AuditLLM: A Compact and Domain-Specialized Large Language Model Family for Intelligent Auditing via Two-Stage Continual Pre-Training

Anonymous ACL submission

Abstract

Domain-specific LLMs have become an increasingly important research issue in recent years and various LLMs has been proposed to specific domains, such as finance, healthcare and legal. However, the current LLMs adopted in auditing faces critical challenges like cloud-API restrictions under data privacy compliance, hardware limitations in deploying trillion-parameter models, and deficiencies in factual accuracy and logical rigor exhibited by general-purpose LLMs in auditing contexts. This paper addresses training LLMs for auditing and proposes a two-phase framework to develop compact, audit-specialized LLMs tailored for Chinese auditing workflows. First, Qwen2.5 is selected as the base model through systematic comparisons of sub-5B parameter architectures. Subsequently, domain-adaptive continual pre-training by a carefully designed data sampling strategy is performed on a curated corpus of Chinese audit texts to inject domain expertise. Finally, multi-task instruction-tuning aligns the model with practical audit requirements. Extensive experiments demonstrate that the proposed framework can significantly improve the performance of domain specific LLMs in audit tasks, enhancing their accuracy and practicality for real-world applications. This study underscores the importance of domain-adaptive pre-training. The source codes, models, and audit-domain dataset are publicly available at https://anonymous.4open.science/r/AuditLLM-E004

1 Introduction

The rapid advancement of large language model (LLM) has revolutionized natural language processing, enabling breakthroughs across diverse domains such as healthcare, finance and legal services (Brown et al., 2020). Increasingly, domain LLMs fine-tuning specialized tasks have emerged to address unique requirements, including compliance analysis, medical diagnostics and contract review.

In audit domain, the evolution of intelligent auditing system has intensified the demand for tailored LLMs capable of handling domain-specific tasks such as regulatory compliance verification, financial anomaly detection and risk assessment (Onwubuariri et al., 2024). Initial attempts to employ LLMs have either relied on cloud-based APIs or focused on developing retrieval-augmented LLM systems for rule-based compliance checking or structured data processing.

However, intelligent audit workflows impose three critical constraints: (1) Uploading enterprise data to cloud services is unsuitable for auditing duo to data privacy and regulatory compliance consideration. (2) Deploying a large-scale model with hundreds of billions of parameters present significant hardware resource challenges for enterprises (Victor et al., 2019). (3)Based on our overall evaluation of current mainstream LLMs, existing general-purposed LLMs have exhibited deficiencies in factual accuracy and logical rigor in their outputs in auditing applications.

This gap underscores the urgent demand for compact yet high-performing language models tailored to modern auditing practices. A related open-source LLM, named AuditWen (Huang et al., 2024) is proposed for auditing by fine-tuning Qwen-7B, which shows significant performance on various of audit NLP tasks compared with the state-of-the-art LLMs. In this study, we focus on continual pre-training of LLMs for Chinese auditing and propose a two-phase framework to develop a compact, audit-specialized LLM. This initiative will lead to the release of a series of auditing-specific LLMs.

First, following comparative benchmarking of sub-5B parameter models, Qwen2.5 was selected as the foundational architecture for subsequent training. Second, continuous pre-training is performed on a curated corpus of Chinese audit-related texts (including normative documents, audit reports and audit cases) to inject domain-specific

knowledge into the base model. Finally, fine-tune the continual pre-trained model with instruction dataset with consisting of multi-audit-specific-tasks. This approach not only preserves the ability to semantically understand unstructured audit data, but also ensures efficient inference of the 5B model on resource-constrained mobile and embedded devices, enabling practical adoption in field auditing workflows. To our knowledge, no prior work has established a localized, parameter-efficient LLM family explicitly designed for intelligent auditing while balancing domain expertise and computational practicality. The contributions of this study are as follows:

(1) An two-stage continual pre-training framework for domain LLM. We systematically explore methodologies for adapting compact models to auditing needs, including a two-stage training framework (i.e., domain-adaptive pre-training followed by task-specified instruction-tuning) and parameter-efficient optimization strategies.

(2) First open-source continual pre-trianed audit LLM family. We introduce the first family of Chinese audit-specialized LLMs with models ranging from 0.5B to 3B parameters, including both base and instruction-tuned variants.These models are designed to bridge the gap between domain expertise and deployability intelligent auditing.

(3) An audit-domain dataset for continual pre-train LLM. To support further research, we will openly release a comprehensive audit-domain dataset for continual pre-training LLMs.

Outstanding performance. Extensive experiments results suggest that continual pre-training enhances domain knowledge absorption, while instruction-tuning aligns models with practical audit workflows.

2 Related Works

Continual pre-train learning. Existing works emphasize the importance of continual pre-training with high-quality knowledge data to enhance language model's performance (Luo et al., 2022; Beltagy et al., 2019), where the typical representative models include BloombergGPT (Wu et al., 2023) and FinBERT (Liu et al., 2020). During the continual pre-training process of LLMs, new datasets from emerging domains (such as the medical field (Yuan et al., 2024)) or those targeting specific tasks (such as event temporal reasoning (Han et al., 2021)) are collected and used to update pre-trained models, rather than re-training them from scratch. Current methods of continual pre-train mainly focused on the effects of parameters and their combination on the train process, such as warm-up strategies (Gupta et al., 2023), learning rate (LR) re-warming (Ibrahim et al., 2024), LR re-decaying (Ibrahim et al., 2024; Raffel et al., 2020), and replay of previous data (Ibrahim et al., 2024). In addition, it becomes available and is a much cheaper and more efficient solution to enable pre-trained LLM to adapt domain-specific knowledge. Xie et al. (2024b) proposed a data selection strategy with just 10% of corpus size that reduces the computational cost of continuous pre-training. Wu et al. (2024) introduced LLaMA Pro, enabling continual LLM pre-training via Transformer block expansion to learn new tasks without catastrophic forgetting. Que et al. (2024) introduced the D-CPT law to minimize domain loss by fixing model sizes and training token volumes, while Gu et al. (2024) proposed critical mixture ratio of general and domain data to trade-off between general and domain-specific capabilities.

SFT for domain LLMs. Supervised fine-tuning (SFT) technology of LLMs for domain tasks has emerged to improve the adaptability of LLMs on domain tasks with lower data volume and training cost. A methodology for obtaining a domain-adapted LLM involves fine-tuning a domain-specific base LLM using specialized domain target tasks, like XuanYuan 2.0 (Xuanyu and Qing, 2023) built upon the BLOOM-176B, Baichuan4-Finance (Zhang et al., 2024) built upon the Baichuan4-Turbo base model. Other studies explore directly fine-tuning general open-source LLMs to adapt downstream tasks. For example, PiXiu (Xie et al., 2023) and FinBen (Xie et al., 2024a) are LLMs specialized in financial domain by fine-tuning LLaMA series LLMs, medical LLMs Huatuo (Wang et al., 2023)and legal LLM ChatLaw (Cui et al., 2023)are also fine-tuned from LLaMA, AuditWen is fine-tuned for auditing from Qwen-7B and the result shows significant performance on various of audit NLP tasks compared with the state-of-the-art LLMs.

To our knowledge, no prior work has systematically developed a compact and audit-specialized LLM family (0.5B–3B parameters) that balances Chinese domain expertise, task adaptability and deployability.

3 Methodology

This study presents a systematic framework designed to enhance LLMs for auditing tasks. It employs a two-stage optimization process, beginning with domain-adaptive continual pre-training (CPT), followed by multi-task supervised fine-tuning (SFT) based on the outcomes of the first stage, as illustrated in Figure 1. Note that, the optimal base model need to be selected from a series candidate LLMs before CPT progress.

The framework is detailed in three key components. Section 3.1 elaborates on the continual pre-training process, which focuses on optimizing the proportion of domain-specific data. Section 3.2 describes the supervised fine-tuning stage which emphasizes multi-task instruction tuning to enhance task-specific capabilities. Section 3.3 introduces the evaluation benchmark used to evaluate the model's performance at each stage. This structured approach ensures that the model is progressively refined to address the unique demands of auditing applications effectively.

3.1 Continual Pre-training

This study collects and constructs four types of datasets, namely, (1)finance domain dataset, (2) laws and regulations dataset from audit domain, (3) Chinese general dataset and (4) English general dataset. We further explore the corresponding matching strategies by systematically integrating the four domain datasets for continual pre-training.

3.1.1 Data Construction for CPT

AuditCorpus. This dataset includes three sources: audit-related regulations, Baidu encyclopedia entries of audit-related concept and audit-related news, with totaling of 500MB. The details of the dataset is: (1)Audit Regulations encompass regulatory documents and standards pertinent to auditing; (2) Baidu Encyclopedia Entries provides encyclopedic knowledge on audit-related entities; (3) News Articles covers audit-related news. The dataset composition strikes a balance between adequately representing audit knowledge and facilitating effective domain adaptation for auditing tasks.

FinCorpus. The pre-training dataset for the financial domain is collected from XuanYuan and is referred to as FinCorpus dataset. XuanYuan is a large-scale text dataset specialized in the financial domain, including listed company announcements, financial news, financial articles and financial exam questions, with totaling of 60GB.

Chinese General Corpus. This dataset is collected from two sources: (1) Wikimedia, a publicly accessible dataset comprising Wikipedia and (2) TigerBot, an open-source Chinese pre-training dataset of containing Chinese books, internet text and encyclopedia. The sources are combined to construct a comprehensive general Chinese pre-training corpus.

English General Corpus. This dataset is also collected from Wikimedia, which provides a vast amount of information across various domains in multiple languages.Here, the English dataset from Wikimedia is selected to construct a comprehensive general English pre-training corpus.

To processing the pre-training data, data chunking is conducted at first to ensure the length of each data segment limited to 4096 tokens with using the Qwen2.5 tokenizer. Furthermore, the Data-Juicer toolkit [1] is employed to sample and clean the pre-training data to obtain high-quality inputs. Specifically, the general pre-training data and the FinCorpus were downsampled to 10M tokens from each to balance the dataset sizes of four corpus. Table 4 provides the details of the data pre-process, including the original and resulting sample sizes of each dataset, along with respective sampling ratios.

3.1.2 Sampling from Each Dataset

In the pre-training process, dataset sampled from different domains exhibit distinct linguistic features, vocabulary distributions and semantic structures. A domain-skewed pre-training corpus risks overfitting to domain-specific patterns, compromising cross-domain generalization. Underrepresented domains in training data impede the model's acquisition of domain-specific patterns, thereby degrading task-specific performance. Therefore, carefully sampling and balancing domain proportions is crucial to optimize the model's cross-domain generalization performance.

Research by (Gururangan et al., 2020) on domain-adaptive pre-training highlights that tailoring the pre-training data distribution to downstream tasks can significantly enhance model performance, further justifying the need for sampling exploration. In this study, to investigate the impact of domain-specific data proportions on continual pre-training, we designed a structured sampling strategy that allows controlled adjustments while maintaining consistency across experiments. The sampling process focuses on two main components: a baseline

[1] https://github.com/modelscope/data-juicer

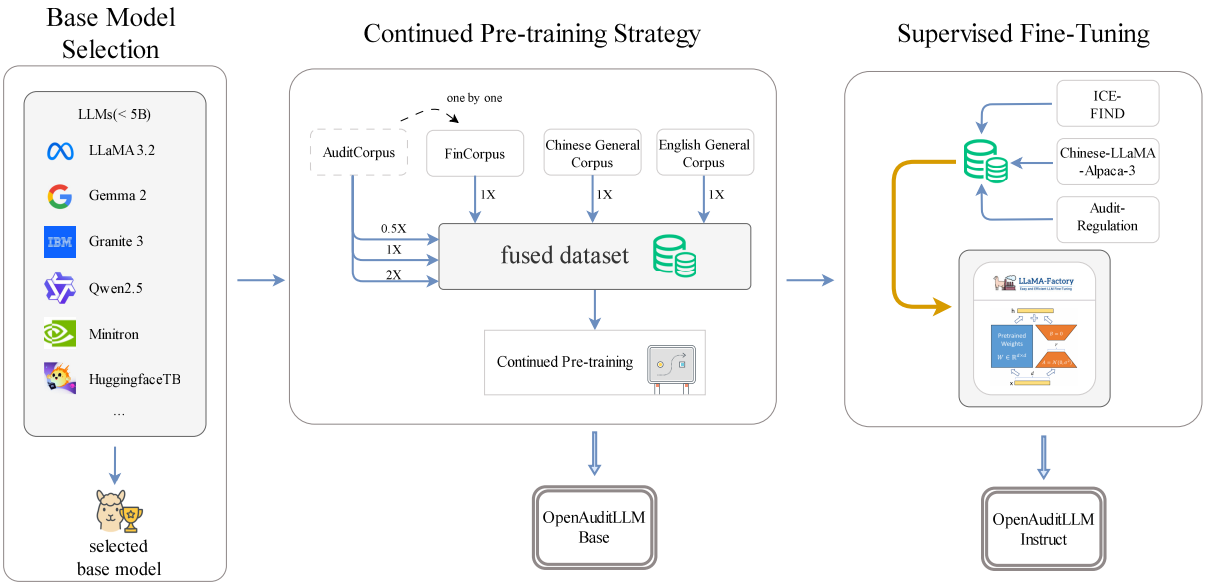

Figure 1: Overview framework of training Audit Language Model (AuditLLM). The framework consists of three phases: (1) Base Model Selection, i.e, select an optimal base model from open-source LLMs; (2) Continual Pre-Training (CPT) with a domain-specific dataset by designing a series of sampling strategies and result in base model, called "OpenAuditLLM Base"; (3) Supervised Fine-Tuning (SFT) the base models by multi-tasking instruction tuning on datasets and result in instruct model, called "OpenAuditLLM Instrust".

setting with equal domain contributions and proportion adjustments for individual domains.

First, a baseline configuration was established by synchronously scaling data from the four domains in equal proportions, namely, finance, English, Chinese and auditing. For example, in the initial setup, each domain contributes 0.0125 billion (B) tokens, totaling 0.05B tokens. This balanced design ensures that the model is exposed to a uniform distribution of knowledge across domains at the starting point. The configurations generated under equal scaling are visualized in Figure 2, where all domain token counts increase synchronously across different configurations.

To further analyze the impact of varying a single domain's representation on the model, we introduced proportion adjustments where one domain was over-sampled or under-sampled while the others remained equally scaled. Over-sampling starts from 0.025B tokens with a step size of 0.025B, and under-sampling starts from 0.00625B tokens with a step size of 0.00625B. This design allows to independently evaluate how increasing or decreasing the data volume of a specific domain affects the model performance.

The specific adjustment for the auditing domain is illustrated in Figure 2. Here, the "Regulation Over" and "Regulation Under" lines represent the cases where the auditing domain was respectively over-sampled and under-sampled, while the rest domains maintained equal proportions. This figure highlights how the auditing data proportion changes across configurations, demonstrating the flexibility of the sampling framework.

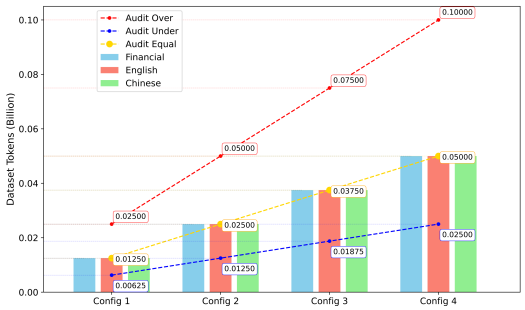

Figure 2: Proportion adjustment for the auditing domain, showing equal-sampling, over-sampling and under-sampling settings relative to other domains.

3.2 Supervised Fine-Tuning

3.2.1 Data Construction for SFT

The supervised fine-tuning (SFT) phase leverages datasets of containing audit & finance domain instructions and general instructions. The datasets are collected to reflect diverse linguistic patterns and domain knowledg relevant to each domain and ensures a balanced representation of the domains.

The details of the instruction fine-tuning datasets

are introduced as follows.

ICE-FIND dataset derived from the ICE-PIXIU framework[2] is selected as the domain-specific instruction dataset. ICE-PIXIU is a comprehensive cross-lingual financial instruction framework of containing 32 tasks with 604k instructions, encompassing various Chinese financial NLP tasks.

Audit dataset is optimized from audit-specific fine-tuning dataset developed by AuditWen[3], which consists of 14 tasks, with totaling of 30,269 instructions.

Chinese-LLaMA-Alpaca-3 dataset [4] is employed as general instructions, which contains kinds of high-quality instruction tasks in both Chinese and English, with a total of 307k instructions.

3.2.2 SFT Data Exploration

We fine-tune the OpenAuditLLM 0.5B model with three distinct datasets, namly the ICE-FIND dataset, the Audit dataset and the Chinese-LLaMA-Alpaca-3 dataset. The performance of the models are evaluated at each training stages to determine the optimal dataset for subsequent fine-tuning of the 1.5B and 3B models.

3.3 Evaluation Benchmark

The auditing domain specific dataset, referred as AuditEva Datasets, is employed to test the model's ability in audit domain. The dataset includes four different classification tasks, namely audit-item entity classification (AIEC), audit-problem entity classification (APEC), audit legal name classification (ALNC) and audit relation classification (ARC). These diverse tasks collectively evaluate the model's robustness and versatility in handling auditing-specific content. Table 5 in Appendices provides a detailed overview of these datasets, including their descriptions and examples categories.

FinanceIQ[5] evaluates the model's understanding of financial concepts, terminology and reasoning, which are also relevant to audit domain. Furthermore, this study employed a standardized evaluation pipeline by combining FinanceIQ and AuditEva dataset, enabling fair comparisons and reflecting the models' performance on auditing-relevant capabilities throughout the experimentation process.

[2] https://github.com/YY0649/ICE-PIXIU

[3] https://github.com/HooRin/AuditWen

[4] https://github.com/ymcui/Chinese-LLaMA-Alpaca-3

[5] https://huggingface.co/datasets/Duxiaoman-DI/FinanceIQ

4 Experimental Results

To evaluate the performance of our approach, comprehensive benchmark is conducted against existing models. The evaluation focused on both base models and instruct models to assess their capabilities across various settings.

4.1 Baseline LLMs selection for continual pre-training

We curated a diverse set of baseline models that encompass various architectures and parameter scales, which are publicly available and constrained to under 5B. This parameter ceiling was chosen to balance computational efficiency with model capacity. Table 1 provides an overview of the evaluated baseline models.

To enables a systematic comparison of their initial performance on auditing-relevant tasks and provides a foundation for identifying the most promising candidates for subsequent experiments, these models were subjected to the evaluation benchmark outlined in Section 3.3. The evaluation focused on accuracy for all tasks. Table 1 demonstrates the evaluation result of different models. The Qwen2.5 series consistently outperformed other candidates across the majority of the benchmark tasks. This superior capability is attributed to their robust generalization across financial and audit contexts, making them well-suited as baseline models for the auditing-focused experiments in this study. Consequently, the Qwen2.5 series are selected as the foundation for continual pre-training and SFT phases.

4.2 Continued Pretraining Result

Experimental Setup. The pre-training process utilized 64 A100 80GB GPUs across 8 nodes, requiring 250 hours for one epoch. The key hyperparameters were set to tackle the research question, with learning rate of 1 x 10^{-5} using the cosine schedule, a weight decay of 0.00001, a warm-up ratio of 0.05, a batch size of 2 per device, and the maximum sequence length of 8,192 tokens.

Figure 3 illustrate the impact of under-sampling and over-sampling strategies on the model performance respectively. The horizontal axis represents the total parameter count of the four datasets (FinCorpus, English General Corpus, Chinese General Corpus, and AuditCorpus), while the vertical axis denotes their proportion configurations.

Results and Analysis. We evaluated multiple data proportion configurations for the Qwen2.5-

Model Size	Model Name	FinanceIQ	AIEC	APEC	ALNC	ARC	Average
<1B	EleutherAI/pythia-410M	24.54	40.56	7.19	5.96	**17.95**	19.24
	HuggingFaceTB/SmolLM2-135M	24.23	41.45	16.99	9.63	17.09	21.878
	HuggingFaceTB/SmolLM2-360M	25.21	41.45	16.34	8.26	14.53	21.158
	Qwen/Qwen2.5-0.5B	**44.03**	**41.78**	**26.80**	**14.68**	14.53	**28.364**
1B∼2B	Meta-Llama/Llama-3.2-1B	26.52	41.45	14.38	15.14	42.74	28.046
	Qwen/Qwen2.5-1.5B	**55.55**	**81.25**	**36.60**	**34.40**	**51.28**	**51.816**
	Internlm/internlm2.5-1.8B	52.56	50.79	33.99	22.94	39.32	39.92
	IBM-granite/granite-3.0-2B-Base	37.25	70.38	28.76	29.82	29.91	39.224
	Google/gemma-2-2B	33.83	51.05	33.99	27.52	5.98	30.474
2B∼5B	Qwen/Qwen2.5-3B	**65.51**	**81.75**	**42.20**	38.53	**52.99**	**56.196**
	meta-llama/Llama-3.2-3B	39.90	68.40	35.29	**46.79**	25.64	43.204
	Nvidia/Minitron-4B-Base	40.71	66.05	36.60	42.66	7.69	38.742

Table 1: Overall performance of different models on the evaluation dataset. Models with sizes below 5B were selected for evaluation. The results indicate that the Qwen series, including Qwen2.5 0.5B, 1.5B, and 3B models, achieve the highest scores on most tasks and exhibit the best overall performance.

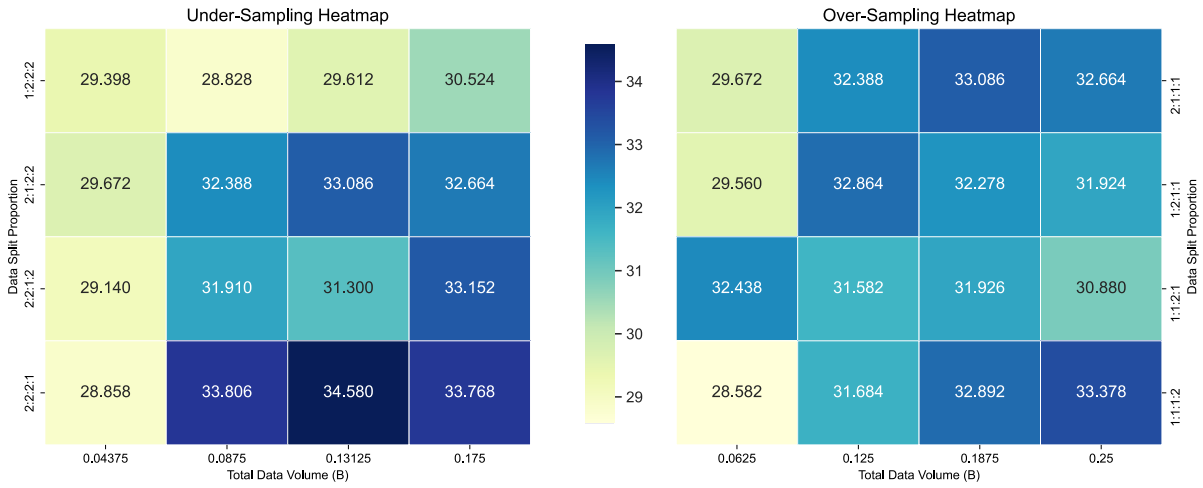

Figure 3: Model Performance with Under-sampling and Over-Sampling Configurations.

0.5B base model, with results presented in Tables 6, 7 and 8 in Appendices. Table 6 indicates that within a certain range, increasing the total parameter count of the pre-training dataset enhances model performance. A heatmap in Figure 5, derived from Tables 7 and 8, visually illustrates the experimental outcomes. By analyzing result, we identified the optimal data mixing strategy, namely 2:2:2:1 of the four dataset, which involves under-sampling the AuditCorpus dataset. Consequently, we adopted a configuration with a total parameter count of 0.175B under-sampling AuditCorpus as the most effective strategy.

To further evaluate the effectiveness of the proposed continued pre-training approach, we conducted a comprehensive comparison between the Qwen2.5 base models and our pre-trained OpenAuditLLM models, as shown in Table 2. The results demonstrate the effectiveness of the optimized data proportion strategy. For instance, OpenAuditLLM-0.5B achieves an average performance of 33.77, surpassing Qwen2.5-0.5B's 32.62 with notable gains in AIEC task (61.09 vs. 56.19). Similarly, OpenAuditLLM-3B outperforms Qwen2.5-3B with an average score of 57.31 compared to 56.07 particularly in ALNC (44.50 vs. 39.45) and ARC tasks (52.99 vs. 47.86). These improvements highlight the consistent benefits of our continued pre-training strategy across different model scales.

Metrics	0.5B		1.5B		3B	
	Qwen2.5	OALLM	Qwen2.5	OALLM	Qwen2.5	OALLM
FinanceIQ	43.12	41.75	56.07	56.27	65.76	65.13
AIEC	56.19	61.09	81.69	81.50	82.84	82.77
APEC	25.49	25.49	37.25	37.91	44.44	41.18
ALNC	16.06	17.43	36.23	39.45	39.45	44.50
ARC	22.22	23.08	52.14	51.28	47.86	52.99
Average	32.62	**33.77**	52.68	**53.28**	56.07	**57.31**

Table 2: Performance comparison between Qwen2.5 and OpenAuditLLM (OALLM) at different parameter scales on financial tasks.

4.3 Instruction Tuning

Experimental Setup. In Stage 2, instruction tuning is conducted to enhance the OpenAuditLLM model's ability to follow task-specific instructions tailored to the auditing domain. The process utilized four NVIDIA A6000 GPUs with mixed-precision training (bf16) to improve computational efficiency and minimize memory usage. The configuration utilizes the AdamW optimizer with critical hyperparameters set to a learning rate of 1e-4, a batch size of 8, a single training epoch, applied LoRA with a rank of 8 and alpha value of 16.

Results and Analysis. The OpenAuditLLM 0.5B model is instruction-tuned using three distinct datasets as described in Section 3.2.1. The baseline used in all comparisons is the Qwen2.5-0.5B Instruct model, while the 20% to 100% stages represent sampling data of different proportions for fine-tuning on the respective datasets. The experimental results are shown in the Figures 4, which denotes that using the ICE-FIND and Chinese-LLaMA-Alpaca-3 datasets leads to suboptimal performance across most evaluation metrics. In contrast, the Audit dataset consistently delivers significant performance improvements with larger training data volumes. Based on these findings, the Audit dataset is selected exclusively for the instruction tuning stage.

By comparing the instruction tuning results, we adopted a 5-shot setting for ARC task and zero-shot for the rest tasks. Here, OpenAuditLLM instruct model denotes supervised fine-tuning of the OpenAuditLLM base model with the Audit dataset. Table 3 shows the overall performane of different models on four tasks. From Table 3, it is evident that the OpenAuditLLM instruct model outperforms the Qwen2.5 instruct model across most metrics and model sizes. Notably, OpenAuditLLM instruct achieves substantial improvements in AIEC, APEC and ALNC across all model sizes, with particularly strong gains in the 1.5B and 3B configurations. While its performance in ARC is less consistent, the overall average scores of OpenAuditLLM instruct surpass those of Qwen2.5 instruct. The overall results demonstrate the effectiveness of instruction tuning with the Audit dataset and highlights the robustness of OpenAuditLLM instruct in handling domain-specific tasks.

Effect of Sampling Strategy. Beyond basic quality, the domain sampling strategy during continued pretraining also plays a critical role. As discussed in Section 3, we introduced two kinds of sampling: equal domain scaling and controlled domain proportion adjustment.The sampling settings (illustrated in Figure 1 and Figure 2) show that both the volume and proportion of domain-specialized data can impact the model's specialization and generalization. Particularly, auditing domain performance was sensitive to the amount of audit dataset: over-sampling auditing data improved performance on auditing-specific tasks but slightly reduced general generalization, while under-sampling had the opposite effect.

These findings emphasize that the quality and the quantity, distribution of data must be carefully designed according to domain-specific application requirements.

5 Conclusion

In this work, we proposed a framework for adapting lightweight large language models (LLMs) to the auditing domain, with a particular emphasis on continual pretraining and task-aware instruction tuning. Our experimental results underscore the efficacy of the two-stage training paradigm, namely contin-

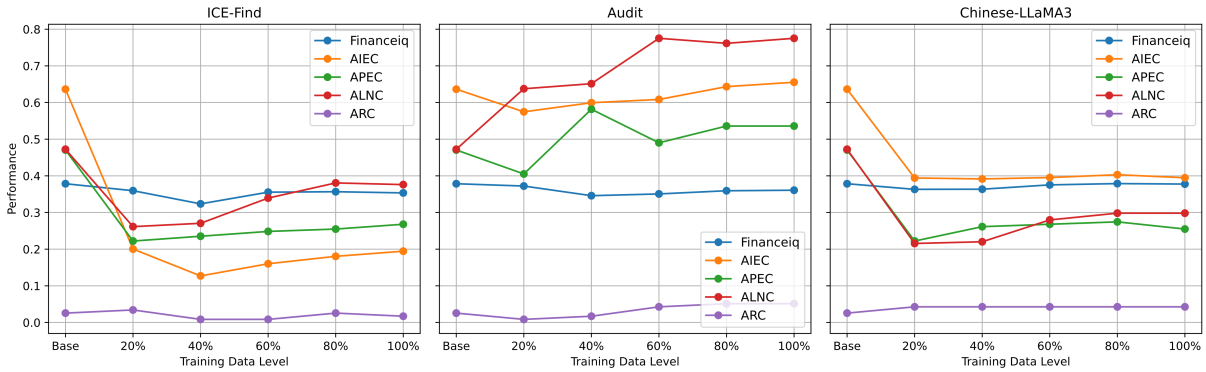

Figure 4: Performance comparison of models fine-tuned with three datasets (ICE-FIND, Audit, Chinese-LLaMA-Alpaca-3) on various tasks. "Base" corresponds to the Qwen2.5 0.5B Instruct model without instruction tuning, while "20%" to "100%" indicate different proportions of instruction tuning data used from each dataset.

Model Size	Model Name	FinanceIQ	AIEC	APEC	ALNC	ARC	Average
0.5B	Qwen2.5 instruct	**37.85**	63.64	47.06	47.25	**19.66**	43.09
	OpenAuditLLM instruct	36.08	**65.54**	**53.59**	**77.52**	10.26	**48.60**
1.5B	Qwen2.5 instruct	**54.51**	55.12	10.46	44.95	52.99	43.61
	OpenAuditLLM instruct	51.82	**60.46**	**49.67**	**78.44**	**56.41**	**59.36**
3B	Qwen2.5 instruct	60.41	53.59	32.68	48.62	**47.86**	48.63
	OpenAuditLLM instruct	**60.66**	**76.10**	**55.56**	**79.36**	**47.86**	**63.91**

Table 3: Performance of Qwen2.5 instruct and OpenAuditLLM instruct across model sizes on five sub-tasks and their average.

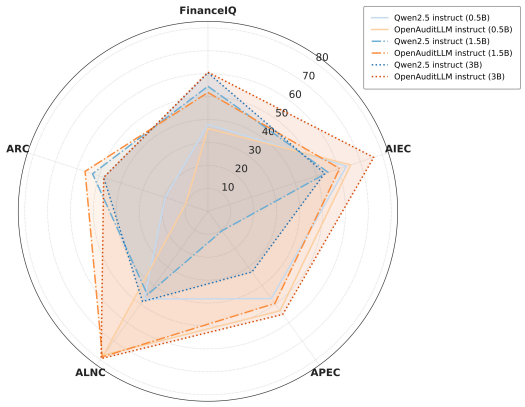

Figure 5: Performance Radar of Supervised Fine-Tuned OpenAuditLLM and Qwen2.5 Models

ual pretraining significantly enhances the model's ability to internalize domain-specific knowledge, and instruction tuning refines the model's capacity to perform realistic auditing tasks. Finally, we obtained the first family of open-sourced Chinese audit-specialized LLMs, covering models with parameter sizes from 0.5B to 3B. In the future, reinforcement learning strategy is considered after the SFT parse to further improve the performance of the OpenAuditLLM.

Limitations

Our approach depends heavily on the availability of high-quality domain-specialized data. While we curated improved datasets for continued pretraining and instruction tuning, the process is labor-intensive and not easily generalizable to other domain or languages without similar data quality.

In addition, our evaluation is constrained to five datasets primarily focused on classification tasks. This narrow scope may not fully capture the broader range of audit-related reasoning and generation capabilities, leaving generalization to other audit tasks an open question. Our future work will incorporate more evaluation tasks related to audit application scenarios.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620, Hong Kong, China.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 33, NEURIPS 2020*, 33.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*, abs/2306.16092.

Jiawei Gu, Zacc Yang, Chuanghao Ding, Rui Zhao, and Fei Tan. 2024. Cmr scaling law: Predicting critical mixture ratios for continual pre-training of language models. *Computing Research Repository*, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing:16143–16162.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model? *CoRR*, abs/2308.04014.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Rujun Han, Xiang Ren, and Nanyun Peng. 2021. Econet: Effective continual pretraining of language models for event temporal reasoning. In *Conference on Empirical Methods in Natural Language Processing*.

Jiajia Huang, Haoran Zhu, Chao Xu, Tianming Zhan, Qianqian Xie, and Jimin Huang. 2024. AuditWen: An open-source large language model for audit. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1351–1365.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *TMLR 2024*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. pages 4513–4519. Special Track on AI in FinTech.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Ebere Ruth Onwubuariri, Beatrice Oyinkansola Adelakun, Omolara Patricia Olaiya, and Joseph Elikem Kofi Ziorklui. 2024. Ai-driven risk assessment: Revolutionizing audit planning and execution. *Finance & Accounting Research Journal*, 6(6):1069–1090.

Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yinghao Ma, Feiyu Duan, ZhiqiBai, JiakaiWang, Yuanxing Zhang, Xu Tan, Jie Fu, Jiamang Wang, Lin Qu, Wenbo Su, and Bo Zheng. 2024. D-cpt law: Domain-specific continual pre-training scaling law for large language models. *NeurIPS 2024*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sanh Victor, Debut Lysandre, Chaumond Julien, and Wolf Thomas. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Obstetric Protocols for Labor Ward Management*.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXivorg*, abs/2304.06975.

Chengyue Wu, Yukang Gan, Yixiao Ge, Zeyu Lu, Jiahao Wang, Ye Feng, Ying Shan, and Ping Luo. 2024. Llama pro: Progressive llama with block expansion. *Annual Meeting of the Association for Computational Linguistics*, pages 6518–6537.

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

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 15 others. 2024a. Finben: A holistic financial benchmark for large language models. *Conference on Neural Information Processing Systems*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *Computing Research Repository*.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2024b. Efficient continual pre-training for building domain specific large language models. *ICLR 2024*.

Zhang Xuanyu and Yang Qing. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 4435–4439, New York, NY, USA. Association for Computing Machinery.

Dong Yuan, Eti Rastogi, Gautam Naik, Sree Prasanna Rajagopal, Sagar Goyal, Fen Zhao, Bharath Chintagunta, and Jeff Ward. 2024. A continued pretrained llm approach for automatic medical note generation. *Computing Research Repository*, pages 565–571.

Hanyu Zhang, Boyu Qiu, Yuhao Feng, Shuqi Li, Qian Ma, Xiyuan Zhang, Qiang Ju, Dong Yan, and Jian Xie. 2024. Baichuan4-finance technical report. *CoRR*, abs/2412.15270.

A Appendix

A.1 Summary of Data Cleaning and Sampling for Pre-training Datasets

Table 4 provides a detailed summary of the data cleaning and sampling processes applied to the pre-training datasets. It includes the dataset type, specific dataset names, original and resulting token counts, and the ratio of retained tokens after cleaning. The "origin tokens" and "res tokens" columns denote the number of tokens before and after processing respectively, while the "ratio" column indicates the proportion of tokens retained.

A.2 Overview of Audit-Related Datasets

Table 5 presents an overview of the audit-related datasets evaluated in this study. The table includes descriptions of each dataset, example entities, and classification choices used for audit tasks. Each entry details the dataset's role in audit-related evaluations, with the "Examples" column providing representative instances and the "Choices" column listing possible categories for annotation or classification.

A.3 Model Performance with Equal-Proportion Datasets

Table 6 shows the performance of models trained on equal-proportion datasets, reporting final scores across different total parameter counts. The table highlights how model performance varies with parameter scale, with scores reflecting accuracy or other relevant metrics.

A.4 Model Performance with Under-Sampling Configurations

Table 7 details the performance of models under under-sampling configurations, presenting scores for different dataset proportions and parameter counts, along with average performance metrics. The table illustrates the impact of reducing dataset sizes on model outcomes.

A.5 Model Performance with Over-Sampling Configurations

Table 8 summarizes the performance of models under over-sampling configurations, showing scores for various dataset proportions and parameter counts, along with averages. The table demonstrates how increasing dataset sizes affects model performance.

A.6 Overview of Language Models from Various Organizations

Table 9 provides an overview of language models developed by various organizations, including model names and their key characteristics. The table outlines critical features such as architecture, training data, or intended applications, facilitating a comparison of models across different providers.

Dataset type	dataset	origin tokens	origin num	res tokens	ratio
AuditCorpus	regulation	0.1B	59737	51860	0.8681
	baike	–	4768	4466	0.9367
	news	–	8807	1573	0.1786
FinCorpus	fincorpus	0.4B	439601	112443	0.2558
Chinese General Corpus	wiki-zh	0.15B	257180	118353	0.4602
	tiger-bot-zh	0.4B	866254	221013	0.2551
English General Corpus	wiki-en	0.15B	336495	316133	0.9395

Table 4: Summary of data cleaning and sampling for pre-training datasets, including original and resulting token counts and ratios.

Dataset	Description	Examples	Choices
Audit-Item Entity	An entity (or phrase) representing an audit item	坐支管理费用审计, 环境治理审计, 应交税金审计	财政审计, 公共工程审计, 海关审计, 金融审计, 民生审计, 税收审计, 资源环保审计
Audit-Problem Entity	An entity (or phrase) expressing an audit doubt	扶持资金管理制度不完善, 自然资源被违法占用, 伪造税务登记证件	财政审计, 公共工程审计, 海关审计, 金融审计, 经济Sass审计, 民生审计, 农业农村审计, 审计共性问题, 企业审计, 税收审计, 资源环保审计
Legal-Name Entity	An entity (or phrase) expressing a legal name used in auditing	中华人民共和国财政违法行为处罚处分条例, 天津市水资源税改革试点实施办法, 中华人民共和国税收征收管理法	财经法规, 财政法规, 个人所得税, 金融综合, 劳动就业, 上市公司, 社会保障, 行业管理, 增值税, 征收管理, 资产评估法规, 资源税, 综合管理, 综合税收政策
Regulation Relation Extraction	Classify the given audit-relevant entity pair to one of the given choices, extracted from a sentence	[规避招标, 招标投标法], [合同履行情况审计, 检查]	审计问题, 审计事项, 审计依据, 审计方法, 审计机构, 审计成果, 被审计单位, 涉及的行业或领域

Table 5: Overview of audit-related datasets evaluated in this study, including descriptions, example entities, and classification choices. Each entry details the dataset's role in audit tasks.

Total Parameter Count (B)	Final Score
0.05	28.362
0.1	29.028
0.15	30.758
0.2	31.976

Table 6: Model performance with equal-proportion datasets, showing final scores for different total parameter counts.

Proportion	Total Parameter Count (B)				Average
	0.04375	0.0875	0.13125	0.175	
1:2:2:2	29.398	28.828	29.612	30.524	29.5905
2:1:2:2	29.672	32.388	33.086	32.664	31.9525
2:2:1:2	29.140	31.910	31.300	33.152	31.3755
2:2:2:1	28.858	33.806	34.580	33.768	32.7530

Table 7: Model performance with under-sampling configurations, showing scores for different proportions and parameter counts, along with averages.

Proportion	Total Parameter Count (B)				Average
	0.0625	0.125	0.1875	0.25	
2:1:1:1	29.672	32.388	33.086	32.664	31.9525
1:2:1:1	29.560	32.864	32.278	31.924	31.6565
1:1:2:1	32.438	31.582	31.926	30.880	31.7065
1:1:1:2	28.582	31.684	32.892	33.378	31.6340

Table 8: Model performance with over-sampling configurations, showing scores for different proportions and parameter counts, along with averages.

Organization	Model(s)	Description
Google	Google gemma-2-2b	Gemma-2-2B is a 2B-parameter model from the Gemma family, designed for efficient performance in NLP tasks.
IBM	ibm-granite granite-3.0-2b-base	Granite-3.0-2B-Base is a 2B-parameter model from IBM's Granite series, optimized for enterprise applications.
xAI	internlm2.5-1.8B	InternLM2.5-1.8B is a 1.8B-parameter model developed by xAI, focusing on efficient and scalable language modeling.
Meta AI	Llama-3.1-Minitron-4B-Depth-Base, Llama-3.1-Minitron-4B-Width-Base, Llama-3.2-1B, Llama-3.2-3B	LLaMA models are from Meta AI, including Minitron variants (4B parameters) optimized for depth and width, and LLaMA-3.2 models with 1B and 3B parameters. The models are known for efficiency in research applications.
NVIDIA	Minitron-4B-Base	Minitron-4B-Base is a 4B-parameter model by NVIDIA, designed for high efficiency in natural language tasks.
EleutherAI	pythia-410m	Pythia-410M is a 410M-parameter model by EleutherAI, developed for research with a focus on transparency and reproducibility.
Hugging Face	SmolLM2-135M, SmolLM2-360M, SmolLM2-1.7B	SmolLM2 series by Hugging Face, with parameter sizes of 135M, 360M, and 1.7B, are optimized for lightweight and efficient language modeling.
Alibaba	Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-3B	Qwen2.5 series by Alibaba, with parameter sizes of 0.5B, 1.5B, and 3B, are designed for robust performance in multilingual and domain-specific tasks.

Table 9: Overview of large language models from various organizations, including model names and their key characteristics.