Sequential fine-tuning framework for multitask learning in financial and regulatory domains

Anonymous ACL submission

Abstract

LLMs excel in general NLP yet struggle in specialized domains such as finance and regulation. This study proposes a sequential fine-tuning framework for multitask learning in a unified LLM, structuring tasks into foundational, question-answering, and stylistic-answer knowledge to mitigate catastrophic forgetting and enhance knowledge transfer. Evaluations on the COLING 2025 Regulations Challenge dataset demonstrate significant improvements, with notable gains in financial QA and MOF license abbreviation recognition. Unlike Chain-of-Thought inference-based methods, this approach integrates reasoning during training, reducing inference costs and improving scalability. While challenges remain with sparse and context-dependent data, the findings highlight structured task sequencing as a promising strategy for domain-adapted LLM.

1 Introduction

Large language models (LLMs), such as Llama (Touvron et al., 2023) and Qwen (Bai et al., 2023), have excelled in various natural language processing (NLP) tasks, including serving as conversational agents. However, they struggle in specialized domains such as finance and regulation, where precise contextual understanding, multistep reasoning, and adaptability are essential. While API-based services such as GPT (Achiam et al., 2023) and Gemini (Reid et al., 2024) ensure accessibility and scalability, their generalized design limits effectiveness in domain-specific applications due to challenges in data privacy, compliance, and customization.

The local fine-tuning enables tailored solutions while maintaining data control, optimizing LLMs for diverse domain-specific tasks with unified reasoning, consistency, and accuracy. Advancements in fine-tuning and prompting enhance LLMs in specialized domains. Prompt engineering structures task-specific instructions for precise, context-aware outputs (Mizrahi et al., 2023; White et al., 2023; Zheng et al., 2024), while Chain-of-Thought (CoT) prompting improves reasoning by guiding models through intermediate steps, particularly in arithmetic and reasoning (Wei et al., 2022; Wang et al., 2023). Fine-tuning methods have advanced to enhance LLMs for downstream tasks while optimizing performance and efficiency. Early approaches, such as full fine-tuning in GPT-3, updated all parameters but incurred high computational costs (Brown et al., 2020). In contrast, Parameter-Efficient Fine-Tuning (PEFT) methods such as adapters and Low-Rank Adaptation (LoRA) modify select parameters, achieving comparable results with significantly lower computational costs (Hu et al., 2023; Liu et al., 2022; Hu et al., 2021).

Despite advancements, adapting a single LLM for multitasking in domain-specific datasets remains challenging due to catastrophic interference, where new tasks degrade prior performance, and overfitting, requiring careful data curation (Goodfellow et al., 2013; Aghajanyan et al., 2020). This study proposes a sequential fine-tuning framework for regulatory and financial domains, structuring tasks by relevance, complexity, and dataset characteristics. Foundational tasks establish domain knowledge, generalized tasks refine response styles, and specialized tasks address complex challenges, ensuring effective knowledge transfer while minimizing interference. By internalizing reasoning patterns, the framework eliminates the need for inference-stage techniques such as CoT prompting. It integrates task-specific prompts, input templates, and sequential training to align model outputs with regulatory and financial requirements. Leveraging diverse datasets, the model learns both factual knowledge and stylistic nuances, improving accuracy, efficiency, and scalability. The key contributions of this paper are:

1. A unified framework that fine-tunes a single LLM for multitasking in financial domains.

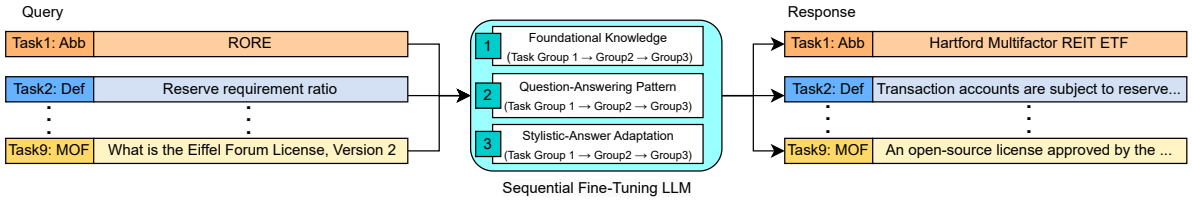

Figure 1: Sequential Fine-Tuning Framework Supporting Nine Regulatory and Financial Tasks

2. Sequential fine-tuning that optimizes task learning based on domain relevance, complexity, and characteristics.

3. Internalized response patterns, ensuring factual accuracy and stylistic coherence through integrated knowledge acquisition and task-specific adaptation.

The paper is structured as follows: task overview, methodology, experimental setup, results, limitations, and conclusions.

2 Task overview

This study utilizes data from the COLING 2025 Regulations Challenge (Wang et al., 2024), which benchmarks LLM performance across nine regulatory and financial tasks. The abbreviation recognition task (Abb) tests acronym identification and expansion; the definition recognition task (Def) focuses on extracting definitions; the Named Entity Recognition (NER) Task identifies and categorizes entities such as organizations, laws, and monetary values. The question answering task (QA) assesses responses to complex legal queries, and the link retrieval task (Link) challenges models to locate legal documents within extensive corpora. The certificate question task (CFA) involves solving CFA and CPA-style multiple-choice questions, while the XBRL analytics task (XBRL) examines data extraction from XBRL filings. The Common Domain Model (CDM) task evaluates fintech industry interoperability standards, and the Model Openness Framework (MOF) licenses task tests understanding of licensing compliance. These tasks collectively evaluate the linguistic, analytical, and reasoning capabilities of LLMs.

3 Methodology

This study proposes a sequential fine-tuning framework to enhance multitasking with a single LLM in financial domains. By structuring tasks in a domain-relevant sequence, the model internalizes foundational knowledge before tackling complex challenges. It also integrates response patterns and stylistic adaptation to reinforce reasoning.

3.1 Sequential Training Strategy

The key innovation lies in leveraging sequential training to mitigate catastrophic interference, enhance task-specific generalization, and optimize knowledge transfer across related tasks. The sequential fine-tuning organizes the training process into three phases: foundational knowledge, question-answering patterns, and stylistic adaptation, ensuring effective knowledge transfer while minimizing catastrophic forgetting.

3.1.1 Foundational Knowledge

The first phase fine-tunes the LLM on essential foundational knowledge, providing a strong base for all tasks. It introduces the model to domain-specific terminology, regulatory concepts, and general task structures. For instance, training on legal definitions, XBRL taxonomies, financial regulation and accounting helps the model develop a baseline understanding necessary for advanced reasoning. This phase is crucial for preventing task-specific overfitting in later fine-tuning stages, ensuring robust knowledge transfer.

3.1.2 Question-Answering Pattern

The second phase focuses on fine-tuning the model for question-answering patterns. This phase enables the model to interpret queries and generate concise, contextually accurate responses. Training data includes question-answer pairs from tasks such as regulatory compliance queries, financial certification exams, and document retrieval. By learning to structure responses effectively, the model gains flexibility in handling various task formats, enhancing reasoning and adaptability. This phase ensures that the model interprets task requirements and provides precise, task-aligned outputs.

3.1.3 Stylistic-Answer Adaptation

The final phase fine-tunes the model to generate responses that adhere to domain-specific stylis-

tic requirements. Regulatory and financial tasks often require structured formats, such as formal language, concise summaries, or well-organized answers. For example, the Definition Recognition Task demands clear, precise definitions, while the Certificate Question Task requires structured multiple-choice responses. This phase ensures the model produces outputs that align with real-world expectations, improving usability and user trust.

3.2 Task Management

Group	Domain	Task	Training size	Metrics
Group 1	CDM	All Required	2,414	BERTscore
	Definition	All Required	1,720	
	QA	All Required	1,349	
	XBRL	Domain and numeric query (D&N que)	723	
	MOF	Detailed QA	424	
	XBRL	XBRL Terminology (Term)	143	
Group 2	Abbreviation	EMIR	210	ROUGE-1
	Abbreviation	Stock Tickers (NYSE)	8,320	
	MOF	License Abbreviations (Lic-abb)	240	
Group 3	NER	EMIR	1,905	F1score
	XBRL	XBRL tag query (Tag que)	1,209	Accuracy
	XBRL	Financial Math (Fin-math)	222	
	CFA	CFA Level 1	1,032	
	Link-Retrieval	All Required	460	
	MOF	License approval (Lic-app)	380	

Table 1: Sequence of tasks in sequential fine-tuning

3.2.1 Task Grouping

Tasks are grouped based on their domain relevance, complexity, and functional characteristics, as outlined in Table 1. The nine regulatory tasks from the Regulations Challenge were organized into three groups based on the metrics used for evaluation. Tasks sharing similar functional attributes but differing in evaluation metrics, such as XBRL Tag Query and XBRL Financial Math, are separated to maintain focus. In contrast, tasks with thematic similarities, such as XBRL Terminology and Definition Recognition, which share thematic elements, are grouped to leverage overlapping knowledge.

3.2.2 Task Ordering

Task ordering was strategically designed to optimize knowledge transfer while minimizing task interference. The sequence began with general tasks that foundational knowledge and broad applicability, leveraging diverse datasets and evaluation metrics such as BERTScore. For example, legal and financial definition tasks were prioritized for their generalizability. Following this, tasks requiring high specificity and precision, such as abbreviation retrieval, which used ROUGE-1 as an evaluation metric, were fine-tuned. Finally, highly specialized tasks such as link retrieval, which rely on explicit memorization and direct dataset references, were trained last. This approach preserved the model's knowledge base and reasoning patterns from earlier phases, ensuring minimal task interference.

3.3 Task-Specific Prompts and Templates

This framework unifies multiple regulatory tasks by incorporating task-specific prompts and input templates, ensuring accurate and consistent responses. These tailored prompts address the unique requirements of each task, enabling efficient handling of diverse regulatory challenges while maintaining coherence. Table 5 outlines the tasks and their corresponding prompts, demonstrating the precision and reliability of this approach.

4 Experimental Setting

This section outlines the datasets, training configurations, and evaluation metrics in this study.

4.1 Datasets

The COLING-2025 Regulations Challenge dataset[1] integrates regulatory data from sources such as EUR-LEX, SEC, Federal Reserve, and XBRL, supporting tasks such as abbreviation recognition, definition extraction, and question answering across domains such as EMIR and U.S. financial laws. Structured for sequential fine-tuning, it captures foundational knowledge, question-answer pairs, and stylistic nuances for specialized tasks. The validation set[2] (Wang, 2024) spans 29 acronyms, 16 stock tickers, 19 definitions, 4 NER samples, 20 QA cases, and 22 link retrieval tasks, along with datasets from XBRL (54 terms, 100 financial math cases), CDM (16 product/process examples), MOF (17 licensing tasks), and CFA (1,032 Flare-CFA samples [3]). The testing set evaluates 444 abbreviations, 162 definitions, 45 NER tasks, 103 QA cases, 161 link retrieval tasks, 391 XBRL terms, 90 financial math cases, and MOF queries (e.g., licenses and approvals). These datasets provide a benchmark for assessing model robustness and accuracy in regulatory and financial contexts.

4.2 Training Setup

The model is fine-tuned using PEFT with LoRA, configured with a rank of 32, a scaling factor of 32, a 5% dropout rate, and 10 epochs. Training employed a per-device batch size of 1, with gradient accumulation over 8 steps and a learning rate

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

[2] https://github.com/Open-Finance-Lab/Regulations_Challenge

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

Method					Overall score	Task1	Task2	Task3	Task4	Task5	Task6	Task7				Task8	Task9		
Model	Task sequence	Know-ledge	Q&A	Style		Abb	Def	NER	QA	Link	CFA	XBRL				CDM	MOF		
												Term	D&N que	Fin-math	Tag que		Lic-Abb	Lic-App	QA
						R1	BERT	F1	BERT	Acc		BERT		Acc		BERT	R1	Acc	BERT
Llama3.1-ins-8B	Baseline				53.83	30.92	85.52	46.32	84.58	7.65	58.81	84.75	85.42	10.61	21.63	83.53	9.93	66.25	77.7
Qwen2.5-ins-7B	Baseline				55.3	27.5	85.79	50.41	85.16	8.75	68.94	84.32	83.45	12.68	31.2	83.14	9.85	66.02	77.02
THaLLE0.1-7B	Baseline				54.83	27.28	86.64	48.53	85.11	8.72	67.8	84.52	82.8	10.18	30.77	83.12	9.85	65.47	76.78
Finetune Qwen2.5-ins-7B	Non-seq				54.65	26.52	81.63	49.05	86.8	8.48	68.26	86.77	81.75	12.75	30.19	79.12	9.87	64.59	79.33
	G1-G2-G3		✓		59.78	45.83	80.41	62.44	78.03	27.1	62.82	77.91	78.26	61.11	27.17	74.41	14.29	68.17	78.94
	G1-G3-G2		✓		57.5	43.13	80.59	59.25	77.06	25.32	58.99	72.35	74.35	57.07	25.93	76.42	13.28	63.34	77.86
	G2-G1-G3		✓		59.14	44.89	81.18	63.78	75.26	26.86	60.98	75.12	78.02	60.18	26.82	76.15	14.29	67.65	76.79
	G2-G3-G1		✓		55.78	27.28	85.37	51.22	86.3	8.72	70.73	85.86	82.7	12.7	30.93	85.63	9.97	66.88	76.68
	G3-G1-G2		✓		55.51	42.09	77.86	57.66	74.11	24.34	56.66	69.19	71.87	54.7	24.93	73.05	12.97	62.14	75.52
	G3-G2-G1		✓		55.79	41.81	78.62	57.06	74.58	24.79	56.46	69.71	72.33	55.77	24.88	74.83	12.97	61.45	75.78
	G1-G2-G3	✓			57.94	44.79	78.08	60.14	74.52	26	60.6	76.33	75.87	58.9	26.34	72.29	13.83	66.18	77.32
	G1-G2-G3			✓	61.69	47.01	82.79	65.44	80.1	27.78	64.51	79.51	80.3	63.74	27.77	78.19	14.88	69.88	81.7
	G1-G2-G3	✓		✓	63.82	47.48	84.38	68.61	84.96	28.79	66.39	81.65	81.55	66.45	28.94	83.1	15.53	72.18	83.52
	G1-G2-G3		✓	✓	61.63	46.93	82.74	65.71	79.99	27.66	64.78	79.05	79.9	63.23	27.93	78.59	14.9	69.46	81.98
	G1-G2-G3	✓	✓	✓	66.42	49.77	87.67	70.66	88.44	30.01	68.72	84.92	84.46	69.77	29.91	86.7	16.12	74.91	87.82

Table 2: Comparison of Fine-Tuning Strategies and Task Orderings on the test Set (%).

of 0.0002, optimized with AdamW. A warm-up phase is applied for stability. Datasets are shuffled with a fixed seed of 42 to ensure reproducibility. Training is conducted on an NVIDIA A6000 GPU over a span of 26 hours, utilizing mixed-precision to enhance computational efficiency.

4.3 Evaluation Metrics

This study evaluates LLM performance across nine regulatory tasks using tailored metrics: mean Accuracy (Acc) for Link Retrieval, CFA, XBRL financial math, XBRL tag query and MOF License OSI Approval; mean ROUGE-1 F1-score (R1) (Lin, 2004) for Abbreviation Recognition and MOF License Abbreviation; mean BERTScore with the roberta-large setting (BERT) (Zhang et al., 2019) for tasks such as Definition Recognition, XBRL terminology and Question Answering; and mean F1-score (F1) for NER. The overall score is the mean of all tasks, weighted equally.

5 Experimental Results and Discussion

Table 2 compares fine-tuning strategies, task ordering, and training phases across nine tasks from the regulations Challenge, highlighting the benefits of sequential fine-tuning over baseline and non-sequential approaches. Zero-shot baseline models struggled with domain-specific tasks, with **Llama3.1-ins**, **Qwen2.5-ins**, and **THaLLE0.1** achieving mean scores of 53.83%, 55.3%, and 54.83%, respectively. Non-sequential fine-tuning slightly improved performance (54.65%) but suffered from catastrophic interference. In contrast, sequential fine-tuning with predefined task ordering (**BERTScore-evaluated tasks** $\rightarrow$ **ROUGE-1 tasks** $\rightarrow$ **precision-based tasks**) achieved a higher mean score of **66.42%**, effectively leveraging foundational knowledge, question-answering patterns, and stylistic adaptation. Notable gains include MOF license QA (87.82% BERT) and abbreviation recognition (87.67% ROUGE-1). The results underscore the importance of task ordering in optimizing knowledge transfer. Incorporating all training phases improved knowledge retention, reasoning flexibility, and stylistic coherence, leading to more consistent performance across tasks.

6 Conclusion

This study presents a sequential fine-tuning framework to enhance LLMs for regulatory and financial multitask learning. By structuring tasks into foundational, generalized, and specialized categories, the framework improves financial question answering and link retrieval while addressing catastrophic forgetting. Unlike inference-based methods, it incorporates question-answer pairs, stylistic adaptation, and reasoning during training, enhancing scalability. Despite challenges with sparse and context-dependent data, the findings underscore the effectiveness of structured task sequencing in developing robust and adaptable LLM applications.

Limitations

Despite notable improvements, certain tasks, such as XBRL Tag Query and MOF License Approval, demonstrate potential for further refinement. Future research may explore data augmentation to enhance dataset diversity, dynamic task ordering that adjusts based on real-time performance metrics, and advanced fine-tuning techniques, such as multi-stage fine-tuning or memory networks, to mitigate catastrophic forgetting. Moreover, hardware constraints present a significant limitation, as the computational demands of training and inference on large models affect scalability and accessibility. Addressing these challenges could further enhance the adaptability and efficiency of LLMs in regulatory and financial domains.

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, Madelaine 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.

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2020. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *ArXiv*, abs/2012.13255.

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.

Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgeting in gradient-based neural networks. *CoRR*, abs/1312.6211.

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.

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.

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.

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.

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, Jaisal Patel, Charlie Shen, Daniel S. Kim, Andy Zhu, Alex Lin, Luca Borella, Cailean Osborne, Matt White, Steve Yang, Kairong Xiao, and Xiao-Yang Liu Yanglet. 2024. A report on financial regulations challenge at COLING 2025. *CoRR*, abs/2412.11159.

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*.

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.

A Appendices

A.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	83.950	85.633	86.077
NER	BERT	31.434	76.113	68.290
QA	BERT	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	81.921	81.465	81.976
MOF (License OSI approval)	Acc	0.000	0.000	0.000
MOF (Detailed QA)	BERT	89.128	87.476	86.854
MOF (License abbreviation)	BERT	14.306	9.607	12.118
Overall score		**44.009**	**53.082**	**52.473**

Table 3: Model performance comparison on the validation set (%)

To assess performance for model selection, we compared Qwen2.5-7B-Instruct[4] (Team, 2024; Yang et al., 2024) with Llama-3.1-8B-Instruct[5] and THaLLE-0.1-7B-fa[6] (Labs et al., 2024) across multiple tasks. Table 3 provides a detailed comparison, showcasing Qwen2.5-7B-Instruct as a strong contender, particularly excelling in reasoning and domain-specific tasks. With its 7 billion parameters, the model maintains an optimal balance between computational efficiency and the ability to handle complex tasks. Given its superior performance and well-balanced architecture, we selected Qwen2.5-7B-Instruct as the base model for fine-tuning across various financial and regulatory tasks.

[4]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[5]meta-llama/Llama-3.1-8B-Instruct
[6]https://huggingface.co/KBTG-Labs/THaLLE-0.1-7B-fa

A.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	86.330	85.525
NER	BERT	76.752	77.463
QA	BERT	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	82.159	82.234
MOF (License OSI approval)	Acc	0.000	0.000
MOF (Detailed QA)	BERT	87.476	86.878
MOF (License abbreviation)	BERT	9.704	20.267
Overall score		**53.491**	**57.016**

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 and input template (detailed in Table 5) against ChatGPT's default system prompt ("You are a helpful assistant") (Zheng et al., 2024), using the same input template from Table 5. Our fine-tuned prompt consistently outperforms the default, raising the overall mean score from 53.49 to 57.02. Notable improvements are seen in acronym abbreviation (32.59 → 66.00), ticker abbreviation (1.33 → 2.27), and link retrieval (28.10 → 33.39), highlighting its effectiveness in handling complex abbreviations and legal linking. Additional gains are observed in NER, XBRL Terminology, and XBRL Tag Query tasks, where the fine-tuned prompt successfully addresses previously unhandled cases. However, Definition, QA, and CFA tasks show minimal improvements, indicating areas for further refinement. Overall, these results confirm that tailored prompt fine-tuning enhances LLM accuracy and reliability, particularly for specialized and complex tasks.

A.3 The task-specific system prompts and input template for fine-tuning the LLM

Table 5 presents the task-specific system prompts and input templates used for fine-tuning the language model.

A.4 Chat template, features, and labels for fine-tuning the LLM

The chat template used for fine-tuning the LLM follows the structure:

```
<|im_start|>system
{system_prompt}<|im_end|>
<|im_start|>user
{user_prompt}<|im_end|>
<|im_start|>assistant
{assistant_response}<|im_end|>
```

The system prompt is provided in Table 5. The user prompt varies based on the training objective. For foundational knowledge and question-answering strategy, it corresponds to the features listed in Table 6. For stylistic-answer adaptation strategy, the user prompt begins with the input template, followed by the appropriate feature from Table 6. The assistant response for each training strategy serves as the corresponding label, as detailed in Table 6.

A.5 Queries and expected responses of a fine-tuned LLM

Table 7 presents examples of queries and expected responses for each task in fine-tuning the LLM.

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:	You are an expert in definition recognition. 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} 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 5: Fine-tune task-specific system prompts and input templates

Training strategy	Features (X)	Label (Y)
Foundational lnowledge	the straight-line method.	The straight-line depreciation method evenly distributes the cost of an asset over its useful life. To determine the annual depreciation expense, the asset's initial purchase price and its estimated salvage value at the end of its lifespan are considered. The difference between these values represents the total depreciable amount, which is then divided by the number of years the asset is expected to be in use. This results in a consistent annual expense, helping businesses systematically allocate costs and assess the asset's declining value over time.
Question-answering pattern	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.How do you calculate the annual depreciation expense? Provide the solution.	Annual Depreciation = (Purchase Price- Salvage Value) / Useful Life = (7229.15 860.73)/2 = 3184.21
Stylistic-answer adaptation	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

Table 6: Examples of features and labels for each training strategy in fine-tuning LLM

Task		Query	Expected response in stylistic-answer format
Abbreviation	Ticker	EIG	Answer: Employers Holdings Inc.
	Acronym	EMIR	Answer: European Market Infrastructures Regulation
Definition		Financial Stability Oversight Council (FSOC) Collaboration	Answer: The Federal Reserve works closely with the FSOC, a body established to identify and monitor systemic risks to the financial system. While FSOC oversees macroprudential risks, the Fed focuses on ensuring financial institutions comply with regulations that maintain system stability.
NER		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.	Answer: {"Organizations":["European Parliament","Council of the European Union"],"Legislations":["Regulation (EU) No 648/2012"],"Dates":["4 July 2012","16 August 2012"],"Monetary Values":[],"Statistics":[]}
QA		What two major segments is the insurance industry divided into?	Answer: Property/casualty insurance
Link retrieval		Advertisement of Membership (Part 328)	Answer: https://www.ecfr.gov/current/title-12/chapter-III/subchapter-B/part-328
CFA		Question: The nominal risk-free rate is best described as the sum of the real risk-free rate and a premium for: A. Maturity, B. Liquidity, C. Expected Inflation	Answer: C. Expected Inflation
XBRL	Terminology	Authoritative Reference	Answer: Citations to specific authoritative accounting literature (pronouncements, standards, rules, and regulations) derived from various authoritative sources (Securities and Exchange Commission, Financial Accounting Standards Board, American Institute of Certified Public Accountants, etc.) and used to help define an element.
	Financial math	A project expects annual cash inflows of $6,000 for 4 years. If the discount rate is 8%, what is the NPV of the project?	Answer: 21462.58
	Domain numeric query	What is the FY2019 fixed asset turnover ratio for Activision Blizzard? Fixed asset turnover ratio is defined as: FY2019 revenue / (average PP&E between FY2018 and FY2019). Round your answer to two decimal places. Base your judgments on the information provided primarily in the statement of income and the statement of financial position.	Answer: 24.26
	XBRL tag query to XBRL reports	What is the US GAAP XBRL tag for Accounts Payable as reported by Verizon Communications Inc for the Fiscal Year ending in FY 2023? (Response format: XBRL tag, e.g., 'Answer: us-gaap:Depreciation')	Answer: us-gaap:AccountsPayableAndAccruedLiabilitiesCurrent
	Financial ratio formula with XBRL tags	What is the value of Chevron Corp's Net Profit Margin for the Fiscal Year ending in FY 2021? (Response format: Percentage, e.g., '45.5%')	Answer: 9.62%
CDM		What Is the Process Model?	Answer: The CDM Process Model has been designed to translate the technical standards that support those industry processes into a standardised machine-readable and machine-executable format. Machine readability and executability is crucial to eliminate implementation discrepancy between market participants and increase interoperability between technology solutions. It greatly minimises the cost of adoption and provides a blueprint on which industry utilities can be built.
MOF	License OSI approval	BSD 3-Clause License	Answer: Yes
	License abbreviation	APL-1.0	Answer: Adaptive Public License 1.0
	Detail QA	What type of license is the Apache License, Version 2.0?	Answer: The Apache License is a permissive free software license written by the Apache Software Foundation.

Table 7: Examples of queries and expected responses