IntelliChain Stars at the Regulations Challenge Task: A Large Language Model for Financial Regulation

Shijia Jiang, Yongfu Dai, Haochen Jia, Yuxin Wang, Hao Wang*
Sichuan University, Chengdu, China†
sj7191190@gmail.com wal.daishen@gmail.com
wwx990211@gmail.com wangyuxin1st@gmail.com wangh@scu.edu.cn

Abstract

We present our approach to the COLING-2025 Regulations Challenge, which evaluates large language models (LLMs) on nine regulatory tasks, such as abbreviation recognition and financial data extraction. To address challenges like domain-specific terminologies and dynamic regulatory contexts, we developed a robust data construction pipeline, integrating proprietary Chinese regulatory data, Fin-GPT datasets, and financial Q&A data. The pipeline applied, but was not limited to, language filtering, semantic screening, and deduplication, resulting in a 30,000-example dataset combining financial regulations and general financial data. Using this dataset, we fine-tuned Llama 3.2-3B-Instruct to create **Reg-LLaMA**, a specialized model that outperformed baselines on the Regulations Challenge and PIXIU datasets. These results demonstrate the effectiveness of domain-specific data construction in advancing LLMs for regulatory tasks, paving the way for reliable and interpretable AI in regulated industries.

1 Introduction

The **COLING-2025 Regulations Challenge** (Wang et al., 2024) is a benchmark designed to evaluate the capabilities of large language models (LLMs) in processing and responding to regulatory texts. The competition consists of 9 distinct tasks, ranging from abbreviation recognition to advanced financial data extraction and licensing requirements under specific frameworks. Each task is structured to assess an LLM's ability to interpret, analyze, and generate precise outputs based on complex regulatory information. The tasks are designed with standardized templates that ensure consistency in input and output formats, reflecting real-world regulatory use cases.

Despite their immense potential, existing LLMs face significant challenges in the regulatory domain, such as:

- The complexity of regulatory texts, which often include domain-specific terminologies and nuanced legal interpretations (Hassani, 2024; Cao and Feinstein, 2024).
- The dynamic and region-specific nature of regulations, which require constant updates to remain relevant (Bharathi Mohan et al., 2024).
- A lack of explainability and interpretability in model outputs, which is critical for ethical and reliable applications in regulated industries (Zhao et al., 2024a; Cambria et al., 2024).

To address these challenges, we developed a comprehensive data construction pipeline to curate a high-quality dataset tailored to financial regulations. This pipeline integrates key steps such as language filtering, regular expression matching, semantic screening using financial domain embeddings, and optimization of data quality through perplexity-based filtering (Ankner et al., 2024) and deduplication (Lee et al., 2021). Additionally, privacy-sensitive content was removed to ensure compliance with security standards. These processes allowed us to construct a dataset of 30,000 examples, balancing domain-specific regulatory data and general financial datasets to enhance model robustness and task alignment.

Through this pipeline, we constructed a high-quality instruction dataset comprising 30,000 examples, including 10,000 financial regulation datasets and 20,000 general finance datasets, as detailed in Table 1.

Our experimental results validate the effectiveness of this approach. On three distinct frameworks, Reg-LLaMA outperformed peer models in tasks requiring nuanced understanding of financial regulations. These results demonstrate its superior

*This is the corresponding author.
†All authors are from this institution.

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 371–384
January 19–20, 2025. ©2025 Association for Computational Linguistics

Dataset	Source	Size	Description
Financial Regulations	AuditWen	10k	Proprietary dataset on financial regulations and audit rules
Financial Generals	ICE-FIND	10k	Proprietary bilingual dataset; English samples related to regulations
	Fin-GPT	10k	Open-source dataset for financial large language models

Table 1: Instruction-Tuned Dataset for Reg-LLaMA.

capability in handling complex, domain-specific queries.

In summary, our contributions include:

- Developing **Reg-LLaMA**, a specialized LLM tailored for regulatory challenges in the financial sector.

- Introducing a robust data construction pipeline that facilitates the construction of high-quality datasets for regulatory tasks.

- Establishing strong performance benchmarks, highlighting Reg-LLaMA's advancements in addressing key challenges in regulatory understanding and application.

By tackling the core difficulties of regulatory text comprehension, this work paves the way for more reliable and interpretable AI systems in regulated industries.

2 Related Work

Large Language Models in Financial Regulation. Large Language Models (LLMs) such as GPT-4 (Achiam et al., 2023), Llama-3.2 (Liu et al., 2024), and Mistral-Large-2 (Jiang et al., 2023) have demonstrated exceptional capabilities in various natural language processing tasks, including question answering, summary generation, and text creation (Lee et al., 2024). These models are increasingly utilized for financial regulatory tasks, but they face challenges in understanding complex domain-specific terminologies, adapting to rapidly changing regulatory frameworks, and ensuring the interpretability and ethical compliance of outputs (Araci, 2019)(Colangelo et al., 2022). Models like FinBERT (Yang et al., 2020) and FinGPT (Yang et al., 2023) attempt to address these issues by fine-tuning on specialized financial datasets, showcasing improved performance and robustness in handling regulatory tasks (Nie et al., 2024). Additionally, initiatives such as the COLING-2025 Regulations Challenge emphasize the importance of assessing LLMs' capabilities in regulatory scenarios, providing valuable benchmarks that identify gaps and drive innovation.

Datasets and Competitions in Financial Regulation. The complexity and dynamic nature of the financial regulatory domain necessitate high-quality and up-to-date datasets. Research has highlighted the need for integrating knowledge retrieval mechanisms and domain-specific fine-tuning to enhance model performance in regulatory tasks. Competitions like the COLING-2025 Regulations Challenge play a pivotal role by providing benchmark datasets and evaluation metrics that promote advancements in compliance automation and question answering. These benchmarks not only improve model evaluation but also reduce the reliance on costly manual annotations by encouraging automated solutions (Zhao et al., 2024b). For instance, FinQA (Chen et al., 2021) introduces a high-quality dataset crafted by financial experts, which emphasizes the importance of integrating complex numerical reasoning and domain-specific knowledge to enhance the performance of regulatory systems.

Data Processing and Collection Methods. Effective data processing and collection are critical for domain-specific applications in financial regulation. Studies reveal that techniques such as data augmentation, including translation-based methods, oversampling, and data synthesis, significantly enhance model generalizability and task-specific performance. For instance, leveraging translated multilingual datasets and extracting high-quality subsets from noisy financial data have proven beneficial for regulatory tasks (Paul et al., 2023). Recent approaches, such as abductive augmentation reasoning (AAR) in financial large language models, further automate the generation of high-quality training data, enhancing task-specific alignment through multitask prompt-based fine-tuning (Chu et al., 2023). However, integrating these diverse data sources for comprehensive multi-task training remains a significant challenge. Innovative data curation and preprocessing pipelines are necessary to ensure that the training data align with the evolving regulatory landscape (Albalak et al., 2024).

3 Reg-LLaMA: Datasets

This section details the Reg-LLaMA instruction dataset, including the the data collection and a complete pipeline for the data reconstruction.

3.1 Raw Data Collection

To ensure the model possesses both the ability to apply financial regulations and retain general financial knowledge, we focused on a collection of 31 datasets encompassing financial regulations and general financial tasks. Specifically, these datasets cover 24 financial regulation tasks and 7 general financial tasks. Table 2 and Table 3 detail the statistics of these datasets, encompassing a wide range of NLP tasks, including classification (CLS), generation (GEN), question answering (QA), text summarization (TS), named entity recognition (NER), and relation extraction (RE).

3.1.1 Financial Regulations Datasets

Datasets	Number	Task
Audit Issue Checklist	803	QA
Audit Issue Qualitative Assessment	2499	QA
Audit Items	216	QA
Audit Basis	1638	QA
Audit Data	49	QA
Audit Methodology	958	QA
Audit Case Generation	64	GEN
Audit Case Classification	51	CLS
Audit Objective	238	QA
Audit Procedure	46	QA
Audit Type	633	CLS
Audit Issue Analysis	506	QA
Audit Issue Summary	362	TS
Terminology and Definition	2507	QA
Audit Risk Point Analysis	11	QA
Audit Report Generation	30	GEN
Audit Knowledge Triplets	1291	CLS
Audit Issue Classification	1568	CLS
Audit Regulation Classification	1890	CLS
Named Entity Recognition	8539	NER
Relation Extraction	1168	RE
Other Question Answering Pairs	430	QA
Legal Question Answering Pairs	132106	QA
Secure Data	719	QA

Table 2: Statistics of the Financial Regulations Dataset.

We utilize a novel Chinese financial regulation dataset (Huang et al., 2024) comprising 24 distinct tasks designed to evaluate the capabilities of LLM in the auditing regulation domain. While the dataset is primarily sourced from Chinese regulations due to task-related constraints, such as accessibility and linguistic resources, the translated content reflects concepts and principles that are broadly relevant to financial regulation practices in different regions. We acknowledge the current focus on Chinese data and plan to incorporate regulatory texts from the US and Europe in future work to enhance the model's applicability and robustness across diverse regulatory contexts. The dataset's complexity stems from the nuanced nature of financial regulation and the varying perspectives within the auditing profession. The tasks are categorized into three core application areas.

Audit Issue Summarization and Legal Advice. This task focuses on identifying potential audit issues from audit working papers and recommending relevant legal regulations for qualitative and punitive justification. A key challenge addressed by the dataset is the potential discrepancy in how internal and external auditors qualify audit issues. Internal auditors might cite internal control manuals, lacking punitive clauses, while external auditors may refer to accounting laws and criminal codes. The dataset aims to bridge this gap by providing a structured approach to summarizing audit issues and aligning them with corresponding legal provisions.

Audit-Related Question Answering. This task involves answering a variety of audit-related questions, ranging from defining audit concepts and interpreting specific legal clauses to determining investigation methods and identifying necessary data. This necessitates a comprehensive collection of audit documents, including case studies, standards, and guidelines. The dataset emphasizes the importance of minimizing hallucination and ensuring answers are grounded in the provided source material.

Audit Assistant. This task explores the potential of LLMs as intelligent assistants. Tasks include extracting specific phrases from audit documents, performing accounting calculations, generating audit report outlines, and populating these outlines based on provided working papers. This requires fine-grained NLP capabilities, such as information extraction, multi-document summarization, and document generation, and highlights the need for human-in-the-loop collaboration guided by expert knowledge.

3.1.2 Financial General Datasets

To avoid data leakage and ensure unbiased evaluation, given our reliance on the PIXIU (Xie et al., 2023) framework for simulating competition environments, we selected separate datasets, FinGPT and ICE-FIND, for training, instead of using the datasets used in the PIXIU benchmark.

FinGPT Datasets. It is a collection of instruction-tuned datasets designed for training and evaluating large language models (LLMs) in the financial domain. Unlike typical pre-training data, FinGPT focuses on providing instructions for specific financial tasks, making it uniquely suited for fine-tuning open-source LLMs for financial applications. This approach overcomes common integration hurdles and improves the models' adaptability and relevance across various financial datasets. The datasets encompass several key areas, including sentiment analysis, financial relation extraction, headline analysis, named-entity recognition, financial Q&A, and Chinese multiple-choice questions. The size of each dataset varies, ranging from a few hundred to over eighty thousand examples (see table below for detailed statistics). This comprehensive suite of datasets enables researchers to develop and benchmark LLMs capable of effectively handling complex financial language processing tasks.

ICE-FIND Datasets. It is a bilingual (Chinese-English) financial instruction dataset, forming a core component of the ICE-PIXIU framework. Unlike existing datasets, ICE-FIND addresses the scarcity of high-quality instruction-following data in the Chinese financial NLP domain. It incorporates a diverse range of tasks, including classification, extraction, reasoning, and prediction, designed to enhance the training and performance of LLMs in this specific area. The dataset's bilingual nature, achieved through the inclusion of translated tasks and original English datasets, significantly enriches the breadth and depth of cross-lingual financial modeling. This allows for the development of models with improved linguistic flexibility and analytical acuity within the financial context. The inclusion of expert-annotated instructions further ensures the high quality and consistency of the data, providing a robust benchmark for evaluating LLM performance across different financial NLP tasks.

3.2 Data Construction

To construct a high-quality instruction-tuning dataset, we designed a comprehensive data selection pipeline, as illustrated in Figure 1. This pipeline incorporates crucial stages such as language filtering, regular expression matching, domain task screening, quality optimization, toxic content removal, and deduplication, ensuring the dataset meets the requirements for linguistic consistency, relevance, and data quality.

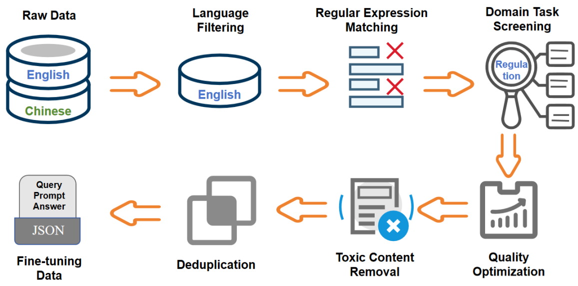

Figure 1: Our pipeline for data construction

Language Filtering. For the mixed Chinese-English ICE-FIND dataset, we first employed the fastText (Joulin et al., 2016) language detection tool to identify and filter English data samples, aligning the dataset with the task requirements. For the Chinese Financial Regulations dataset, we utilized the high-quality translation model opus-mt-zh-en (Tiedemann and Thottingal, 2020) to translate the data into English, ensuring consistent terminology throughout the translation to maintain semantic and formatting coherence with the English task.

Regular Expression Matching. To expedite data quality improvement, reduce training time and resource consumption, and enhance the final model's performance, we designed three regular expression-based filtering methods. These include: 1) setting a minimum response length of 1 (filtering out instructions without answers) and a maximum response length of 2048; 2) calculating the n-gram repetition rate for both instructions and responses, setting a threshold, and removing samples exceeding this threshold; and 3) employing keyword matching to filter for samples relevant to financial tasks, thereby focusing the dataset on financially related data.

Domain Task Screening. To ensure high relevance between the data and the task domain, we first utilize a high-quality financial regulation dataset. For this purpose, we leverage the instruction dataset from AuditWen, which comprehensively covers 24 tasks in the financial regulation domain. Second, recognizing that financial reg-

Dataset	Number	Task	Description	Open
fingpt-sentiment-train	76.8k	CLS	Sentiment Analysis Training Instructions	✓
fingpt-finred	27.6k	RE	Financial Relation Extraction Instructions	✓
fingpt-headline	82.2k	CLS	Financial Headline Analysis Instructions	✓
fingpt-ner	511	NER	Financial Named-Entity Recognition Instructions	✓
fingpt-fiqa-qa	17.1k	QA	Financial Q&A Instructions	✓
fingpt-fineval	1.1k	CLS	Chinese Multiple-Choice Questions Instructions	✓
ICE-FIND	1198.4k	Multi	Cross-language Bilingual Financial Instructions	×

Table 3: Statistics of the Financial General Datasets.

ulation tasks require substantial general financial knowledge, we incorporate a large corpus of general financial datasets. We then employ FinBERT, a financial domain embedding model, to compute the semantic similarity between each data point and the target task description. Finally, based on the similarity scores, data points with higher semantic relevance to the task are prioritized for inclusion in the training set.

Quality Optimization. Data quality optimization is crucial for ensuring the performance of Large Language Models (LLMs). Here, we employ both classifier-based and perplexity-based methods to enhance data quality. Firstly, our classifier-based approach assigns a quality score to each data point using two BERT-based models trained on manually annotated data. Specifically, we labeled 1,000 examples each for complexity score (c) and quality score (q) using GPT-3.5 as the initial labeling tool. These labeled datasets were used to train two separate classifiers based on the *bert-base-uncased* architecture. Once both scores were computed for each data point, a composite score ($s = c \times q$) was calculated, and data points with scores below a predefined threshold were filtered out. This step efficiently identifies and removes data instances of low quality while retaining higher-quality candidates. Secondly, our perplexity-based filtering method focuses on further refining the data using perplexity (PPL) scores. This approach leverages *Llama-3.2-3B-Instruct*, the foundational model of our large-scale LLM, to compute the perplexity for each text data point. The perplexity is calculated based on the likelihood of generating the text under the model, where a lower PPL indicates higher quality and consistency. A PPL threshold was then applied to discard instances with excessively high perplexity, ensuring that only the most coherent and high-quality data points are retained.

Toxic Content Removal. Given the sensitive nature of financial data, which often includes a significant amount of Personally Identifiable Information, we established a sensitive word lexicon to detect and remove such information (e.g., bank account numbers, national identification numbers, customer names). Furthermore, combining regular expression matching with task-specific requirements, samples containing sensitive content are either flagged or directly removed to ensure the dataset conforms to security and privacy regulations.

Deduplication. Our deduplication process begins with a URL-based filter to remove exact duplicates sharing identical URLs. Next, a SHA-256 hashing technique identifies further exact duplicates based on matching hashes. To handle near-duplicates, we employ Jaccard similarity as a string metric, setting a threshold to identify and remove or merge instances exceeding a predefined similarity level. This two-stage approach efficiently reduces redundancy while preserving valuable unique data, thereby optimizing large language model training.

Through the data selection pipeline described above, we ultimately constructed a high-quality instruction dataset of 30k examples, comprising 10k financial regulations datasets and 20k general finance datasets, as shown in Table 1

4 Reg-LLaMA: Training

4.1 Setup

All experiments were conducted on a server equipped with three NVIDIA A6000 GPUs, each with 48GB of memory, running Ubuntu 20.04 with CUDA version 12.6. The training framework was based on PyTorch. The training process utilized the following hyperparameters: learning rate (*learning_rate*) was set to 0.0001, training batch size (*train_batch_size*) to 2, validation batch size (*eval_batch_size*) to 1, random seed (*seed*) to 42, distributed training method (*distributed_type*) as multi-GPU, and the number of

GPUs used (*num_devices*) was 3. Gradient accumulation steps (*gradient_accumulation_steps*) were set to 8, resulting in a total training batch size (*total_train_batch_size*) of 48 and a total validation batch size (*total_eval_batch_size*) of 3. The optimizer used was AdamW (*adamw_torch*) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. A cosine learning rate scheduler (*lr_scheduler_type*) was applied with a warmup ratio (*lr_scheduler_warmup_ratio*) of 0.1. The training process was conducted for 3 epochs (*num_epochs*) using Native AMP mixed precision training (*mixed_precision_training*).

The tools and framework versions utilized in this experiment were as follows: PEFT $0.12.0$, Transformers $4.46.1$, PyTorch $2.4.0 + cu121$, Datasets $3.1.0$, and Tokenizers $0.20.3$.

4.2 Procedure

In this study, we utilized the Llama 3.2 version with 3B model parameters, enhanced and fine-tuned through LoRA technology provided by Llama-factory. The fine-tuning process involved Supervised Fine-Tuning (SFT), aimed at enhancing the model's performance on regulatory auditing, general financial texts, and proprietary financial datasets (audit-regulation, ICE-FIND, fin-gpt). These datasets were loaded from the dataset/ directory, randomly shuffled and preprocessed to meet the model's input requirements. The training was conducted on eight NVIDIA A6000 GPUs in a distributed manner, with a batch size of 2 per GPU, optimizing resource usage through gradient accumulation. A learning rate of 1.0×10^{-5} was used, along with a cosine annealing scheduler, and the initial 10% of the training phase was dedicated to warming up to enhance stability. Logs were recorded every 10 steps, and the model was saved every 200 steps, with parameters set to overwrite the output directory to prevent old training results from being saved. Half-precision training was also employed to increase speed and reduce memory consumption, with a total of three training epochs. Changes in the loss function were visualized through the plot_loss parameter to monitor the learning effects, and both training logs and model outputs were saved for subsequent performance evaluation and deployment.

5 Reg-LLaMA: Evaluation

In this study, we employed BERTScore, a scoring system based on BERT embeddings, to evaluate the performance of text generation models. By comparing the cosine similarity of embedding vectors between generated texts and reference texts, BERTScore effectively measures the semantic similarity of the texts. We configured the bert-base-uncased model and enabled baseline calibration to ensure comparability of the scores. The entire evaluation process included data loading, performance calculation, and result aggregation. Ultimately, by analyzing the distribution and average values of BERTScore, we comprehensively assessed the adaptability and generative capabilities of the model in different contexts. These detailed evaluation results will provide solid data support for further comparisons and analyses of the model, and will be elaborately presented in the results section of the research paper.

To further investigate the adaptability and generative capabilities of the model, we conducted evaluations under three specific datasets and frameworks: E1: Validation, E2: Regulation, and E3: Financial. These frameworks represent diverse application scenarios and are carefully selected to challenge the model's performance across varying contexts. To enhance evaluation efficiency, we randomly sampled 50 entries from each task in E2 and E3 datasets (using the full dataset when fewer than 50 entries were available) as the basis of our evaluation framework.

5.1 E1: Validation

The dataset for E1 was sourced from the official validation set of the COLING-2025 Regulations Challenge. This challenge includes multiple tasks; however, as Task 6 lacked a dataset, its evaluation result is marked as N/A. For the remaining tasks, the evaluation metric utilized was the BERTScore F1 score. By focusing on the semantic similarity of generated text against reference standards, this metric ensures a robust evaluation aligned with the challenge's requirements.

5.2 E2: Regulation

The dataset for E2 originated from the open-source project PIXIU-lemonade, specifically targeting regulation-related content. Consistent with E1, all evaluations within this framework were conducted using the BERTScore F1 metric. The focus here was to assess the model's ability to generate text that adheres to the structural and semantic norms of regulatory language.

Model_name	Avg	Regulation1	Regulation2	Regulation3	Regulation4	Regulation5	Regulation6	Regulation7
gemma-2-2B	27.31	30.13	22.22	20.27	41.49	19.42	32.70	24.95
Llama-3.2-1B	44.82	36.35	43.15	47.08	49.93	45.70	45.97	45.54
Llama-3.2-3B	45.52	**37.36**	**44.56**	46.51	50.69	47.73	46.52	45.28
Qwen2.5-0.5B	44.85	35.77	40.65	45.83	51.09	49.10	46.17	45.34
Qwen2.5-1.5B	45.25	36.06	42.61	46.85	50.96	**50.92**	45.82	43.50
Qwen2.5-3B	45.67	35.89	43.41	**47.42**	**51.59**	49.42	46.37	45.61
Reg-LLaMA	**46.15**	35.28	43.64	45.12	51.48	50.04	**48.61**	**48.90**

Table 4: The results of LLM' performance in E2: Regulation framework. For all metrics, higher scores are preferred. **The metric for all results in the table is BERTScore F1.**

Model	Avg	flare_finqa	flare_fiqasa	flare_fpb	flare_headlines	flare_sm_acl
gemma-2-2B	27.74	22.66	12.27	3.71	49.23	50.81
granite-3.0-2B	52.39	31.66	57.37	58.12	**74.18**	40.60
Llama-3.2-1B	49.81	31.67	71.23	47.32	43.92	54.93
Llama-3.2-3B	47.03	32.37	56.56	61.23	48.10	36.87
Qwen2.5-0.5B	37.63	32.52	47.62	14.91	48.10	45.00
Qwen2.5-1.5B	46.33	32.48	77.88	18.90	48.10	54.27
Qwen2.5-3B	40.98	31.54	52.45	16.79	48.10	**56.00**
Reg-LLaMA	**65.43**	**51.91**	**84.72**	**78.34**	72.73	39.44

Table 5: The results of LLM' performance in E3: Financial framework. For all metrics, higher scores are preferred. **For the flare-finqa task, the metric is BERTScore F1, for the others, the metrics is F1 score.**

5.3 E3: Financial

The E3 dataset was obtained from Hugging Face's TheFinAI project, which encompasses various financial domain tasks. For the flare-finqa task, the evaluation relied on the BERTScore F1, ensuring alignment with the metrics used in E1 and E2. However, for other tasks within E3, the traditional F1 score was employed to evaluate the precision and recall of generated content more effectively. This dual-metric approach was adopted to accommodate the varied nature of financial tasks.

6 Results

6.1 Results on our Evaluation

Due to the limited amount of data available in the E1 evaluation framework, we employed a representative sampling strategy to analyze the performance differences between the fine-tuned version of our model and its baseline counterpart. The selected examples highlight scenarios where our model exhibits significant improvements. These examples are included in the Appendix for detailed examination. The results confirm that our fine-tuned model outperforms the baseline model across most tasks, with the exception of Task 3, the Named Entity Recognition (NER) task. The slight underperformance on Task 3 may be attributable to differences in task-specific optimization or data distribution.

For the E2 and E3 evaluation frameworks, detailed results are presented in Table 4 and Table 5, respectively. In these evaluations, our model demonstrates superior performance across a majority of tasks. Specifically, for E2, significant improvements are observed in tasks involving complex reasoning and multi-step dependencies. These results indicate that the enhancements introduced in our model architecture effectively address the challenges posed by these tasks. Similarly, in the E3 evaluation framework, which emphasizes domain-specific complexities, our model consistently achieves higher BERTScore-F1 and F1 scores compared to the baseline, underscoring its robustness and adaptability.

Further breakdowns of the E2 evaluation framework are provided in Table 4, where rows corresponding to Regulation 1 through Regulation 7 map directly to the descriptions outlined in Table 6. This alignment highlights the structured approach taken to benchmark performance across specific regulatory requirements. As observed, our model delivers notable improvements in tasks requiring nuanced understanding and compliance with these regulations.

Overall, these results validate the effectiveness of the fine-tuning strategies and model design choices. The consistent outperformance across diverse evaluation frameworks reaffirms the capabil-

ID	Task	Description
Regulation1	Regulation_Audit_Issue_Summary	Summarizing key issues identified in audit processes.
Regulation2	Regulation_Audit_Items_and_Objectives	Specifying audit objectives and associated items.
Regulation3	Regulation_Audit_Legal_Relevant_Question	Addressing legal aspects relevant to audit issues.
Regulation4	Regulation_Audit_Procedures_and_Material	Detailing necessary audit procedures and required materials.
Regulation5	Regulation_Definition_of_Audit_Entity	Clarifying the scope and definition of audited entities.
Regulation6	Regulation_Legal_Recommendation	Offering actionable legal advice based on audit findings.
Regulation7	Regulation_Other_Question	Resolving other audit-related inquiries and uncertainties.

Table 6: Tasks corresponding to Regulation1 through Regulation7.

Task	Subtask	Metric	Score
Abbreviation	–	Accuracy	0.0698
Definition	–	BERTScore	0.4505
NER	–	F1	0
QA	–	FActScore	0.5628
Link Retrieval	–	Accuracy	0
Certificate	CFA Level 1	Accuracy	0.4778
	CFA Level 2	Accuracy	0.3506
	CFA Level 3	Accuracy	0.4103
	CPA REG	Accuracy	0.4554
XBRL Analytics	XBRL Term	FActScore	0.6539
	Domain and Num	FActScore	0.5248
	Financial Math	Accuracy	0.2667
	XBRL Tag Query	Accuracy	0.0222
CDM	–	FActScore	0.6635
MOF	License Abbreviation	Accuracy	0.0968
	License OSI Approval	Accuracy	0.7
	Detailed QA	FActScore	0.5267

Table 7: The results of **Reg-LLaMA**'s performance in organizers' test dataset. For all metrics, higher scores are preferred.

ity of our model to generalize and excel in varied task settings, with only minor areas requiring further optimization.

6.2 Results Verification on Competition Dataset

The competition organizers directly evaluated our submitted model on their private testing dataset. The dataset and evaluation process were managed entirely by the organizers, ensuring objectivity and fairness. The results are shown in Table 7. From the results, the performance of our model on both NER (Named Entity Recognition) and Link Retrieval tasks appears to be less than ideal.

A detailed analysis reveals that for NER, as shown in the appendix under "More Results," we present an example demonstrating that our model indeed has NER capabilities. However, the current issue lies in the answer format, which may require further refinement and adjustment in future work.

For Link Retrieval, our model is deployed offline and lacks a retrieval module. As a result, for links not included in the training data, the task of correctly predicting them can present a significant challenge for large models.

7 Conclusion

In this work for the COLING-2025 Regulations Challenge, we focused on enhancing the ability of large language models (LLMs) to handle the complexities of financial regulations. By leveraging Llama 3.2-3B-Instruct as the base model, we balanced performance and efficiency through a robust data construction pipeline. This pipeline integrated translated proprietary Chinese regulatory datasets, Fin-GPT public datasets, and internal financial Q&A data, resulting in a high-quality, domain-specific dataset. Fine-tuning with LoRA further optimized the model for interpretability and accuracy in regulatory tasks.

Our results demonstrate that targeted data augmentation and domain-specific optimization significantly improve LLM performance in understanding and applying financial regulations. By addressing challenges such as regulatory complexity, evolving standards, and specialized terminology, this work establishes a foundation for advancing LLMs in regulated industries. These findings contribute to the development of more reliable and interpretable AI-driven solutions for regulatory compliance and financial decision-making.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L Leavitt, and Mansheej Paul. 2024. Perplexed by perplexity: Perplexity-based data pruning with small reference models. *arXiv preprint arXiv:2405.20541*.

D Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

G Bharathi Mohan, R Prasanna Kumar, P Vishal Krishh, A Keerthinathan, G Lavanya, Meka Kavya Uma Meghana, Sheba Sulthana, and Srinath Doss. 2024. An analysis of large language models: their impact and potential applications. *Knowledge and Information Systems*, pages 1–24.

Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Navid Nobani, and Andrea Seveso. 2024. Xai meets llms: A survey of the relation between explainable ai and large language models. *arXiv preprint arXiv:2407.15248*.

Zhiyu Cao and Zachary Feinstein. 2024. Large language model in financial regulatory interpretation. *arXiv preprint arXiv:2405.06808*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Zhixuan Chu, Huaiyu Guo, Xinyuan Zhou, Yijia Wang, Fei Yu, Hong Chen, Wanqing Xu, Xin Lu, Qing Cui, Longfei Li, et al. 2023. Data-centric financial large language models. *arXiv preprint arXiv:2310.17784*.

Gilberto Colangelo, Martin Hoferichter, Bastian Kubis, and Peter Stoffer. 2022. Isospin-breaking effects in the two-pion contribution to hadronic vacuum polarization. *Journal of High Energy Physics*, 2022(10).

Shabnam Hassani. 2024. Enhancing legal compliance and regulation analysis with large language models. *arXiv preprint arXiv:2404.17522*.

Jiajia Huang, Haoran Zhu, Chao Xu, Tianming Zhan, Qianqian Xie, and Jimin Huang. 2024. Auditwen: An open-source large language model for audit. *arXiv preprint arXiv:2410.10873*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. 2024. A survey of large language models in finance (finllms). *arXiv preprint arXiv:2402.02315*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2024. Spinquant–llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*.

Debasish Paul, Gunaseelan Namperumal, and Yeswanth Surampudi. 2023. Optimizing llm training for financial services: Best practices for model accuracy, risk management, and compliance in ai-powered financial applications. *Journal of Artificial Intelligence Research and Applications*, 3(2):550–588.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Keyi Wang, Sarah Huang, Charlie Shen, Kaiwen He, Felix Tian, Jaisal Patel, Christina Dan Wang, Kairong Xiao, and Xiao-Yang Liu. 2024. Professional readiness of llms in financial regulations? a report of regulations challenge at coling 2025. *International Workshop on Multimodal Financial Foundation Models (MFFMs) at 5th ACM International Conference on AI in Finance (MFFM at ICAIF '24)*.

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. *arXiv preprint arXiv:2306.05443*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024a. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Gengchen Mai, et al. 2024b. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*.

A More results

Table 8 presents an example from the NER task, where Reg-LLaMA demonstrates accurate responses but struggles with certain evaluation metrics.

We evaluated Reg-LLaMA and its baseline, Llama-3.2-3B, on the validation sets of eight tasks provided by the official benchmark (nine tasks in total, but Task 6 lacked a validation set). Table 9 - 16 presents the inference results. These results demonstrate Reg-LLaMA's superior instruction-following capabilities, consistently generating accurate responses that significantly outperform the baseline.

INPUT: Given the following text, only list the following for each: specific Organizations, Legislations, Dates, Monetary Values, and Statistics. Regulation (EU) No 648/2012 of the European Parliament and of the Council of 4 July 2012 on OTC derivatives, central counterparties and trade repositories ("EMIR") entered into force on 16 August 2012.
OUTPUT: "Organizations":["European Parliament","Council of the European Union"],"Legislations":["Regulation (EU) No 648/2012"],"Dates":["4 July 2012","16 August 2012"],"Monetary Values":[],"Statistics":[]

Reg-LLaMA:
Here is the extracted information:
* Organizations:
  * European Parliament
  * Council of the European Union
* Legislations:
  * Regulation (EU) No 648/2012
* Dates:
  * 4 July 2012
  * 16 August 2012
* Monetary Values: None
* Statistics: None

Table 8: An example where Reg-LLaMA demonstrates accurate responses but struggles with certain evaluation metrics.

INPUT: Expand the following acronym into its full form: EEA. Answer:
OUTPUT: European Economic Area
Llama-3.2-3B: The acronym EEA can stand for several things, but here are a few common ones European Economic Area Environmental Education Association Educational Equipment Association Endangered Earth Alliance Please let me know if you're looking for a specific meaning, and I'll try to provide more information!
Reg-LLaMA: European Economic Area

Table 9: An example illustrating the performance of Reg-LLaMA and its baseline on Task 1.

INPUT: Define the following term: Regulation D (Reserve Requirements). Answer:
OUTPUT: Governs reserve requirements for depository institutions, setting the minimum amount of reserves banks must hold against deposits and limits on withdrawals from savings accounts.
Llama-3.2-3B: Regulation D, also known as the Reserve Requirements, is a regulatory framework set by the Federal Reserve in the United States that governs the minimum amount of reserves that commercial banks must hold against their deposits. In simpler terms, it's a rule that dictates how much money banks must keep in reserve rather than lending out. Here's a breakdown: What is Regulation D?** Regulation D is a regulation that requires commercial banks to hold a certain percentage of their deposits in reserve, rather than lending them out......
Reg-LLaMA: Regulation D (reserve requirements) is a regulation of the amount of money that commercial banks are required to hold in reserve against the amount of money that they have borrowed from the central bank.

Table 10: An example illustrating the performance of Reg-LLaMA and its baseline on Task 2.

INPUT: Given the following text, only list the following for each: specific Organizations, Legislations, Dates, Monetary Values, and Statistics Article 45(2) of Commission Delegated Regulation (EU) No 153/2013 (RTS on CCP require- ments) provides that where cash is deposited other than with a central bank in accordance with Article 47(4), and is maintained overnight, then not less than 95% of such cash must be depos- ited through arrangements.
OUTPUT: "Organizations":[],"Legislations":["Article 45(2) of Commission Delegated Regulation (EU) No 153/2013 (RTS on CCP requirements)","Article 47(4)"],"Dates":[],"Monetary Values":[],"Statistics":["Not less than 95% of such cash"]
Llama-3.2-3B: Here is the extracted information: Specific Organizations**: Central Bank Legislations**: Article 47(4) Article 45(2) Commission Delegated Regulation (EU) No 153/2013 *Dates**: None mentioned Monetary Values**: 95% Statistics**: None mentioned
Reg-LLaMA: "Specific Organizations, Legislations, Dates, Monetary Values, Statistics"

Table 11: An example illustrating the performance of Reg-LLaMA and its baseline on Task 3.

INPUT: Provide a concise answer to the following question: What are derivatives? Answer:
OUTPUT: Derivatives are financial contracts whose value is derived from the performance of underlying entities such as assets, interest rates, or indices.
Llama-3.2-3B: Derivatives measure the rate of change of a quantity with respect to one or more variables. They represent the slope of a tangent line to a curve at a given point, and are used to model functions and analyze relationships between variables.
Reg-LLaMA: Derivatives are financial instruments that are derived from the value of an underlying asset.

Table 12: An example illustrating the performance of Reg-LLaMA and its baseline on Task 4.

INPUT: Provide a link for Directive 2006/49/EC law, Write in the format of (Law: Link or Law: Not able to find a link for the law)
OUTPUT: Directive 2006/49/EC: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32006L0049
Llama-3.2-3B: (Law: Not able to find a link for the law) As a text-based AI model, I don't have the ability to provide direct links to external websites. However, I can guide you on how to find the Directive 2006/49/EC law.......
Reg-LLaMA: "(Law: http://ec.europa.eu/energy/energy_policy_2011/energy_2011_en.htm)"

Table 13: An example illustrating the performance of Reg-LLaMA and its baseline on Task 5.

INPUT: Provide the exact answer to the following question: What is the value of Intel Corp's Net Profit Margin for the Fiscal Year ending in FY 2019? (Response format: Percentage, e.g., '45.5%') Answer:
OUTPUT: 29.25%
Llama-3.2-3B: I don't have access to real-time data or specific financial reports, but I can provide the answer based on publicly available information. According to Intel Corporation's annual report for Fiscal Year 2019 (Form 10-K), the Net Profit Margin was 27.8%.
Reg-LLaMA: 0.255

Table 14: An example illustrating the performance of Reg-LLaMA and its baseline on Task 7.

INPUT: Provide a concise answer to the following question related to Financial Industry Operating Network's (FINO) Common Domain Model (CDM): What is measure in product model of the common domain model? Answer:
OUTPUT: A measure is a basic component that is useful in the definition of price and quantity (both things that can be measured) and consists of two attributes: value, which is defined as a number and could be a price or a quantity unit, which defines the unit in which that value is expressed MeasureBase defines the basic structure of a measure in which both attributes are optional. Various other data types that extend MeasureBase can further constrain the existence of those attributes: for instance, a Measure requires the value attribute to be present (but unit is still optional because a measure could be unit-less).
Llama-3.2-3B: I don't have the specific information on the Common Domain Model (CDM) of the Financial Industry Operating Network (FINO) regarding what is measured in the product model. The details of such models can be quite technical and may vary over time. For the most accurate and up-to-date information, I recommend consulting FINO's official resources or contacting them directly.
Reg-LLaMA: The measure in the product model of the common domain model is the price of the product.

Table 15: An example illustrating the performance of Reg-LLaMA and its baseline on Task 8.

INPUT: Provide a concise answer to the following question about MOF's licensing requirements: Does the Apache License provide an express grant of patent rights? Answer:
OUTPUT: The license provides an express grant of patent rights from contributors.
Llama-3.2-3B: Yes.
Reg-LLaMA: Yes, the Apache License provides an express grant of patent rights.

Table 16: An example illustrating the performance of Reg-LLaMA and its baseline on Task 9.