the limb. According to the latest judicial practice, which of the following statements is correct? (A)Although B sustained severe injury, the fact that B accidentally fell into the roadside construction pit constitutes a self-responsible act on the part of the victim; therefore, A is liable only for the ordinary injury offense under Article 277(1) of the Criminal Code for striking B's head. (B)Because B's right lower leg merely sustained a fracture with destruction, it does not satisfy the definition of "destruction or serious impairment of the function of one or more limbs" constituting a serious injury under Article 10(4)(4) of the Criminal Code. (C) A should have had objective foreseeability that B might fall into the construction pit and thereby suffer destruction of the right lower leg; thus, A bears liability for the offense of causing serious injury under the latter part of Article 277(2) of the Criminal Code. (D) Because B's serious injury did not result from the aggravation of the head contusion, the offense of causing serious injury under the latter part of Article 277(2) of the Criminal Code is not established. Correct answer: (C).

Both the base Breeze-7B model and the bz5k variant frequently selected option (D), reflecting a misunderstanding of Taiwan's established doctrine that the victim's panic-driven flight does not constitute a self-responsible interruption of causation. Under contemporary Taiwanese criminal jurisprudence the aggravated injury remains objectively foreseeable to the initial assailant and therefore attributable to him under Article 277(2) of the Criminal Code. This example illustrates how large-scale jurisdiction-specific MCQ fine-tuning enables the model to internalize nuanced doctrinal concepts—such as the scope of risk and objective foreseeability—that are rarely captured by general-purpose LLMs.

Beyond the specific examples, a broader pattern emerges from our qualitative inspection. The models—particularly the base Breeze-7B and the bz5k variant—tend to exhibit substantially higher error rates on questions that depend on (i) Taiwan-specific statutory provisions, and (ii) multi-step causal attribution under contemporary Taiwanese criminal jurisprudence. Items requiring precise knowledge of local statutory wording (e.g., the threshold definitions of "serious injury" under Criminal Code Article 10 or institutional authority allocations under the organizational law of the courts) frequently led to incorrect answers by non-specialised models, suggesting that general LLM pretraining corpora contain insufficient coverage of Taiwanese doctrinal sources.

Similarly, questions involving complex causation analysis—such as differentiating between victim self-responsibility, independent intervening acts, and results that remain within the objectively foreseeable scope of risk—proved particularly challenging. These topics involve doctrinal nuances that are heavily shaped by Taiwan's case law and academic commentary. The bz70k model performs notably better in these scenarios, indicating that large-scale instruction tuning on jurisdiction-grounded MCQs provides the model with access to precisely those doctrinal subtleties that generic LLMs lack.

Taken together, these patterns demonstrate that errors cluster around areas where local legal knowledge or multi-stage legal reasoning is required—precisely the domains in which jurisdiction-specific fine-tuning is most impactful.

6. Conclusion

6.1. Performance of Fine-tuned Small LLMs

The inference results across the Breeze model family—from bz5k to bz70k—demonstrate the decisive influence of both data volume and domain specificity on model performance in legal tasks. The

Taiwanese Bar Examination (TBE) dataset, in particular, functions as a highly practical source of localized legal knowledge, providing a rigorous benchmark for evaluating how effectively AI systems can operate within jurisdiction-specific contexts. This aligns with a broader imperative in AI development: the need for datasets embedded in concrete legal and cultural environments to evaluate the extent to which LLMs can meaningfully adapt to localized conditions.

The strong performance of bz70k on the TBE benchmark shows that, given sufficient targeted instruction tuning, small LLMs can attain substantial proficiency in legal reasoning, statutory interpretation, and doctrinal alignment. This finding has direct implications for the deployment of AI in legal settings, where accuracy, jurisdictional compliance, and the capacity to reflect professional reasoning practices are essential. The progression from bz5k to bz70k underscores the value of incremental fine-tuning strategies and highlights the necessity of aligning training data with the substantive demands of legal tasks.

More broadly, this study affirms the central premise that data quantity—when paired with appropriate data quality—plays a critical role in shaping the semantic competence of LLMs. This observation echoes the title of the paper, "Quantity Affects Quality," by illustrating how substantial and well-curated fine-tuning inputs can significantly enhance model outputs, even when the baseline model is comparatively modest. Ultimately, high-quality, jurisdiction-grounded training data enables LLMs to acquire the doctrinal precision and contextual awareness necessary for real-world legal applications.

Our improvement plan has met the original research objectives:

1. **Localized and Personalized LLMs.** *Digital Sovereignty and Localization.* The project successfully incorporates jurisdiction-specific legal knowledge into LLMs through targeted fine-tuning. This approach aligns with the global movement toward digital sovereignty, in which regions or institutions regain control over the informational and operational parameters of their AI systems. Personalizing LLMs to reflect local legal standards enhances both practical utility and doctrinal fidelity, ensuring that the generated analyses and responses remain legally sound and contextually appropriate.
2. **Data Volume and Model Performance.** The study confirms that expanding the volume of fine-tuning data produces substantial gains in model performance, particularly for smaller architectures that lack the inherent advantages of models such as GPT-3.5 or GPT-4o. Larger and well-curated datasets allow these smaller models to close the performance gap, challenging the assumption that scale alone determines capability. This finding underscores the iterative nature of AI development: future advances in base models can likely benefit from the same domain-specific fine-tuning strategies examined here.
3. **Practical Application and Instructional Method.** *Alignment with Real-World Use Scenarios.* By using direct question input as the primary instruction mechanism, the project mirrors real-world legal inquiry patterns. This alignment improves practical usability, enabling the model to generate responses that more closely reflect the tasks and reasoning structures encountered in actual legal practice.

6.2. Research Limitations and Future Prospects

1. **Technological and Resource Limitations.** Constraints on computing resources limited our ability to experiment with larger base models, a common challenge in contemporary LLM research. Because larger models demand substantial computational infrastructure, future work would benefit from institutional partnerships or shared-resource models that enable experimentation with more advanced architectures. Such collaborations could accelerate progress and foster a more equitable research ecosystem.
2. **Expanding the Scope of Legal Informatics.** Although MCQs offer a structured and scalable foundation for modeling legal knowledge, they capture only a subset of legal reasoning. Doctrinal analysis, argumentation, and judgment-like reasoning remain more complex tasks that require deeper semantic understanding. Future research should therefore expand beyond MCQs to

incorporate higher-order legal tasks—such as argument mining, statutory interpretation across contexts, or analysis of legal policy debates—that more closely approximate human legal reasoning.

3. **Improving Evaluation Methods.** Existing evaluation protocols may not fully capture the reproducibility, normative coherence, or deliberative quality required for legal applications. Effective legal AI must produce consistent and defensible outputs across varied conditions. Thus, more robust evaluation frameworks are essential—potentially involving standardized datasets, metrics that reflect legal reasoning quality rather than accuracy alone, and systematic cross-validation procedures. Human-in-the-loop assessments and explainability metrics should likewise play a larger role in future evaluations.

The principle that "Quantity Affects Quality" highlights the transformative influence of data scale and quality on LLM development. Our results demonstrate that large-scale fine-tuning on jurisdictional MCQs significantly enhances the semantic precision of small LLMs, particularly in statutory interpretation and norm recognition. Yet this achievement also exposes a fundamental limitation: specialization may narrow generality. When optimized to excel at multiple-choice reasoning, models can lose robustness in broader or multilingual settings, as reflected in our comparative performance on MMLU. This trade-off—between localization and generalization—remains a core challenge in legal NLP and cannot be fully resolved under current training paradigms.

Future research must proceed along three complementary directions. Technologically, modular fine-tuning and mixture-of-experts architectures may allow specialized legal reasoning modules to coexist with general-purpose capabilities. Methodologically, expanding beyond MCQs to integrate diverse legal tasks—such as argument extraction, contract-clause consistency checking, multilingual statutory alignment, or legal discourse analysis—can enrich the semantic landscape of fine-tuning. Evaluatively, new benchmarks are needed to assess legal reasoning quality, doctrinal soundness, and explainability, rather than accuracy alone.

In sum, while jurisdictional MCQs provide a cost-efficient and semantically rich foundation for training localized legal LLMs, they also reveal the enduring tension between domain specialization and broad adaptability. Addressing this tension will require sustained innovation in training strategies, cross-jurisdictional collaboration, and the development of evaluation frameworks that respect both local legal sovereignty and the need for interoperability in global AI research.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Jenkins J. What can information technology do for law. Harvard J Law Technol. 2008;21:589.

[2] Waltl B, Zec M, Matthes F. A data science environment for legal texts. In: Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL). 2015. p. 193-194.

[3] Conrad JG, Al-Kofahi K, Zhao Y, Karypis G. Effective document clustering for large heterogeneous law firm collections. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2005. p. 177-187.

[4] Pacheco HR, Pavez MM. Contemporary challenges in legal informatics: Workshop INJU. In: 2016 11th Iberian Conference on Information Systems and Technologies (CISTI). 2016. p. 1-2.

[5] Lachmayer F, Cyras V. Visualization of legal informatics. J Vis Law. 2021;3:3-10.

[6] Sharma S, Gamoura S, Prasad DM, Aneja A. Emerging legal informatics towards legal innovation: Current status and future challenges and opportunities. Legal Inf Manage. 2021;21:218-235.

[7] Katz D, Dolin R. Legal informatics. Cambridge: Cambridge University Press; 2021.

[8] Šavelka J, Ashley KD. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. Front Artif Intell. 2023;6:1279794. doi:10.3389/frai.2023.1279794.

[9] Sun Z. A short survey of viewing large language models in legal aspect. ArXiv. 2023;abs/2303.09136. doi:10.48550/arXiv.2303.09136.
[10] Shaghaghian S, Feng L, Jafarpour B, Pogrebnyakov N. Customizing contextualized language models for legal document reviews. In: 2020 IEEE International Conference on Big Data (Big Data). 2020. p. 2139-2148. doi:10.1109/BigData50022.2020.9378201.
[11] Zhang D, Petrova A, Trautmann D, Schilder F. Unleashing the power of large language models for legal applications. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 2023. doi:10.1145/3583780.3615993.
[12] Trozze A, Davies TP, Kleinberg B. Large language models in cryptocurrency securities cases: Can ChatGPT replace lawyers? ArXiv. 2023;abs/2308.06032. doi:10.48550/arXiv.2308.06032.
[13] Elwany E, Moore DA, Oberoi G. BERT goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding. ArXiv. 2019;abs/1911.00473.
[14] Shui R, Cao Y, Wang X, Chua T. A comprehensive evaluation of large language models on legal judgment prediction. ArXiv. 2023;abs/2310.11761. doi:10.48550/arXiv.2310.11761.
[15] Robinson J, Rytting C, Wingate D. Leveraging large language models for multiple choice question answering. ArXiv. 2022;abs/2210.12353. doi:10.48550/arXiv.2210.12353.
[16] Zhang Z, Lei L, Wu L, Sun R, Huang Y, Long C, Liu X, Lei X, Tang J, Huang M. SafetyBench: Evaluating the safety of large language models with multiple choice questions. ArXiv. 2023;abs/2309.07045. doi:10.48550/arXiv.2309.07045.
[17] Bitew SK, Deleu J, Develder C, Demeester T. Distractor generation for multiple-choice questions with predictive prompting and large language models. ArXiv. 2023;abs/2307.16338. doi:10.48550/arXiv.2307.16338.
[18] Ministry of Examination, Taiwan. "Exam Question and Answer Search System," *Ministry of Examination*, Available at: https://wwwq.moex.gov.tw/exam/wFrmExamQandASearch.aspx?y=2012&e=101120, accessed on August 8, 2024.
[19] Nay JJ, Karamardian D, Lawsky S, Tao W, Bhat MM, Jain R, Lee AT, Choi JH, Kasai J. Large language models as tax attorneys: A case study in legal capabilities emergence. ArXiv. 2023;abs/2306.07075. doi:10.48550/arXiv.2306.07075.
[20] Phogat KS, Harsha C, Dasaratha S, Ramakrishna S, Puranam SA. Zero-Shot Question Answering over Financial Documents using Large Language Models. ArXiv. 2023;abs/2311.14722. doi:10.48550/arXiv.2311.14722.
[21] Kojima T, Gu S, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. ArXiv. 2022;abs/2205.11916. doi:10.48550/arXiv.2205.11916.
[22] Cheng D, Huang S, Bi J, Zhan YW, Liu J, Wang Y, Sun H, Wei F, Deng D, Zhang Q. UPRISE: Universal Prompt Retrieval for Improving Zero-Shot Evaluation. ArXiv. 2023;abs/2303.08518. doi:10.48550/arXiv.2303.08518.
[23] Ni Y, Jiang S, Wu X, Shen H, Zhou Y. Evaluating the robustness to instructions of large language models. ArXiv. 2023;abs/2308.14306. doi:10.48550/arXiv.2308.14306.
[24] Xu C, Sun Q, Zheng K, Geng X, Zhao P, Feng J, Tao C, Jiang D. WizardLM: Empowering large language models to follow complex instructions. ArXiv. 2023;abs/2304.12244. doi:10.48550/arXiv.2304.12244.
[25] Hugging Face. "Open LLM Leaderboard MMLU," *Hugging Face Blog*, Available at: https://github.com/huggingface/blog/blob/main/open-llm-leaderboard-mmlu.md, accessed on August 8, 2024.
[26] Liu Yu-Wei. "LLM Model Evaluation," *GitHub Repository*, Available at: https://github.com/LiuYuWei/llm_model_evaluation, accessed on August 8, 2024.
[27] MediaTek Research. "Breeze-7B-Base-v1_0," *Hugging Face Model*, Available at: https://huggingface.co/MediaTek-Research/Breeze-7B-Base-v1_0, accessed on August 8, 2024.
[28] Pei-Yuan Liu. "MMLU Dataset," *Kaggle Dataset*, Available at: https://www.kaggle.com/datasets/peiyuanliu2001/mmlu-dataset, accessed on August 8, 2024.
[29] iKala. "TMMLUPlus Dataset," *Hugging Face Dataset*, Available at: https://huggingface.co/datasets/ikala/tmmluplus, accessed on August 8, 2024.