Multi-perspective Analysis of Large Language Model Domain Specialization: An Experiment in Accounting Audit Procedures Generation

Yusuke Noro
yusukenoro83@gmail.com

Abstract

Two major domain specialization approaches for Large Language Models (LLMs), fine-tuning and In-Context Learning (ICL), have been compared across various domains. While prior research has examined the similarities and differences between these approaches in task-specific capabilities, less is known about how they affect the feature of the generated text itself. To address this research gap, we conducted an experimental study using Accounting Audit Procedures Generation (AAPG) task, a highly specialized task requiring expert accounting knowledge. This task provides a practical testbed for a multi-perspective analysis of domain specialization due to its technical complexity and the large gap between general and domain expert knowledge. The results show consistent differences in output characteristics across models when comparing fine-tuning, ICL, and their combined approaches.

1 Introduction

Domain specialization, which adapts general-purpose LLMs to domain-specific contextual data and domain objectives, has been developed across various specialized fields such as healthcare and finance (Ling et al., 2024; Lee et al., 2019; Yang et al., 2023; Li et al., 2023; Singhal et al., 2022). Two widely used approaches for domain specialization of LLMs are fine-tuning and prompt augmentation. Fine-tuning is a method that performs additional training to adapt pre-trained LLMs to specific tasks or domains. Prompt augmentation encompasses ICL (few-shot prompting), which incorporates a small number of examples in prompts during inference, and Retrieval Augmented Generation (RAG), which dynamically integrates external knowledge into LLMs.

Recent studies have shown that ICL and RAG can achieve performance comparable to fine-tuning (Ovadia et al., 2024; Soudani et al., 2024; Bassamzadeh and Methani, 2024). On the other hand, other research suggests that RAG does not serve as a complete substitute for fine-tuning but rather complements it, with the combined application of both methods yielding enhanced performance (Balaguer et al., 2024).

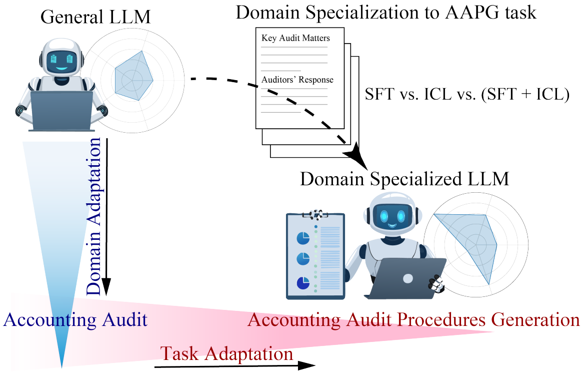


Figure 1: The focus of the experimental design. Domain specialization encompasses domain adaptation and task adaptation. The domain specialization approaches were compared on the AAPG task, a highly specialized niche domain task that provides substantial potential for improvement over general LLMs.

While prior research has examined the similarity and difference between these approaches in task-specific capabilities, less is known about how they affect the characteristics of the generated text itself. Therefore we pose a key research question: **"Do fine-tuning and prompt augmentation develop distinct capabilities in open-ended question answering scenarios, and does their combination produce additive effects or simply complement their separate effects?"** However, evaluating this hypothesis presents a methodological challenge. Existing Long-Form Question Answering (LFQA) datasets presented limitations for this purpose, as general-purpose LLMs already perform well on several evaluation dimensions, making it difficult to observe meaningful differences between

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 17660–17682
November 4-9, 2025 ©2025 Association for Computational Linguistics

specialization approaches. Evaluation of the *ground truth* in some LFQA datasets are shown in Appendix B.4.

To address this challenge, we conducted an experimental study using AAPG task, a highly specialized task requiring expert knowledge. The task is based on data describing the actual company conditions and concurrent procedures conducted by the accounting audit experts. The specialized nature of this task enabled us to create domain-specific growth potential for LLMs and investigate the characteristics of this methodology.

Specifically, we analyze the textual properties of outputs generated through Supervised Fine-Tuning (SFT), ICL and their combined approaches. The evaluation framework uses multi-perspective criteria by LLM-as-a-judge, such as comprehensiveness, specificity, and relevance.

2 Related Work

2.1 Fine-tuning vs. Prompt Augmentation

Several studies have compared the effectiveness of fine-tuning and prompt augmentation in enhancing the capabilities of LLMs. Ovadia et al. (2024) evaluated fine-tuning and RAG across five tasks from the MMLU (Massive Multitask Language Understanding) benchmark—anatomy, astronomy, biology, chemistry, and temporal reasoning—showing that RAG is equivalent to or sometimes outperforms fine-tuning. Soudani et al. (2024) categorized Wikipedia-based tasks into different popularity tiers and showed RAG's superiority for less common Wikipedia topics. Other studies have also compared fine-tuning and RAG or ICL in various settings (Mosbach et al., 2023; Alghisi et al., 2024; Bassamzadeh and Methani, 2024). Additionally, Balaguer et al. (2024) showed that combining fine-tuning with RAG improves performance when applied to agricultural data.

Nevertheless, existing studies primarily focused on overall performance comparisons between SFT and prompt augmentation or examine differences using simple metrics. In contrast, our research introduces an interpretable multi-perspective evaluation of the specific textual properties induced by SFT, ICL, and their combined approaches.

2.2 Application of Language Model in Auditing

The application of language models in auditing has been explored, particularly in areas such as information extraction and verification. Biesner et al. (2022) leveraged Sentence-BERT to match financial report paragraphs with checklist items. Eulerich et al. (2024) evaluated ChatGPT's performance on professional accounting certification exams, while Huang et al. (2025) developed and assessed LLM adaptations specifically for the accounting audit domain.

For practical applications, researchers have explored the application of LLMs to human-LLM collaboration in audit work (Gu et al., 2024) and the extraction of audit evidence and the verification of consistency (Li et al., 2024).

These research has primarily focused on relatively mechanical tasks such as information extraction and simple verification procedures. In contrast, the task introduced in our study focuses on a more complex challenge: the generation of audit procedures. This task demands advanced expertise and judgment, representing a markedly different application of LLMs compared to prior work in auditing.

3 Methods

Figure 1 illustrates the focus of the experiments in this paper. This research investigates the domain specialization process for the highly specialized domain and task of accounting audit field, specifically accounting audit procedures generation.

The background of domain specialization is detailed in Appendix B.4. For reproducibility, the code, prompt and dataset used in this study, along with details of the experimental settings, are available at `https://github.com/nororo/AAPG-task`.

3.1 Dataset

For the AAPG task, Key Audit Matters (KAMs) data, containing descriptions of audit matters and auditors' responses to them, can be easily extracted, making them valuable as high-quality question-and-answer sets. The dataset used in this paper consists of audit reports from securities reports with fiscal year-ends between March 31, 2021, and March 31, 2024. These reports, which were submitted up to

July 2024, were obtained via the EDINET API [1].

From the dataset dated March 31, 2024, we randomly sampled 500 audit reports, which contained a total of 607 KAMs, for the evaluation split. We used 8,350 KAMs from the remaining dataset for the training split, of which 90% were used for SFT training and 10% were used for validation monitoring. Further preprocessing steps are described in Appendix C.

3.2 Models

The selected models represent the widely-adopted open-weight models across various foundational capabilities: Qwen2-7B[2] (Yang et al., 2024), Llama-3.1-8B[3] (Grattafiori et al., 2024), and Llama-3.1-Swallow-8B-Instruct-v0.1[4] (Fujii et al., 2024; Okazaki et al., 2024). These models were chosen based on their high performance on the Japanese benchmark of the Swallow evaluation project[5]. The knowledge cutoff date for Llama-3.1 was December 31, 2024. While the knowledge cutoff date for Qwen2 and Swallow are undisclosed, Qwen2 was released on June 6, 2024, and Swallow was trained on synthetic data from Gemma-2, which was released on June 27, 2024. Since the earliest submission date of audit reports with KAMs in the evaluation dataset was May 31, 2024, based on these dates, the likelihood of data leakage appears minimal.

3.3 Supervised Fine-tuning

From the audit reports, we extracted the descriptions of consideration items and corresponding auditor responses for each KAM, using them as input and output for LLMs, respectively. We investigated with two approaches for the LoRA weights:

(1) Using weights from models fine-tuned with instruction tuning (Supervised Fine-tuning on an Instruction-Tuned model: SFT-IT). Fine-tuning was initiated from the instruction-tuned model. The LoRA parameters were trained using Equation (9) in Appendix A.

(2) Using weights from pre-instruction-tuned models and adding a Chat Vector (SFT on a base model with a Chat Vector: SFT-CV). In training phase, $\theta_{\text{new}} = (A, B)$ were updated in

$$W \leftarrow W_{\text{base}} + BA, \tag{1}$$

while in the inference phase, the estimated parameters $\hat{B}\hat{A}$ of $\hat{\theta}_{\text{new}}$ were added to the instruction tuned model:

$$W_{\text{eff}} \leftarrow W_{\text{instruct}} + \hat{B}\hat{A}. \tag{2}$$

This approach is analogous to adding a chat vector (Huang et al., 2024). Specifically, the transformation applied in Equation 2 is equivalent to adding to Equation 1. This adjustment modifies the base model by incorporating instruction-tuned parameters, similar to how chat vectors adjust model weights to encode conversational behaviors.

Given computational resource limitations, the SFT in this paper utilized QLoRA (Dettmers et al., 2023).

3.4 In-Context Learning

ICL does not involve updating the model parameters. Instead, ICL provides the model with a prompt that includes a few demonstration examples. Given a query input x_{query}, the inference is performed as follows:

$$\hat{y} = M_{\theta}(x_{\text{query}}, D_{\text{demo}}), \tag{3}$$

where $D_{\text{demo}} = \{(x_j, y_j)\}_{j=1}^{k}$ denotes a set of k selected demonstration examples. ICL enables the model to adapt its behavior in inference time by leveraging the context provided by these examples.

In many of the studies referred to in Section 2.1, the demonstration examples in ICL are expected to provide only information for task adaptation. On the other hand, in this study, the demonstration examples also serve as injected knowledge similar to those in RAG, providing essential guidance for generating audit procedures, which are influenced by relevant audit standards and audit firm policies.

3.5 Few-shot Selection for ICL

While research suggests that selecting examples more similar to the input is beneficial for few-shot sample strategies (Liu et al., 2022), various approaches have been proposed for few-shot selection. These include studies highlighting the importance of diversity (Chang et al., 2021), studies demonstrating performance gains from

[1] https://disclosure.edinet-fsa.go.jp/EKW0EZ0015.html

[2] Model weights are available at https://huggingface.co/Qwen/Qwen2-7B

[3] Model weights are available at https://huggingface.co/blog/llama31

[4] Model weights are available at https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-v0.1

[5] https://swallow-llm.github.io/index.ja.html

incorporating unrelated documents (Cuconasu et al., 2024; Zhang et al., 2024), and findings indicating that random sampling can yield comparable results (Cegin et al., 2024). To evaluate whether appropriate few-shot examples could effectively substitute SFT, we experimented with several sampling strategies.

We investigated configurations with 1, 2, 5, 10, and 20 examples, in which the maximum size of examples is due to computational resource constraints. We also examined three selection strategies for demonstration examples: (1) random selection, (2) selection based on the nearest example, and (3) a hybrid approach. The hybrid approach first selects the most similar example to the input, then iteratively selects the remaining k-1 examples that maximize distance from previously selected examples. All similarity calculations were based on the descriptions of the KAMs, which correspond to questions in question-and-answer sets. The nearness was computed based on the cosine similarity of the sentence embeddings of KAM descriptions using multilingual E5 (Wang et al., 2024). Multilingual E5 demonstrates high performance in the Japanese version of MTEB[6] while being multilingual and open source.

Based on the evaluation across different configurations in Section 4.2, we selected the best-performing setup for subsequent comparison with supervised fine-tuning approaches.

3.6 Supervised Fine-tuning with Few-Shot: SFT-FS

Retrieval augmented fine-tuning (RAFT) (Zhang et al., 2024) is an approach that combines prompt augmentation and fine-tuning. RAFT combines questions with either relevant documents containing correct answers or unrelated distractor documents for fine-tuning, aiming to improve robustness against retriever errors.

In this research, we applied a framework similar to RAFT. We performed supervised fine-tuning with few-shot (SFT-FS) using prompts in the ICL context.

Specifically:

$$\theta_{\text{new}}^* = \arg\min_{\theta_{\text{new}}} \mathcal{L}(\theta_{\text{frozen}}, \theta_{\text{new}}; D_{\text{train}}, D_{\text{demo}}). \tag{4}$$

For a proportion p of the data, D_{demo} is defined as follows:

$$D_{\text{demo}} = (x, y) \in D_{\text{nearest}}, \tag{5}$$

while for the remaining $(1 - p)$ proportion of the data, D_{demo} is defined as follows:

$$D_{\text{demo}} = (x, y) \in D_{\text{farthest}}. \tag{6}$$

Zhang et al. (2024) showed that including distractor documents during fine-tuning can improve accuracy in certain cases. In our 1-shot setting, we examined $p = 0.5$ and $p = 1$, and $p = 0.5$ demonstrated better performance. This results are shown in Appendix D.

3.7 Prompt

Without prompt expansion, we used the following simple prompt (the prompts were originally written in Japanese.):

As an auditor, you are provided with the following audit considerations.
Please plan the corresponding audit responses in Japanese.
{INSERT DESCRIPTION OF KAM}

For ICL inference and SFT-FS training, we used the following prompt with demonstration examples:

As an auditor, when given audit considerations, you are required to plan corresponding audit procedures.
Example 1
Given the following considerations:
Considerations:
{INSERT DESCRIPTION OF KAM (example)}
The corresponding audit procedures are as follows:
Audit Procedures:
{INSERT AUDITORS' RESPONSE (example)}
Example 2
...
Please plan the corresponding audit responses in Japanese as shown above.
{INSERT DESCRIPTION OF KAM}

3.8 Evaluation Metrics

To analyze the differences in generation behavior, we employed a multi-perspective evaluation approach. In particular, we evaluated textual properties using four perspectives: accuracy,

[6] https://github.com/sbintuitions/JMTEB/blob/main/leaderboard.md

comprehensiveness, relevance, and specificity. These metrics are derived from the requirements of accounting audit procedures. In accounting audits, auditors emphasize the comprehensiveness of audit evidence obtained through audit procedures and their relevance to examination items (IAASB and IFAC, 2024a; JICPA, 2024; PCAOB, 2004). Additionally, they require specificity in documenting procedures in audit working papers (IAASB and IFAC, 2024b; JICPA, 2022; PCAOB, 2010).

3.8.1 Accuracy

We employed an evaluation approach based on question-answer pairs generated from ground truth data (Deutsch et al., 2021; Wang et al., 2020). First, we extracted audit procedures as bullet points from the audit procedures in the evaluation data (ground truth). For each audit procedure, we created evaluation instances by masking one technical term by GPT-4o-2024-08-06.

When evaluating the generated audit procedures, we assessed whether the masked terms in the audit procedures could be predicted by referring to the generated audit procedures. This prediction task was performed using GPT-4o-mini-2024-07-18, and the "accuracy" score is defined as the average ROUGE-F1 scores (Lin, 2004) in the evaluation data.

3.8.2 LLM-as-a-judge Evaluation

To assess comprehensiveness, specificity, and relevance, we adopted the LLM-as-a-judge approach. The evaluation prompts were created with reference to AzureML Model Evaluation (Microsoft, 2023) but were also refined for the accounting audit domain; presented in Appendix H.

Comprehensiveness, specificity, and relevance were evaluated using 5-point scales. **Comprehensiveness** was assessed by measuring the extent to which generated procedures covered ground truth content (including similar or abstracted content). **Specificity** scores were assigned based on the clarity and precision of the generated audit procedures, with points deducted for ambiguity. **Relevance** was assessed based on whether the generated audit procedures aligned with the given considerations.

For the evaluation models, we used GPT-4-turbo-2024-04-09 for comprehensiveness and specificity, as GPT-4-turbo is commonly used for LLM-as-a-judge tasks and has a high correlation with human evaluation (Gu et al., 2025). Relevance scores by GPT-4-turbo were consistently inflated, making evaluation of domain specialization difficult: when we evaluated responses generated by vanilla model of Llama-3.1-8B-Instruct, almost all samples received a score of 5. Therefore, we used GPT-4o-2024-08-06 to assess relevance.

The sensitivity analysis of these evaluation metrics is presented in Appendix E.

3.8.3 Normalization and Comparison

To evaluate the relative performance gain compared to the vanilla LLMs, each evaluation metric score is normalized to a range from 0 to 1 using the following min-max normalization: The minimum value is set to the baseline model's score, while the maximum value is 1 for accuracy and 5 for other metrics. The average scores were calculated from the normalized values. For instance, if vanilla LLM scores 1, SFT scores 4, ICL scores 3, and the maximum possible score is 5, the SFT normalized score is $(4-1)/(5-1) = 0.75$ and the ICL normalized score is $(3-1)/(5-1) = 0.5$. In spite of the normalization, the comparison of interest was tested using a paired t-test with family-wise error (FWE) correction for comparisons across multiple perspectives. The raw evaluation scores are shown in Appendix G.

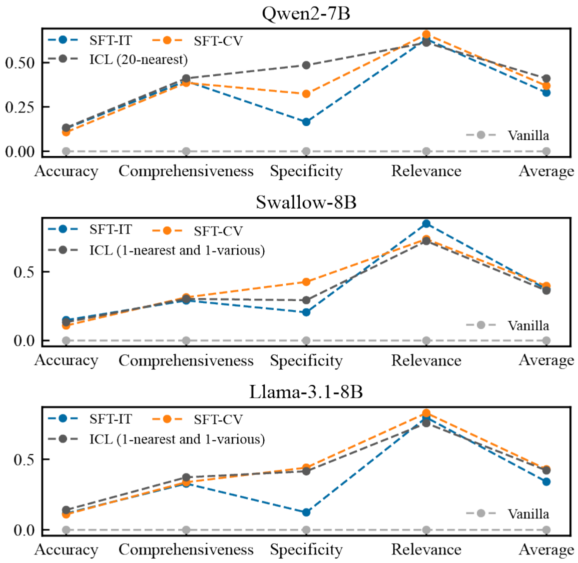
4 Results

4.1 Experiment 1: SFT vs. ICL

Figure 2 compares the improvement scores relative to the vanilla LLMs as the baseline for SFT and ICL approaches applied to Qwen2, Swallow, and Llama-3.1.

For SFT, we evaluated both SFT-IT and SFT-CV approaches. For ICL, we selected the best-performing few-shot selection method, hereafter referred to as ICL (optimal strategy). This corresponds to "20-nearest" for Qwen2 and "1-nearest and 1-diverse" for Swallow and Llama-3.1. These results are presented in Section 4.2.

SFT and ICL demonstrated improvements over the baseline of vanilla LLMs, across all four metrics, indicating the growth potential of LLMs on AAPG tasks. The comparison between SFT-IT and ICL revealed distinct performance variations across different metrics. While both approaches showed comparable improvements in **accuracy** and **comprehensiveness**, SFT-IT showed less improvement than ICL in **specificity**, but



	Accuracy	Comprehensiveness	Specificity	Relevance
SFT-IT vs. ICL	-	ICL *	ICL *	SFT-IT *
SFT-CV vs. ICL	ICL *	-	-	SFT-CV *
SFT-IT vs. SFT-CV	SFT-IT *	-	SFT-CV *	-

Figure 2: Comparison of SFT and ICL performance in zero-shot setting. *Top*: normalized score improvements from vanilla LLMs (vanilla LLMs = 0, gray baseline). *Bottom*: winner of the comparison with consistency between models. "-" indicates inconsistent results; "*" indicates statistical significance (P < .05, FWE corrected).

SFT approaches consistently outperformed ICL in **relevance** across the models. SFT-CV also demonstrated higher improvements in **relevance** compared to ICL, while showing lower improvements in **accuracy** but comparable improvements in **comprehensiveness**. These results suggest differences in capability development between domain specialization approaches.

4.2 Experiment 2: Selection Strategy of Demonstration Examples in ICL

In order to select the ICL (optimal strategy), we conducted two experiments. First, we examined the effect of selecting the number of demonstration examples, ranging from 1 to 20, as shown in Figure 3. The results suggest that increasing k does not always lead to better performance. Qwen2 achieved maximum performance at $k = 20$, while Swallow and Llama-3.1 performed best at $k = 2$.

Second, we also examined strategies for selecting demonstration examples (Figure 4): random selection, selection based on nearest examples, and selection based on both the nearest and diverse examples.

The results varied across models: nearest selection was the most effective for Qwen2, while a combination of the nearest and diverse selection strategies yielded the best results for Swallow and Llama-3.1.

4.3 Experiment 3: Combination of SFT and ICL

For SFT-IT and SFT-CV, responses are generated using prompts that include a single demonstration example during inference. The same prompting approach is also applied to SFT-FS. Parameter p is set to 0.5 for SFT-FS as it showed better performance than $p = 1$ (see Appendix D for details).

First, we compared SFT-IT and SFT-CV with 1-nearest shot to 0-shot. Figure 5 illustrates the performance improvements over the baseline, where the baseline is defined as the model's output with a single demonstration example in the prompt. Compared to methods without few-shot prompting, improvements in **accuracy** were observed across SFT-IT and SFT-CV. SFT-IT and SFT-CV also improved in **comprehensiveness**. These findings suggest that the behavior acquired by combining SFT and ICL is not a simple union but creates additive effects. On the other hand, the results for **specificity** and **relevance** were inconsistent, and additive effects were not observed across all evaluation aspects.

Second, the hybrid approaches (SFT-IT, SFT-CV, and SFT-FS) were compared to ICL (Figure 6), following the same methodology as in section 4.1, in which ICL (optimal strategy) was employed. Comparing SFT-based methods with few-shot prompting to the ICL (optimal strategy), SFT-IT and SFT-CV showed superior improvements in **accuracy** and **relevance**, while SFT-CV also excelled in **comprehensiveness**. However, all SFT methods still underperformed ICL (optimal strategy) in **specificity** (Figure 6, bottom table upper row).

Among the hybrid approaches, no statistically significant differences were observed across models regarding **accuracy** and **comprehensiveness**; however, SFT-FS consistently demonstrated greater improvements in **relevance** compared to other methods. Additionally, SFT-IT showed less improvement in **specificity** than other approaches (Figure 6, bottom table lower row).

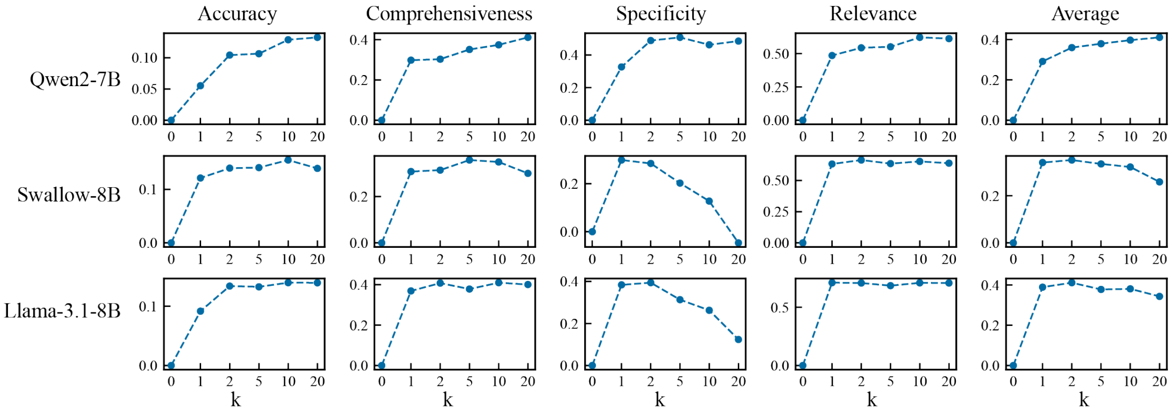


Figure 3: Experiment on the effect of ICL in selecting the optimal k-value.

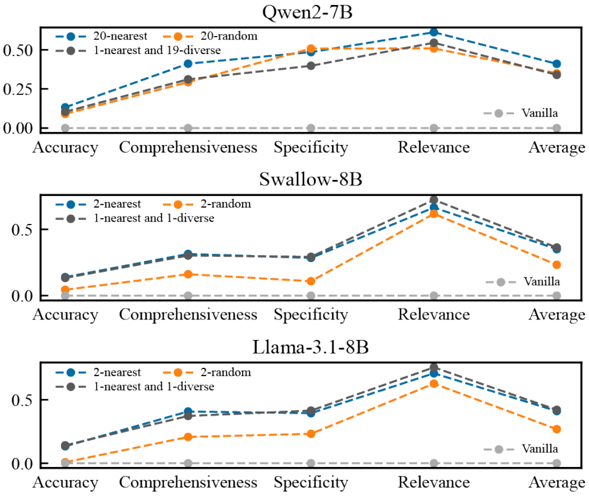
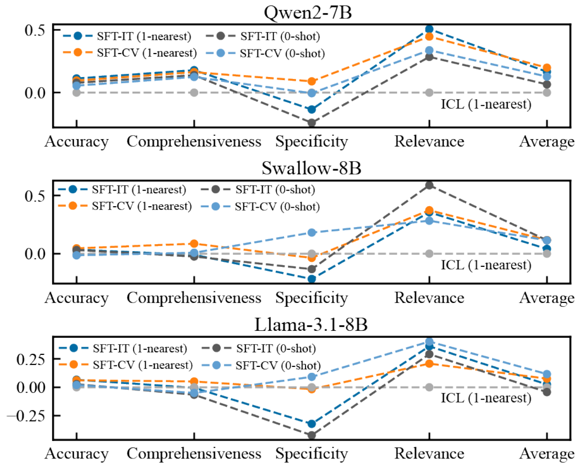


Figure 4: Normalized score improvement across different demonstration selection strategies for ICL. Vanilla LLM performance is normalized to 0 (gray baseline).



	Accuracy	Comprehensiveness	Specificity	Relevance
SFT-IT (1-nearest vs. 0-shot)	1-nearest *	1-nearest *	-	-
SFT-CV (1-nearest vs. 0-shot)	1-nearest *	1-nearest *	-	-

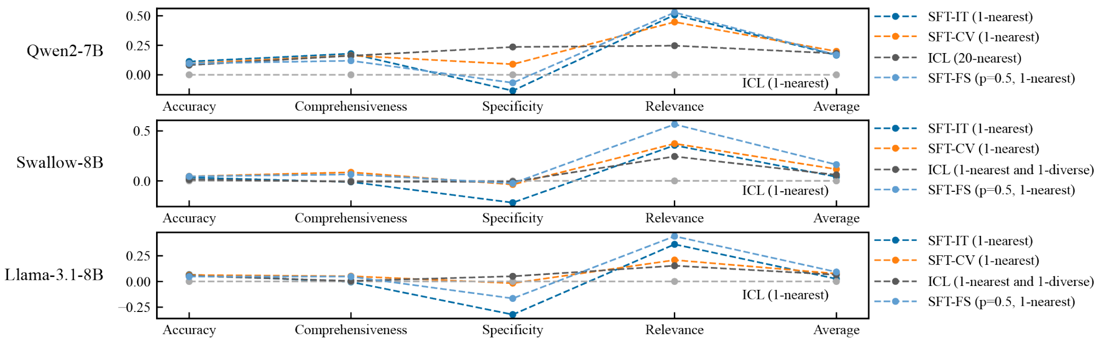
Figure 5: The score improvement in combined approaches against 0-shot in SFT-IT and SFT-CV. *Top*: normalized score improvements from vanilla LLMs (vanilla LLM with 1-nearest-shot = 0, gray baseline). *Bottom*: winner of the comparison between 1-nearest-shot and 0-shot.

4.4 Experiment 4: Experiments in other LFQA Datasets

In order to examine the generalizability of the results shown above, we conducted experiments with Qwen2-7B on other LFQA datasets, MilkQA and cMedQA2 (see Appendix B.4), which are expected to have relatively high domain specificity. Table 1 shows the experimental results for the MilkQA dataset. Regarding **accuracy**, similar to results in the AAPG task, SFT-CV and ICL achieved comparable scores, and the combination of these methods showed additive effects of improved accuracy. On the other hand, for SFT-IT, the combined approaches did not yield significant improvements compared to ICL alone. For **comprehensiveness**, we observed an additive effect of SFT-CV and ICL, similar to the AAPG task.

However, **specificity** and **relevance** tended to deteriorate with SFT-based methods. Nevertheless, as with the AAPG task, performance degradation was reduced in SFT-CV compared to SFT-IT.

Table 2 presents the experimental results for the cMedQA2 dataset. Regarding accuracy, SFT-IT and SFT-CV showed improvements comparable to ICL (with SFT-IT showing slightly better improvement), but combination did not produce additive effects. For the other three metrics besides accuracy, ICL showed only minimal score improvements, while other methods tended to



	Accuracy	Comprehensiveness	Specificity	Relevance
SFT-IT (1-nearest) vs. ICL	SFT-IT (1-nearest) *	-	ICL *	SFT-IT (1-nearest) *
SFT-CV (1-nearest) vs. ICL	SFT-CV (1-nearest) *	SFT-CV (1-nearest) *	ICL	SFT-CV (1-nearest) *
SFT-FS (p=0.5, 1-nearest) vs. ICL	-	-	ICL *	SFT-FS (p=0.5, 1-nearest) *
SFT-CV (1-nearest) vs. SFT-IT (1-nearest)	-	-	SFT-CV (1-nearest) *	-
SFT-FS (p=0.5, 1-nearest) vs. SFT-IT (1-nearest)	-	-	SFT-FS (p=0.5, 1-nearest) *	SFT-FS (p=0.5, 1-nearest) *
SFT-FS (p=0.5, 1-nearest) vs. SFT-CV (1-nearest)	-	SFT-CV (1-nearest)	-	SFT-FS (p=0.5, nearest 1-shot) *

Figure 6: Comparison between the hybrid approaches and ICL (optimal strategy). *Top*: normalized score improvements from vanilla LLMs between SFT-IT, SFT-CV, SFT-FS, and ICL with the 1-nearest-shot inference (vanilla LLM with 1-nearest-shot = 0, gray baseline). *Bottom*: winner methods with consistency across the models (upper row: SFT with 1-nearest-shot vs. ICL (optimal strategy); lower row: comparison between hybrid approaches).

	Accuracy	Comprehensiveness	Specificity	Relevance
Vannila	0.214	2.21	4.26	4.32
ICL (20-nearest)	0.234	2.46	**4.59**	**4.56**
SFT-IT	0.225	2.24	3.42	3.61
SFT-CV	0.231	2.42	4.21	4.08
SFT-IT (1-nearest)	0.236	2.32	3.44	3.62
SFT-CV (1-nearest)	**0.242**	**2.51**	4.44	4.30

Table 1: Results of replication experiments with Qwen2 on the MilkQA dataset. Scores are presented without normalization.

	Accuracy	Comprehensiveness	Specificity	Relevance
Vannila	0.236	2.97	4.81	4.55
ICL (20-nearest)	0.269	**3.05**	**4.82**	**4.58**
SFT-IT	**0.277**	2.70	3.06	3.47
SFT-CV	0.260	2.95	4.38	4.26
SFT-IT (1-nearest)	0.268	2.62	3.27	3.62
SFT-CV (1-nearest)	0.261	2.90	4.34	4.27

Table 2: Results of replication experiments with Qwen2 on the cMedQA2 dataset. Scores are presented without normalization.

deteriorate. Additionally, SFT-based methods showed performance degradation. As with the AAPG and MilkQA, the performance degradation was reduced in SFT-CV compared to SFT-IT.

5 Discussion

5.1 SFT vs. ICL

This study analyzed the features of text generated by the various domain-specialized LLMs. SFT and ICL showed almost equivalent improvements in **average** scores, which was consistent with task-based analysis in the previous research (Ovadia et al., 2024; Soudani et al., 2024; Bassamzadeh and Methani, 2024; Mosbach et al., 2023). These results, showing that their combination approaches improved more in some metrics, also align with Balaguer et al. (2024).

On the other hand, the comparison between SFT-based approaches and ICL revealed distinct performance variations in generated text across different metrics. For **accuracy**, while zero-shot SFT-based methods and ICL showed similar improvements from vanilla, hybrid approaches demonstrated additive improvements beyond using either method alone. For **comprehensiveness**, similar results were observed, but only for SFT-CV. In contrast, regarding **specificity**, SFT-based methods consistently underperformed compared to ICL, and even hybrid approaches underperformed relative to ICL. Additionally, for **relevance**, SFT-based methods consistently improved over ICL, but the combined methods did not show consistent improvements over zero-shot SFT.

An important consideration is that experiments conducted on other datasets showed limited generalizability of these findings. In MilkQA dataset, While **accuracy** and **comprehensiveness** showed similar trends, we could not observe similar patterns for **specificity** and **relevance**. Based on the preliminary study shown in Appendix B.4, *ground truth* evaluation scores for these metrics were lower than those for vanilla LLMs, suggesting that improvements in SFT-based methods were limited by the quality of the training data. In the cMedQA2 dataset, the results observed in AAPG could not be replicated. This may suggest limited potential for domain specialization, possibly due to the dataset's reliance on publicly available online platforms (Zhang et al., 2018), which could have been included in the LLM's pre-training data.

5.2 Explanations of the Performance Differences

The performance differences between SFT and ICL can be attributed to several factors. We classified evaluation rationales for deductions made by LLM-as-a-judge into into several categories to investigate potential differences between SFT and ICL scoring results. In terms of **specificity**, SFT-IT demonstrated more frequent negative rationales about the target of the audit procedure compared to ICL. Regarding relevance, we observed fewer negative rationales for audit procedures in the relatively challenging topic of accounting estimates. Furthermore, while **comprehensiveness** scores of SFT-IT and ICL showed minimal overall differences, SFT-IT exhibited more observations of negative rationales related to IT or internal controls compared to ICL. Additional details of the results are shown in Appendix F. These findings suggest that SFT-based methods, when compared to ICL, demonstrate superior improvement in selecting issues directly corresponding to the question (matters under consideration), while tending to provide insufficient or ambiguous descriptions of supplementary matters.

Moreover, we conducted a further ablation study to understand ICL's domain specialization. To understand the mechanism behind ICL's performance in **specificity**, we investigated whether relevant descriptions in the context enhance accuracy. Using k=20 nearest ICL (specificity score = 4.738) as baseline, we performed ablation studies that disrupted relationships between relevant passages. The results showed that when providing only shuffled nouns from ICL context, specificity dropped to 4.663 (-0.074), while shuffled sentences resulted in a smaller decrease to 4.719 (-0.018). Notably, disrupting input-output correspondence through shuffling actually improved the score slightly to 4.747 (+0.0095). These findings indicate that ICL in this study operates primarily through knowledge injection from relevant contextual information rather than through pattern recognition of input-output correspondences, distinguishing it from traditional in-context learning mechanisms.

5.3 Implications for Applications to Accounting Audit

This research demonstrated that different domain specialization methods exhibit distinct patterns in generation behavior. When creating domain-specialized LLMs, methods should be selected according to the desired output features. For instance, SFT-FS, which demonstrates high relevance, is suitable for creating audit procedure drafts. In contrast, SFT-IT, which demonstrates high comprehensiveness, is more effective for checking the completeness of human-designed audit procedures. Moreover, when the reliability of the training data is questionable, ICL can mitigate the risk of performance degradation. Alternatively, when SFT approaches are preferred, SFT-CV effectively minimizes performance degradation.

6 Conclusion

The comparison of domain specialization methods for AAPG tasks revealed that ICL and SFT exhibit distinct characteristics in their generation. SFT demonstrated a greater improvement in relevance, while ICL showed a greater increase in specificity. The hybrid approaches of ICL and SFT outperformed the individual methods, suggesting an additive effect between the two approaches. The different hybrid methods of SFT and ICL also exhibited varying patterns of capability acquisition. These findings provide insights into the potential differences in domain specialization.

7 Limitations

This study has the following limitations: (1) The experiments were conducted with a narrow focus on audit procedures generation as the domain-specific target, which constitutes a highly specialized domain and task. While some results are also demonstrated in other LFQA datasets, it presents replication challenges for the key differences of SFT and ICL with higher-quality question-answering datasets. (2) This experiment focused on LLMs with model sizes of approximately 7-8B parameters, and it remains unclear whether similar results would be obtained with smaller or larger model sizes. For example, (Soudani et al., 2024) obtained different conclusions with relatively smaller model sizes. In particular, since model size affects the susceptibility to catastrophic forgetting (Ramasesh et al., 2022), further experiments with different model sizes are necessary. (3) Since fine-tuning employs PEFT (QLoRA), it remains unclear whether these results can be similarly reproduced in full-parameter models or PEFT without quantization. Moreover, the range of k in ICL ablation study is also limited by computational resources. (4) Validation of LLM-as-a-judge through sensitivity analysis alone does not guarantee reliability. For example, there may be a gap between the evaluation metrics and what we actually perceive.

8 Ethical Considerations

Although this research provides several insights into the application of domain-specialized LLMs to accounting audit, it is essential to consider the potential risks and practical implications for audit procedures, as accounting audit work is a highly regulated field with significant social responsibility. Additionally, the experiments in this study do not directly address real-world audit procedure tasks, as audit procedures are determined according to audit firm policies and auditing standards and need to be verified by audit professionals. Care must be taken when applying the results of this study to real-world audit procedures.

Acknowledgments

We express our gratitude to the anonymous reviewers for their valuable and constructive feedback.

References

Simone Alghisi, Massimo Rizzoli, Gabriel Roccabruna, Seyed Mahed Mousavi, and Giuseppe Riccardi. 2024. Should we fine-tune or rag? evaluating different techniques to adapt llms for dialogue.

Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, Rafael Padilha, Morris Sharp, Bruno Silva, Swati Sharma, Vijay Aski, and Ranveer Chandra. 2024. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *Preprint*, arXiv:2401.08406.

Nastaran Bassamzadeh and Chhaya Methani. 2024. A comparative study of dsl code generation: Fine-tuning vs. optimized retrieval augmentation. *Preprint*, arXiv:2407.02742.

David Biesner, Maren Pielka, Rajkumar Ramamurthy, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, and Rafet Sifa. 2022. Zero-shot text matching for automated auditing using sentence transformers. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1637–1642.

Jan Cegin, Branislav Pecher, Jakub Simko, Ivan Srba, Maria Bielikova, and Peter Brusilovsky. 2024. Use random selection for now: Investigation of few-shot selection strategies in llm-based text augmentation for classification. *Preprint*, arXiv:2410.10756.

Ernie Chang, Xiaoyu Shen, Hui-Syuan Yeh, and Vera Demberg. 2021. On training instance selection for few-shot neural text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 8–13, Online. Association for Computational Linguistics.

Marcelo Criscuolo, Erick Rocha Fonseca, Sandra Maria Aluísio, and Ana Carolina Sperança-Criscuolo. 2017. MilkQA: a dataset of consumer questions for the task of answer selection. In *Proceedings of the 6th Brazilian Conference on Intelligent Systems (BRACIS)*, volume 1, pages 354–359, Uberlândia, Brazil. IEEE.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 719–729, New York, NY, USA. Association for Computing Machinery.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Nobushige Doi, Yusuke Nobuta, and Takeshi Mizuno. 2024. Measuring semantic similarity in japanese key audit matters. In *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 468–475.

Marc Eulerich, Aida Sanatizadeh, Hamid Vakilzadeh, and David A. Wood. 2024. Is it all hype? ChatGPT's performance and disruptive potential in the accounting and auditing industries. *Review of Accounting Studies*, 29(3):2318–2349.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: long form question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3558–3567. Association for Computational Linguistics.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, page (to appear), University of Pennsylvania, USA.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen

Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Hanchi Gu, Marco Schreyer, Kevin Moffitt, and Miklos Vasarhelyi. 2024. Artificial intelligence co-piloted auditing. *International Journal of Accounting Information Systems*, 54:100698.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Jiajia Huang, Maowei Jiang, and Haoran Zhu. 2025. Audit-ft at the regulations challenge task: An open-source large language model for audit. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 335–348.

Shih-Cheng Huang, Pin-Zu Li, Yu-chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tsai, and Hung-yi Lee. 2024. Chat vector: A simple approach to equip LLMs with instruction following and model alignment in new languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10943–10959, Bangkok, Thailand. Association for Computational Linguistics.

IAASB and IFAC. 2024a. International standard on auditing 230: Audit documentation.

IAASB and IFAC. 2024b. International standard on auditing 500: Audit evidence.

JICPA. 2022. JICPA auditing standards committee statement no. 500: Audit evidence.

JICPA. 2024. JICPA auditing standards committee statement no. 230: Audit documentation.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

V. I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions and reversals. *Doklady. Akademii Nauk SSSR*, 163(4):845–848.

Huaxia Li, Marcelo Machado de Freitas, Heejae Lee, and Miklos Vasarhelyi. 2024. Enhancing continuous auditing with large language models: Ai-assisted real-time accounting information cross-verification.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Preprint*, arXiv:2303.14070.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhagash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, Carl Yang, and Liang Zhao. 2024. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *Preprint*, arXiv:2305.18703.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Microsoft. 2023. Azureml metrics python package. [Online; accessed 2025-02-15].

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. 2024. Building a large japanese web corpus for large language models. In *Proceedings of the First Conference on Language Modeling*, COLM, page (to appear), University of Pennsylvania, USA.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 237–250, Miami, Florida, USA. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

PCAOB. 2004. AS 1215: Audit documentation.

PCAOB. 2010. AS 1105: Audit evidence.

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. 2022. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *Preprint*, arXiv:2212.13138.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Fine tuning vs. retrieval augmented generation for less popular knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, page 12–22. ACM.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-source financial large language models. *FinLLM Symposium at IJCAI 2023*.

S. Zhang, X. Zhang, H. Wang, L. Guo, and S. Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. RAFT: Adapting language model to domain specific RAG. In *First Conference on Language Modeling*.

A Background

To understand the experimental settings and discuss them using consistent terminology, we introduce the following definitions:

A.1 Domain and Task

Based on Pan and Yang (2010), we define domain as $\mathcal{D} = (\mathcal{X}, P(X))$, where $\mathcal{X}$ represents the input text for LLMs and $P(X)$ denotes the marginal probability distribution of input text, with $X \in \mathcal{X}$. In domain adaptation, the target domain is denoted $\mathcal{D}_t = (\mathcal{X}_t, P_t)$. Task is defined as $\mathcal{T} = (\mathcal{Y}, \mathcal{L})$, where $\mathcal{Y}$ represents the output space of LLMs and $\mathcal{L}$ represents the objective function. The target task is denoted as $\mathcal{T}_t = (\mathcal{Y}_t, \mathcal{L}_t)$.

A.2 Supervised Fine-tuning with LoRA (Low-Rank Adaptation)

We consider updating the parameters θ of the pre-trained model M_θ using the training dataset $D_t = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathcal{X}_t \times \mathcal{Y}_t$ from the target domain.

LoRA freezes most of the pre-trained parameters and introduces small additional parameters in a low-rank decomposition for training. For a weight matrix W from the pre-trained parameters θ, LoRA adds a low-rank decomposition:

$$W \leftarrow W + BA, \tag{7}$$

where $A \in \mathbb{R}^{r \times d_{in}}$, $B \in \mathbb{R}^{d_{out} \times r}$ are the newly introduced LoRA parameters (rank r), W is the original pre-trained matrix, which remains frozen [7]. Let $\theta_{\text{new}} = (A, B)$ denote the trainable LoRA parameters, and let θ_{frozen} represent the frozen pre-trained parameters. The model can be written as:

$$M_{\theta_{\text{frozen}}, \theta_{\text{new}}}. \tag{8}$$

We only optimize over $\theta_{\text{new}} = (A, B)$ by minimizing task-specific objective $\mathcal{L}_t$:

$$\theta^*_{\text{new}} = \arg\min_{\theta_{\text{new}}} \mathcal{L}_t(\theta_{\text{frozen}}, \theta_{\text{new}}; D_{\text{train}}). \tag{9}$$

B Analysis of Other LFQA datasets

Beyond the AAPG task, we conducted experiments evaluating LLM-generated text from multiple perspectives for question answering tasks using other publicly available LFQA datasets.

B.1 MilkQA

MilkQA (Criscuolo et al., 2017) comprises consumer questions and expert answers from the dairy sector of Embrapa (a Brazilian agricultural research company), collected by their customer service department between 2003 and 2012, representing real-world dairy consultation scenarios. From the 2,657 question-answer pairs available, we used 1,412 pairs ranging from 50 to 1,000 words. We allocated 20% of the dataset for evaluation and the rest for training or development. Prompts for inference and evaluation were created in Portuguese.

B.2 cMedQA2

The cMedQA2 dataset consists of questions and answers collected from a Chinese online health consultation website, covering symptom descriptions, disease diagnosis and treatment, medication use, and psychological consultations, containing approximately 54,000 questions and more than 101,000 answers (Zhang et al., 2018). Due to the high computational cost of LLM-as-a-judge evaluation, this research focuses on relatively high-quality data. Specifically, we narrowed down to answers provided by multiple physicians with a minimum similarity threshold with other answers (correlation coefficient of embedding vectors projected by multilingual-e5 being at least 0.9), and answers between 100-200 characters in length, resulted in 9,936 pairs. We designated 5% of the dataset for evaluation and the remainder for training or development. Prompts for inference and evaluation were created in Chinese, and character segmentation for ROUGE-F1 score calculation utilized the Rouge-Chinese library[8].

B.3 ELI5-Category

The ELI5 Category dataset (ELI5-Category) is a more recent, categorized English question-answer dataset that, while smaller than the original ELI5 dataset (Fan et al., 2019), features content collected from Reddit's "r/explainlikeimfive" subreddit from January 2017 to June 2021. It consists of fact-based questions requiring extended responses, along with their corresponding responses. For this research, we excluded the 'Repost' category, which contains duplicate content and randomly selected 5% (452 instances) from question-answer pairs under 1,000

[7] This is simplified by omitting the scaling term.

[8] https://github.com/Isaac-JL-Chen/rouge_chinese

words as evaluation data.

B.4 Evaluation of the *Ground Truth* Answers

Table 3 shows the comparisons using four metrics comparing the responses generated by Qwen2-7B, Swallow-8B, and Llama-3.1-8B with the ground truth on the evaluation split of the dataset for the AAPG task. For the other datasets, we also performed a similar evaluation using the same four metrics, comparing responses generated by Qwen2 and Llama-3.1 with the ground truth. The results indicate that for the AAPG task, the ground truth received higher evaluations across all four metrics compared to outputs from vanilla LLMs, suggesting the potential for domain specialization. In contrast, for the other datasets, both Qwen2 and Llama-3.1 generated responses that scored higher than the ground truth in terms of specificity and relevance. These datasets are therefore not suitable for analyzing improvements in domain-specialized LLMs.

C Dataset and Pre-processing

AAPG dataset was extracted from of audit reports from securities reports with fiscal year-ends between March 31, 2021, and March 31, 2024, submitted up to July 2024, obtained from the Electronic Disclosure for Investors' NETwork (EDINET) site [9]. Data can be accessed through the EDINET API [10]. The data was provided in eXtensible Business Reporting Language (XBRL) format and was parsed using the Arelle library[11]. Considering computational resources, from 13,878 audit reports with 17,326 KAMs we selected cases where the token size of KAM consideration descriptions was below 768 tokens and the auditors' response descriptions below 1024 tokens. The training data pool, consisting of data prior to March 31, 2024, contained in 9,566 KAMs after excluding KAMs from the same submitters included in the evaluation data (500 audit reports with 607 KAMs).

Due to minimal annual updates in KAM descriptions, we excluded similar KAMs from the training data (Doi et al., 2024). For each submitter, we calculated the Levenshtein distance (Levenshtein, 1965) with previously submitted KAMs, excluding past KAMs if the distance was less than 200. This resulted in 8,350 KAMs as training cases.

The following preprocessing was performed: (1) HTML parsing and normalization, (2) converting verb endings from past tense to regular form and (3) converting auditor response descriptions to markdown format using llama-3.1-8B-Instruct to reduce evaluation variance due to formatting differences.

For SFT training, the 8,350 KAMs were divided into training and validation data at a ratio of 90% and 10%. The training was conducted with 4-bit quantization, LoRA rank of 16, learning rate of 2e-5, and batch size of 2. Training was conducted for up to 6 epochs and selected the model with the lowest loss on the validation data, and computations were performed using NVIDIA A6000, taking approximately 2 days for each fine-tuning process.

D Comparison between $p = 1$ and $p = 0.5$ of SFT-FS Training

In the RAFT paper (Zhang et al., 2024), the proportion of training instances including oracle documents varies by dataset, with settings of $p = 0.4$, $p = 0.6$, and $p = 1$. Therefore, we compared cases of $p = 0.5$ and $p = 1$ in SFT-FS. The results are shown in Figure 7. Consistently across models, $p = 0.5$ showed improved Specificity scores compared to $p = 1$; however, the improvement margin was not statistically significant. Furthermore, while Qwen2 and Llama-3.1 demonstrated improved Relevance scores, Swallow showed no significant difference and showed a lack of consistency across models. On the other hand, the accuracy scores were higher with $p = 1$ compared to $p = 0.5$.

E Sensitivity Analysis of LLM-as-a-Judge

In this study, we use LLM-as-a-Judge for evaluating comprehensiveness, specificity and relevance of generated audit procedures. Since the evaluation prompt is original to this task, we validate its effectiveness in LLM-as-a-judge.

First, to check **comprehensiveness** sensitivity of the LLM evaluator, we prepare synthetic audit procedures for the KAM by intentionally including procedures from the ground truth when generating audit procedures with GPT-4o-2024-08-06. Second, to check **specificity** sensitivity of the LLM evaluator, we prepare a synthetic diluted KAM

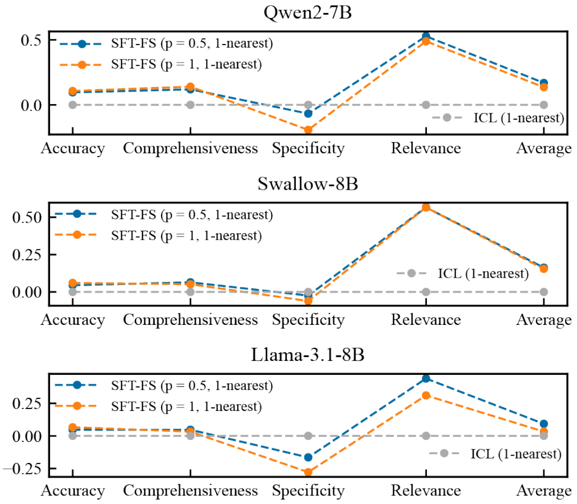
[9] http://disclosure.edinet-fsa.go.jp/

[10] https://disclosure.edinet-fsa.go.jp/EKW0EZ0015.html

[11] https://arelle.org/arelle/

Task	Model (or ground truth)	Accuracy	Comprehensiveness	Specificity	Relevance
AAPG	Qwen2	0.239	2.641	4.491	4.091
	Swallow	0.237	2.652	4.417	3.956
	Llama-3.1	0.241	2.496	4.361	3.768
	ground truth	0.867	5.000	4.581	4.813
MilkQA	Qwen2	0.214	2.212	4.268	4.325
	Llama-3.1	0.215	2.098	4.561	4.462
	ground truth	0.893	4.996	3.02	3.519
cMedQA2	Qwen2	0.236	2.975	4.814	4.555
	Llama-3.1	0.224	2.486	4.452	4.251
	ground truth	0.916	4.967	3.018	3.635
ELI5	Qwen2	0.255	2.541	4.829	4.627
	Llama-3.1	0.250	2.560	4.866	4.771
	ground truth	0.829	5.000	3.929	4.301

Table 3: Multi-perspective score of vanilla model and ground truth for long-form question answering task.



	Accuracy	Comprehensiveness	Specificity	Relevance
(p=1) vs. (p=0.5)	(p=1) *	-	(p=0.5)	-

Figure 7: Difference between selection of parameters in SFT-FS training phase. *Top*: normalized score improvements from vanilla LLMs (vanilla LLM with 1-nearest-shot = 0, gray baseline). *Bottom*: winner with consistency between models.

	Comprehensiveness
Without ground truth procedure	3.599
Include one ground truth procedure	3.691
Onclude all ground truth procedures	4.909

	Specificity
Plain	4.855
Diluted with another KAM	4.563
Diluted with 3 other KAM	4.420

	Relevance
Without unrelated KAM	4.393
Include one procedure for an unrelated KAM	3.845
Include all procedures for an unrelated KAM	2.434

Table 4: Sensitivity analysis of LLM-as-a-judge evaluation. *Top*: comprehensiveness changes by inserting ground truth procedures. *Middle*: specificity changes by diluting with other KAMs. *Bottom*: Relevance changes by inserting procedures for unrelated KAMs.

by integrating a KAM with its k-nearest neighbor KAMs (k = 1 to 3) as noise into a single KAM using GPT-4o with the prompt "Summarize the following several audit considerations and summarize them as one generalized consideration." Finally, to check **relevance** sensitivity of the LLM evaluator, we prepare synthetic audit procedures for the KAM by intentionally including procedures from other unrelated KAMs when generating audit procedures. Table 4 shows scores decrease by these synthetic KAMs.

	vanilla	ICL	SFT-IT	SFT-CV	SFT-IT (1-nearest)	SFT-CV (1-nearest)	SFT-FS (1-nearest)
IT and System Controls	233.6	169.2	217.3	186.2	185.4	169.2	193.3
Tax Effect Accounting and Deferred Tax Assets	248.5	157.9	152.8	159.6	153.9	136.8	159.3
Involvement and Verification of External Experts	417.4	265.8	310.0	274.1	260.8	255.7	261.4
Assessment of Internal Control Design and Operation	529.6	240.8	252.0	286.7	286.8	238.7	284.6
Evaluation of Construction Cost Estimates	484.0	327.2	345.1	331.9	315.2	286.0	314.3
Business Plan and Future Cash Flow Analysis	381.4	295.3	305.6	302.0	325.2	295.8	267.5
Audit Data Analysis	443.9	246.9	257.7	261.6	227.0	234.5	249.2
Business Plan Evaluation and Performance Analysis	490.6	311.7	296.4	293.9	271.6	295.8	262.5
Asset Impairment Testing	450.1	272.4	227.5	254.1	258.0	264.2	226.6
Sales Transaction Verification	557.8	413.5	437.6	420.9	408.7	389.5	411.8

	SFT-CV/Vanilla - ICL/Vanilla	SFT-CV(1-nearest)/Vanilla - ICL/Vanilla	SFT-CV(1-nearest)/Vanilla - SFT-CV/Vanilla
IT and System Controls	0.073	0.000	**-0.073**
Tax Effect Accounting and Deferred Tax Assets	0.007	**-0.085**	**-0.092**
Involvement and Verification of External Experts	0.020	-0.024	-0.044
Assessment of Internal Control Design and Operation	0.087	-0.004	**-0.091**
Evaluation of Construction Cost Estimates	0.010	**-0.085**	**-0.095**
Business Plan and Future Cash Flow Analysis	0.017	0.001	-0.016
Audit Data Analysis	0.033	-0.028	-0.061
Business Plan Evaluation and Performance Analysis	-0.036	-0.032	0.004
Asset Impairment Testing	-0.040	-0.018	0.022
Sales Transaction Verification	0.013	-0.043	-0.056

Table 5: Topic-specific deductions regarding **comprehensiveness**. *Top*: distribution of deductions from LLM-as-a-judge across topics. Smaller values indicate fewer deductions. *Bottom*: improvement and comparison of topic-specific deductions regarding comprehensiveness. For ratios, values less than 1 mean fewer deductions than Vanilla, indicating improvement. For differences, negative values indicate improvement compared to the reference. Topics with relatively large differences are highlighted in bold if they show improvement compared to the reference, or underlined if the reference shows greater improvement.

	vanilla	ICL	SFT-IT	SFT-CV	SFT-IT (1-nearest)	SFT-CV (1-nearest)	SFT-FS (1-nearest)
Sales and performance forecasts by management	85.9	55.3	71.2	60.3	75.4	52.4	64.4
Standards and verification in internal controls	138.7	69.3	110.0	75.6	89.5	82.9	86.3
Contents of specific transactions	73.1	63.4	117.6	53.6	88.8	50.7	78.0
Impairment recognition and profitability	118.5	68.4	87.4	57.4	77.0	64.0	87.5
Insufficient coverage of the mentioned scope	115.8	80.1	95.5	65.5	102.7	80.1	91.3
Future forecasts and cash flow	74.9	42.9	43.3	35.7	62.4	49.2	47.3
Recoverability of deferred tax assets	66.6	53.9	81.2	40.0	72.7	66.2	62.8
Evaluation of business plan assumptions	123.9	62.2	74.4	67.7	71.1	64.8	76.0
Data analysis methods	126.2	65.0	80.6	78.9	86.7	78.7	59.3
External factors, market influences, and risks	97.8	59.3	94.0	78.5	105.1	84.5	81.0

	ICL / Vanilla	SFT-IT / Vanilla	SFT-IT(1-nearest)/Vanilla	SFT-IT/Vanilla - ICL/Vanilla	SFT-IT(1-nearest)/Vanilla - SFT-IT/Vanilla
Sales and performance forecasts by management	0.644	0.829	0.877	0.185	0.048
Standards and verification in internal controls	0.500	0.793	0.645	0.294	**-0.148**
Contents of specific transactions	0.868	1.610	1.216	0.741	**-0.394**
Impairment recognition and profitability	0.577	0.737	0.650	0.160	-0.088
Insufficient coverage of the mentioned scope	0.692	0.825	0.887	0.133	0.062
Future forecasts and cash flow	0.573	0.578	0.833	0.005	0.255
Recoverability of deferred tax assets	0.809	1.218	1.091	0.409	**-0.127**
Evaluation of business plan assumptions	0.502	0.601	0.574	0.099	-0.027
Data analysis methods	0.515	0.639	0.687	0.124	0.048
External factors, market influences, and risks	0.606	0.962	1.075	0.356	0.113

Table 6: Topic-specific deductions regarding **specificity**.

	vanilla	ICL	SFT-IT	SFT-CV	SFT-IT (1-nearest)	SFT-CV (1-nearest)	SFT-FS (1-nearest)
Revenue recognition (period attribution, existence)	102.0	80.4	116.4	82.9	104.8	110.4	97.8
Impairment assessment and cash flow	119.6	77.0	30.0	43.1	44.3	55.1	30.1
Evaluation of estimate appropriateness	169.8	87.6	42.5	62.0	59.9	55.3	34.7
Internal controls and validity of revenue recognition	117.2	50.6	54.9	31.6	36.4	39.6	38.7
Verification of sales/revenue existence	61.5	41.3	44.5	40.5	30.6	43.9	27.3
Audit reports and specific documentation of issues	952.4	59.3	28.2	53.1	28.4	19.7	22.4
Tax effect accounting and deferred tax asset valuation	111.1	46.9	31.4	44.7	30.0	36.0	26.0
Inventory valuation and impairment assessment	69.7	29.1	21.0	28.5	17.3	27.1	15.0
Accuracy of sales data and internal controls	80.0	42.1	38.8	47.1	35.8	32.4	33.1
Assumption evaluation and risk management process	95.0	41.7	31.8	36.0	28.8	40.5	20.4

	ICL / Vanilla	SFT-IT / Vanilla	SFT-CV / Vanilla	SFT-IT/Vanilla - ICL/Vanilla	SFT-CV/Vanilla - ICL/Vanilla
Revenue recognition (period attribution, existence)	0.788	1.141	0.813	0.353	0.025
Impairment assessment and cash flow	0.644	0.251	0.360	**-0.393**	**-0.283**
Evaluation of estimate appropriateness	0.516	0.251	0.365	**-0.265**	**-0.151**
Internal controls and validity of revenue recognition	0.432	0.469	0.269	0.037	**-0.163**
Verification of sales/revenue existence	0.672	0.725	0.659	0.053	-0.013
Audit reports and specific documentation of issues	0.062	0.030	0.056	-0.033	-0.006
Tax effect accounting and deferred tax asset valuation	0.422	0.283	0.402	**-0.139**	-0.020
Inventory valuation and impairment assessment	0.418	0.302	0.409	**-0.116**	-0.009
Accuracy of sales data and internal controls	0.526	0.485	0.590	-0.041	0.063
Assumption evaluation and risk management process	0.439	0.335	0.378	**-0.104**	-0.060

Table 7: Topic-specific deductions regarding **relevance**.

F Topic-Based Analysis of LLM-as-a-Judge Evaluation Differentials

In LLM-as-a-judge evaluations, assessment scores are generated following the output of judgment rationales.

This section interprets score differentials for the LLM-as-a-judge evaluation metrics of comprehensiveness, specificity, and relevance highlighted in our main text. Specifically, we utilized GPT-4o-mini to extract rationales of deduction from the judgment rationales for each evaluation sample. These extracted deduction comments were classified into ten topics using Latent Dirichlet Allocation (LDA), and each evaluation sample's deductions were categorized into these ten topics according to topic weights.

Regarding **comprehensiveness**, both SFT-CV and ICL demonstrated comparable improvements, with further performance enhancement observed when hybridizing these approaches, indicating an additive effect. According to the deduction topics, the comparison between SFT-CV and ICL showed that ICL demonstrated relatively greater improvement in "IT and System Controls" and "Assessment of Internal Control Design and Operation." Since IT and internal control procedures serve as indirect verification methods, they are more susceptible to comprehensiveness critiques, suggesting that the examples provided in ICL offered an advantage (Table 5).

Furthermore, hybrid approaches compensated for areas where SFT-CV showed relatively minor improvement, such as "IT and System Controls" and "Assessment of Internal Control Design and Operation," achieving levels comparable to ICL. Simultaneously, "Tax Effect Accounting and Deferred Tax Assets" and "Evaluation of Construction Cost Estimates" showed enhancement. While the improvement differential between SFT-CV and ICL for these topics was not substantial, the hybrid approach demonstrated improvement from both SFT-CV and ICL perspectives, which demonstrates additive effects. These topics involve the evaluation of accounting estimates, which are challenging areas for audit procedure planning (Table 5).

For **specificity**, the improvement of SFT-IT was less pronounced than ICL's, and even the combination of SFT-IT with 1-shot did not reach ICL's level of improvement (Table 6 top). Topic-specific comparison revealed that the primary differences in improvement magnitude between SFT-IT and ICL were most evident in "Standards and verification in internal controls," "Contents of specific transactions," "Recoverability of deferred tax assets," and "External factors, market influences, and risks." This suggests that SFT-IT relatively lacked specific descriptions regarding audit procedure targets. While "External factors, market influences, and risks" showed improvement in three approaches outside of one hybrid method, it still did not attain ICL's level (Table 6 bottom).

Regarding **relevance**, SFT-IT and SFT-CV demonstrated more substantial improvement than ICL (Table 7 top). Topic-specific analysis indicated common differentials from ICL in "Impairment assessment and cash flow" and "Evaluation of estimate appropriateness." These topics represent challenging areas for proposing audit procedures related to accounting estimates (Table 7 bottom).

G Raw Evaluation Scores

Normalized increase of evaluation score have already shown and discussed in the main text, however to provide objective viewpoints raw scores of each evaluation metric is also shown in Table 8 for Qwen2-7B, 9 for Swallow-8B, and 10 for Llama-3.1-8B. Notably, the performance of SFT on Swallow-8B is higher than that of Llama-3.1, which is the base model of Swallow. This indicates that domain adaptation to Japanese language leads to accompanying domain specialization in Japanese-specific expert tasks, such as audit procedures generation task.

H The Prompts Used for LLM-as-a-Judge Evaluation

Prompt for LLM-as-a-judge evaluation is shown in the following.

	Accuracy	Comprehensiveness	Specificity	Relevance	Normalized Average Increase
Vanilla	0.239	2.641	4.491	4.091	0.000
ICL 1-nearest	0.281	3.344	4.657	4.532	0.291
ICL 2-nearest	0.318	3.356	4.740	4.585	0.360
ICL 5-nearest	0.320	3.470	4.750	4.591	0.379
ICL 10-nearest	0.337	3.524	4.727	4.656	0.397
ICL 20-nearest	0.340	3.611	4.738	4.647	0.410
ICL 1-nearest and 19-diverse	0.318	3.376	4.694	4.586	0.340
ICL 20-random	0.307	3.333	4.750	4.554	0.350
SFT-IT	0.337	3.575	4.575	4.666	0.331
SFT-CV	0.320	3.551	4.656	4.690	0.369
SFT-IT (1-nearest)	0.362	3.641	4.611	4.769	0.392
SFT-CV (1-nearest)	0.350	3.614	4.688	4.741	0.415
SFT-FS (p=1)	0.358	3.577	4.591	4.761	0.372
SFT-FS (p=0.5)	0.350	3.542	4.634	4.779	0.392

Table 8: Raw scores and normalized average score increases of domain specialized **Qwen2-7B**.

	Accuracy	Comprehensiveness	Specificity	Relevance	Normalized Average Increase
Vanilla	0.237	2.652	4.417	3.956	0.000
ICL 1-nearest	0.330	3.376	4.591	4.616	0.340
ICL 2-nearest	0.344	3.390	4.583	4.649	0.351
ICL 5-nearest	0.344	3.493	4.535	4.619	0.334
ICL 10-nearest	0.355	3.473	4.492	4.638	0.321
ICL 20-nearest	0.343	3.357	4.390	4.623	0.258
ICL 1-nearest and 1-diverse	0.339	3.362	4.588	4.710	0.363
ICL 2-random	0.271	3.031	4.481	4.600	0.233
SFT-IT	0.350	3.336	4.537	4.842	0.373
SFT-CV	0.320	3.389	4.666	4.725	0.396
SFT-IT (1-nearest)	0.353	3.359	4.502	4.753	0.340
SFT-CV (1-nearest)	0.360	3.516	4.577	4.759	0.393
SFT-FS (p=1)	0.370	3.460	4.567	4.834	0.404
SFT-FS (p=0.5)	0.360	3.481	4.582	4.834	0.409

Table 9: Raw scores and normalized average score increases of domain specialized **Swallow-8B**.

	Accuracy	Comprehensiveness	Specificity	Relevance	Normalized Average Increase
Vanilla	0.241	2.496	4.361	3.768	0.000
ICL 1-nearest	0.311	3.420	4.606	4.644	0.389
ICL 2-nearest	0.343	3.517	4.613	4.641	0.411
ICL 5-nearest	0.342	3.443	4.561	4.613	0.378
ICL 10-nearest	0.347	3.522	4.529	4.643	0.381
ICL 20-nearest	0.347	3.499	4.440	4.641	0.343
ICL 1-nearest and 1-diverse	0.348	3.428	4.626	4.699	0.421
ICL 2-random	0.247	3.015	4.509	4.540	0.269
SFT-IT	0.329	3.316	4.440	4.748	0.341
SFT-CV	0.324	3.341	4.643	4.787	0.429
SFT-IT (1-nearest)	0.356	3.413	4.479	4.773	0.380
SFT-CV (1-nearest)	0.353	3.501	4.600	4.718	0.424
SFT-FS (p=1)	0.356	3.471	4.497	4.755	0.389
SFT-FS (p=0.5)	0.343	3.494	4.541	4.801	0.413

Table 10: Raw scores and normalized average score increases of domain specialized **Llama-3.1-8B**.

[Prompt for Evaluating **Comprehensiveness** (the original prompts were written in Japanese)]

```
Please evaluate the comprehensiveness of the provided answer and assign a score according to the following instructions.
## Evaluation Criteria
Comprehensiveness is measured by how much of the content listed in the correct answer is included in the predicted response. The more elements from the correct answer that are included or similarly expressed in the predicted response, the higher the score.
## Scoring Scale
"5": "All elements listed in the correct answer are included in the predicted response with similar content"
"4": "Most elements listed in the correct answer are included in the predicted response with similar content"
"3": "About half of the elements listed in the correct answer are included in the predicted response with similar content"
"2": "Only a small portion of the elements listed in the correct answer are included in the predicted response with similar content"
"1": "None of the elements listed in the correct answer are included in the predicted response with similar content"
## Notes
- Evaluate only from the perspective of comprehensiveness. For example, do not consider the appropriateness of the audit procedures themselves or the specificity of the description.
- If the predicted response includes abstracted versions of elements listed in the correct answer, consider those elements as included.
- First provide step-by-step logical reasoning, then answer in the specified format.
## Response format
### Reasoning
Step-by-step logical reasoning
### Conclusion
{s̈corë:̈(integer from 1 to 5)}̈
## Evaluation Example
### Correct Answer
Apple characteristics:
1. Red or green skin
2. Sweet taste
3. Rich in dietary fiber
4. Contains Vitamin C
### Predicted Response
Apples have red skin and are sweet, delicious fruits. They are also considered good for health.
### Evaluation Result
#### Reasoning
Color and taste are mentioned, but nutritional aspects (dietary fiber, Vitamin C) are not mentioned. There is a general reference to health benefits, but it lacks specificity.
#### Conclusion
{s̈corë: 3}
Based on these instructions, please evaluate the predicted response against the provided correct answer and assign an appropriate score.
## Correct Answer
{INSERT GROUND TRUTH AUDIT PROCEDURES}
## Predicted Response
{INSERT GENERATED AUDIT PROCEDURES}
```

[Prompt for Evaluating **Specificity** (the original prompts were written in Japanese)]

```
You are an evaluator of responses in accounting audits. Specificity is measured by how well individual situations in the consideration items are reflected in the audit procedures of the predicted response. Please evaluate based on the following criteria.
## Evaluation Criteria
Reflection of consideration items: The higher the score, the more the predicted response covers the characteristics and concerns presented in the consideration items, and the more specific and feasible the proposed audit procedures are.
## Scoring Scale
"5": "Reflects all individual situations shown in the consideration items, the description of audit procedures is specific, and there are no ambiguous points."
"4": "Reflects about 90% of individual situations shown in the consideration items with specific audit procedures, but there is one ambiguous point."
"3": "Reflects most individual situations shown in the consideration items with specific audit procedures, but there are two or more ambiguous points."
"2": "Partially reflects the individual situations shown in the consideration items, but there are ambiguous points in the description of audit procedures."
"1": "Only partially reflects the individual situations shown in the consideration items in the predicted audit procedures, and the description of audit procedures is not specific."
## Notes - Evaluate only from the perspective of specificity of description. Do not consider the comprehensiveness of the described audit procedures or their relevance to the risks mentioned in the consideration items.
- First, extract the individual situations shown in the consideration items, then examine step by step whether they are specifically reflected in the description of audit procedures. Finally, answer in the specified format.
## Response format
### Reasoning
Step-by-step logical reasoning
### Conclusion
{s̈core:̈(integer from 1 to 5)̈}
## Example Evaluation
### Consideration Items
Revenue is recognized based on acceptance criteria, but there is a risk that the period attribution of sales at the end of the month is inappropriate.
### Predicted Response
Verify that the sales recording date at the end of the month matches the date of the supporting documentation received from the customer.
### Evaluation Results
#### Reasoning
The mention of the end of the month partially reflects the individual situation, but the supporting documentation mentioned is not specific, and there is room for improvement, such as specifying acceptance documents, etc.
#### Conclusion
{s̈core:̈ 2}
Based on the above instructions, please evaluate the provided correct answer and predicted response and assign an appropriate score.
## Predicted Response
{INSERT GENERATED AUDIT PROCEDURES}
```

[Prompt for Evaluating **Relevance** (the original prompts were written in Japanese)]

You are a grader of responses in accounting audits. Scoring for relevance is based on whether the predicted audit procedures address the issues described in the matters for consideration. The comprehensiveness of addressing the issues stated in the matters for consideration is not taken into account.

The score is determined according to the following evaluation scale:

Evaluation Scale

"5": "All of the predicted audit procedures directly address the matters stated in the considerations, and there is no room for improvement in terms of relevance."

"4": "About 90% of the predicted audit procedures address the matters stated in the considerations, but one procedure has low relevance."

"3": "The majority of the predicted audit procedures address the matters stated in the considerations, but two or more procedures have low relevance."

"2": "Some of the predicted audit procedures address the matters stated in the considerations, but many procedures have low relevance."

"1": "None of the predicted audit procedures have high relevance to the matters stated in the considerations."

Notes

- Please evaluate only from the perspective of relevance. Do not consider the comprehensiveness or specificity of the described audit procedures.
- Please examine the relevance of each predicted audit procedure to the matters for consideration step by step, and then respond in the specified format.

Response format

Reasoning

Step-by-step logical reasoning

Conclusion

{s̈core:̈(integer from 1 to 5)̈}

Evaluation Example

Matters for Consideration

Revenue is recognized based on acceptance criteria, but there is a risk that the period attribution of revenue in the final month is inappropriate.

Predicted Audit Procedures

For sales transactions recorded in the final month, verify the recording date against the acceptance date on the acceptance document received from the customer.

For sales transactions recorded in the final month, verify the recorded amount against the amount on the acceptance document received from the customer.

Evaluation Results

Rationale

While verifying the recording date against the acceptance document date addresses the issue stated in the matters for consideration, verifying the amount does not address the stated issue.

Conclusion

{s̈core:̈ 3}

Predicted Response

{INSERT GENERATED AUDIT PROCEDURES}