Continual Pre-Training is (not) What You Need in Domain Adaptation

Pin-Er Chen F10142001@NTU.EDU.TW
Da-Chen Lian D08944019@NTU.EDU.TW
Jou-An Chi JAJOANNE.CHI88@GMAIL.COM
Shu-Kai Hsieh SHUKAIHSIEH@NTU.EDU.TW
Sieh-Chuen Huang SCHHUANG@NTU.EDU.TW
National Taiwan University
Taipei,Taiwan

Hsuan-Lei Shao HLSHAO@TMU.EDU.TW
Taipei Medical University
Taipei, Taiwan

Jun-Wei Chiu CLCHIU@NVIDIA.COM
Yang-Hsien Lin YANGHSIENL@NVIDIA.COM
Zih-Ching Chen VIRGINIAC@NVIDIA.COM
Cheng-Kuang Lee CKL@NVIDIA.COM
Eddie TC Huang TZUNGCHIH@NVIDIA.COM
Simon See SSEE@NVIDIA.COM
NVIDIA AI Technology Center, NVIDIA Corporation
Santa Clara, CA, USA

Editors: Hung-yi Lee and Tongliang Liu

Abstract

The recent advances in Legal Large Language Models (LLMs) have transformed the landscape of legal research and practice by automating tasks, enhancing research precision, and supporting complex decision-making processes. However, effectively adapting LLMs to the legal domain remains challenging due to the complexity of legal reasoning, the need for precise interpretation of specialized language, and the potential for hallucinations. This paper examines the efficacy of Domain-Adaptive Continual Pre-Training (DACP) in improving the legal reasoning capabilities of LLMs. Through a series of experiments on legal reasoning tasks within the Taiwanese legal framework, we demonstrate that while DACP enhances domain-specific knowledge, it does not uniformly improve performance across all legal tasks. We discuss the trade-offs involved in DACP, particularly its impact on model generalization and performance in prompt-based tasks, and propose directions for future research to optimize domain adaptation strategies in legal AI.

Keywords: Large Language Models (LLMs); Legal AI; Domain Adaptation; Continual Pre-Training; Legal Reasoning

1. Introduction

The advent of legal AI represents a significant transformation in the delivery of legal services and the execution of legal research. As AI technologies advance, they offer unprecedented capabilities for automating routine tasks, enhancing the precision of legal research, and facilitating complex decision-making processes. Legal LLMs, in particular, enable large-scale legal text processing and democratize access to legal knowledge (Lai et al., 2023). However, integrating legal-domain knowledge into LLMs poses significant challenges, particularly in ensuring ethical use, maintaining transparency in decision-making, and addressing concerns about bias. Among these challenges, the most crucial is enhancing the reasoning capabilities of LLMs (Almeida et al., 2024). Legal reasoning is a complex process that involves interpreting statutes, case laws, and regulations, and applying them to specific *facts*. Unlike other forms of logical reasoning, legal reasoning demands an understanding of the precise and normative meanings of legal language, which is often highly specialized and context-dependent (Bongiovanni et al., 2018).

Given these challenges, Domain-Adaptive Continual Pre-Training (DACP) offers a promising solution for improving LLMs' legal reasoning and reducing the occurrence of hallucinations. It is noteworthy that most existing research on DACP in the legal domain has been conducted within Anglo-American legal systems. This paper seeks to break new ground by focusing on the Taiwanese-Mandarin legal system, which has been adapted from and influenced by the Continental legal system.

The rest of the paper is organized as follows. We briefly review the literature on legal reasoning and related evaluation benchmarks. In Section 3, we present our models, which have been trained on legal data. Section 4 describes the construction of our benchmark, specifically designed to assess the legal reasoning capabilities of LLMs within the Taiwanese legal framework. The benchmark consists of multiple tasks: single-multiple choice questions (4.1), argument-based decision-making in legal symposia (4.2), and essay questions (4.3). The experiment results of the LLMs are illustrated respectively. Section 5 concludes the paper, and Section 6 discusses limitations and future work.

2. Related Work

Domain-Adaptive Continual Pre-Training. Building on the success of English LLMs, many studies explored language- and task-specific continual pre-training (Cui et al., 2024; Guo and Hua, 2023; Zhao et al., 2024; Zheng et al., 2023). Domain-Adaptive Continual Pre-Training (DACP) extends a general-purpose model with large-scale domain-specific unlabeled data (Gururangan et al., 2020; Jin et al., 2022; Shi et al., 2024), aligning it more closely with in-domain distributions (Wu et al., 2022; Xie et al., 2023; Çağatay Yıldız et al., 2024). Well-known examples include BioBERT (Lee et al., 2019), PubMedBERT (Gu et al., 2021), BloombergGPT (Wu et al., 2023), and EcomGPT-CT (Ma et al., 2023).

Despite notable gains in some settings, DACP often introduces trade-offs, such as general knowledge forgetting or reduced performance in general NLP tasks like NER or long-context understanding (Chen et al., 2020; Ma et al., 2023; Yang et al., 2024; Zhang et al., 2023; Zheng et al., 2023). Its benefits are thus inconsistent across task types. For example, Cheng et al. (2023) found that DACP strengthens domain knowledge but it unexpectedly reduces performance on prompting tasks across various domains, while Xie et al. (2023) showed that FinPythia outperforms Pythia in some financial tasks but underperforms in sentiment analysis. Similarly, smaller domain-specific models such as PARAMANU-AYN (Niyogi and Bhattacharya, 2024) can achieve comparable results at lower cost, further questioning the universal advantage of DACP.

Evaluation Benchmarks Recent advances in LLMs have underscored the need for refined benchmarks to assess domain-specific capabilities. For Chinese language evaluation, benchmarks such as C-Eval (Huang et al., 2023), CMMLU (Li et al., 2024), and TMMLU+ (Tam et al., 2024) extend MMLU (Hendrycks et al., 2021) to Simplified and Traditional Chinese, testing multiple-choice reasoning across diverse academic disciplines. In the legal domain, benchmarks including LawBench (Fei et al., 2023), LAiW (Dai et al., 2024), LegalBench (Guha et al., 2023), and DISC-LawLLM-Eval (Yue et al., 2023) evaluate legal knowledge and reasoning through tasks such as judicial exams, legal application, and argument generation. While these resources mark substantial progress, they remain dominated by classification-style questions, leaving the evaluation of complex, generative legal reasoning an open challenge for future research.

3. Models and Methods

This section describes the training pipeline and models used in our study. We follow a three-stage process—domain-adaptive continual pre-training, instruction tuning, and preference alignment—to develop the Llawa models, and we compare them against two LoRA-based baselines.[1]

For Llawa, we used Llama3-TAIDE-LX-8B-Chat-Alpha1(TAIDE)[2] as our base model for continuously pre-training Llawa. It was continuously pre-trained from Meta's Llama-3-8B (Meta, 2024)[3] on 43B tokens of Traditional Chinese that reflect the linguistic and cultural characteristics of Taiwan. We chose this model with the belief that a domain-adapted, culturally aware model would be beneficial to downstream tasks involving local law.

We carry out three stages of training:

1. We perform Domain-Adaptive Pre-Training on legal documents from Taiwan. This step helps the base model learn knowledge relevant to Taiwan's legal system.

2. We perform full-parameter instruction tuning on several law-related tasks. This teaches the model to answer legal questions helpfully.

3. We perform preference alignment on the instruction-tuned model from the previous stage. Preference alignment often helps improve the model's output in becoming more helpful (e.g., more informative, less harmful)

3.1. Stage 1: Domain-Adaptive Pre-Training in the Legal Domain

Information regarding the pre-training datasets can be found in Table 1. *Taiwan Law* contains publicly available data from Judicial Yuan,[4] including laws and regulations, as well as court documents. *German Law* is a subset of the MultiLegalPile,[5] which is a

[1] The models described in this paper are available on our Hugging Face repository: https://huggingface.co/lopentu.

[2] https://huggingface.co/taide/Llama3-TAIDE-LX-8B-Chat-Alpha1

[3] https://huggingface.co/collections/meta-llama/meta-llama-3-66214712577ca38149ebb2b6

[4] https://www.judicial.gov.tw/tw/mp-1.html

[5] https://huggingface.co/datasets/joelniklaus/Multi_Legal_Pile_Commercial

Dataset	Size (GB)	Tokens (B)
Taiwan Law	112.36	56
German Law	41.23	41
Self-Curated	10.37	10

Table 1: Datasets used for pre-training. Token counts are calculated using the base model's tokenizer.

multilingual legal dataset containing case law, contracts, legislation, and others. We use the **de** subset that contains all German language data. This was added because of the influence Germany's legal system has had on Taiwan's legal system (Zhang, 2019).

Self-Curated contains data such as knowledge graphs from ConceptNet[6] for English, German, and French and the Chinese Buddhist Electronic Texts Association[7] (CBETA). The considerations for selecting these data include enhancing logical reasoning, providing reference knowledge for civil law systems, and familiarizing the model with Classical Chinese (since the legal language style in Chinese closely resembles Classical Chinese).

Pre-training for Llawa is done on 2 x NVIDIA DGX H100 nodes (16 H100 GPUs). These resources are accessed via the Taipei-1 supercomputer located in the Kaohsiung Software Park. We use NVIDIA NeMo[8] and NVIDIA NeMo Framework Launcher[9] to run our training scripts. We use FP8 mixed-precision training to take advantage of the memory savings and faster training afforded by the hardware. We use a tensor parallel size of 2. We trained for one epoch, using a global batch size of 224 and a sequence length of 8192. Training took approximately 8 days. We started with a learning rate of 1e-4 and gradually decreased it using a cosine decay schedule.

3.2. Stage 2: Instruction Tuning

Instruction tuning serves to bridge the gap between pre-training's next token prediction task on unlabeled data and the task of being a helpful model to users. This entails training the model on (`INSTRUCTION`, `OUTPUT`) pairs (Zhang et al., 2024).

3.2.1. Llawa Training Details

Full-parameter instruction tuning is performed on 4 x NVIDIA RTX 6000 Ada using Axolotl (Axolotl AI, 2024). We train two instruction-tuned versions of Llawa.

The first version, Llawa-TC-YZL-Instruct, is trained in two stages. For the first stage, the base Llawa model is trained on a cleaned version of TaiwanChat (Lin and Chen, 2023).[10] This stage is meant to allow the model to learn general instruction following capabilities. TaiwanChat is an instruction dataset comprising other instruction datasets, which are translated to Traditional Chinese using GPT-3.5-Turbo. We use a cleaned

[6] https://huggingface.co/datasets/conceptnet5/conceptnet5

[7] https://www.cbeta.org/

[8] https://github.com/NVIDIA/NeMo

[9] https://github.com/NVIDIA/NeMo-Framework-Launcher

[10] https://huggingface.co/datasets/yentinglin/TaiwanChat

version that removes duplicates and malformed conversations (e.g., conversations ending with the user instead of an assistant's response). We train on this dataset for three epochs. The second stage further trains the model on legal-related tasks, the types of which can be found in Table 2. This stage can be seen as specializing the model to answer more legal-oriented questions. We train for two epochs, which is when validation loss fails to improve.

The second version, Llawa-TCxYZL-Instruct, is fine-tuned in one stage by combining TaiwanChat and our legal dataset and shuffling the resulting dataset. Training on both the general instruction dataset and the legal instruction dataset simultaneously can reduce the likelihood of the model forgetting any generalizable skills that would be beneficial for completing any tasks. We train for two epochs, which is when validation loss does not continue to improve.

3.2.2. Bllawa & Blawstral Training Details

Bllawa and Blawstral are trained by applying Low-Rank Adaptation (LoRA, Hu et al. 2021) on 1 x NVIDIA A100 80GB to all linear layers ($r = 64$, $\alpha = 128$, $lr = 5 \times 10^-5$). We use the Unsloth (Han and Han, 2024) library, which uses custom-written kernels that reduce memory usage and training time. We train for three epochs on our legal dataset.

3.3. Stage 3: Preference Alignment

Preference alignment is often included as a post-training step. The purpose of this stage is to have the model learn desirable responses and behaviors from the knowledge it has obtained in previous training steps. For example, while a coding model should understand common programming mistakes, it should output high-quality, correct code in normal circumstances (Rafailov et al., 2024).

We experiment with two preference alignment methods: Direct Preference Optimization (DPO) and Odds Ratio Preference Optimization (ORPO). DPO is often used because of its simplicity compared to other methods, such as reinforcement learning from human feedback (RLHF), by optimizing a simple binary cross entropy objective between preferred and dispreferred responses (Rafailov et al., 2024).

While DPO requires instruction-tuning beforehand, ORPO performs model alignment during the supervised fine-tuning stage by including an odds ratio-based penalty to the negative log-likelihood loss (NLL), thus simplifying post-training further (Hong et al., 2024). While ORPO is usually used from a base model before instruction-tuning, we wish to keep the number of potential confounding factors to a minimum (i.e., the order and number of post-training steps), and so perform preference alignment with ORPO after instruction-tuning. We select Llawa-TCxYZL-Instruct as our base model due to its superior performance on our evaluation benchmarks. In lieu of human annotators, we synthetically construct preference pairs from the training and validation datasets by designating the ground truth as the preferred response and the base model's output for the same prompt as the dispreferred response.

Preference optimization was conducted using 2 x NVIDIA A100 80GB GPUs. We use TRL's (von Werra et al., 2020)[11] implementation of DPO and ORPO. We set DPO's and

[11] We use v0.9.6.

ORPO's beta parameters to 0.01 and 0.1, respectively. We train for a maximum of three epochs or until loss on the evaluation dataset converges.[12]

3.4. Baseline Models

Besides training Llawa, we instruction-tune two additional models using LoRA:

1. **Bllawa**, fine-tuned from Meta-Llama-3-8B-Instruct[13]

2. **Blawstral**, fine-tuned from Mistral-Nemo-Instruct-2407[14]

Bllawa and Blawstral are trained by applying LoRA on 1 x NVIDIA A100 80GB to all linear layers ($r = 64$, $\alpha = 128$, $lr = 5 \times 10^-5$). We use the Unsloth (Han and Han, 2024) library, which uses custom-written kernels that reduce memory usage and training time. We train for three epochs on our legal dataset.

The purpose of training these two models is to see if performing only LoRA can achieve comparable results. LoRA is a parameter-efficient method of fine-tuning a model and is more accessible to those with fewer resources. Furthermore, TAIDE was continuously pre-trained from Meta-Llama-3-8B. By fine-tuning Llama-3-8B-Instruct on our legal tasks, we can observe how the additional knowledge gained from TAIDE's pre-training and our pre-training stages benefit the final model.

Task	Data Source	Instances	Input	Output
A	Bar and Judicial Exam	662	Multiple-choice question with 4 options	Answer (*sg.*)
B	Taiwan Jurist Journal	242	Multiple-choice question with 4-6 options	Answer (*sg./pl.*)
C	Legal Symposium of Taiwan High Court	1854	Issue for discussion with multiple arguments	Final argument
D	Bar and Judicial Exam	40	Hypothetical legal scenario	Essay

Table 2: Overview of Tasks A, B, C, and D. *sg.* represents an answer with a single option, while *pl.* represents an answer with multiple options.

[12] Details for preference training can be found in the supplementary materials: https://osf.io/sxjbr/?view_only=59bc169477ae4eea84dab4b5092d2046

[13] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[14] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

Model	**Task A**	**Task B**	**Task C**	**Task D**
gpt-4-turbo	53.93	43.80	49.43	80.85
Mistral-Nemo-Instruct-2407	27.19	30.99	51.28	**74.26**
Blawstral	**37.76**	**37.19**	**56.54**	62.03
Meta-Llama-3-8B-Instruct	**38.82**	**38.84**	51.64	**68.25**
Bllawa	36.56	32.64	**52.18**	51.42
Llama-3-TAIDE-LX-8B-Chat-Alpha1	24.92	**33.47**	51.40	77.62
Llawa-TC-YZL-Instruct	25.38	30.17	**53.49**	0
Llawa-TCxYZL-Instruct	**28.55**	**33.47**	53.07	0
Llawa-TCxYZL-ORPO	25.38	29.34	45.67	0
Llawa-TCxYZL-DPO	24.47	28.51	43.94	0

Table 3: Model performance of the three tasks. The metric is accuracy (%) for Tasks A, B, and C. Task D used GPT-4o as the evaluator in which each model's response was graded against the reference answer.

4. Experiments

We create four legal reasoning tasks as datasets,[15] detailed in Table 2. We conduct a comparison between our models and other open-source models. For each model, we employ greedy decoding for generation. The maximum input token length is 7192, with prompts exceeding this limit truncated on the right. All models are assessed in one-shot settings, with an example question-answer pair following the instructions.

4.1. Tasks A&B: Multiple Choice Questions

Table 3 summarizes model performance across all tasks. For Tasks A and B, gpt-4-turbo leads among all models on multiple-choice question answering. While fine-tuning is expected to enhance task-specific performance, our results show a more nuanced pattern. Notably, Meta-Llama-3-8B-Instruct surpasses Bllawa on both tasks despite the latter being fine-tuned for them, whereas Blawstral (fine-tuned from Mistral-Nemo-Instruct-2407) demonstrates marked improvement compared to its base model.

The performance drop from Meta-Llama-3-8B-Instruct to Bllawa may stem from the former's extensive post-training processes (SFT, rejection sampling, DPO), which yield stronger reasoning and test-taking skills. Our fine-tuning process may have inadvertently over-specialized Bllawa, improving in-domain generation but weakening general evaluative abilities. This highlights an open question regarding the optimal balance of domain-specific fine-tuning versus retention of general reasoning skills.

Llawa-TCxYZL-Instruct, adapted from Llama-3-TAIDE-LX-8B-Chat-Alpha1, shows slightly higher accuracy on Task A and comparable results on Task B. Since its gains over the base model are marginal, we examined the effects of preference optimization (DPO

[15] The data for Tasks A, B, and D were sourced from Taiwan's Bar and Judicial Exam, while the data for Task C were gathered from the Taiwan High Court website (https://tph.judicial.gov.tw/tw/np-1131-051.html).

and ORPO). Using ground-truth outputs as “preferred” and model responses as “rejected” did not improve performance.

Possible explanations include: (1) suboptimal preference pair design, (2) low data diversity causing overfitting, and (3) insufficient hyperparameter tuning due to limited resources and time. Future research should explore these factors to clarify the impact of preference optimization in fine-tuning and address the discrepancies observed here.

Input		**Output**	
Issue	Opinions	Final Argument	Review/Research Opinion
After the commencement of a juvenile court trial, can the victim of a crime in a juvenile case file a supplementary civil lawsuit to seek damages?	Opinion A: According to Article 1 of the Juvenile Delinquency Act, which states... Opinion B: The nature of juvenile corrective measures is fundamentally different...	Opinion B is adopted. The Juvenile Delinquency Act explicitly enumerates the provisions of the Code of Criminal Procedure that apply (such as in Articles 16 and 24 of the Juvenile Delinquency Act). Since there is no provision in the Juvenile Delinquency Act for applying the supplementary civil lawsuit procedures of the Code of Criminal Procedure, they cannot be applied.	Judicial Yuan Second Division Research Opinion: The research opinion agrees with the discussion result and finds Opinion B to be correct. However, the issue originally referred to ”the crime victim in a juvenile case,” which is a misnomer and should be corrected to ”the victim in a juvenile corrective measure case.” As for the crime victim in a juvenile criminal case, they may still file a supplementary civil lawsuit.

Table 4: Model input and output for Task C. This is an instance of legal symposium: Judicial Yuan Opinion No. 059 (1980). For brevity, the details of the two opinions are omitted. The full text with English translation is in Appendix G.

4.2. Task C: Argumentation in Legal Symposium

For Task C, we prompt the LLMs to select a specific stance based on the arguments within a legal symposium.

4.2.1. Legal Symposium

Taiwan’s annual legal symposium is organized by the High Court, High Administrative Court, and Intellectual Property Court. The case-handling process generally follows four steps: (1) the proposing court raises an issue and presents opinions (e.g., Opinion A and B); (2) its opinion forms the preliminary discussion result; (3) the High Court panel issues a review opinion, possibly adding new views (e.g., Opinion C); (4) a discussion result is determined by voting among court, prosecutor, and lawyer representatives.

Outcomes are classified as preliminary discussion results, review opinions, and discussion results.[16] These outcomes are non-binding for the Supreme Court but provide references for judicial practice, legal research, and education.

4.2.2. Task Description and Results

For the input, we provide the models with information about a legal symposium: its metadata, the legal issue being discussed, and the arguments (Opinion A and B), as shown in

[16] **Preliminary discussion result**: the proposing court’s original position. **Review opinion**: the High Court panel’s assessment, which may support, revise, or add new perspectives. **Discussion result**: the final vote of symposium participants, which task C models should match.

Table 4. The task requires generating the *discussion results*, which involve nuanced outcomes—such as partially adopting an argument, affirming or denying a claim, reserving judgment, or referring the case to a higher court—depending on the context and number of arguments.

As shown in Task C of Table 3, Blawstral achieves the best performance, followed by Llawa-TC-YZL-Instruct and Llawa-TCxYZL-Instruct. Bllawa also surpasses its base model, Meta-Llama-3-8B-Instruct, indicating its strength in generating final opinions from legal arguments. In contrast, gpt-4-turbo, Mistral-Nemo-Instruct-2407, and Meta-Llama-3-8B-Instruct show lower accuracy, possibly due to limited exposure to Taiwanese legal data and the inherent complexity of symposium discussions, where factual bases are often incomplete or conflicting (Bongiovanni et al., 2018).

In actual symposia, the final discussion results are from group discussion and voting, and even a selected opinion may be revised. Task C evaluates model outputs only against the final discussion result, without considering the intermediate review process, which is a limitation to be addressed in future work.

4.3. Task D: Essay Questions in the Bar and Judicial Exams

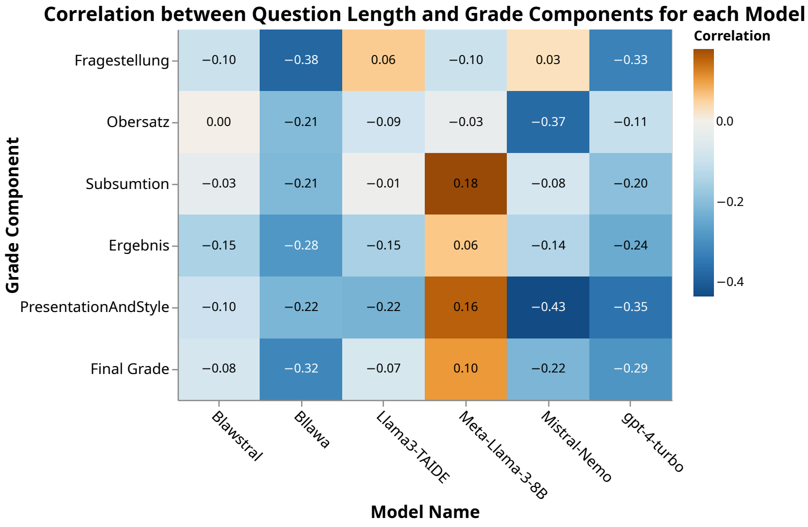

Figure 1: Correlation between question length and the grade components for each model. The six components include Fragestellung, Obersatz, Subsumtion, Ergebnis (four components of the Juristisches Gutachten legal reasoning method), presentation and style, and the final grade.

For Task D, we use criminal law essay questions from the bar and judicial exam dataset. Unlike previous benchmarks, this task is difficult to evaluate due to the open-ended nature of essay responses. It examines both the effect of question length on model performance and the models' ability to handle complex legal reasoning.

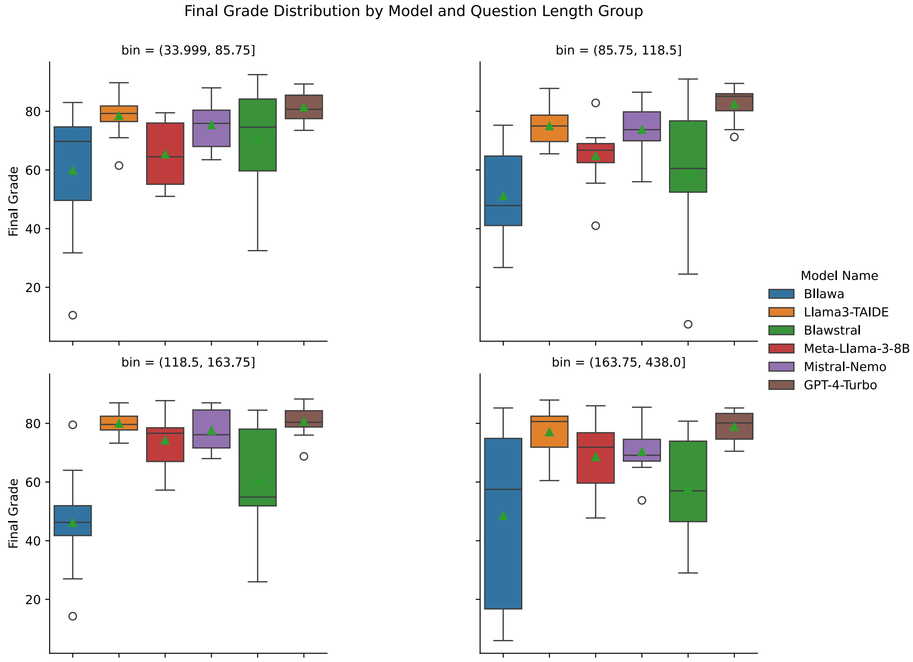

Figure 2: Model performance varies as question length increases for many of the models.

GPT-4-Turbo is prompted to segment the reference answers into four components,[17] following the German Juristisches Gutachten format for legal reasoning. The segmented reference answers are then reviewed by law school experts. Next, we divide the dataset into four subsets by question length and randomly select ten questions from each, totaling forty questions. Each tested model generates answers using the four-part Gutachten format. Finally, GPT-4o evaluates chunk-level similarity and legal reasoning quality against the expert-viewed reference answers. The instruction prompt is shown in Appendix F.

Figure 1 shows that the correlation between question length and score components across models is relatively weak (−0.43 to 0.18). As illustrated in Figure 2, question length slightly affects certain models (e.g., Bllawa) but has minimal influence on others such as GPT-4-Turbo. Mann–Whitney U tests reveal significant performance differences among most model pairs (13/15, $p<0.05$) with large effect sizes (9/15, $d>0.8$), while ANOVA results indicate no significant within-model effect of question length ($p>0.05$). Robustness, however, varies considerably (GPT-4-Turbo: 3.2% vs. Bllawa: 18.9% drop), suggesting that model architecture and training largely determine resilience to question length variation.[18]

In Task D, GPT-4-Turbo scored highest (80.85), reflecting its strong instruction-following and multi-step reasoning abilities. TAIDE followed (77.62), benefiting from both

[17] The four components include: Fragstellung (problem statement), Obersatz (general principle or major premise), Subsumtion (application of law to facts), and Eergebnis (conclusion). This can be analogous to the IRAC reasoning steps proposed in the LegalBench benchmark: Issue, Rule, Application, and Conclusion, which is a framework American legal scholars use to execute legal reasoning. A detailed description is shown in Appendix E.

[18] Detailed statistics available in the supplementary materials: https://osf.io/sxjbr/?view_only=59bc169477ae4eea84dab4b5092d2046

general reasoning and familiarity with Taiwanese legal texts. By contrast, Llawa variants failed to generalize, producing repetitive or incoherent text and scoring near zero. General instruction-tuned models performed moderately (Mistral-Nemo-Instruct: 74.26; Meta-Llama-3-8B-Instruct: 68.25), while LoRA-based models underperformed (Blawstral: 62.03; Bllawa: 51.42). Overall, Task D favors models with strong multi-step reasoning and structured generation, while excessive domain-specific tuning may reduce flexibility and degrade performance.

As summarized in Table 3, the strong performance of Meta-Llama-3-8B-Instruct across most tasks (except Task C) suggests that extensive post-training for general reasoning and instruction-following can outweigh domain-specific continual pre-training. The degradation observed in Bllawa supports this view: excessive in-domain fine-tuning does not necessarily lead to improvement and may harm general reasoning skills. This supports our main argument that DACP is not always the most optimal strategy.

5. Conclusion

This paper examined the effect of Domain-Adaptive Continual Pre-Training (DACP) on Legal LLMs across general and complex reasoning tasks in Taiwanese Mandarin. While DACP enhances domain-specific reasoning by integrating legal knowledge, its benefits are task-dependent. It can strengthen specialized reasoning yet diminish the model's performance on prompt-based or general tasks, indicating a trade-off between specialization and generalization. Future work should explore hybrid strategies, such as combining DACP with task-specific fine-tuning or meta-learning, to balance legal expertise with broader reasoning skills. Improved evaluation benchmarks are also needed to better reflect the diversity and complexity of legal reasoning tasks.

6. Limitations

A limitation of our study is the potential for data contamination in the pre-training corpora. In the legal domain, some overlap between training and evaluation data is inevitable, as foundational public documents like statutes and court judgments are ubiquitous. To mitigate this risk, however, we took several precautions. A substantial portion of our continual pre-training corpus consists of non-public legal documents, which are inherently separate from public benchmarks. Moreover, we deliberately sourced our evaluation data from distinct, non-public datasets, thereby minimizing data leakage and preserving the credibility of our results. The *optimal mixture ratio* between the general corpus and the legal-domain corpus remains an open question, which Que et al. (2024) also highlighted.

Our fine-tuning methodology introduces another limitation. Our approach to generating preference data deviates from the standard practice of using human annotators. We recognize that using model-generated outputs as rejected responses, while scalable, is a potential constraint as it may not reflect the same error patterns or quality distribution found in human-curated data.

A key limitation is our evaluation methodology. Traditional metrics like BLEU or ROUGE are insufficient for capturing the nuances of flexible legal reasoning. While we use GPT-4o for certain tasks, we recognize that using LLMs as evaluators can introduce potential biases and lacks full transparency. This challenge persists even with the development of more comprehensive benchmarks. Therefore, the creation of specialized, human-aligned evaluation frameworks for advanced legal tasks remains a critical need for the field.

Acknowledgments

We thank Angle Publishing for providing access to educational materials, the NVIDIA-NTU Artificial Intelligence Joint Research Center for providing access to the Taipei-1 supercomputer, and the NTU Research Center for Digital Law.

References

Guilherme F.C.F. Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. Exploring the psychology of llms' moral and legal reasoning. *Artificial Intelligence*, 333:104145, August 2024. ISSN 0004-3702. doi: 10.1016/j.artint.2024.104145. URL http://dx.doi.org/10.1016/j.artint.2024.104145.

Axolotl AI. axolotl-ai-cloud/axolotl, 9 2024. URL https://github.com/axolotl-ai-cloud/axolotl.

Giorgio Bongiovanni, Gerald Postema, Antonino Rotolo, Giovanni Sartor, Chiara Valentini, and Douglas Walton. *Handbook of legal reasoning and argumentation*. Springer, 2018.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.634. URL https://aclanthology.org/2020.emnlp-main.634.

Daixuan Cheng, Shaohan Huang, and Furu Wei. Adapting large language models via reading comprehension. *arXiv preprint arXiv:2309.09530*, 2023.

Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for chinese llama and alpaca, 2024. URL https://arxiv.org/abs/2304.08177.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. Laiw: A chinese legal large language models benchmark, 2024. URL https://arxiv.org/abs/2310.05620.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. Lawbench: Benchmarking legal knowledge of large language models, 2023. URL https://arxiv.org/abs/2309.16289.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, October 2021. ISSN 2637-8051. doi: 10.1145/3458754. URL http://dx.doi.org/10.1145/3458754.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, 2023. URL https://arxiv.org/abs/2308.11462.

Zhen Guo and Yining Hua. Continuous training and fine-tuning for domain-specific language models in medical question answering, 2023. URL https://arxiv.org/abs/2311.00204.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740.

Michael Han and Daniel Han. Unsloth. https://github.com/unslothai/unsloth, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model, 2024. URL https://arxiv.org/abs/2403.07691.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models, 2023. URL https://arxiv.org/abs/2305.08322.

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. Lifelong pretraining: Continually adapting language models to emerging corpora. In Angela Fan, Suzana Ilic, Thomas Wolf, and Matthias Gallé, editors, *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 1–16, virtual+Dublin, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bigscience-1.1. URL https://aclanthology.org/2022.bigscience-1.1.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. Large language models in law: A survey, 2023. URL https://arxiv.org/abs/2312.03718.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz682. URL http://dx.doi.org/10.1093/bioinformatics/btz682.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese, 2024. URL https://arxiv.org/abs/2306.09212.

Yen-Ting Lin and Yun-Nung Chen. Taiwan LLM: bridging the linguistic divide with a culturally aligned language model. *CoRR*, abs/2311.17487, 2023. doi: 10.48550/ARXIV.2311.17487. URL https://doi.org/10.48550/arXiv.2311.17487.

Shirong Ma, Shen Huang, Shulin Huang, Xiaobin Wang, Yangning Li, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. Ecomgpt-ct: Continual pre-training of e-commerce large language models with semi-structured data, 2023. URL https://arxiv.org/abs/2312.15696.

Meta. Introducing Meta Llama 3: The most capable openly available LLM to date, 4 2024. URL https://ai.meta.com/blog/meta-llama-3/. Accessed: 2024-08-15.

Mitodru Niyogi and Arnab Bhattacharya. Paramanu-ayn: Pretrain from scratch or continual pretraining of llms for legal domain adaptation?, 2024. URL https://arxiv.org/abs/2403.13681.

Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yinghao Ma, Feiyu Duan, Zhiqi Bai, Jiakai Wang, Yuanxing Zhang, Xu Tan, Jie Fu, Wenbo Su, Jiamang Wang, Lin Qu, and Bo Zheng. D-cpt law: Domain-specific continual pre-training scaling law for large language models, 2024. URL https://arxiv.org/abs/2406.01375.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey, 2024. URL https://arxiv.org/abs/2404.16789.

Zhi-Rui Tam, Ya-Ting Pai, Yen-Wei Lee, Sega Cheng, and Hong-Han Shuai. An improved traditional chinese evaluation suite for foundation model. *arXiv preprint arXiv:2403.01858*, 2024.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023. URL https://arxiv.org/abs/2303.17564.

Tongtong Wu, Massimo Caccia, Zhuang Li, Yuan Fang Li, Guilin Qi, and Gholamreza Haffari. Pretrained language model in continual learning: A comparative study. In Yann LeCun, editor, *International Conference on Learning Representations 2022*. OpenReview, 2022. URL https://openreview.net/group?id=ICLR.cc/2022/Conference,https://iclr.cc/Conferences/2022.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. Efficient continual pre-training for building domain specific large language models, 2023. URL https://arxiv.org/abs/2311.08545.

Qimin Yang, Rongsheng Wang, Jiexin Chen, Runqi Su, and Tao Tan. Fine-tuning medical language models for enhanced long-contextual understanding and domain expertise, 2024. URL https://arxiv.org/abs/2407.11536.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. Disc-lawllm: Fine-tuning large language models for intelligent legal services, 2023. URL https://arxiv.org/abs/2309.11325.

Alex Zhang. Taiwan Legal Research, 7 2019. URL https://www.nyulawglobal.org/globalex/Taiwan1.html. Accessed: 2024-08-15.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey, 2024. URL https://arxiv.org/abs/2308.10792.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters, 2023. URL https://arxiv.org/abs/2305.12002.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. Llama beyond english: An empirical study on language capability transfer, 2024. URL https://arxiv.org/abs/2401.01055.

Ziqiang Zheng, Jipeng Zhang, Tuan-Anh Vu, Shizhe Diao, Yue Him Wong Tim, and Sai-Kit Yeung. Marinegpt: Unlocking secrets of ocean to the public, 2023. URL https://arxiv.org/abs/2310.13596.

Çağatay Yıldız, Nishaanth Kanna Ravichandran, Prishruit Punia, Matthias Bethge, and Beyza Ermis. Investigating continual pretraining in large language models: Insights and implications, 2024. URL https://arxiv.org/abs/2402.17400.