FinMind-Y-Me at the Regulations Challenge Task: Financial Mind Your Meaning based on THaLLE

Pantid Chantangphol* , **Pornchanan Balee*** , **Kantapong Sucharitpongpan*** ,
Chanatip Saetia and **Tawunrat Chalothorn**
Kasikorn Labs Co., Ltd., Kasikorn Business-Technology Group, Thailand
{pantid.c,pornchanan.b,kantapong.s,chanatip.s,tawunrat.c}@kbtg.tech
*These authors contributed equally to this work.

Abstract

This paper presents our submission to the COLING 2025 regulation challenge, focusing on nine tasks in the regulatory and financial domains. The challenge aims to advance large language models beyond general-purpose capabilities, adapting them for regulatory and financial tasks using a unified framework of task-specific prompts and input templates. We propose a sequential fine-tuning approach that integrates reasoning-based training, tailored system prompts, and Chain-of-Thought (CoT) inference to optimize task-specific performance. This method improves accuracy and reliability across diverse tasks. Notably, CoT inference demonstrates exceptional effectiveness in handling complex scenarios and tasks requiring specific answer patterns, such as named entity recognition and financial calculations. Our model achieved an overall score of 54.801%, ranking 1st among all teams and becoming the top performer in the challenge. These results highlight the effectiveness of sequential fine-tuning, advanced reasoning techniques, and fine-tuned prompts in improving performance and scalability for complex regulatory and financial applications.

1 Introduction

The COLING 2025 regulations challenge is a rigorous initiative designed to advance the capabilities of large language models (LLMs) in understanding and processing complex regulatory and financial documents. This challenge comprises nine carefully crafted tasks that target critical aspects of regulatory text comprehension and practical application, such as deciphering domain-specific acronyms, extracting definitions, identifying named entities, answering intricate regulatory queries, and performing advanced analytics on financial filings. While LLMs such as GPT (Achiam et al., 2023), Llama (Touvron et al., 2023), Gemini (Reid et al., 2024), and Qwen (Bai et al., 2023) have demonstrated remarkable versatility across general natural language processing tasks, they often falter in specialized domains such as regulation and finance. These fields demand deep reasoning, multistep problem-solving, and precise contextual understanding—capabilities that traditional LLMs, optimized for straightforward, one-step responses, frequently lack. Furthermore, their propensity to hallucinations exacerbates their limitations, particularly when confronted with tasks involving complex calculations, nuanced regulatory language, or sophisticated financial analyses.

This paper presents a novel framework that enables a single LLM to effectively manage multitasking across various regulatory and financial domains. The framework addresses a range of specialized tasks, These tasks collectively enable the model to navigate the complexities of regulatory and financial domains. Collectively, these tasks require the model to demonstrate both the knowledge and capabilities needed to navigate the complexities of regulatory and financial domains, and each task demands precise management of domain-specific contexts and information.

Our approach integrates Unified Modeling (Zha et al., 2023) with Task-Specific Prompts (Zhou et al., 2022; Zhang et al., 2023) and Input Templates (Kojima et al., 2022), tailoring the focus and contextual comprehension of the model for each task to ensure coherent and relevant responses to regulatory and financial challenges. To optimize the learning and performance of the model, we employ Sequential Fine-Tuning (Lialin et al., 2023), where the model is progressively trained on tasks in a specific sequence. This approach leverages prior knowledge while minimizing the risk of catastrophic forgetting. To enhance the model's reasoning capabilities, we introduce Reasoning-Based Training, which enables more logical analysis and interpretation of complex datasets by leveraging prior reasoning. During inference, we utilize

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 349–362
January 19–20, 2025. ©2025 Association for Computational Linguistics

Chain of Thought (CoT) prompting (Wang et al., 2022), which guides the model through a step-by-step logical reasoning process. This method breaks down complex queries into manageable components, ensuring accurate and contextually relevant responses.

By integrating these techniques, our approach significantly improves the performance of LLMs in handling regulatory and financial tasks, surpassing traditional direct-response methods. This contribution advances LLMs for specialized applications, opening new avenues for LLMs in complex and regulated environments. Building on this foundation, the main contributions of this paper are as follows:

1. A unified framework for adapting a single LLM to multitask effectively across diverse regulatory and financial domains.

2. Integration of Task-Specific Prompts and Input Templates within a unified model, ensuring coherent, contextually relevant, and task-oriented responses.

3. Implementation of Sequential Fine-Tuning, where the model is trained progressively on tasks in a defined sequence, leveraging prior knowledge while mitigating catastrophic forgetting.

4. Introduction of Reasoning-Based Training to enhance the capability of model to logically analyze and interpret complex datasets.

5. Application of CoT prompting during inference to guide the model through step-by-step logical reasoning, resulting in more accurate and contextually aligned outputs.

The remainder of the paper is organized as follows: Section 3 discusses related works; Section 4 presents the methodology; Section 5 outlines the experimental setup; Section 6 details the results; Section 7 addresses the limitations; and Section 8 concludes the paper.

2 Task overview

The COLING 2025 Regulations Challenge comprises nine complex tasks aimed at evaluating diverse skills required for processing regulatory and financial texts. The Abbreviation Recognition Task tests a model's ability to identify and expand acronyms prevalent in regulatory documents, emphasizing domain-specific terminology understanding. The Definition Recognition Task involves accurately extracting definitions from dense legal and financial texts, demanding precise contextual comprehension. The Named Entity Recognition (NER) Task focuses on identifying and categorizing entities such as organizations, laws, dates, and monetary values, requiring high accuracy in structured data extraction. The Question Answering Task challenges models to provide precise answers to intricate legal questions, testing their ability to interpret both explicit and implicit content. The Link Retrieval Task assesses models' efficiency in locating specific legal documents, necessitating adept navigation through extensive regulatory corpora. The Certificate Question Task evaluates the capability of LLMs to solve multiple-choice questions from professional financial certification exams, such as the Chartered Financial Analyst (CFA) and Certified Public Accountant (CPA) exams, highlighting their analytical proficiency in meeting global certification standards and achieving examination success. The XBRL Analytics Task examines a model's ability to extract and analyze financial data from eXtensible Business Reporting Language (XBRL) filings, showcasing technical expertise in handling financial data formats. The Common Domain Model (CDM) Task focuses on understanding the Fintech Open Source Foundation's standards for financial industry interoperability. Lastly, the Model Openness Framework (MOF) Licenses Task evaluates models on licensing requirements, emphasizing regulatory compliance understanding. Collectively, these tasks represent a rigorous challenge, demanding advanced linguistic, analytical, and reasoning skills.

3 Related Work

3.1 Task-Specific Prompts

The prompt engineering (Mizrahi et al., 2023) has emerged as a critical skill for effectively utilizing LLMs. By providing structured instructions, prompts guide LLMs to adhere to predefined rules and align with specific task requirements (White et al., 2023). Recent studies (Zheng et al., 2024) emphasize the importance of designing prompts that are tailored to the nuances of each task. This task-specific prompt engineering approach enables models to focus on task-relevant features, resulting in enhanced performance on the given tasks.

3.2 Chain of Thought prompting

The CoT prompting (Wang et al., 2023) refers to the sequence of intermediate natural language reasoning steps that lead to the final output. Chain-of-thought prompting (Wei et al., 2022) enhances the reasoning capabilities of LLMs. Not only does it facilitate reasoning explanations, but it also enables sequential thinking, resulting in more natural and coherent answers. Experimental results (Wei et al., 2022) show that CoT prompting improves performance across various arithmetic, commonsense, and symbolic reasoning tasks. Moreover, this prompting approach requires only a small training dataset, learning effectively from just a few examples. This work (Wei et al., 2022) demonstrates the exceptional ability of CoT prompting to handle a variety of tasks.

3.3 Fine-Tuning LLMs techniques

Fine-tuning LLMs focusing on adapting pre-trained models to specific downstream tasks. Traditional full fine-tuning approaches, as demonstrated in GPT-3 (Brown et al., 2020), involve updating all model parameters, enabling high task performance but at significant computational and memory costs. To address these limitations, Parameter-Efficient Fine-Tuning (PEFT) methods have emerged, such as adapters (Hu et al., 2023; Liu et al., 2022), which optimize only a small subset of parameters while keeping the majority of the pre-trained weights frozen. Among these, Low-Rank Adaptation (LoRA) (Hu et al., 2021) has gained prominence for its ability to achieve competitive performance by training low-rank matrices added to frozen weight layers, significantly reducing memory and compute requirements. These techniques collectively highlight the trade-offs between resource efficiency and performance, driving advancements in scalable fine-tuning for large-scale models.

4 Methodology

Our methodology leverages four complementary strategies to enhance LLMs for regulatory and financial tasks: sequential fine-tuning to gradually build domain knowledge, task-specific prompts to align inputs and outputs effectively, reasoning-based training to improve logical problem-solving, and chain-of-thought prompting to ensure precise, template-aligned answers through structured reasoning.

4.1 Sequential Fine-Tuning

Group	Domain	Task	Training size	Metrics
Group 1	XBRL	Financial Math	222	Accuracy
Group 2	CDM	All Required	2,414	Factscore
Group 3	MOF	Detailed QA	424	Factscore
Group 4	Definition	All Required	1,720	BERTscore
	XBRL Term	XBRL Terminology	143	Factscore
Group 5	QA	All Required	1,349	Factscore
Group 6	XBRL	XBRL Tag Query	7,209	Accuracy
Group 7	NER	EMIR	1,905	F1score
Group 8	CFA	CFA Level 1	1,032	Accuracy
Group 9	MOF	License Abbreviations	240	Accuracy
Group 10	Abbreviation	EMIR	210	Accuracy
Group 11	Abbreviation	Stock Tickers (NYSE)	8,320	Accuracy
Group 12	Link-Retrieval	All Required	460	Accuracy

Table 1: Sequence of tasks in sequential fine-tuning

Sequential fine-tuning is a strategic approach that incrementally enhances a capability of LLMs by adapting it to a series of tasks in a predefined order. This method builds on knowledge from earlier tasks to improve performance on subsequent tasks, enabling a comprehensive understanding of complex domains such as regulation and finance. In our framework, tasks are grouped by domain relevance and complexity.

As outlined in Table 1, The nine regulatory tasks were organized into 12 groups based on evaluation metrics, domain-specific importance, and functional characteristics. Tasks within the same domain but evaluated using different metrics, such as XBRL Tag Query and XBRL Financial Math, were assigned to separate groups. Conversely, tasks from distinct domains with similar functional attributes, such as XBRL Terminology and Definition Tasks, were grouped together.

The sequence of tasks for sequential fine-tuning was carefully organized based on the specificity of the data and the type of responses required. The process began with foundational tasks, such as Financial Math, to build a strong base of knowledge. Even though these tasks required precise answers, the responses followed clear patterns of calculation and reasoning. Subsequently, specialized tasks were prioritized for fine-tuning based on their generalizability, the adaptability of evaluation metrics (e.g., BERTScore and FactScore), and training dataset size. For instance, question-answering tasks in the CDM and MOF domains, which are more specialized, were fine-tuned next. The responses for these tasks could take various forms, offering flexibility in how they were answered. Evaluation metrics such as FactScore were used to assess their effectiveness and ensure adaptability. After that, tasks requiring more specific and precise responses, such as those within the Definition domain, were addressed. These tasks involved generating detailed descriptions where precise word choice was crucial. BERTScore was employed to ensure accuracy and prevent unintended changes to the intended meaning. Finally, tasks demanding highly specific responses and significant memorization, such as abbreviation retrieval and link retrieval, were fine-tuned in the final stages. These tasks relied on explicit recall and often involved retrieving responses directly from specialized datasets.

By layering learning in a systematic sequence, the model achieves robust supervised fine-tuning while addressing challenges such as imbalanced datasets and task-specific skill demands, including calculation, analysis, and memorization. This approach enables insights gained from simpler tasks to inform and enhance solutions for more advanced challenges

4.2 Unified Modeling with Task-Specific Prompts and Input Template

This approach integrates multiple regulatory tasks into a cohesive model framework. Using task-specific prompts and input templates ensures that each task is addressed with a focused contextual understanding. These prompts serve as tailored instructions, guiding the model in interpreting inputs and generating accurate responses. This structured design enables the model to handle diverse regulatory tasks efficiently while maintaining consistency and coherence. Table 7 details the tasks and their corresponding prompts. Each prompt is designed to meet the specific requirements of its task, ensuring precise and reliable output. This unified framework combines task-specific customization with a scalable and adaptable architecture, making it suitable for various regulatory domains.

4.3 Reasoning-Based Training

Reasoning-based training enhances the ability of LLMs to analyze and interpret complex regulatory data by integrating logical reasoning into the training process, as demonstrated in Table 8. This approach departs from traditional methods that rely solely on the final answer as the labeled response, instead prioritizing the reasoning process during training. By focusing on problem-solving steps, it fosters a more nuanced understanding of financial and regulatory content, enabling the generation of accurate and contextually relevant responses. Table 8 provides illustrative examples of training data, contrasting reasoning-based and final-answer-focused approaches in financial and regulatory tasks. Each question is accompanied by a step-by-step explanation of the reasoning process, offering clarity and structure. This systematic approach enables models to decompose complex tasks into transparent and reliable steps, thereby enhancing their interpretability and trustworthiness.

4.4 Chain of Thought Prompting in Inference

CoT prompting enables models to generate responses through a step-by-step logical progression during inference, breaking down complex queries into manageable parts rather than relying solely on a single system prompt. The CoT methodology in this work, as detailed in , comprises two key steps to ensure structured and precise reasoning. First, a task-specific system prompt, guides the model to decompose complex queries into logical, sequential components, establishing a clear framework for logical analysis and problem-solving. Second, a refinement prompt captures the exact context of the query and specifies the desired answer pattern. Logical coherence is verified at each step, ensuring that reasoning remains accurate and well-structured. The final response is generated after confirming logical correctness and alignment with task-specific requirements. This two-step CoT process ensures accuracy and delivers well-structured, reasoned answers, especially for tasks involving regulatory analysis, complex decision-making, or multi-faceted data interpretation.

5 Experiment setup

5.1 Model selection

Task	Metrics	Llama3.1-ins	Qwen2.5-ins	THaLLE0.1
Abbreviation (Ticker)	R1	1.658	1.323	5.051
Abbreviation (Acronym)	R1	29.070	32.298	51.810
Definition	BERT-R	83.950	85.633	86.077
NER	BERT-R	31.434	76.113	68.290
QA	BERT-R	86.119	85.700	85.692
Link Retrieval	Acc	6.533	27.814	21.847
CFA Level 1	Acc	58.624	67.966	66.860
XBRL (Terminology)	R1	82.540	80.599	82.218
XBRL (Domain-Numeric Query)	R1	81.464	79.713	80.421
XBRL (Financial Math)	R1	0.813	1.276	0.743
XBRL (Tag Query)	R1	12.573	79.254	57.143
CDM	BERT-R	81.921	81.465	81.976
MOF (License OSI Approval)	Acc	0.000	0.000	0.000
MOF (Detailed QA)	BERT-R	89.128	87.476	86.854
MOF (License Abbreviation)	BERT-R	14.306	9.607	12.118
Overall	**Overall**	**49.347**	**58.162**	**58.113**

Table 2: Model performance Comparison (%)

To evaluate performance for model selection, we compared the Qwen2.5-7B-Instruct[1] (Team,

[1] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

2024; Yang et al., 2024) model with Llama-3.1-8B-Instruct[2] and THaLLE-0.1-7B-fa[3] (Labs et al., 2024) across multiple tasks. Table 2 presents a detailed comparison, highlighting the competitive performance of Qwen2.5-7B-Instruct, particularly in reasoning and domain-specific tasks. Its balanced architecture, with 7 billion parameters, effectively handles complex tasks while remaining computationally efficient. Based on its superior performance and the optimal balance between size and capability, we selected Qwen2.5-7B-Instruct as the base model for fine-tuning across various regulatory tasks.

5.2 Metrics

This study evaluates LLM performance across nine regulatory tasks using specific metrics. the experiment 5.1, the experiment 6.1 and the experiment 6.2 assess tasks as follows: Link Retrieval, MOF License OSI Approval, and CFA are evaluated using mean Accuracy (Acc) ; Abbreviation Recognition and MOF License Abbreviation use the mean ROUGE-1 F1-score (R1) (Lin, 2004); Definition Recognition, Question Answering, XBRL Term, XBRL Domain and Numeric Query, MOF License Detail Query, and Common Domain Model Analysis are assessed with mean BERTScore using the roberta-large setting (BERT-R) (Zhang et al., 2019); and Named Entity Recognition (NER) is evaluated by mean F1-score.

The experiment 6.3, conducted by the organizers following the evaluation framework in (Wang et al., 2024), uses different metrics: mean Accuracy for classification tasks (e.g., abbreviation, link retrieval, certification exams, XBRL Financial Math, XBRL Tag Query, MOF License Abbreviations, and MOF License OSI Approval), mean BERTScore with the bert-base-uncased setting (BERT-B) for semantic similarity in definitions, mean F1-score (F1) for NER, and FactScore (Min et al., 2023) for factual correctness in QA, XBRL, and MOF tasks.

The overall score is calculated as a weighted average, with each task contributing 10%, except for CFA, which is weighted at 20%, ensuring a balanced evaluation framework.

[2]meta-llama/Llama-3.1-8B-Instruct

[3]https://huggingface.co/KBTG-Labs/THaLLE-0.1-7B-fa

5.3 Dataset and data collection

5.3.1 Training

The training dataset for the COLING-2025 regulations challenge [4] was carefully curated to encompass key regulatory domains. It integrates data from leading finance and compliance sources listed at the challenge website [5], including EUR-LEX, ESMA, SEC, Federal Reserve, FDIC, and XBRL. The dataset spans tasks such as abbreviation recognition, definition extraction, and question answering, covering areas such as EMIR, U.S. financial laws, and accounting. This dataset provides a robust foundation for training a unified LLM capable of independently handling diverse regulatory tasks.

5.3.2 Validation

The validation set [6] (Wang, 2024), provided by the organizers of the COLING-2025 Regulations Challenge, covers a wide range of essential regulatory tasks with diverse samples. It includes 29 acronym examples from EMIR, U.S. financial laws, and other sources, 16 stock tickers, 19 definitions, 4 NER samples, and 20 QA cases covering topics such as securities, exchanges, the Federal Reserve, and accounting. Link retrieval tasks feature 22 samples, while the XBRL dataset comprises 54 terms, 100 financial math cases, and additional queries. The CDM dataset includes 16 examples focused on products, events, and processes, and the MOF dataset offers 17 samples for licensing tasks and QA. Additionally, the CFA dataset, derived from the Flare-CFA corpus [7], contributes 1,032 samples, enhancing the scope of evaluation for regulatory and financial text analysis. This comprehensive validation set ensures a thorough evaluation across complex regulatory domains.

5.3.3 Testing

The testing set[8](Wang, 2024), also curated by the COLING-2025 Regulations Challenge organizers, focuses on benchmarking model performance under diverse regulatory scenarios with a larger and more varied set of examples. It comprises 444 abbreviation cases and 162 definition tasks to assess terminology and contextual understanding, alongside 45 NER samples and 103 QA cases for evalu-

[4]https://coling2025regulations.thefin.ai

[5]https://coling2025regulations.thefin.ai/dataset

[6]https://github.com/Open-Finance-Lab/Regulations_Challenge_COLING_2025/tree/main/validation

[7]https://huggingface.co/datasets/ChanceFocus/flare-cfa

[8]https://github.com/Open-Finance-Lab/Regulations_Challenge_COLING_2025/tree/main/testing

ating entity recognition and information retrieval. The link retrieval section includes 161 samples, while the XBRL dataset is robust, featuring 391 terminology samples, 90 tag-to-report tasks, and 89 domain numeric queries, emphasizing its utility for structured data reasoning. Additionally, the testing set covers 90 financial math problems, 110 CDM queries targeting specific processes, 59 MOF detail queries, 31 MOF license abbreviations, and 50 MOF license approval samples. This dataset is designed to challenge models comprehensively, evaluating their robustness and accuracy across varied regulatory and financial contexts.

5.4 Implemetation

In this fine-tuning setup, several key configurations are designed to optimize performance and efficiency. Supervised Fine-Tuning is applied to guide the model in adapting to task-specific requirements. LoRA (Hu et al., 2021) is employed with a rank of 32, a scaling factor of 32, and a dropout rate of 5%, as inspired by (Labs et al., 2024). These settings enable the model to adapt to new tasks by focusing on low-rank adjustments in specific projection layers, such as query, key, and value projections, without updating all model weights. The training dataset is shuffled with a fixed seed (42) to ensure reproducibility and balanced sampling. Each sequence in the dataset is repeated for 10 epochs, inspired by (Shu et al., 2024), to maximize learning opportunities.

The training process is managed with a per-device batch size of 1 and gradient accumulation steps set to 8, effectively simulating larger batch sizes by accumulating gradients over multiple steps before updating the model weights (Labs et al., 2024). A learning rate of 0.0002 (Shu et al., 2024), is applied with the AdamW optimizer (Loshchilov and Hutter, 2017) to ensure stable and precise updates. The learning rate is scheduled to start gradually with a warm-up phase for better stability during initial training (Labs et al., 2024). Regular checkpoints preserve progress, and metrics are logged periodically to monitor performance. Mixed-precision training, leveraging bfloat16 precision, is enabled to improve computational efficiency, and padding is handled using the end-of-sequence token for consistency. Additionally, loss masking selectively applies loss to task-specific components, ensuring prompts and outputs for each task are fine-tuned without overwriting shared knowledge (Labs et al., 2024).

Furthermore, PEFT methods, specifically low-rank decomposition, minimize computational and memory costs by freezing most model parameters while adapting task-specific components through low-rank matrices. This significantly reduces the number of trainable parameters, lowering computational and storage overhead (Labs et al., 2024). The model is trained and evaluated on an NVIDIA A6000 GPU, leveraging its computational power and memory for efficient fine-tuning and inference. This setup supports mixed-precision operations, gradient accumulation, and low-rank adaptation, optimizing task-specific performance by balancing computation, memory, and stability.

6 Experimental Results and Discussion

6.1 Comparison of non-sequential and sequential fine-tuning approaches

Task	Metric	Non-sequential	Sequential
Abbreviation (Ticker)	R1	6.648	1.333
Abbreviation (Acronym)	R1	59.674	32.588
Definition	BERT-R	87.300	86.330
NER	BERT-R	74.171	76.752
QA	BERT-R	87.203	86.384
Link Retrieval	Acc	23.941	28.095
CFA Level 1	Acc	47.290	68.508
XBRL (Terminology)	R1	82.408	81.333
XBRL (Domain-Numeric Query)	R1	84.978	80.415
XBRL (Financial Math)	R1	1.103	1.289
XBRL (Tag Query)	R1	85.000	80.000
CDM	BERT-R	82.655	82.159
MOF (License OSI Approval)	Acc	0.000	0.000
MOF (Detailed QA)	BERT-R	88.294	87.476
MOF (License Abbreviation)	BERT-R	13.733	9.704
Overall	**Overall**	**48.663**	**59.731**

Table 3: Comparison of non-sequential and sequential fine-tuning performance on the validation set (%).

Table 3 presents an experiment comparing sequential fine-tuning, which follows the order specified in Table 1, with traditional non-sequential fine-tuning, where all datasets are combined into a single set for training. Sequential fine-tuning significantly improves overall performance, increasing the mean score from 48.66 (non-sequential) to 59.73. Notable gains are observed in tasks involving financial concepts (e.g., the CFA task) and link retrieval, demonstrating the effectiveness of this approach in these areas. However, performance declines in tasks such as abbreviation tickers, acronym validation, and certain XBRL queries, potentially due to overfitting or complexities introduced by sequential fine-tuning. Tasks with very low or zero performance further suggest issues with task formulation. In summary, while sequential fine-tuning offers substantial benefits in specific domains, its varied impact across tasks highlights the importance of adopting tailored fine-tuning strategies to optimize performance across diverse requirements.

6.2 Comparison of default Prompt and our fine-tune system prompt

Task	Metric	Default	Our
Abbreviation (Ticker)	R1	1.333	2.273
Abbreviation (Acronym)	R1	32.588	66.004
Definition	BERT-R	86.330	85.525
NER	BERT-R	76.752	77.463
QA	BERT-R	86.384	86.384
Link Retrieval	Acc	28.095	33.394
CFA Level 1	Acc	68.508	68.508
XBRL (Terminology)	R1	81.333	82.397
XBRL (Domain-Numeric Query)	R1	80.415	79.869
XBRL (Financial Math)	R1	1.289	1.548
XBRL (Tag Query)	R1	80.000	82.500
CDM	BERT-R	82.159	82.234
MOF (License OSI Approval)	Acc	0.000	0.000
MOF (Detailed QA)	BERT-R	87.476	86.878
MOF (License Abbreviation)	BERT-R	9.704	20.267
Overall	**Overall**	**59.731**	**64.720**

Table 4: Comparison of Default Prompt and Our Fine-Tune System Prompt on the validation set (%).

Table 4 compares the performance of our fine-tuned system prompt, detailed in Table 7, with ChatGPT's default system prompt ('You are a helpful assistant') (Zheng et al., 2024). Our fine-tuned prompt consistently outperforms the default across most tasks, increasing the overall mean score from 59.73 to 64.72. Significant improvements are observed in tasks such as acronym abbreviation (32.59 to 66.00), ticker abbreviation (1.33 to 2.27), and link retrieval (28.10 to 33.39), demonstrating its effectiveness in handling complex abbreviations and legal linking. Further gains are noted in NER, XBRL Terminology, and XBRL Tag Query tasks, where the fine-tuned prompt addresses previously unhandled cases. However, tasks such as Definition, QA, and CFA show minimal improvements, indicating areas for further optimization. Overall, these results confirm that tailored prompt fine-tuning enhances model accuracy and reliability, particularly for specialized and complex tasks.

6.3 Comparison of direct-response and COT-based inference with Training Variants

Table 5 contrasts direct-response inference, utilizing a system prompt (Table 7), with the proposed COT-based inference, which incorporates both a system and refinement prompt (as detail in the Section 4.4), across various training configurations. Direct-response inference achieves a mean score of 64.72, while COT-based methods demonstrate superior performance, with non-explanatory COT scoring 66.98 and reasoning-based COT achieving 68.23. COT inference methods yield significant performance improvements in complex tasks such as NER, MOF License OSI Approval and XBRL Financial Math, demonstrating their capability in step-by-step analysis and producing responses in the desired format. Reasoning-based training further enhances performance in XBRL Terminology and Financial Math tasks, underscoring the advantages of structured reasoning. In summary, reasoning-enhanced COT inference offers significant improvements in model performance across diverse, specialized tasks, emphasizing its effectiveness and adaptability.

6.4 Comparison of our model with baseline

Table 6 compares the performance of our model against leading baselines on the testing set, conducted by the organizers following the evaluation framework in (Wang et al., 2024). Our model achieves an overall score of 54.801%, outperforming Llama 3.1 8B (53.572%) and demonstrating competitive performance across tasks. Our model outperforms best in the Definition task, achieving a score of 58.49%, which is higher than GPT-4o (55.2%), Mistral Large 2 (53.38%), and Llama 3.1 8B (51.3%). It also achieves the highest score in NER at 71.74%, surpassing GPT-4o (71.08%) and other baselines. Additionally, our model demonstrates strong performance in QA (86.09%), outperforming most baselines and closely approaching GPT-4o. It also excels in MOF (Detailed QA and License OSI Approval) and shows robust results in XBRL (Domain-Numeric Query). However, areas such as Abbreviation and Link Retrieval highlight improvement opportunities, where GPT-4o and Mistral Large 2 outperform. Overall, our model provides robust performance, particularly in knowledge-intensive and domain-specific tasks, while maintaining computational efficiency.

7 Limitations and Future Work

The primary challenge of this research is to develop a single LLM capable of effectively multitasking across nine distinct regulatory and financial tasks through fine-tuning while maintaining versatility, domain expertise and efficient knowledge transfer. The LLM must perform these tasks simultaneously without any performance degradation, mitigate task interference, and manage specialized terminologies

Task	Metric	Direct-response Inference	COT-based Inference	
			Non-explanatory-based Training	Reasoning-based Training
Abbreviation (Ticker)	R1	2.273	3.835	3.992
Abbreviation (Acronym)	R1	66.004	63.705	63.653
Definition	BERT-R	85.525	85.392	85.290
NER	BERT-R	77.463	92.074	92.712
QA	BERT-R	86.384	86.319	87.513
Link Retrieval	Acc	33.394	52.272	53.825
CFA Level 1	Acc	68.508	68.702	68.716
XBRL (Terminology)	R1	82.397	84.275	86.107
XBRL (Domain-Numeric Query)	R1	79.869	80.034	81.610
XBRL (Financial Math)	R1	1.548	37.667	39.097
XBRL (Tag Query)	R1	82.500	82.500	82.532
CDM	BERT-R	82.234	82.204	82.096
MOF (License OSI Approval)	Acc	0.000	100	100
MOF (Detailed QA)	BERT-R	86.878	87.199	87.590
MOF (License Abbreviation)	BERT-R	20.267	16.477	16.687
Overall	**Overall**	**64.720**	**66.977**	**68.227**

Table 5: Comparison of Our Fine-Tune System Prompt and COT-based Inference Methods on the validation set (%).

Task	Metric	FinMind-Y-Me	Llama 3.1 8B	GPT-4o	Mistral Large 2
Abbreviation	Acc	20.95	23.2	37.84	22.3
Definition	BERT-B	58.49	51.3	55.2	53.38
NER	F1	71.74	63.52	71.08	70.62
QA	FactScore	86.09	80.79	88.42	82.63
Link Retrieval	Acc	23.6	43.48	20.5	58.75
Certificate (CFA Level 1)	Acc	48.89	51.11	68.89	68.89
Certificate (CFA Level 2)	Acc	46.75	40.26	57.14	55.84
Certificate (CFA Level 3)	Acc	44.87	41.03	65.38	64.1
Certificate (CPA REG)	Acc	47.52	40.59	71.29	64.36
XBRL (Terminology)	FactScore	63.27	70.83	85.03	82.21
XBRL (Domain-Numeric Query)	FactScore	66.36	58.45	58.51	68.31
XBRL (Financial Math)	Acc	64.44	76.67	88.42	74.44
XBRL (Tag Query)	Acc	26.67	16.67	77.78	86.67
CDM	FactScore	85.28	79.8	88.2	86.32
MOF (License OSI Approval)	Acc	74.0	72.0	96.0	44.0
MOF (Detailed QA)	FactScore	80.75	69.56	81.56	82.29
MOF (License Abbreviations)	Acc	3.23	12.9	19.35	12.9
Overall	**Overall**	**54.801**	**53.572**	**63.567**	**62.489**

Table 6: Performance Comparison of our model with baseline Across Tasks on the testing set (%)

and context shifts. However, several limitations hinder its effectiveness. These include suboptimal performance in link retrieval due to generating links from queries rather than directly accessing a database; difficulties in abbreviation expansion caused by context-dependent ambiguities; inaccuracies in answering certification questions stemming from misinterpretation; and challenges with XBRL and MOF subtasks resulting from insufficient data availability.

These limitations underscore the need for more comprehensive, diverse and contextually relevant datasets, improved fine-tuning approaches, and the development of advanced reasoning strategies. Future research should aim to broaden the range of regulatory and financial tasks to enhance the versatility and scalability of the LLM. Efforts should also focus on automating prompt engineering to reduce reliance on manual design and explore advanced reasoning methods, such as reinforcement learning with human feedback. Furthermore, optimizing task sequences and addressing challenges such as computational resource demands, data dependencies, and processing costs are vital to improving system robustness and adaptability within dynamic regulatory and financial environments.

8 Conclusion

This study presents a unified modeling framework that integrates task-specific prompts, input templates, and sequential fine-tuning to improve performance in regulatory and financial tasks on the COLING2025 regulation challange. Sequential fine-tuning demonstrates improvements in areas such as financial computations, though its variable impact underscores the importance for tailored strategies. Fine-tuned system prompts outperform standard prompts, while reasoning-based training and Chain-of-Thought prompting further boost performance. Our model achieved an overall score of 54.801% across all tasks, the highest among all participants, securing first place in the financial regulation competition and demonstrating excellence across all nine tasks. Future work should focus on broadening task coverage, automating prompt engineering, refining sequential fine-tuning, and exploring hybrid models to enhance scalability and adaptability in dynamic regulatory contexts.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Made laine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Is abella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Jo hannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *ArXiv*, abs/2304.01933.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

KBTG Labs, Danupat Khamnuansin, Atthakorn Petchsod, Anuruth Lertpiya, Pornchanan Balee, Thanawat Lodkaew, Tawunrat Chalothorn, Thadpong Pongthawornkamol, and Monchai Lertsutthiwong. 2024. Thalle: Text hyperlocally augmented large language extension – technical report.

Vladislav Lialin, Namrata Shivagunde, Sherin Muckatira, and Anna Rumshisky. 2023. Relora: High-rank training through low-rank updates. In *International Conference on Learning Representations*.

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.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *ArXiv*, abs/2205.05638.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Moran Mizrahi, Guy Kaplan, Daniel Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2023. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem W. Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomás Kociský, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, J Christopher Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Ying-Qi Miao, Lukás Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontan'on, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, A.E. Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Venkatesh Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matt Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara N. Sainath, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela de Castro Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, S'ebastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Joshua Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost R. van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya B Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance

Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, S'ebastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Michael B. Chang, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravichandra Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Luvci'c, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjosund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos L. Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Zhufeng Pan, Zachary Nado, Stephanie Winkler, Dian Yu, Mohammad Saleh, Lorenzo Maggiore, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Chung-Cheng Chiu, Zoe C. Ashwood, Khuslen Baatarsukh, Sina Samangooei, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxi aoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabriel Barth-Maron, Craig Swanson, Dominika Rogozi'nska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Ren shen Wang, Dave Lacey, Anastasija Ili'c, Yao Zhao, Woohyun Han, Lora Aroyo, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Raphael Lopez Kaufman, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, T. Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anais White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Kiran Vodrahalli, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, cCauglar Unlu, David Reid, Zora Tung, Daniel F. Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Rui Zhu, Ricardo Aguilar, Mai Gim'enez, Jiawei Xia, Olivier Dousse, Willi Gierke, Soheil Hassas Yeganeh, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Daniel Niels Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Daniel Toyama, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nicholas Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, Donghyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alexey Yakubovich, Nilesh Tripuraneni, James Manyika, Haroon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Anna Bulanova, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Cl'ement Farabet, Pedro Valenzuela, Quan Yuan, Christoper A. Welty, Ananth Agarwal, Mianna Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Elnaz Davoodi, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, A. Ya. Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Alejandro Lince, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiří ima, Anna Koop, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Kalpesh Krishna, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas Fitzgerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Ilia Shumailov, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Katherine Lee, S. Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel El Kaed, Jing Li, Jakub Sygnowski, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Poder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Junwen Bai, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, Oriol Vinyals, and Alexandra Chronopoulou. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the us legal system. In *International Conference on Information and Knowledge Management*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. In *Conference on Empirical Methods in Natural Language Processing*.

Keyi Wang. 2024. Regulations challenge coling 2025. `https://github.com/Open-Finance-Lab/Regulations_Challenge_COLING_2025`.

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)*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *ArXiv*, abs/2302.11382.

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, 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, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Text alignment is an efficient unified model for massive nlp tasks. *ArXiv*, abs/2307.02729.

Jingwei Zhang, Saarthak Kapse, Ke Ma, Prateek Prasanna, Joel H. Saltz, Maria Vakalopoulou, and Dimitris Samaras. 2023. Prompt-mil: Boosting multi-instance learning schemes via task-specific prompt tuning. *ArXiv*, abs/2303.12214.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *ArXiv*, abs/2211.01910.

A Appendices

A.1 The task-specific system prompts for fine-tuning models

The table 7 offers a structured overview of input templates defined by the organizers and our fine-tuned system prompts.

A.2 Examples of non-explanatory and reasoning-based data for financial and regulatory tasks

The table 8 provides the distinction between non-explanatory responses and reasoning-based responses for fine-tuning LLMs.

A.3 Inference strategies with Chain of Thought prompting

The table 9 outlines task-specific strategies for using CoT prompting to improve inference across various financial and regulatory tasks.

Task	Input Templates	System Prompt
Abbreviation	"Expand the following acronym into its full form: acronym. Answer:"	You are an expert in abbreviation-expanded-form matching for financial regulation. Analyze and expand the following acronym into its official full form. Provide the most accurate expansion only.
Definition	"Define the following term: regulatory term or phrase. Answer:"	Define the following term while categorizing it into regulatory or financial domains (e.g., Federal Reserve Regulations, Accounting). Provide the definition clearly and concisely.
NER	"Given the following text, only list the following for each: specific Organizations, Legislations, Dates, Monetary Values, and Statistics: input text."	You are an expert in Name entity recognition. Extract and classify entities such as Organizations, Legislations, Dates, Monetary Values, and Statistics from the given text. Return the output in JSON format with proper labels.
QA	"Provide a concise answer to the following question: detailed question? Answer:"	You are an expert in regulations and finance. Provide precise and accurate answers to detailed questions about regulatory practices or laws based on the provided query.
Link Retrieval	"Provide a link for ... law, Write in the format of ("Law: Link" or "Law: Not able to find a link for the law")"	You are an expert in link retrieval. Provide a link for the specified regulation based on its name and format. Ensure the URL follows the correct structure (e.g., EUR-Lex). Return only the link or specify if unavailable.
CFA	"(This context is used for the question that follows: context). Please answer the following question with only the letter and associated description of the correct answer choice: question and answer choices. Answer:"	You are a financial expert tasked with solving a certificate exam question. Break down the query logically, analyze each answer choice, and provide the best answer based on regulations or financial principles.
XBRL	"Provide the exact answer to the following question: detailed question? Answer:"	You are an expert in eXtensible Business Reporting Language (XBRL). Provide precise answers to detailed questions about financial data using eXtensible Business Reporting Language. Address areas such as definitions, calculations, or US GAAP tags systematically.
CDM	"Provide a concise answer to the following question related to Financial Industry Operating Network's (FINO) Common Domain Model (CDM): detailed question? Answer:"	You are an expert in Common Domain Model (CDM). Provide accurate and precise responses to questions related to the CDM within the financial and fintech context. Break down terms or processes where applicable.
MOF	"Provide a concise answer to the following question about MOF's licensing requirements: detailed question? Answer:"	You are an expert in Model Openness Framework (MOF). Answer queries about license requirements, OSI approval, or abbreviations with precision and clarity. Provide only the relevant details.

Table 7: Fine-tune task-specific system prompts

User prompt	Non-explanatory response	Reasoning response
An asset with a purchase price of $7229.15 and a salvage value of $860.73 is depreciated over 2 years using the straight-line method. What is the annual depreciation expense?	**Answer**: $\$3184.21$	**Solution**: Annual Depreciation = (Purchase Price - Salvage Value) / Useful Life $= (7229.15 - 860.73)/2 = 3184.21$ **Answer**: $\$3184.21$
An asset with a purchase price of $4754.66 and a salvage value of $396.31 is depreciated over 9 years using the sum-of-years'-digits method. What is the depreciation expense for year 6?	**Answer**: $\$387.41$	**Solution**: Depreciation for year 6 = (Purchase Price - Salvage Value) * Remaining Useful Life / Sum of Years' Digits $(= 4754.66 - 396.31) * 4/(9 * (9 + 1)/2) = 387.41$ **Answer**: $\$387.41$
What is the effective annual interest rate of a 14.21% nominal rate compounded 2 time(s) per year?	**Answer**: 14.71%	**Solution**: Effective Rate = (1 + Nominal Rate / Periods) $^{Periods} - 1$= (1 + 0.1421 / 2) $^{2} - 1 = 0.1471 =$ 14.71% **Answer**: 14.71%

Table 8: Examples of non-explanatory and reasoning-based data for financial and regulatory tasks

Task	Chain of Thought Process	System Prompt	User Prompt
Abbreviation	Identify abbreviations related to finance and regulations. Analyze the context of each abbreviation and determine its full expanded form based on common financial and regulatory usage.	Step1: "Identify the abbreviations in the domain of regulations and finance, match each abbreviation with its expanded form."	Step1: "abbreviation as fullquestion answer only fullquestion stands for ... and focus on the one most relevant to the domain of regulations and finance."
	Cross-check the abbreviation context from the previous step and match it with the single, most relevant expanded definition. Extract the exact full name or phrase without any extra explanation.	Step2: "Match an abbreviation with its expanded form."	Step2: "From this response response, extract only the full form of the abbreviation and extract only one answer."
Definition	Categorize financial and regulatory terms into their respective categories based on common industry standards or classification systems. Use logical categorization methods.	Step1: "Categorize the following regulatory and financial term or phrase into one of the categories: Federal Reserve Regulations, European Market Infrastructures Regulation, Securities and Exchanges or Accounting and Auditing. Answer only with the category."	Step1: "Term or phase as question"
	Based on the assigned category, determine the definition of the financial or regulatory term. Use established definitions from financial research and regulatory analysis.	Step2: "Provide the definition of the following regulatory and financial term or phrase in category category. Answer as: The term [term] means..."	Step2: "Term as question"
	Analyze the definition and distill the core meaning into the most concise response. Ensure no extraneous context or explanation is included.	Step3: "Correctly define a regulatory term or phrase."	Step3: "From this response response, extract only the meaning of the definition and extract only one answer."
NER	This step involves extracting and categorizing entities (e.g., organizations, legislations, dates, monetary values, statistics) from the provided financial text. All entities should be properly labeled and organized into a structured JSON format to ensure consistency and accuracy.	Step1: "You are tasked with extracting specific entities from financial text. Your job is to identify and classify the following entities: - Organizations - Legislations - Dates - Monetary Values - Statistics After identifying each entity in the text, return the results in the following JSON format. Make sure to follow the structure strictly and provide the correct labels for each entity type. Each entity type should be in its own list, even if there is only one entity for that type.	Step1: Given the following financial text, extract only the following entities: Organizations, Legislations, Dates, Monetary Values, and Statistics. Text: question Please return the results in the JSON format specified by the system.
QA	Analyze the provided financial or regulatory question in detail. Employ systematic reasoning, utilizing domain expertise and logical inference to ensure accuracy.	Step1: "You are an expert in regulations and finance. Ensure the output matches the correct answer to a detailed question about regulatory practices or laws."	Step1: "Question as question"
Link Retrieval	Categorize the provided financial or regulatory query into predefined legal categories. The classification should help pinpoint the most applicable legal category.	Step1: "Categorize the following regulatory and financial questions into one of the categories: Federal Reserve Regulations, European Market Infrastructures Regulation, The Federal Deposit Insurance Corporation, or Securities and Exchange Commission. Answer only with the category."	Step1: "Term or phase as question, answer as category"
	Identify and provide the most accurate legal reference link based on the classification derived from Step 1. The link should correspond to the relevant law or regulation context.	Step2: "Ensure the provided link is accurate and corresponds to the relevant law in the category response1, focusing specifically on the most applicable law in the domain of regulations and finance."	Step2: "Please provide the law related to: question"
CFA	Carefully analyze the CFA exam question by breaking it down into its key financial components. Clearly outline the reasoning process and draw on formulas, definitions, and financial concepts as needed.	Step1: "You are a financial expert. Please read the following certificate exam question carefully, analyze the key components, and answer the question step by step. Break down any complex terms or procedures and provide a clear, concise final answer. If applicable, use formulas, examples, or definitions to support your response. Be sure to verify the accuracy of your answer once completed."	Step1: "question as question"
	After detailed analysis, select the most accurate answer choice (A, B, or C) based on logical reasoning. The response should focus only on the final correct choice without unnecessary explanation.	Step2: "You are a financial expert tasked with carefully reading, analyzing, and answering the following certificate exam question. Please follow the steps below:"	Step2: "Your task is to carefully read the certificate exam question as question, analyze it step-by-step, and provide your answer as responseexplain. Select the most accurate answer from the choices provided, listed as choices. Only answer with A, B, or C. Do not provide any other response."
XBRL	Logical reasoning to identify and categorize the provided XBRL context using the five focus areas (definitions, numeric queries, domain analysis, etc.).	Step1: "Provide precise answers to detailed questions about financial data extraction and application using XBRL (eXtensible Business Reporting Language) filings, a standardized digital format for sharing and analyzing financial information. This task covers five areas: defining XBRL terms, domain-specific queries, financial math, numeric queries, and providing the correct US GAAP XBRL tags (e.g., US GAAP XBRL tag for revenue should be answered asusgaap :RevenueFromContractWithCustomerExcludingAssessedTax'. Ensure responses strictly match the correct answer without additional explanation.When answering questions about XBRL, it's essential to follow a structured approach. Here's how to methodically address these types of questions:"	Step1: "Question as question"
	Execution of extraction and application logic using the structured reasoning methodology for context-specific results (e.g., matching correct US GAAP tags).	Step2: "You are a financial expert tasked with carefully reading, analyzing, and answering the following eXtensible Business Reporting Language. Please follow the steps below:"	Step2: "Your task is to read the eXtensible Business Reporting Language XBRL question question and find the final answer based on the explanation provided response. Provide only the final answer,final answer is ..."
CDM	Addressing CDM inquiries from the Fintech Open Source Foundation, applying logical mapping to provide relevant responses for complex financial modeling or structured analysis.	Step1: "Deliver precise responses to questions about the Fintech Open Source Foundationś FINOS Common Domain Model CDM)."	Step1: "Question: question"
MOF	Licensing logic for MOF compliance focusing on financial license inquiries or compliance context by narrowing domain relevance.	Step1: "Deliver precise responses to questions concerning the requirement of license under the Model Openness Framework."	Step1: "Question: question"

Table 9: Chain of Thought strategies and refinement prompting for financial and regulatory tasks