EasyChair Preprint

№ 15242

Quantity Affects Quality: Instruction Fine-Tuning on LLM's Multiple-Choice Question Abilities

Hsuan-Lei Shao, Wei-Hsin Wang and Sieh-Chuen Huang

October 18, 2024

Quantity Affects Quality: Instruction Fine-Tuning on LLM's Multiple-choice Question Abilities

Hsuan-Lei SHAO
Graduate Institute of Health and Biotechnology Law,
Taipei Medical University
ORCiD ID:0000-0002-7101-5272

Wei-hsin WANG
College of Law,
National Taiwan University

Sieh-chuen HUANG
College of Law,
National Taiwan University
ORCiD ID:0000-0003-3571-5236
schhuang@ntu.edu.tw

Abstract

This paper discovered the potential of instruction fine-tuning to significantly performance of large language models (LLMs) on legal multiple-choice questions (MCQs) abilities. By manipulating the volume of training data, we aim to demonstrate a strong correlation between the quantity of data used in fine-tuning can lift LLM's quality, paving the way for LLMs in specific task (ex: legal knowledge). We compared Breeze-7B (based on Mistral-7B) and its fine-tuned version. Adding more MCQs data can enhance their abilities, there are two models: the first is adding 5,000 new samples(bz5k), and the second is 70,000(bz70k). We compare these with the general baseline model, GPT-3.5, GPT-4o, and one traditional Mandarin LLM(TAME). Then, the MCQs dataset of the MMLU, TMMLU, and the 2023 Taiwanese Bar Examination be evaluated. We find that fine-tuning LLMs might degrade its original capabilities little. However, surpassing a specific data volume can markedly enhances the model's effectiveness. This balance ensures that while the LLM's proficiency in specialized legal domains is enhanced. Practically speaking, we developed a legal MCQ-specific LLM that demonstrated the benefits of model customization. For specialized applications, smaller-scale, personalized LLMs can be developed with reduced training costs, making advanced legal tools more accessible and adaptable to specific knowledge areas or unique legal frameworks. This approach also addresses concerns about digital sovereignty by aligning the model's functionalities with jurisdiction-specific legal regulations.

1 Introduction

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

Legal informatics, an interdisciplinary field that merges legal studies with information technology, has evolved significantly since its inception. It began primarily with the automation of legal documentation and progressed to more complex applications, including data management and electronic access to statutes and case law. This development was spurred by the legal profession's need to manage increasing volumes of information and the desire for more efficient legal processes[1, 2]

In the early days, legal informatics focused on creating databases for case law and legislation, facilitating quicker and more reliable access to legal resources. As technology advanced, the field expanded to include tools for legal analysis, document automation, and even predictive technologies that could forecast litigation outcomes[3, 4]. Furthermore, visualization techniques have become an important methodological step in translating legal texts into formal languages, bridging the gap between human understanding and machine processing [5].

Recent advancements in artificial intelligence have further propelled legal informatics towards innovation. The collaboration between legal, computational, and data science communities aims to build innovative legal models to improve the existing legal system[6]. Comprehensive overviews of legal informatics, such as the work by Katz and Dolin [7], provide valuable insights into real-world applications like document review and online dispute resolution.

1.2 Challenges in Implementing LLMs Across Diverse Legal Systems

The integration of Large Language Models (LLMs) in legal practices has shown promising results in areas such as document drafting and legal research. However, applying general LLMs faces significant challenges in specialized fields like legal regulation, where every country may have distinct laws and regulations. This specificity requires LLMs to understand and adapt to diverse legal frameworks, a task that general models are currently ill-equipped

to handle[8, 9].

The risk of losing legal diversity is significant, particularly for smaller countries or unique cultural contexts. These regions often have legal nuances that are not well represented in the vast data pools used to train standard LLMs. This phenomenon, known as "sovereignty AI," highlights the need for models that respect and incorporate different jurisdictions' legal sovereignty and specificities [10, 11]. To address this, there is a growing push for developing customized LLMs that are trained on localized data sets, ensuring that the legal advice and documentation generated are relevant and compliant with local laws[12, 13].

We utilized a "Mandarin version" derivative of the Mistral-7B model, named Breeze-7B, which was further enhanced through prompt finetuning. This process involved the integration of an additional set of MCQs and answers aimed at improving the model's capabilities in legal contexts. Then, we developed two variations of this model: one fine-tuned with 5,000 new samples (bz5k) and another with 70,000 new questions (bz70k). Our findings reveal a nuanced interplay specializing legal knowledge in the performance of LLMs. Specifically, the models Breeze-7B, bz5k and bz70k showed that finetuning with insufficient data volumes can indeed degrade the model's original capabilities, negatively impacting the architecture designed for knowledge tasks. Conversely, when the data volume surpassed a certain threshold (as with bz70k), the model's effectiveness significantly improved in legal knowledge but can influent other tack performance.

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 in assisting with legal education 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 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 datasets (like 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

We use a basic method in legal LLMs–multiple choice questions (MCQs). The MSQ plays a critical role in legal education and professional assessments. 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].

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].

3.2 Model Instructure Finetuning: 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 finetuning 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 selects the correct answer and understands the context and details associated with each option, enhancing its ability to handle similar questions effectively. For example:


```
{
    "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 1: Instructure Finetuning Datasets Structure

3.3 Evaluation Design

These models were benchmarked against a general baseline model, GPT-3.5, the more advanced GPT-4, and a traditional Mandarin Large Language Model (TAME)[28]. The datasets employed for evaluation included the Multimodal Legal Understanding (MMLU)[29], the Taiwanese Multimodal Legal Understanding (TMMLU)[30], and the 2023 Taiwanese Bar Examination questions[18].

The evaluation of these models was conducted using two distinct methodologies to assess their performance in legal multiple-choice question scenarios:

- Probability Selection Method for MMLU Dataset: This method involves the extraction of probabilities corresponding to the multiple-choice options 'A', 'B', 'C', and 'D' using a specific function designed to interact with the LLM's output layers. The option with the highest probability is selected as the model's response. This approach is feasible with available LLM configurations where computational costs are within manageable limits[20] . In practical, we used the code provide by the MMLU dataset directly.

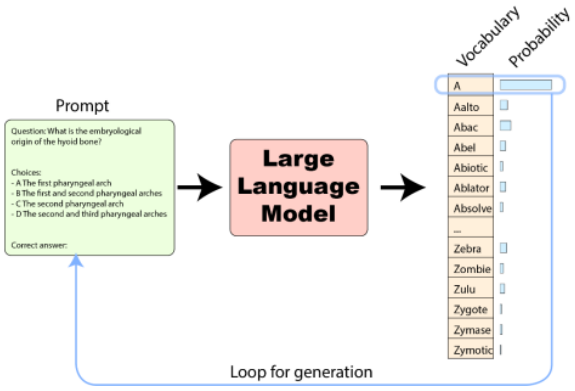

Figure 2: Probability Selection Method[25]

- Prompt-Based Zero-Shot Evaluation: The second evaluation method utilizes a zero-shot approach through direct prompting, which can be applied not only to our custom models but also in conjunction with external services such as the OpenAI API. This enables us to include and assess the performance of other models like GPT-3.5 and GPT-4 in a straightforward and practical manner, leveraging their built-in capabilities without additional fine-tuning[21].

Our approach to evaluating prompts involves a straightforward method where we directly input the question, for example,*"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"*. We then expect it to respond with "(A) Yes". Therefore, we extract the first option that appears in the output (A, B, C, or D). If none of these options (A, B, C, or D) appear in the output, we default to "C" for consistency across all models.

Given the availability of correct answers, we are able to calculate the performance of all models on MCQs. In this study, we employ "accuracy" as the criterion to assess the efficacy of each model. This metric allows us to quantitatively evaluate how well the models are performing in selecting the correct responses from the provided options[22].

Table 1: Comparing Different Finetuning Quantity Effects

Dataset\Model	Breeze	bz5k	bz70k
TMMLU(Law)	0.407	0.401	**0.486**
TMMLU(Enginnering)	**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"

4 Research Result and Discussion I: Finetuning Quantity Effect

4.1 Probability Selection Evaluation

The table "Comparing Different Finetuning Quantity Effects" showcases the impact of varying quantities of data used in finetuning 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 finetuning 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 finetuning 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 finetuning can be beneficial, excessive finetuning may lead to diminishing returns or negative transfer, where too much specificity detracts from the model's general applicability. This demonstrates that if a LLM performs better in one language, it often performs worse in another. This may be related to the parameters of individual tokens, where finetuning can detrimentally affect the original linguistic structure of the LLM.

3. The Newest Local Knowledge: TBE

   Performance on the Taiwanese Bar Examination dataset improves significantly with the highest data volume (bz70k), moving 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

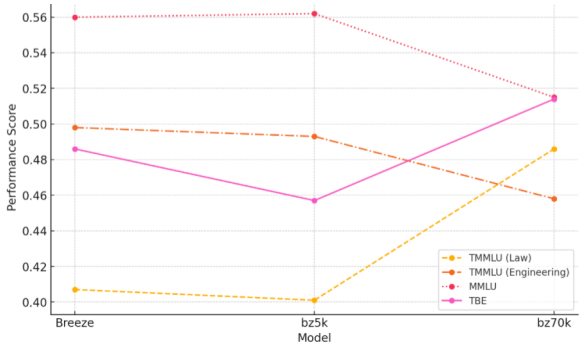

Figure 3: Model Quantity with Each Evaluation Dataset

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 finetuning 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 finetuning, potentially due to overfitting or less relevant finetuning data for engineering-specific content.

3. MMLU (dashed line): Performance slightly increases with moderate finetuning (bz5k) but decreases with extensive finetuning (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 finetuning (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 finetuning 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 finetuning to optimize performance without compromising the model's generalization capabilities. This insight is crucial for applying LLMs in specialized fields where accuracy and specificity are paramount.

5 Research Result and Discussion II: Prompt Evaluation

5.1 Prompt Evaluation

In the second phase, we input the MCQs string by the API directly (refer to 3.3. Evaluation Design), which allows us to pull other outside LLMs to compare.

Table 2: Comparing Different Finetuning Quantity Effects by Prompting Input

Dataset\Model	Breeze	bz5k	bz70k	TAME	GPT-3.5	GPT-4o
TMMLU(administrative_law)	0.250	0.380	**0.580**	0.480	0.336	0.650
MMLU	0.320	0.540	0.590	0.470	**0.660**	0.860
TBE	0.106	0.423	**0.640**	0.390	0.423	0.680

The table provides comparative performance data for different language models on MCQs across three datasets: TMMLU (administrative law, subcategory 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 because its performance is too strong, making it only possible for us to attempt to approach its performance; moreover, because its size is much larger, it is not comparable to our approximately 7B parameter model. We can see the trends on the Breeze-series and other LLMs:

1. Impact of Finetuning Method: Because our instruct fine-tuning itself has enough MCQs diversity, the bz70k model can achieve high performance when we ask it directly. It happens on TAME's performance, which is on other fields of TMMLU performances are better than bz70k. Only because it was not "familiar" with the instructions.

2. Quantity Affects Quality: This table clearly illustrates that finetuning 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.

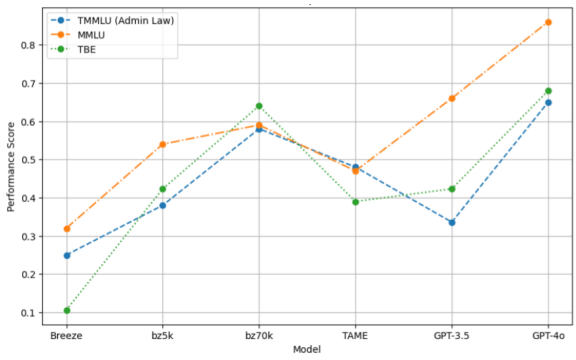

Figure 4: More Model Performance with Each Evaluation Dataset

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 finetuning 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 finetuning 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 finetuning 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.

6 Conclusion

6.1 Finetuned Small LLMs Can Battle

The fine-tuned trend observed from Breeze through bz5k to bz70k highlights the significant role that the volume and specificity of finetuning data play in enhancing model performance across diverse legal datasets. We particularly emphasize the representativeness and practicality of the Taiwanese Bar Examination (TBE) dataset as a significant representative of local knowledge, underscoring its value for testing the efficacy of AI models in handling specific legal contexts relevant to Taiwan. This aligns with the broader need in AI development for datasets deeply embedded in particular legal and cultural environments, thus serving as practical tools for evaluating how well AI technologies can adapt to localized conditions.

The success of bz70k in accurately handling the TBE dataset indicates that with sufficient targeted training, LLMs can achieve high levels of proficiency in legal reasoning and analysis. This is promising for deploying AI in legal practices, where accuracy, understanding of local laws, and practical applicability are paramount. This progression supports the effectiveness of incremental finetuning strategies and emphasizes the necessity of aligning model training with the specific demands of the tasks and datasets to optimize performance in specialized applications like law.

In other words, the central assertion of the text is that the quantity of data, particularly when it is of high quality, plays a crucial role in enhancing the performance of AI models. This principle is reflected in the paper's title, "Quantity Affects Quality," which posits that substantial inputs in terms of data can translate into significant improvements in the output capabilities of a model, even if the initial performance of the model is not particularly strong. While the quantity of the finetune data is highlighted as a key factor, the quality of this data is equally important. High-quality data for finetuning ensures that the model learns relevant and accurate information, which is crucial for effectively applying the model in real-world scenarios.

Our improvement plan has achieved the original research targets:

1. Localized and Personalized LLMs: Digital Sovereignty and Localization: The project successfully integrates localized legal knowledge into LLMs through finetuning processes. This approach aligns with the growing demand for digital sovereignty, where regions or organizations dictate the informational and operational contours of the technology they deploy. Personalizing LLMs to reflect local legal standards and knowledge bases enhances their practicality and relevance, ensuring that the generated content and advice are legally sound and contextually appropriate.

2. Data Volume and Model Performance: The research confirms that significant enhancements in model performance can be achieved by increasing the volume of training data used during finetuning. This finding is crucial for smaller models, which might not start out with the computational or architectural advantages of larger models like GPT-3.5 or GPT-4.0. By effectively using larger datasets for finetuning, these smaller models can bridge the performance gap, challenging the notion that bigger is always better. The acknowledgment that future advancements in base models could lever-

age similar finetuning strategies underscores the iterative nature of AI development. This forward-looking perspective encourages continuous adaptation and enhancement as newer and more robust technologies emerge.

3. Practical Application and Instructional Method: Alignment with Use Scenarios: By employing direct input of questions as instructions, the project aligns model usage with real-world application scenarios, particularly in legal practices. This method improves the practical usability of the LLM, as it mimics the actual inquiries and tasks that users would perform, thus providing more accurate and contextually relevant responses.

6.2 Research Limitations and Future Prospects:

1. Technological and Resource Limitations: The limitation in experimenting with larger models due to equipment constraints is a common challenge in computational research. Larger models typically require substantial computational resources, which may not be readily available to all research institutions. Seeking partnerships and support can facilitate access to more advanced computational resources, enabling the exploration of larger, more capable models that may offer enhanced performance and new capabilities. The pursuit of collaborative efforts with institutions that have the necessary infrastructure can accelerate research and development efforts, pooling resources for mutual benefit.

2. Expanding the Scope of Legal Informatics: While MCQs are a common format for testing and training AI systems due to their structured nature, legal reasoning represents a more complex challenge that involves understanding nuances and making judgments similar to those a human lawyer would make. Expanding LLM research to include these aspects can significantly impact the legal profession by providing tools that can assist with more sophisticated tasks. Legal Reasoning and AI: Future research could focus on enhancing the capabilities of LLMs to handle complex legal reasoning and argumentation, potentially revolutionizing how legal analysis and advice are delivered.

3. Improving Evaluation Methods: Current evaluation methods might not adequately capture the universality and reproducibility necessary for legal applications. Legal AI systems must produce consistent and reliable results under various conditions to be truly effective and trustworthy. Therefore, developing more robust evaluation frameworks that can more accurately assess the effectiveness of AI in legal contexts is essential. This might involve creating standardized datasets, developing new metrics that better reflect the quality of legal reasoning, or adopting more rigorous cross-validation techniques to ensure the AI's decisions are sound and defensible.

The concept that "Quantity Affects Quality" underlines the transformative potential of data volume and quality in AI development. It suggests that strategic finetuning with carefully selected data can significantly uplift even underperforming models, broadening the scope for AI enhancements and applications across various industries. This principle not only guides practical AI development strategies but also sets a foundation for future research into effective and efficient AI training methodologies.

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] Yenting Lin. "Llama-3-Taiwan-8B-Instruct," *Hugging Face Model*, Available at: https://huggingface.co/yentinglin/Llama-3-Taiwan-8B-Instruct, accessed on August 8, 2024.

[29] Pei-Yuan Liu. "MMLU Dataset," *Kaggle Dataset*, Available at: https://www.kaggle.com/datasets/peiyuanliu2001/mmlu-dataset, accessed on August 8, 2024.

[30] iKala. "TMMLUPlus Dataset," *Hugging Face Dataset*, Available at: https://huggingface.co/datasets/ikala/tmmluplus, accessed on August 8, 2024.