Surpassing Human Counterparts: A Breakthrough Achievement of Large Language Models in Professional Tax Qualification Examinations in China

1st Lifeng Xu*
360 AIResearch
Beijing, China
xulifeng@emails.bjut.edu.cn

1st Chuanrui Hu*†
360 AIResearch
Beijing, China
huchuanrui_1206@163.com

2rd Hua Zhang
360 AIResearch
Beijing, China
zhanghua3@360.cn

3th Jiahui Zhai
360 AIResearch
Beijing, China
zhaijiahui@emails.bjut.edu.cn

4th Wei Tang
ZSTAX
Beijing, China
tangwei@zstax.com

5th Yuchen Li
ZSTAX
Beijing, China
liyuchen1993@gmail.com

6th Zhao Peng
ZSTAX
Beijing, China
pengzhao@zstax.com

7th Qiuwu Chen
AIGCode
Beijing, China
chenqiuwu@aigcode.net

8th Shiyu Sun
ZSTAX
Beijing, China
sunshiyu@zstax.com

9th Ao Ji
360 AIResearch
Beijing, China
aoji_tjut@outlook.com

10th Yin Sun
AIGCode
Beijing, China
sunyin@aigcode.net

11th Zimou Liu
AIGCode
Beijing, China
liuzimou@aigcode.net

12th Su Wen
AIGCode
Beijing, China
wensu@aigcode.net

13th Liao Bin
AIGCode
Beijing, China
binliao@aigcode.net

***Abstract*—In recent times, the proliferation of Large Language Models (LLMs) has catalyzed advancements across various professional domains. Despite these strides, the attainment of accreditation in human professional examinations by an LLM remains elusive. Presently, this study showcases the inaugural LLM in China to achieve qualification in professional examinations, having achieved a score of 62 points in the 2023 tax qualification exam and 80 points in the 2022 tax qualification exam, surpassing the performance of human counterparts who scored 60 points. Distinguished by a unique training methodology, this work departs from conventional professional domain training. Our approach involves the initial fine-tuning of a multi-task complex model followed by the refinement of a single-task model. This methodology proves markedly more effective than direct single-task model fine-tuning. Furthermore, within the professional domain, we introduce strategic techniques that significantly enhance the LLMs' proficiency in generating responses. Additionally, this study contributes a comprehensive set of solutions tailored to the tax law domain, which can be extrapolated to other analogous domains. These solutions offer novel insights for the successful integration of LLMs into specific professional contexts. Our findings not only underscore the potential of LLMs in professional examinations but also offer practical guidelines for the effective deployment of LLMs in specialized domains, thereby fostering a new paradigm for domain-specific application.**

***Index Terms*—Tax Exam LLM, Cascade Training, Calculation Ability, Knowledge Comprehension Ability**

*These authors contributed equally to this work.
†Corresponding author.

I. Introduction

Since the emergence of ChatGPT, it has provided a new direction for generative language modeling. It has also inspired excellent generalized LLMs such as ChatGLM [1], BaiChuan2 [2], Qwen [3], etc., which have shown excellent capabilities in basic tasks. At the same time, LLM is especially important for domain-specific research, which will create an LLM with excellent domain performance. Currently, there are also some domain LLMs, such as EduChat [4], ChatLaw [5], ChatHome [6], Medical: HuaTuo [7], Zhongjing [8], and Finance: XuanYuan [9]. There is no LLM in the field of tax law, and we are the first ones to get the LLM for the qualification-type exam in China.

The Tax Exam LLM is different from other domain LLMs in that it requires sufficient knowledge comprehension as well as strong computational power to answer questions correctly. The LLMs trained in other domains learn enough domain knowledge to improve the answering effect of the LLMs, however, the data of the tax exam is less compared to other domains, and the learning resources are limited. At present, the commonly trained LLMs often fail to solve correctly once the domain knowledge questions involving multi-step computation are involved, and the fundamental reason is that the computational ability of the model is insufficiently generalized. Therefore, the training process of the LLM for the tax exam is mainly two major parts: knowledge comprehension

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00243

and logical calculation ability.

For knowledge comprehension, the model is expanded based on real tax law questions by randomly disrupting the options of the original questions and adding a certain proportion of negative sample size to ensure that the content of the options rather than the options is learned, making the model strongly robust. For complex problems, similar questions and answers are retrieved by building a question bank as reference knowledge, and then the original questions and reference knowledge are merged. By training QA pairs with such complex data, the model's understanding of complex questions is improved.

The main contributions of this paper are as follows:

1) Cascade Training: By constructing a multi-task SFT corpus to fine-tune an all-purpose tax LLM, and based on the all-purpose tax LLM in the second fine-tuning of a single task, we can get the LLM with the stronger effect of a single task.
2) Some tricks to improve model comprehension as well as logical calculations.
3) We propose a set of training frameworks to realize generalized solutions for domain-specific landings.

II. Related Work

Nowadays, with the rapid development of LLM, there is the openai team's ChatGPT and Meta's open source large language model LLaMA [10], the domestic open source model currently has ChatGLM [1], Baichuan [2], Qwen [3], Skywork [11], and so on have already achieved a more extensive influence in the general-purpose large model. Such generalized LLM expresses strong performance in content understanding and language generation by pre-training on a large number of texts and instructions.

Although the current general-purpose LLM can show superior results in various basic domains, the text generated by the LLM does not have specialist knowledge. When applied to pendant domains, LLM-generated text lacks terminology and has short and inaccurate answers. To solve this problem, several LLMs have emerged that have been fine-tuned with vertical domain knowledge, including Zhongjing [8], HuaTuo [7], ChatDoctor [12], DoctorGLM [13] in the medical field, FinGPT [14], XuanYuan [9], FLANG [15] in the financial field, EduChat [4] in the education field, and ChatLaw [5], Lawyer LLaMA [16] and DISC-LawLLM [17] in the legal field. In the field of e-commerce, there is EcomGPT [18], and in the field of self-media, there is MediaGPT [19], and so on. From the above study, it is shown that all the LLMs of major pendant domains are stronger than the general LLM in pendant knowledge Q&A. Currently, there is a lack of LLM for pendant domains that pass human professional exams, so we aim to build the first LLM to get a tax accountant qualification certificate.

Compared with other domain LLMs, the domain exam LLM does not have enough datasets but involves a wide variety of exams, including single-choice, multiple-choice, short-answer, computational, and comprehensive analysis questions. Compared with other domain LLMs, the tax exam LLM requires more accurate computation and stronger knowledge comprehension ability and has to be fine-tuned to produce a high-accuracy exam LLM based on the limited exam data and a wide range of question tasks. Therefore, the solution to the tax exam LLM is to improve the model's computational ability and knowledge comprehension ability.

TABLE I
COMPARED WITH ALL-IN-ONE LLM AND SINGLE-TASK LLM.

Model	Single ACC	Multiple ACC	Overall ACC
All-In-One LLM	0.84	0.69	0.78
Single LLM	**0.87**	**0.73**	**0.81**

III. Tax exam LLM

The tax domain is different from other general vertical domains, general vertical domains may only need to carry out the knowledge quiz, such as medical, education, government affairs, transportation, and other domains. These domains do not involve specialized computational capabilities, but the tax domain LLM not only needs to carry out a specialized knowledge quiz but also has the corresponding computational process along with the knowledge quiz process. The computation process not only involves single-step but also multi-step computation and the statistical database found that multi-step computation accounts for the majority of the questions, so the tax model needs to have complex computation ability and strong knowledge comprehension ability. Next, the three main focuses of the tax exam LLM will be introduced in detail, section III-A is the dataset, section III-C is the tax calculation ability, and section III-D is the tax knowledge comprehension ability, and section III-E is the tax question knowledge library. The structure of the Tax Exam LLM is shown in Fig. 1.

A. Datasets

Less And More The main training data of the Tax Exam LLM comes from the annual tax exam questions. The tax exam data contains the following five subjects: Tax Law (I), Tax Law (II), Laws Related to Tax-Related Services, Finance and Accounting, and Tax-Related Services Practice. These five subjects contain five types of questions such as single choice, multiple choice, calculation, general analysis, and short answer. The total number of tax exam data collected is 6000, and it can be noticed that with a very small amount of data, there is an unbalanced distribution of data for each type of subject. Overall, the amount of data on tax exam questions is small, and yet the types of tasks are numerous. This makes training an LLM of a professional tax exam a challenge. The small amount of data and the variety of question types are the drawbacks of the current professional exam field.

B. Cascade Training

All-In-One LLM And Single-Task LLM The training process of the tax exam grand model is different from that of a general pendant model, which is to build an SFT corpus, fine-tune the pre-trained model, and then measure it. The training process of the tax exam LLM is to build a multi-task SFT

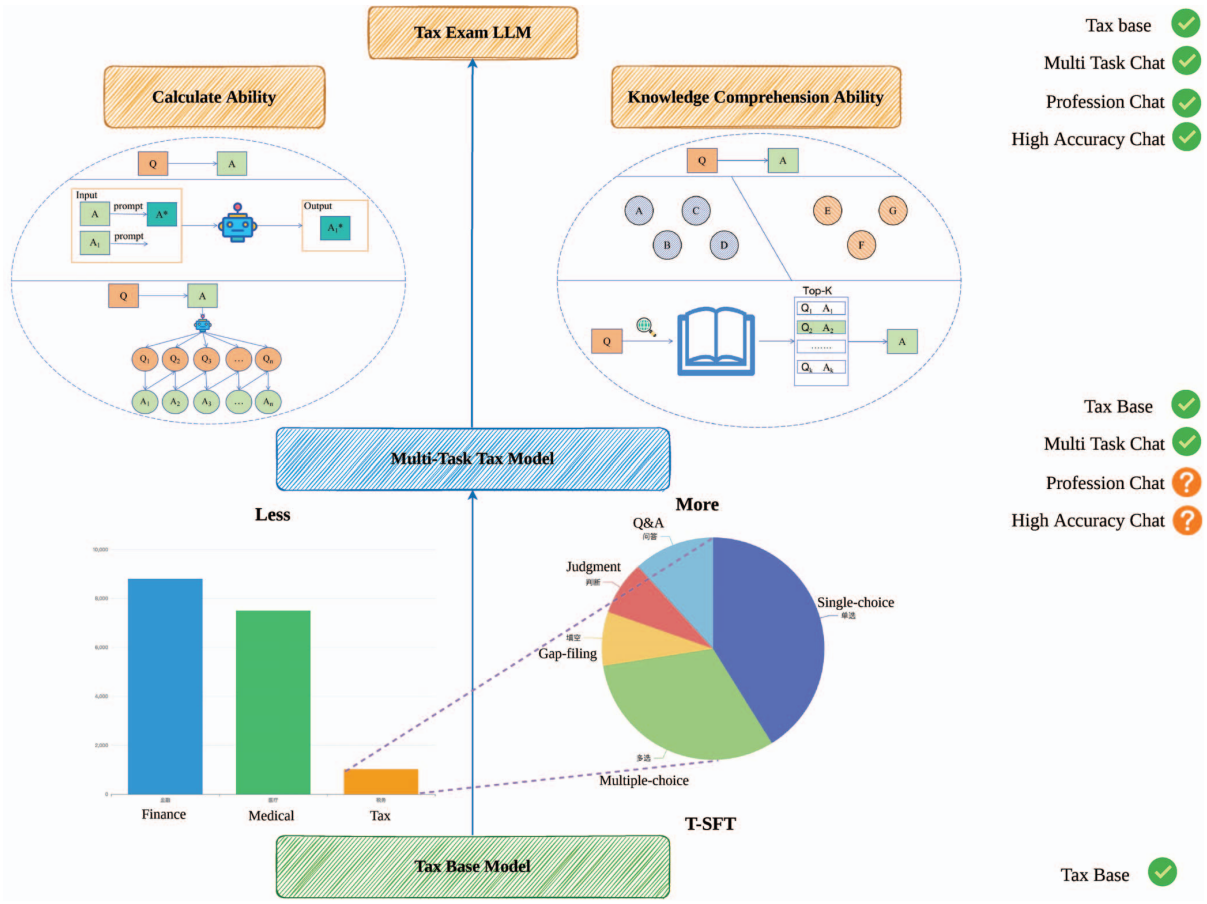

Fig. 1. Tax LLM Model Training Process: The Base Model was fine-tuned with T-SFT data to get the Multi-Task Tax Model. After that, the calculation power and knowledge understanding ability of the model were further improved through several tricks, and the Tax Exam LLM was obtained after fine-tuning

corpus to fine-tune an all-purpose tax LLM first and then fine-tune the all-purpose tax LLM for the multiple-choice SFT corpus individually to fine-tune the multiple-choice tax LLM, which is tested on multiple-choice questions with higher effect than that of the direct fine-tuning of multiple-choice questions. It is proved that the model is first fine-tuned based on complex tasks and then fine-tuned for single tasks, which will be more effective than direct fine-tuning for single tasks. The complete training process is shown in Fig. 2. A comparison of the effects is shown in Table I.

C. Tax Calculation Ability

To break through the tax calculation skills, you first need to understand the general form of the question content of the calculation task, based on the question and the answer, you will learn that the question will contain information such as numbers and tax keywords involved, but not all the numbers will be used in the calculation, there are certain interfering numbers. As shown in table III. Finally, the question should be solved based on the final problem. Therefore, three innovations are proposed to improve the computational ability of the tax exam model.

COT Formatting Solution Steps Since the solution steps of the questions are non-stepwise, the solution steps with the chain of thought (COT) [20] are obtained by designing the prompt template that enables the model to answer the stepwise steps, and using the answers of the questions and the designed prompt as input for ChatGPT to answer. The generated steps are cleaned to eliminate non-stepwise, as well as incorrectly answered Q&A pairs. The generated answers are shown in table I in the appendix.

Step By Step Disassembly For complex problems that require multi-step answers, we break the problem down into multiple sub-questions, and by answering the sub-questions, we further solve the final problem. We refer to the Least-to-Most Prompting [21]. So train a question disassembly model, which disassembles the question to get multiple sub-questions, by answering the sub-questions sequentially, splicing the answer with the previous step, and asking the next question again until the final question is solved. Therefore, we

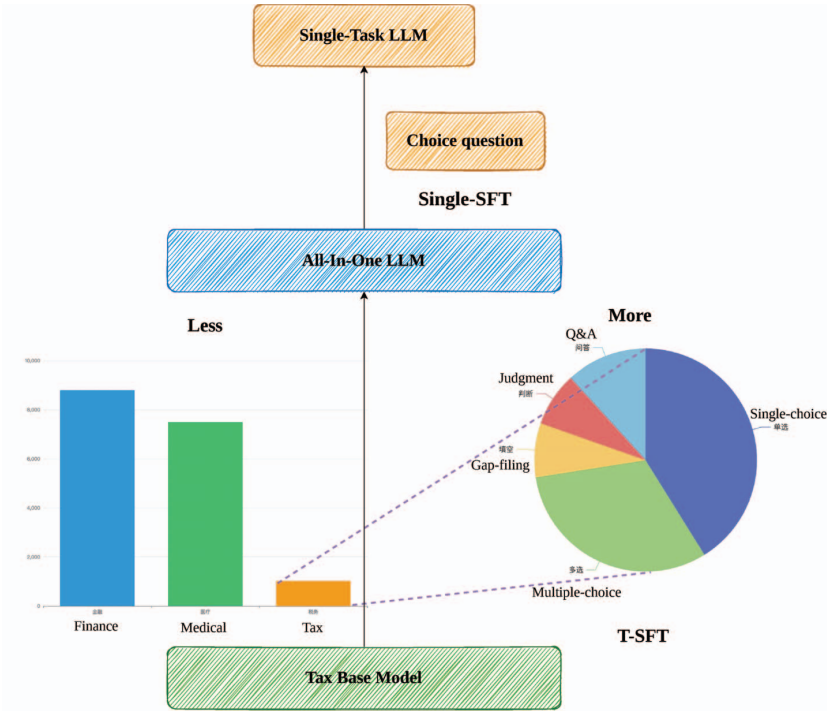

Fig. 2. Cascade Training: The All-In-One LLM is fine-tuned by multi-task data, and then fine-tuned by single-task data twice to obtain the most powerful Single-Task LLM.

TABLE II
TAX EXAM TEST RESULTS.

Model	Tax Law (I)	Tax Law (II)	Finance and Accounting	Tax-Related Services Practice	Total Score
360 Brain	0.22	0.18	0.15	0.2	0.18
All-In-One LLM	0.65	0.62	0.55	0.68	0.62
Single LLM	**0.86**	**0.84**	**0.72**	**0.87**	**0.82**

TABLE III
EXAMPLES OF TAX CALCULATION QUESTIONS.

Chinese	Translate to English
Question: 某出口公司（增值税一般纳税人）适用增值税退（免）税政策，2021年11月从生产企业购进一批高档化妆品，取得增值税专用发票注明价款20万元，增值税2.6万元，当月该批化妆品全部出口取得销售收入35万元，已知高档化妆品的消费税税率为15%，请问该出口公司出口化妆品应退的消费税合计为多少? **Answer**: 高档化妆品消费税税率=15%, 出口化妆品应退消费税=20*15%=3(万元).	**Question**: An export company (general VAT taxpayer) applies the VAT refund (exemption) policy, in November 2021 from the production enterprise to purchase a batch of high-grade cosmetic products, to obtain VAT invoices indicating the price of 200,000 yuan, 26,000 yuan of value-added tax. The batch of cosmetics exported in the same month to obtain sales revenue of 350,000 yuan, known as high-grade cosmetics consumption tax rate of 15%. The export company that exports cosmetics should be refunded the consumption tax total for how much? **Answer**: High-grade cosmetics consumption tax rate = 15%, export cosmetics should be refunded consumption tax = 20 * 15% = 3 (million yuan).

design the corresponding question disassembly SFT corpus, which has the input as the question content and the output as the disassembled sub-questions. The disassembled sub-questions are asked to ChatGPT by designing a specific prompt combined with the answers to the questions so that it can generate multiple sub-questions based on the answers to the questions. The corresponding questions disassembling the SFT corpus are shown in table II in the appendix.

SFT calculates corpus The corpus for tax calculation skill enhancement contains the following elements, the COT format of the real question corpus, the multi-round conversation corpus constructed from the multi-step disassembled questions and the answers to multiple sub-questions, and the stepwise concatenation of the answers to multiple sub-questions to form the corresponding COT formatted corpus. The content of the tax SFT calculation corpus is shown in table I, table III, and table IV in the appendix.

D. Tax Knowledge Comprehension

Options Knowledge Learning For tax knowledge comprehension, the common training method for domain modeling to master knowledge comprehension is to construct a large number of high-quality tax SFT precursors and fine-tune the Base model to master tax knowledge, but testing on multiple-choice questions reveals that its accuracy can only reach **25%**, which is consistent with the effect of blind guessing and does not use tax knowledge. Therefore, we switched to a different way of thinking and trained the questions and answer options directly as the SFT corpus and the fine-tuned model were tested with an accuracy rate of up to **60%**. Since the model output is only the options, to verify that the model learns the content of the options instead of the options, the order of the options is disrupted, and the verification is performed again, and it is found that the accuracy rate can still be maintained. Therefore, it is demonstrated that directly training questions with option answers will be a breakthrough in solving domain knowledge multiple-choice questions with large models in the future.

Negative Samples With a stable accuracy rate, to improve the accuracy of the model again, drawing on the solution of robustness in the recommendation algorithm [22], by adding negative samples to the options of the questions, the negative samples are derived from the correct options under the similar subjects, the effect is found to be improved by 7% compared to the previous one after testing, and the current accuracy rate is up to 67%, which proves that randomly increasing the negative samples can further improve the accuracy of the model. The corpus of adding negative samples is shown in table V in the appendix.

Ref Training Although the introduction of negative samples can improve the accuracy of knowledge questions, there are still some questions that fail to learn the knowledge. We borrowed a Zero-shot [23] approach. Therefore, we consider improving the model accuracy to 80% after Ref-SFT fine-tuning by introducing similar questions as references and combining them with real questions as SFT corpus. It confirms that the model effect can be further improved by Ref-Training for the knowledge that cannot be learned.

E. Tax Question Knowledge Library

Tax Question Retrieval Augmented Generation The model was able to solve most of the questions with improved computational skills and knowledge understanding but still continued to answer a small number of questions incorrectly. From the output of the model, we found that there are errors in proprietary tax rates, incorrect tax formulas, and a lack of understanding of background knowledge in the answering process. To address the above problems, we build a professional tax question vector database, which is derived from previous years' tax questions and answers as well as COT answers constructed by tax law experts, and recall the relevant solution processes by searching the questions and the vector database. The recalled solution process is used as the reference content and then combined with the questions for the model to answer. After testing, the accuracy rate is found to be improved again, proving that the problem of insufficient model capability can be successfully solved by retrieval and generation with the help of the vector database in the case of insufficient knowledge of the LLM.

Tax Vector Database The data in the Vector Database is primarily derived from previous years' tax questions and answers, as well as the COT Q&A corpus written by tax experts. To make the model answer form consistent, it is necessary to unify the COT formatting of the answers to the real questions to keep the same form as the expert answers. The answers to tax questions are formatted and unified by calling ChatGPT and giving several examples of tax experts' answers as references.

IV. Experiment

To train a large model for tax exams, we use a powerful NVIDIA 8*A100 80GB GPU with DeepSpeed distributed training framework for single-computer multi-card fine-tuning training. We use Qlora to fine-tune the 360 brain model by adjusting the lora_rank to make its model learn more parameters. To balance the training cost, we use fp16 accuracy and ZeRo-2, gradient accumulation strategy, limiting the input length to 4096 and use paged_adamw_32bit optimizer with the dropout set to 0.05. We keep 10% of the training set for validation, and the model reaches convergence after 10 epochs. The final model training parameters are shown in Table IV. The test results are shown in Table II.

V. Conclusion

In this paper, we propose the tax exam LLM, which becomes the first large language model in China to obtain the professional exam certification. We propose the method of fine-tuning the all-in-one model and then fine-tuning the single-task data to build a more outstanding model for a single task, which is far more effective than direct single-task fine-tuning. We add some effective tricks in the training process to further improve the accuracy of the model. Finally, we propose

TABLE IV
HYPER-PARAMETER SETTINGS.

Hyper parameter	Value
Precision	fp16
Epochs	10
Batch Size	8
Learning rate	2e-4
gradient_accumulation_steps	4
max_seq_length	4096
lora_rank	64
lora_alpha	16
lora_dropout	0.05
lr_scheduler_type	constant_with_warmup
warmup_steps	3000
optim	paged_adamw_32bit

a complete solution process for professional domains, which provides a solution for landing large models in professional domains.

REFERENCES

[1] Zeng A, Liu X, Du Z, et al. "Glm-130b: An open bilingual pre-trained model[J]," arXiv preprint arXiv:2210.02414, 2022.

[2] Yang A, Xiao B, Wang B, et al. "Baichuan 2: Open large-scale language models[J]," arXiv preprint arXiv:2309.10305, 2023.

[3] Bai J, Bai S, Chu Y, et al. "Qwen technical report[J]," arXiv preprint arXiv:2309.16609, 2023.

[4] Dan Y, Lei Z, Gu Y, et al. "Educhat: A large-scale language model-based chatbot system for intelligent education[J]," arXiv preprint arXiv:2308.02773, 2023.

[5] Cui J, Li Z, Yan Y, et al. "Chatlaw: Open-source legal large language model with integrated external knowledge bases[J]," arXiv preprint arXiv:2306.16092, 2023.

[6] Wen C, Sun X, Zhao S, et al. "ChatHome: Development and Evaluation of a Domain-Specific Language Model for Home Renovation[J]," arXiv preprint arXiv:2307.15290, 2023.

[7] Wang H, Liu C, Xi N, et al. "Huatuo: Tuning llama model with chinese medical knowledge[J]," arXiv preprint arXiv:2304.06975, 2023.

[8] Yang S, Zhao H, Zhu S, et al. "Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue[J]," arXiv preprint arXiv:2308.03549, 2023.

[9] Zhang X, Yang Q. "Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters[C]," //Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 2023: 4435-4439.

[10] Touvron H, Lavril T, Izacard G, et al. "Llama: Open and efficient foundation language models[J]," arXiv preprint arXiv:2302.13971, 2023.

[11] Wei T, Zhao L, Zhang L, et al. "Skywork: A more open bilingual foundation model[J]," arXiv preprint arXiv:2310.19341, 2023.

[12] Yunxiang L, Zihan L, Kai Z, et al. "Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge[J]," arXiv preprint arXiv:2303.14070, 2023.

[13] Xiong H, Wang S, Zhu Y, et al. "Doctorglm: Fine-tuning your chinese doctor is not a herculean task[J]," arXiv preprint arXiv:2304.01097, 2023.

[14] Yang H, Liu X Y, Wang C D. "FinGPT: Open-Source Financial Large Language Models[J]," arXiv preprint arXiv:2306.06031, 2023.

[15] Shah R S, Chawla K, Eidnani D, et al. "When flue meets flang: Benchmarks and large pre-trained language model for financial domain[J]," arXiv preprint arXiv:2211.00083, 2022.

[16] Huang Q, Tao M, An Z, et al. "Lawyer LLaMA Technical Report[J]," arXiv preprint arXiv:2305.15062, 2023.

[17] Yue S, Chen W, Wang S, et al. "Disc-lawllm: Fine-tuning large language models for intelligent legal services[J]," arXiv preprint arXiv:2309.11325, 2023.

[18] Li Y, Ma S, Wang X, et al. "EcomGPT: Instruction-tuning Large Language Model with Chain-of-Task Tasks for E-commerce[J]," arXiv preprint arXiv:2308.06966, 2023.

[19] Wang Z. "Mediagpt: A large language model target chinese media[J]," arXiv preprint arXiv:2307.10930, 2023.

[20] Wei J, Wang X, Schuurmans D, et al. "Chain-of-thought prompting elicits reasoning in large language models[J]," Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[21] Zhou D, Schärli N, Hou L, et al. "Least-to-most prompting enables complex reasoning in large language models[J]," arXiv preprint arXiv:2205.10625, 2022.

[22] Fei H, Tan S, Guo P, et al. "Sample optimization for display advertising[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management," 2020: 2017-2020.

[23] Kojima T, Gu S S, Reid M, et al. "Large language models are zero-shot reasoners[J]. Advances in neural information processing systems," 2022, 35: 22199-22213.