Domain*o*1s: Guiding LLM Reasoning for Explainable Answers in High-Stakes Domains

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are widely applied to downstream domains. However, current LLMs for high-stakes domain tasks, such as financial investment and legal QA, typically generate brief answers without reasoning processes and explanations. This limits users' confidence in making decisions based on their responses. While original CoT shows promise, it lacks self-correction mechanisms during reasoning. This work introduces Domain*o*1s, which enhances LLMs' reasoning capabilities on domain tasks through supervised fine-tuning and tree search. We construct CoT-stock-2k and CoT-legal-2k datasets for fine-tuning models that activate domain-specific reasoning steps based on their judgment. Additionally, we propose Selective Tree Exploration to spontaneously explore solution spaces and sample optimal reasoning paths to improve performance. We also introduce PROOF-Score, a new metric for evaluating domain models' explainability, complementing traditional accuracy metrics with richer assessment dimensions. Extensive experiments on stock investment recommendation and legal reasoning QA tasks demonstrate Domain*o*1s's leading performance and explainability. Our code is available at `https://anonymous.4open.science/r/Domaino1s-006F/`.

1 Introduction

In specific domains such as finance (Xing, 2024; Jeong, 2024; Cheng et al., 2024c), law (Cheong et al., 2024; Colombo et al., 2024), and biomedicine (Labrak et al., 2024; Wang et al., 2023a), Large Language Models (LLMs) are widely used for tasks like recommendation (e.g., stock investment recommendation (Koa et al., 2024; Qin et al., 2024; Takayanagi et al., 2023)) and question answering (e.g., legal reasoning QA (Guha et al., 2024; Wang et al., 2023b; Ujwal et al., 2024)). However, popular approaches mainly

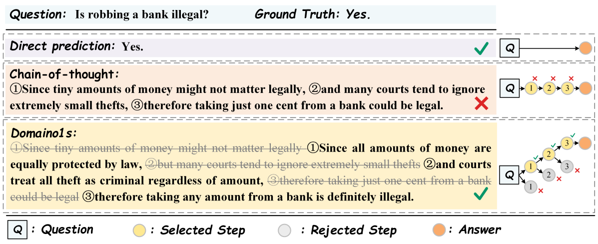

Figure 1: Comparison of Domain*o*1s and other paradigms on a demonstrative example. Domain*o*1s expands reasoning paths and obtains optimal ones through tree search.

adopt direct prediction paradigms that immediately generate brief answers to questions (Cheng et al., 2024a,c; Yue et al., 2023), leading to answers lacking explainability. In practical applications within high-stakes domains like finance and law, users may not trust results lacking explainability (Biran and McKeown, 2017) to guide decision-making. While Chain-of-Thought (CoT) reasoning demonstrates the ability to enhance models' step-by-step thinking and domain problem solving (Li et al., 2024b; Jiang and Yang, 2023; Miao et al., 2024) and provides explainable reasoning processes, its single-pass generated reasoning chains lack error correction mechanisms. If errors occur in early reasoning steps, the model continues reasoning along the flawed path, affecting the subsequent reasoning process, as shown in Figure 1. This poses challenges for solving domain tasks, as flawed reasoning processes may introduce legal and ethical risks.

Recently introduced o1-like models (OpenAI, 2024; OpenO1 Team, 2024; Zhao et al., 2024), with their exceptional reasoning capabilities, demonstrate powerful performance surpassing reasoning methods like CoT in mathematics, physics, and coding. Compared to LLMs using CoT, o1-like models feature longer reasoning chains and reasoning time. They are considered to perform multi-stage reasoning rather than generating complete reasoning chains in single-pass, which enhances the accuracy of LLM reasoning. However, despite high-stakes domains requiring high-quality reasoning, extending o1-like models' capabilities to these domains remains an unexplored research gap.

In this paper, we design Domain$o$1s to provide explainable answers for high-stakes domain problems. Domain$o$1s includes two model variants, Domain$o$1s-finance and Domain$o$1s-legal. As shown in Figure 1, Domain$o$1s can (1) perform autonomous step-by-step reasoning, and (2) expand reasoning paths through tree search to obtain optimal ones. To achieve (1), we use GPT-4o (OpenAI, 2024) to generate CoT data and construct CoT-stock-2k and CoT-legal-2k datasets for supervised fine-tuning. During dataset construction, we employ 26 special tokens (e.g., <SUMMARY>) to prompt GPT-4o to distinguish different steps in the reasoning process explicitly. In the supervised fine-tuning process, we remove these special tokens from the answers, enabling the model to autonomously select and organize intermediate steps in the reasoning chain. To achieve (2) during answer generation, we introduce a novel Selective Tree Exploration method to find the optimal reasoning paths. This method uses the average perplexity of tokens in each reasoning step to decide whether to explore new paths and select the best path. Compared to traditional search methods (Weng et al., 2022; Jiang et al., 2023; Chen and Liu, 2024), Selective Tree Exploration balances search performance and time cost. We evaluate Domains$o$1s on stock investment recommendation (Koa et al., 2024) and legal reasoning QA (Guha et al., 2024) datasets. Unlike most domain benchmarks (Koa et al., 2024; Yang et al., 2022; Guha et al., 2024), we point out that focusing solely on answer accuracy makes it difficult to determine whether models properly reason through given contexts rather than relying on shortcuts or overfitting. We emphasize the necessity of evaluating domain models' explainability and introduce a new evaluation metric PROOF-Score (Principled rating for reasoning completeness, domain safety, and factual accuracy) to fill this gap. Results show that Domain$o$1s improves reasoning accuracy while providing high-quality, explainable reasoning processes. Our contributions are:

• Domain$o$1s is proposed for explainable answers, with two model variants.

• CoT-stock-2k and CoT-legal-2k datasets are constructed for fine-tuning. Selective Tree Exploration is proposed as a reasoning path search method that balances performance and time cost.

• PROOF-Score is proposed to evaluate the explainability of domain model answers, introducing a new perspective for domain model evaluation.

• Domain$o$1s achieves leading performance, demonstrating the effectiveness of its reasoning capabilities in solving high-stakes domain tasks.

2 Related Works

2.1 LLMs for Specific Domains

LLM applications in specific domains typically follow three approaches: training from scratch, fine-tuning, and prompt learning. While training from scratch (e.g., BloombergGPT (Wu et al., 2024)) shows promising results, it requires significant computational resources and data (Yang et al., 2023; Ling et al., 2023; Xie et al., 2023b). Fine-tuning emerges as a cost-effective alternative, with researchers using GPT-4 (Li et al., 2024a) or low-cost automated methods (Cheng et al., 2024c; Koa et al., 2024) to generate fine-tuning data. Prompt learning methods enhance model capabilities without parameter modification through template engineering or knowledge retrieval (Li et al., 2023; Cui et al., 2023; Huang et al., 2023), such as CoT (Wei et al., 2022) reasoning. o1-like models are typically constructed to equip LLMs with CoT reasoning capabilities through fine-tuning, followed by multi-pass search to obtain better reasoning paths.

2.2 Single-Pass vs. Multi-Pass

Prompt-based methods like CoT (Wei et al., 2022; Zhang et al., 2022; Lyu et al., 2023) improve single-pass reasoning through better prompt templates. However, errors in intermediate reasoning steps can propagate through the chain. In contrast, search-based methods explore multiple reasoning paths in the solution space, treating each reasoning step as a node in the tree, and selecting the optimal path to improve reasoning quality (Qi et al., 2024).

2.3 Sampling Reasoning Paths

Research on mathematical reasoning (Brown et al., 2024; Wang et al., 2024) indicates that sampling different reasoning paths can improve performance compared to greedy one-time decoding. Best-of-N search (Weng et al., 2022; Jiang et al., 2023) generates N complete answers, allowing LLM to select the best response based on final results, but may miss high-quality intermediate reasoning steps (Xie et al., 2024; Chen et al., 2024b). Sentence-level

Example 1: Stock Investment Recommendation	***Answer from Qwen-2.5***	***Answer from Domaino1s-finance***
Question: For a given set of facts, estimate their overall impact on Apple's stock price movement. Answer with Positive/Negative. Here are the facts: 2022-09-29 - Bank of America downgrades Apple from "Buy" to "Neutral." - Apple's stock plummets, down by over 5% on the day. ⋯ ***Ground Truth: Positive.***	The overall impact of the given facts on Apple's stock price movement is **predominantly negative**. The facts indicate a series of downgrades, warnings about demand, negative market reactions, and other unfavorable factors, such as stock prices hitting new lows and market performance drags. **These have had a negative impact on Apple's stock price**. **Negative** ✗	<SUMMARY> ⋯ We will examine several factors to drive the analysis, including historical performance, valuation metrics, market position, strategic initiatives, investor sentiment, risks, and recent performance. </SUMMARY> <HISTORICAL CONTEXT> Historically, Apple's stock has shown resilience but faced challenges this year, down 22% by mid-September, impacting overall market sentiment. ⋯ </HISTORICAL CONTEXT> ⋯ <INVESTOR SENTIMENT> Investor sentiment appears mixed, with notable trading activity in options reflecting apprehension amid stage-four challenges. ⋯ </INVESTOR SENTIMENT> ⋯ <OVERALL IMPACT> Despite the confluence of negative operational, strategic, and market pressures, the overall impact on Apple appears tilted towards a positive outlook. ⋯ **Price Movement: Positive.** </OVERALL IMPACT> ✓
Example 2: Legal Reasoning QA	***Answer from Qwen-2.5***	***Answer from Domaino1s-legal***
Question: Here is a service agreement: **If you are a consumer**, the laws of the country in which you reside will apply to any claim ⋯ and you may resolve your claim in any competent court in that country that has jurisdiction over the claim.⋯ As a **Facebook user who lives in Canada**, can I sue Facebook in Canada? ***Ground Truth: Yes.***	If you are a consumer and reside in a country other than the United States, any claim against Facebook must be resolved in the competent court in the country where you reside. **However**, **for non-consumer cases**, including those from Canadian residents, the claim must be resolved exclusively in the U.S. District Court for the Northern District of California or a state court located in San Mateo County, California. **No.** ✗	<SUMMARY> ⋯ For non-consumers, disputes must be resolved in U.S. courts. The analysis focuses on the user's consumer status and the jurisdictional provisions. </SUMMARY> <REASONING> 1. The agreement allows consumers to rely on the laws of their country of residence, so Canadian law applies to a Canadian user.⋯ </REASONING> <EVIDENCE> - The agreement states: "If you are a consumer, the laws of the country in which you reside will apply" and "you may resolve your claim in any competent court in that country." ⋯ </EVIDENCE> <CONCLUSION> A Canadian Facebook user classified as a consumer can sue Facebook in Canada. The agreement permits disputes to be resolved in local courts, and Canadian law supports this. **The answer is: Yes.** </CONCLUSION> ✓

Figure 2: Comparison between the base model Qwen-2.5-Instruct (Qwen-Team, 2024) and Domain*o*1s. The base model shows notable reasoning errors. In contrast, Domain*o*1s breaks problems into multiple reasoning steps and reaches well-supported conclusions through systematic analysis. Details in Appendix C.

Beam Search (Chen and Liu, 2024) generates multiple candidate sentences, selects the best one, and iteratively continues this process, but may get stuck in local optima. Stage-level Beam Search (Xu et al., 2024) offers a compromise by generating and selecting optimal candidates for each reasoning step rather than sentences.

3 Methodology

In this section, we first present the formal definition of LLM-based multi-step reasoning. Then, we introduce Domain*o*1s from two aspects: enhancing reasoning capabilities and solution expansion & sampling. For aspect 1, Domain*o*1s facilitates a progressive reasoning process. For aspect 2, Domain*o*1s improves reasoning performance through tree search to obtain optimal reasoning paths. A comparison of reasoning examples with the base model is shown in Figure 2.

3.1 Preliminaries

For a given question q, the solution process can be decomposed into multiple reasoning steps. Consider a complete solution consisting of up to T reasoning steps. The state S_t comprising all reasoning steps from step 0 to t can be represented as:

$$S_t = \{s_0, s_1, \dots, s_t\}, 0 \leq t < T, t \in \mathbb{Z}, \quad (1)$$

where s_t represents the t-th reasoning step, state S_t represents the collection of reasoning processes from step 0 to t. An action $a_t (0 \leq t < T-1)$ is defined as choosing the next reasoning step s_{t+1}. The LLM constitutes a policy model, where the transition $f(S_{t+1}|a_t, S_t)$ from one state to the next is implemented by auto-regressively generating s_{t+1} through the input sequence. To guide the LLM in selecting more reasonable subsequent reasoning step s_{t+1}, a value function $V(s_{t+1})$ is defined to evaluate the expected return of LLM's strategy.

3.2 Enhancing Reasoning Capabilities

To enhance Domain*o*1s's reasoning capabilities in high-stakes domains (finance and legal), we employ supervised fine-tuning to let the model generate CoT-style responses. Since existing domain datasets or databases lack the detailed reasoning processes required for training Domain*o*1s models, we constructed two new datasets, CoT-stock-2k and CoT-legal-2k, using the training sets from stock investment recommendation (Koa et al., 2024) and legal reasoning QA (Guha et al., 2024; Li et al., 2022) datasets respectively. The construction details are as follows:

Stock Investment Recommendation. Contains price data and tweet information from the top 5 stocks across 11 industries during 2020-2022. The task is to predict stock price movement (positive or negative) for the next trading day based on facts extracted from tweets over the past 5 days. Due to the high volume of daily tweets, we fine-tuned Qwen-2.5-Instruct (Qwen-Team, 2024) to generate daily tweet summaries. We utilized GPT-4o (OpenAI, 2024) to generate CoT data, explicitly prompting it to decompose the answer generation process into 10 structured reasoning steps, including market factors (Fama and French, 1993), company strategies (Porter and Kramer, 1985), and investor sentiment (Baker and Wurgler, 2006):

• **Summary**: Extract key facts from tweets about question q, identify main analysis focus. • **Historical context**: Review historical performance and market context. • **Valuation**: Assess current valuation metrics (e.g., P/E, price targets, market views). • **Market size and dominance**: Evaluate company's industry standing and influence. • **Strategic initiatives**: Review recent strategic moves (partnerships, innovation) and growth potential. • **Investor sentiment**: Gauge investor mood through trading patterns and market discussion. • **Risks and concerns**: Identify key investor concerns and risk factors. • **Recent performance**: Analyze recent price movements and drivers. • **Consolidation**: Review financial/stock structure changes (buybacks, profitability). • **Overall impact**: Synthesize all analysis points, clearly indicate overall impact, and provide a final prediction (positive or negative) for the next trading day's stock price.

Legal Reasoning QA. Includes legal reasoning questions across multiple categories such as legal rule application, reasoning, and legal question classification, presented as multiple choice or true/false questions. We utilized GPT-4o (OpenAI, 2024) to generate CoT data, explicitly prompting it to decompose the answer generation process into 4 structured reasoning steps:

• **Summary**: Extract key points from question q and identify analysis focus. • **Reasoning**: Apply step-by-step logic to reach answers. • **Evidence**: Systematically present supporting text and verify reasoning. • **Conclusion**: Synthesize the analysis and state the final answer.

When explicitly prompting GPT-4o to generate multiple structured reasoning steps, we require the model to use special tokens (e.g., <SUMMARY>) for segmentation. However, we want Domaino1s to organize and initiate necessary steps independently during reasoning to maintain general capabilities. Therefore, we remove all special tokens from the answers during supervised fine-tuning. After training, the model activates each reasoning step based on its own judgment.

3.3 Solution Expansion & Sampling

After supervised fine-tuning, the model can output responses in CoT format. To further enhance the model's reasoning abilities, we enable the model to

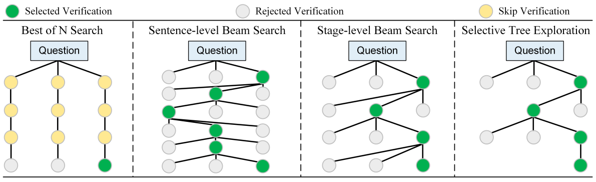

Figure 3: Solution expansion & sampling illustration. Best-of-N search generates N complete responses and selects the best one; Sentence-level Beam Search generates multiple candidates for each sentence and selects the best one; Similarly, Stage-level Beam Search generates multiple candidates for each reasoning step and selects the best one. In contrast, our Selective Tree Exploration dynamically expands each reasoning step node, explores multiple reasoning steps as candidates only when necessary, and selects the best option at each step. Our method balances search performance and computational time overhead.

explore the solution space, and autonomously expand and sample reasoning paths. During sampling, we introduce $V(s_{t+1})$ to evaluate the expected return of reasoning step s_{t+1}. Although $V(s_{t+1})$ can be constructed through direct introduction or training of step-level reward models (Chen et al., 2024a; Xie et al., 2024; Xu et al., 2024), this creates additional model training and inference overhead. In our implementation, we use a more direct but effective approach. We introduce the perplexity p of LLM when generating s_{t+1} to serve as $V(s_{t+1})$:

$$p = \exp(-\frac{1}{N}\sum_{i=1}^{N}\log(\frac{e^{z_{i,k}}}{\sum_{j=1}^{M} e^{z_{i,j}}})), \quad (2)$$

where N represents the number of tokens in s_{t+1}, $z_{i,k}$ is the logit value of the actually generated token k at position i, $z_{i,j}$ is the logit value of candidate token j at position i, M is the vocabulary size representing the number of all candidate tokens, and $\frac{e^{z_{i,k}}}{\sum_{j=1}^{M} e^{z_{i,j}}}$ is the softmax probability of the actually generated token. Overall, we propose Selective Tree Exploration for solution expansion & sampling, following these phases:

(1) Calculate the perplexity value p of tokens at the generation step.

(2) If $p \geq \theta$ (θ is the sampling threshold), regenerate the step until $p < \theta$ or reach the maximum regeneration count K (i.e., maximum beam size). If p of all K generations are no less than threshold θ, greedily sample the candidate with minimum p from the K candidates.

(3) Continue to generate the next step based on the selected step, repeat phases **(1)-(3)** until the complete answer is generated.

As shown in Figure 3, compared to Best-of-

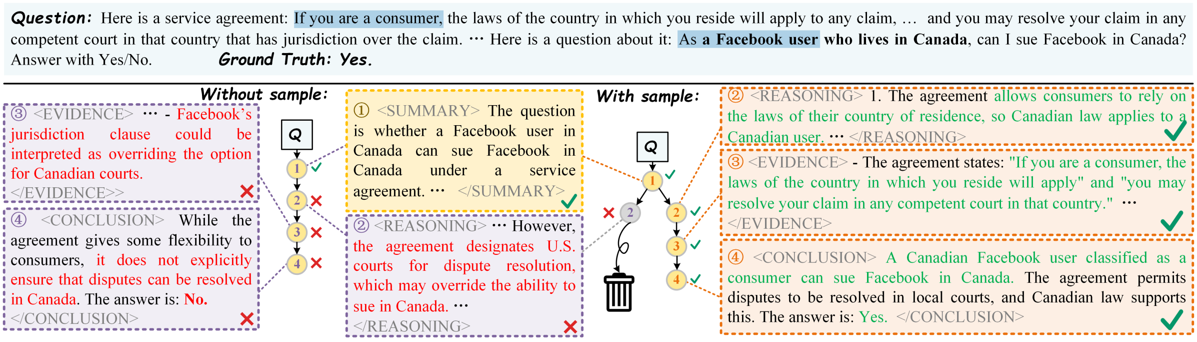

Figure 4: The role of solution expansion & sampling. Intermediate steps in single inference (without sample) may contain errors, while solution expansion & sampling can find better reasoning paths.

N Search (Weng et al., 2022; Jiang et al., 2023), Sentence-level Beam Search (Chen and Liu, 2024) and Stage-level Beam Search (Xu et al., 2024), Selective Tree Exploration balances search performance and time cost. When θ is set to 0, Selective Tree Exploration becomes Stage-level Beam Search as it explores K paths at each step. When θ is set to an extremely large value, Selective Tree Exploration degenerates into CoT with a single inference chain. In other cases, Selective Tree Exploration only expands reasoning paths when necessary, which reduces unnecessary overhead.

To illustrate the role of solution expansion & sampling, as shown in Figure 4, when inference without sampling, although the model generates the reasoning process, errors in intermediate steps (starting from <REASONING>) lead to error accumulation, ultimately resulting in incorrect results. Through exploration and expansion of solution paths, better reasoning paths can be found, leading to more accurate results.

4 Experiments

In this section, we evaluate the performance of Domain$o$1s on stock investment recommendation and legal reasoning QA tasks. Our work aims to address the following questions: **RQ1**: How does Domain$o$1s perform in answer accuracy compared to other LLM methods? **RQ2**: What are the limitations of accuracy-based evaluation metrics in domain tasks, and how can we better evaluate model performance? **RQ3**: How do fine-tuning and solution expansion & sampling help improve the performance of Domain$o$1s?

4.1 Experimental Settings

Baselines. To validate Domain$o$1s's performance on high-stakes domain tasks, we compare it with general purpose LLMs and domain LLMs trained or fine-tuned with domain data.

General Purpose LLMs: We choose Qwen-2.5-Instruct (Qwen-Team, 2024) and Llama-3-Instruct (AI@Meta, 2024) as general purpose LLM baselines due to their remarkable performance on many downstream tasks. We also select OpenO1-Llama and OpenO1-Qwen (OpenO1 Team, 2024) as representatives of o1-like model baselines.

Financial Domain LLMs: Finance-LLM (Cheng et al., 2024c), Finance-Chat (Cheng et al., 2024c), Finance-Llama-3 (Cheng et al., 2024b), FinGPT-Forecaster (Yang et al., 2023), Llama-2-taiwan-btc (Lanz, 2024), and SEP (Koa et al., 2024).

Legal Domain LLMs: Open-Australian-Legal-LLM (Butler, 2023), DISC-LawLLM (Yue et al., 2023), Law-LLM (Cheng et al., 2024c), Law-Chat (Cheng et al., 2024c), and Lawma (Dominguez-Olmedo et al., 2024).

Datasets. For the stock investment recommendation task, we select the stock prediction dataset provided by Koa et al. (Koa et al., 2024). This dataset contains price data and tweet information for the top 5 stocks from 11 industries during 2020-2022, comprising 7,866 test question entries. The task is constructed to predict whether a stock will rise or fall on the next trading day based on facts contained in tweets from the previous 5 days. Any neutral answers are considered incorrect. Due to the large volume of daily tweets, we fine-tune Qwen-2.5-Instruct (Qwen-Team, 2024) to generate daily tweet summaries and apply these summaries as input for all models.

For the legal reasoning QA task, we select LegalBench (Guha et al., 2024), a dataset composed of numerous legal QA datasets and benchmarks. LegalBench includes 5 categories of legal tasks. We select three reasoning-related categories: Rule-application/Rule-conclusion, Interpretation, and Rhetorical-understanding, encompassing 9 datasets

Model	Model size	Interpretation					Rule-application/ Rule-conclusion	Rhetorical-analysis			Avg.
		CC	CAUD	MAUD	PP	IP	PJ	Scalr	TTP	TTD	
Qwen-2.5-Instruct	7B	86.36	80.08	78.75	52.38	48.12	64.00	78.98	99.07	76.96	73.86
Llama-3-Instruct	8B	85.86	81.20	67.43	61.63	50.37	54.00	75.83	**100.00**	78.18	72.72
OpenO1-Llama	8B	85.10	81.31	74.54	62.36	50.37	60.00	80.03	91.52	77.58	73.65
OpenO1-Qwen	7B	84.85	80.13	79.11	59.27	48.87	66.00	80.38	88.78	76.64	73.78
Open-Australian-Legal	1.5B	0.00	0.00	1.20	17.64	1.50	22.00	0.00	0.00	0.00	4.70
DISC-LawLLM	13B	50.00	32.98	64.77	48.09	19.55	56.00	70.05	5.60	20.60	40.85
Law-LLM	7B	10.86	1.59	30.87	3.05	2.26	0.00	58.49	8.41	13.33	14.32
Law-Chat	7B	80.30	**82.31**	39.75	51.69	33.83	48.00	76.36	54.21	52.73	57.69
Lawma	8B	47.73	34.14	69.93	53.31	47.37	36.00	78.46	6.54	26.67	44.46
Domain-CoT-legal	7B	87.88	80.59	**80.47**	65.81	50.37	70.00	86.69	94.40	77.58	77.09
Domain$o$1s-legal	7B	**88.64**	81.76	80.33	**66.54**	**52.63**	**72.00**	**88.97**	95.33	**78.78**	**78.33**

Table 1: Model accuracy (%) on legal reasoning QA tasks. Avg. represents the mean accuracy across all tasks.

Model	Model Size	Accuracy	MCC
Qwen-2.5-Instruct	7B	51.18	-0.017
Llama-3-Instruct	8B	51.41	0.017
OpenO1-Llama	8B	50.87	0.014
OpenO1-Qwen	7B	51.02	0.010
Finance-LLM	7B	48.05	-0.075
Finance-Chat	8B	47.16	-0.004
Finance-Llama-3	8B	49.03	-0.047
FinGPT	7B	46.13	0.016
Llama-2-taiwan-btc	7B	50.66	-0.002
SEP	7B	48.35	0.018
Domain-CoT-finance	7B	51.52	0.020
Domain$o$1s-finance	7B	**51.98**	**0.021**

Table 2: Model accuracy (%) and MCC on stock investment recommendation tasks.

with a total of 35,053 test questions. Question types include true/false and multiple-choice questions.

Implementation Details. In this work, our Domain$o$1s is developed based on Qwen-2.5-Instruct (Qwen-Team, 2024). During the fine-tuning phase for enhancing reasoning capabilities, we set the learning rate, epoch, batch size, gradient accumulation, and maximum tokens length to 5e-5, 120, 2, 2, and 2048 respectively. The θ and K in the sampling process are set to 1.1 and 2 respectively. The experimental hardware, software, and other configuration details can be found in Appendix A.

4.2 Prediction Performance (RQ1)

In this section, we compare Domain$o$1s with relevant baselines to evaluate the answer accuracy.

Table 1 and Table 2 report the quantitative results for legal reasoning QA and stock investment recommendations tasks respectively. For all models where answers cannot be directly parsed from responses, we use GPT-3.5-turbo-16k (Ouyang et al., 2022) to extract the chosen options from responses for fair comparison. Additionally, given that not all stock price movements are necessarily caused by the provided text, accuracy results may not fully indicate a model's reasoning capabilities, as they include some random guesses for non-informative text (Koa et al., 2024). Following stock prediction research (Ding et al., 2015; Feng et al., 2018), we also calculate the Matthews Correlation Coefficient (MCC) as an evaluation metric, which considers the ratios of true and false positives and negatives (Chicco and Jurman, 2020; Chicco et al., 2021). We observe that Domain$o$1s outperforms its base model Qwen-2.5-Instruct on almost all tasks, despite being fine-tuned on only a small amount of data. Moreover, Domain$o$1s and Domain-CoT (model with reasoning-enhanced fine-tuning, without solution expansion & sampling) achieve the best accuracy or MCC on nearly all tasks, even surpassing LLMs that are carefully designed and trained on domain datasets, especially on legal reasoning tasks as shown in Table 1. Although these legal LLMs learn domain knowledge through pre-training or fine-tuning, they lack the reasoning capability to derive correct answers, in contrast to our models. We also analyze the reasoning chain length and inference time of Domain$o$1s and baselines, see Appendix D.

4.3 Explainability Evaluation Pipeline (RQ2)

In previous research, most domain tasks use accuracy as the primary evaluation metric (Koa et al., 2024; Yang et al., 2022; Guha et al., 2024). This evaluation metric makes it difficult to distinguish between models that truly understand and reasonably utilize context and those that simply rely on partial text or overfit on pre-trained domain knowledge (Zhang et al., 2024; Bordt et al., 2024). We sample two subsets from the test sets of stock investment recommendation and legal reasoning QA, with details available in Appendix F.

Model		Finance-Llama-3			Domain$o$1s-finance		
Tweets		Response					
Pos.	Neg.	Pos.	Neg.	Acc	Pos.	Neg.	Acc
0.1	0.9	0.018	0.982	50.45	0.173	0.827	50.91
0.3	0.7	0.182	0.818	48.64	0.391	0.609	49.55
0.5	0.5	0.545	0.455	48.18	0.527	0.473	49.09
0.7	0.3	0.882	0.118	51.36	0.777	0.223	50.91
0.9	0.1	1.000	0.000	50.00	0.882	0.118	50.45

Table 3: Model's accuracy and prediction ratios for Positive (Pos.) and Negative (Neg.) of the stock trend under varying proportions of Pos. and Neg. tweets in stock investment prediction tasks.

Question: Does Microsoft sometimes enable educational institutions that provide their students with access to Microsoft services to monitor students activities on such services? **Grund Truth: Yes.**

Law-Chat (Acc: 26.2%):

Microsoft **does not allow** educational institutions to monitor student activities on their services. While Microsoft provides tools to help educational institutions manage student accounts and usage, these tools are for management purposes only, not for monitoring. Educational institutions may have their own policies to monitor student activities on their devices, but this is separate from Microsoft's services. ✗

Domaino1s-legal (Acc: 3.6%):

<SUMMARY>
Answering this requires access to Microsoft's official documentation regarding their educational institution policies and monitoring permissions, which I currently cannot verify. Without access to Microsoft's official policies, **I cannot make claims about their institutional monitoring capabilities**.
</SUMMARY> ✓

Figure 5: Models responses and accuracy on legal reasoning QA tasks after removing key conditions.

Stock		Legal	
Qwen-2.5-Instruct	6.281	Qwen-2.5-Instruct	3.428
Llama-3-Instruct	6.129	Llama-3-Instruct	3.417
OpenO1-Llama	6.212	OpenO1-Llama	6.554
OpenO1-Qwen	6.227	OpenO1-Qwen	6.588
Finance-LLM	6.023	Open-Australian-Legal	5.152
Finance-Chat	5.583	DISC-LawLLM	0
Finance-Llama-3	5.965	Law-LLM	3.838
FinGPT	3.413	Law-Chat	3.339
Llama-2-taiwan-btc	0	Lawma	0
SEP	6.182		
Domain$o$1s-finance	**6.359**	**Domain$o$1s-legal**	**6.677**

Table 4: Comparison of explanation quality (PROOF-Score) between Domain$o$1s and baselines. For models that generate responses containing no explanations, their PROOF-Scores are set to 0.

In the stock investment recommendation task, stock tweets are manually classified into Positive and Negative tweets and combined in different ratios as model inputs. We compare Domain$o$1s-finance with Finance-Llama-3. As shown in Table 3, when the Positive:Negative ratio of tweets is 0.5:0.5, models' responses maintain a similar 1:1 ratio between Positive and Negative predictions. However, when either Positive or Negative tweets dominate the input, Finance-Llama-3 typically ignores tweets with the opposite sentiment and bases its answer solely on the majority sentiment. In contrast, Domain$o$1s-finance still considers the minority sentiment tweets and generates answers by comprehensively evaluating all tweets. However, both models achieve similar accuracy, making it challenging to determine through accuracy metrics alone whether the models truly understand and reasonably utilize the context in the inputs, rather than overfitting or hallucinating. For the legal reasoning task subset, key conditions are removed from the question text, making it impossible to answer the tasks correctly. As shown in Figure 5, although Law-Chat achieves higher accuracy than Domain$o$1s-legal, its answers are mostly random responses generated from overfitted legal knowledge, while Domain$o$1s-legal refuses to answer due to the absence of necessary reasoning conditions, resulting in an accuracy close to 0. This indicates that accuracy alone is insufficient to determine whether models blindly overfit using domain knowledge to generate irrelevant answers.

For high-stakes domain tasks such as stock investment recommendations and legal reasoning QA, non-transparent text comprehension or inadequate reasoning processes may lead to wrong conclusions or generate advice that violates ethical or legal principles. To evaluate the explainability of domain model responses, we propose a new evaluation metric called PROOF-Score (Principled rating for reasoning completeness, domain safety, and factual accuracy). PROOF-Score uses GPT-4o (OpenAI, 2024) to generate a score from 1 to 7 for response, considering three aspects:

- **Reasoning Completeness (RC)**: Evaluates the completeness and logical coherence.
- **Domain Safety (DS)**: Measures the safety and appropriateness in specific domains.
- **Factual Accuracy (FA)**: Evaluates the factual accuracy of statements.

Detailed prompts can be seen in Appendix E. Here, we define:

$$\text{PROOF-Score} = \frac{RC + DS + FA}{3}. \quad (3)$$

Table 4 shows PROOF-Scores of models on two tasks. Domain$o$1s achieves the highest scores on both tasks, even though we do not train specifically for these three metrics. This indicates Domain$o$1s can inherently consider these factors to generate better responses. We also observe that even when a model's response is incorrect in terms of results, GPT-4o may still give a high PROOF-Score because these responses contain clear and reasonable logic. This may be inappropriate for tasks requiring strict accuracy, where prediction accuracy should be considered the primary metric. However, for

Method	**Acc**	**time(s)**
w/o Sample	86.69	8.35
Best-of-N Search	87.56	40.26
Sentence-level Beam Search	84.93	334.20
Stage-level Beam Search	88.44	133.68
Selective Tree Exploration	**89.14**	**15.18**

Table 5: Accuracy (%) and average inference time comparison between our Selective Tree Exploration and other search methods on the Scalr dataset. Our method (with $\theta = 1.1$) outperforms other approaches under the same beam size settings.

Method	K	**Acc**	**time(s)**
w/o Sample	1	86.69	8.35
Selective Tree Exploration	2	88.97	**24.88**
	3	89.14	45.77
	4	89.84	72.55
	5	**90.01**	93.95

Table 6: Accuracy (%) and average inference time of Domain$o1$s-legal on the Scalr dataset under different beam size K settings. θ is set to 1.05.

Method	θ	**Acc**	**time(s)**
w/o Sample	10000	86.69	8.35
Selective Tree Exploration	1.4	87.21	**8.40**
	1.3	87.91	8.44
	1.2	88.97	10.16
	1.1	89.14	15.18
	1.0	**89.49**	51.03

Table 7: Accuracy (%) and average inference time of Domain$o1$s-legal on the Scalr dataset under different sampling threshold θ settings. K is set to 3.

tasks lacking standard answers or without unique correct answers (e.g., long-term investment advice, asset allocation recommendations), using PROOF-Score becomes effective in evaluating the explainability of model responses.

4.4 Ablation Study (RQ3)

In this section, we evaluate the impact of fine-tuning and solution expansion & sampling on Domain$o1$s's performance. We primarily focus on accuracy metrics in this section, while presenting explainability analysis in Appendix G.

Enhancing Reasoning Fine-tuning. As shown in Table 1 and Table 2, Domain-CoT represents the model configuration using only reasoning-enhanced fine-tuning without solution expansion & sampling. Compared to the base model Qwen-2.5-Instruct, Domain-CoT achieves performance improvements on almost all datasets, which demonstrates that reasoning-enhanced fine-tuning improves the model's reasoning capabilities on domain tasks.

Solution Expansion & Sampling. Table 5 shows the performance comparison on Scalr (a dataset in LegalBench) between best-of-N search (Weng et al., 2022; Jiang et al., 2023), Sentence-level Beam Search (Chen and Liu, 2024), Stage-level Beam Search (Xu et al., 2024), and our Selective Tree Exploration. The baseline search methods use the setup from Xu et al. (Xu et al., 2024), which uses the policy model to evaluate the relative quality of reasoning chains or steps, in contrast to our perplexity-based approach. Results demonstrate that under the same beam setting of $K = 3$, Selective Tree Exploration achieves comparable or better performance compared to all baseline approaches (with and without search) while requiring less computational time for inference than other search methods.

To better illustrate the effectiveness of our Selective Tree Exploration as exploration paths increase, we evaluate model performance under different settings of K and θ on the Scalr dataset. As shown in Table 6, using Selective Tree Exploration brings performance improvements compared to methods without sampling ($K = 1$). Model accuracy improves as K increases, indicating that our Selective Tree Exploration is scalable. As shown in Table 7, model accuracy improves as θ decreases, as this similarly expands the paths explored by Selective Tree Exploration. However, both increasing K and decreasing θ lead to longer inference time. Due to computational resource constraints, we only set $K = 2$, $\theta = 1.1$. However, we demonstrate that increasing beam size K and decreasing sampling threshold θ will lead to performance improvements.

5 Conclusion & Future Works

In this work, we introduce Domain$o1$s and its two model variants for finance and legal domains, guiding LLMs towards explainable high-stakes domain answers. We construct two datasets to fine-tune Qwen-2.5-Instruct and propose Selective Tree Exploration for enabling LLMs to perform multi-stage reasoning. The superior performance on datasets demonstrates Domain$o1$s's exceptional potential in high-stakes domains.

In future work, we plan to build larger training datasets to enhance domain models' reasoning abilities. We also plan to create Domain$o1$s variants using domain-specific pre-trained base models to better solve tasks requiring domain expertise.

6 Limitations

Despite the promising results achieved by Domain$o$1s, there are some limitations. First, while our Selective Tree Exploration method effectively balances search performance and computational costs, the additional inference time required for tree exploration may impact the model's real-time application scenarios, such as in situations requiring high response speed. Second, although we construct high-quality CoT datasets using GPT-4o, the relatively small size of training data (2,000 examples each for finance and legal domains) may limit the model's ability to handle extremely rare or complex domain-specific cases. Additionally, while PROOF-Score provides a comprehensive evaluation framework, research on using LLMs as judges suggests that further refinement and elaboration of evaluation metrics may be beneficial (Gu et al., 2024). Finally, our current implementation focuses on stock recommendation and legal reasoning tasks, and the generalizability of our approach to other domain applications requires further investigation. These limitations point to promising directions for future research, such as optimizing inference efficiency, expanding training datasets, and extending the framework to broader domain applications.

7 Ethical Considerations

In this section, we discuss several important ethical considerations regarding the training, deployment, and use of Domain$o$1s.

7.1 Fairness and Accessibility

We recognize that the computational resources required for training and inference of large language models (LLMs) and tree search exploration may limit accessibility for researchers and practitioners with fewer resources. To address this, we will open-source our implementation and provide efficient variants that can run on consumer-grade hardware. Additionally, we will release the training datasets (CoT-stock-2k and CoT-legal-2k) to enhance reproducibility and facilitate broader participation in this research direction.

7.2 Potential Risks in Financial and Legal Applications

For financial applications, we acknowledge that Domain$o$1s-finance's advice, while explainable, should be viewed as restricted investment references. To mitigate potential risks:

- We explicitly state that Domain$o$1s-finance's outputs should serve as one of many considerations when users make actual investment decisions.
- We implement safety checks in the Domain Safety (DS) metric of PROOF-Score to detect potentially harmful or high-risk advice.
- We emphasize the importance of human oversight and professional judgment in interpreting model reasoning.

For legal applications, Domain$o$1s-legal is intended to assist rather than replace legal professionals. To mitigate potential risks:

- We explicitly state that Domain$o$1s-legal is proposed as a support tool rather than a substitute for professional legal advice.
- We detect responses that contradict legal facts by evaluating the Factual Accuracy (FA) metric of PROOF-Score.
- We emphasize the importance of human oversight and professional judgment in interpreting model reasoning.

7.3 Privacy and Data Security

We have taken multiple measures to protect privacy and ensure data security:

- Our datasets have been carefully screened and curated to exclude sensitive personal information.
- The model's inference process is designed to focus on public information.
- Implement rate limiting and access controls after model and dataset open-sourcing to prevent potential misuse.

7.4 Environmental Impact

We acknowledge the environmental impact of training and running large language models. To minimize this:

- Our proposed Selective Tree Exploration method is designed to improve computational efficiency and reduce inference overhead.
- We provide guidance on optimal hyperparameter settings and encourage the selection of hyperparameter configurations that balance computational costs with model performance to reduce unnecessary computation.

Through these considerations and safeguards, we aim to ensure Domain$o$1s makes positive contributions to the field while minimizing potential risks and negative impacts. We encourage ongoing dialogue with stakeholders and welcome community feedback to further improve the ethical implementation of our technology.

References

AI@Meta. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Malcolm Baker and Jeffrey Wurgler. 2006. Investor sentiment and the cross-section of stock returns. *The journal of Finance*, 61(4):1645–1680.

Or Biran and Kathleen R. McKeown. 2017. Human-centric justification of machine learning predictions. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1461–1467. ijcai.org.

Sebastian Bordt, Harsha Nori, Vanessa Rodrigues, Besmira Nushi, and Rich Caruana. 2024. Elephants never forget: Memorization and learning of tabular data in large language models, 2024. *URL https://arxiv. org/abs/2404.06209*.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. In *Submitted to The Thirteenth International Conference on Learning Representations*. Under review.

Umar Butler. 2023. Open australian legal llm.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024a. Alphamath almost zero: process supervision without process. *arXiv preprint arXiv:2405.03553*.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024b. Step-level value preference optimization for mathematical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7889–7903, Miami, Florida, USA. Association for Computational Linguistics.

Zheng Chen and Zhejun Liu. 2024. Sentence-level heuristic tree search for long text generation. *Complex & Intelligent Systems*, 10(2):3153–3167.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024a. Instruction pre-training: Language models are supervised multitask learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2550, Miami, Florida, USA. Association for Computational Linguistics.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024b. Instruction pre-training: Language models are supervised multitask learners. *arXiv preprint arXiv:2406.14491*.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024c. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. 2024. (a) i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2454–2469.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13.

Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. 2021. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, 14:1–22.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

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

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.

Ricardo Dominguez-Olmedo, Vedant Nanda, Rediet Abebe, Stefan Bechtold, Christoph Engel, Jens Frankenreiter, Krishna Gummadi, Moritz Hardt, and Michael Livermore. 2024. Lawma: The power of specialization for legal tasks. *arXiv preprint arXiv:2407.16615*.

Eugene F Fama and Kenneth R French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56.

Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. 2018. Enhancing stock movement prediction with adversarial training. *arXiv preprint arXiv:1810.09936*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

Cheonsu Jeong. 2024. Fine-tuning and utilization methods of domain-specific llms. *arXiv preprint arXiv:2401.02981*.

Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 417–421.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. Tigerscore: Towards building explainable metric for all text generation tasks. *Transactions on Machine Learning Research*.

Kelvin JL Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. 2024. Learning to generate explainable stock predictions using self-reflective large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4304–4315.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

David Lanz. 2024. Llama2-7b-taiwan-btc-qlora. https://huggingface.co/DavidLanz/llama2_7b_taiwan_btc_qlora.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Jonathan Li, Rohan Bhambhoria, and Xiaodan Zhu. 2022. Parameter-efficient legal domain adaptation. *arXiv preprint arXiv:2210.13712*.

Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, and Jun Huang. 2024b. AlphaFin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 773–783, Torino, Italia. ELRA and ICCL.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *ArXiv*, abs/2303.14070.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun-Qing Li, Hejie Cui, Tian yu Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Carl Yang, and Liang Zhao. 2023. Beyond one-model-fits-all: A survey of domain specialization for large language models. *ArXiv*, abs/2305.18703.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*.

Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Pajaree Krisanapan, Yeshwanter Radhakrishnan, and Wisit Cheungpasitporn. 2024. Chain of thought utilization in large language models and application in nephrology. *Medicina*, 60(1):148.

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

OpenAI. 2024. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/. [Accessed 19-09-2024].

OpenO1 Team. 2024. Open-o1: A model matching proprietary power with open-source innovation. https://github.com/Open-Source-O1/Open-O1. [Accessed 19-11-2024].

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Michael E Porter and Mark R Kramer. 1985. Advantage creating and sustaining superior performance. *Competitive Advantage; Harvard Busines School Publishing: Brighton, MA, USA*.

Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2024. Mutual reasoning makes smaller llms stronger problem-solvers. *arXiv preprint arXiv:2408.06195*.

Chuan Qin, Jun Chang, Wenting Tu, and Changrui Yu. 2024. Followakoinvestor: Stock recommendation by hearing voices from all kinds of investors with machine learning. *Expert Systems with Applications*, 249:123522.

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

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Shah. 2020. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426.

Takehiro Takayanagi, Kiyoshi Izumi, Atsuo Kato, Naoyuki Tsunedomi, and Yukina Abe. 2023. Personalized stock recommendation with investors' attention and contextual information. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3339–3343.

Utkarsh Ujwal, Sai Sri Harsha Surampudi, Sayantan Mitra, and Tulika Saha. 2024. " reasoning before responding": Towards legal long-form question answering with interpretability. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4922–4930.

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

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.

Steven Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dmitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023b. MAUD: An expert-annotated legal NLP dataset for merger agreement understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16369–16382, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2022. Large language models are better reasoners with self-verification. *arXiv preprint arXiv:2212.09561*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2024. Bloomberggpt: A large language model for finance, 2023. *ArXiv preprint: https://arxiv. org/pdf/2303.17564. pdf*.

Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. 2023a. The wall street neophyte: A zero-shot analysis of chatgpt over multimodal stock movement prediction challenges. *arXiv preprint arXiv:2304.05351*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023b. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*.

Frank Xing. 2024. Designing heterogeneous llm agents for financial sentiment analysis. *ACM Transactions on Management Information Systems*.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979.

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

Linyi Yang, Jiazheng Li, Ruihai Dong, Yue Zhang, and Barry Smyth. 2022. Numhtml: Numeric-oriented hierarchical transformer model for multi-task financial forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11604–11612.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Mengqi Zhang, Xiaotian Ye, Qiang Liu, Pengjie Ren, Shu Wu, and Zhumin Chen. 2024. Uncovering overfitting in large language model editing. *arXiv preprint arXiv:2410.07819*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*.

A Experimental Setup

All experiments are conducted using an AMD EPYC 7H12 64-Core processor as CPU and four 48GB NVIDIA RTX 6000 Ada GPUs. For each variant of Domain$o$1s, fine-tuning takes approximately 48 GPU hours per run. The system environment uses CUDA version 12.4, Python version 3.10.15, PyTorch version 2.5.1, and transformers version 4.45.2. The random seed is set to 42.

We employ LoRA (Low-Rank Adaptation) for fine-tuning. The base model is Qwen2.5-7B-Instruct. We use the qwen template with Flash Attention enabled. The training dataset is preprocessed using 16 workers with a maximum sequence length of 2,048 tokens.

The LoRA hyperparameters are set as follows: rank = 8, alpha = 16, and dropout = 0, targeting all model layers. For optimization, we use the AdamW optimizer with a learning rate of 5e-5 and cosine learning rate scheduling. The training runs for 120 epochs. We employ mixed-precision training using bfloat16 format.

The batch size is set to 2 per device with a gradient accumulation of 2 steps, effectively creating a batch size of 16 ($2 \times 2 \times 4$ GPUs). Gradient clipping is applied with a maximum norm of 1.0. The model checkpoints are saved every 100 steps, with loss logging occurring every 5 steps.

For experiments using accuracy or MCC as metrics in Tables 1, 2 and other related figures or tables, to ensure a fair comparison with our Domain$o$1s, we fine-tune the baselines using the training sets of corresponding tasks. During fine-tuning, unlike the CoT data used to train Domain$o$1s, we train the baselines with direct prediction-style answers. Therefore, the fine-tuning data remains consistent between baselines and Domain$o$1s, with only different answer formats. For experiments on explainability metrics, inference time and reasoning chain length in Tables 4, 8, 9, 10, 11, 12 and other related figures or tables, the baselines are not trained on any of our datasets, ensuring they generate answers in their originally designed output formats for fair comparison of model explainability, inference time and reasoning chain length.

B CoT Data Generation

Figure 6 and Figure 7 are the prompt templates for instructing GPT-4o to generate responses in CoT format.

Figure 8 is the prompt template for instructing GPT-4o and Qwen-2.5-Instruct to generate tweet summaries.

Prompt for stock data generation

For a given set of facts, estimate their overall impact on {stock_name}'s stock price movement. You need to generate intermediate steps following this process to arrive at the correct answer:

<SUMMARY> Summarize all facts, clearly identify the core focus and research subject of the analysis, outline the overall background and target scenario. Also indicate what the following analysis should focus on. </SUMMARY>

<HISTORICAL CONTEXT> Analyze the impact of historical background and long-term performance, especially regarding past stock performance and market environment. </HISTORICAL CONTEXT>

<VALUATION> Analyze current stock valuation levels and their impact on investor decisions. Include P/E ratios, analyst price targets, and market views on valuation. </VALUATION>

<MARKET SIZE AND DOMINANCE> Emphasize the company's position and scale in the industry or market, analyze its market influence. </MARKET SIZE AND DOMINANCE>

<STRATEGIC INITIATIVES> Analyze recent strategic actions, including partnerships, innovations, patents, and their potential impact on future growth. Discuss major technological, business, or collaborative actions. Do recent strategic actions support long-term growth? Do these actions enhance company competitiveness? </STRATEGIC INITIATIVES>

<INVESTOR SENTIMENT> Analyze investor sentiment towards the stock, including trading activity, options performance, and market discussions. Discuss investor behavior (such as trading volume, attention) and market sentiment. </INVESTOR SENTIMENT>

<RISKS AND CONCERNS> Analyze risk factors currently concerning investors, including high valuations, market volatility, or external risks. </RISKS AND CONCERNS>

<RECENT PERFORMANCE> Analyze recent stock price performance and its driving factors, including price milestones, short-term trends, and market performance. </RECENT PERFORMANCE>

<CONSOLIDATION> Analyze adjustments in company financials or stock structure, such as stock buybacks, profitability, or long-term return trajectory. </CONSOLIDATION>

<OVERALL IMPACT> Synthesize all analysis points, clearly indicate the direction of overall impact (positive or negative, cannot be neutral), and provide final conclusions. Give your response in this format: Price Movement: Positive/Negative. </OVERALL IMPACT>

Here are the facts:
{facts}

In your generation process, do not use '*', '#', or '-'. Only respond using the above special tags and their generated content in this format:

<SUMMARY></SUMMARY>\n\n<HISTORICAL CONTEXT></HISTORICAL CONTEXT>\n\n<VALUATION></VALUATION>\n\n<MARKET SIZE AND DOMINANCE></MARKET SIZE AND DOMINANCE>\n\n<STRATEGIC INITIATIVES></STRATEGIC INITIATIVES>\n\n<INVESTOR SENTIMENT></INVESTOR SENTIMENT>\n\n<RISKS AND CONCERNS></RISKS AND CONCERNS>\n\n<RECENT PERFORMANCE></RECENT PERFORMANCE>\n\n<CONSOLIDATION></CONSOLIDATION>\n\n<OVERALL IMPACT></OVERALL IMPACT>

Figure 6: Prompt template for stock investment recommendation.

Prompt for legal data generation

Given a legal question, you need to follow this process to generate intermediate steps and arrive at the correct answer:

<SUMMARY>Summarize the materials, text, or paragraphs provided in the question. Also identify what the subsequent analysis needs to focus on in order to answer the question</SUMMARY>

<REASONING>Generate step-by-step analysis to answer the question based on the provided text</REASONING>

<EVIDENCE>List evidence from the presented text that supports the reasoning, and reflect on whether the answers in the reasoning section are correct</EVIDENCE>

<CONCLUSION>Provide a brief summary and draw the final conclusion</CONCLUSION>

Here are the question:
{question}

In your generation process, do not use '*', '#', or '-'. Only respond using the above special tags and their generated content in this format:

<SUMMARY></SUMMARY>\n\n<REASONING></REASONING>\n\n<EVIDENCE></EVIDENCE>\n\n<CONCLUSION></CONCLUSION>

Figure 7: Prompt template for legal reasoning QA.

Prompt for tweets summary generation

Below are multiple tweets about {stock} from {date}, separated by '---':

{' --- '.join(texts)}

Please provide a brief, factual summary of the key events, news, and developments mentioned in these tweets. Focus only on objective facts and events, without any analysis or sentiment. Format the summary as bullet points. Keep each point concise and avoid repetition.

Requirements:
1. Only include events and facts that were explicitly mentioned
2. Remove any duplicated information
3. Keep each bullet point under 15 words
4. Do not include any subjective analysis or market sentiment
5. Focus on company/stock related events only
You only need to generate your summary.

Figure 8: Prompt template for tweet summarization.

Model	time(s)	Length
Qwen-2.5-Instruct	6.91	166.2
Llama-3-Instruct	**5.08**	95.2
OpenO1-Llama	20.27	454.8
OpenO1-Qwen	21.09	465.1
Finance-LLM	14.70	131.3
Finance-Chat	14.52	130.8
Finance-Llama-3	5.34	13.5
FinGPT	6.29	14.2
Llama-2-taiwan-btc	13.63	41.3
SEP	13.182	119.6
Domain-CoT-finance	18.37	**512.1**
Domain*o*1s-finance	27.38	509.8

Table 8: Inference time and reasoning chain length on stock investment recommendation tasks.

Model	time(s)	Length
Qwen-2.5-Instruct	0.65	1.3
Llama-3-Instruct	0.71	1.5
OpenO1-Llama	8.82	261.2
OpenO1-Qwen	9.53	265.9
Open-Australian-Legal	8.33	263.5
DISC-LawLLM	2.38	9.8
Law-LLM	4.18	97.4
Law-Chat	0.64	1.2
Lawma	**0.63**	1.0
Domain-CoT-finance	8.17	268.5
Domain*o*1s-legal	13.54	**269.8**

Table 9: Inference time and reasoning chain length on legal reasoning QA tasks.

C Answer Demonstration

Figures 9 and 10 demonstrate complete question-answering examples for stock investment recommendation and legal reasoning QA tasks using Domain*o*1s and the base model Qwen-2.5-Instruct. Domain*o*1s does not explicitly output special tokens (e.g., <SUMMARY>), but reason according to the structured reasoning process constructed in the CoT-stock-2k and CoT-legal-2k datasets.

As shown in Figure 9, Qwen-2.5-Instruct reaches an incorrect answer by focusing only on partial information (the Negative parts) while ignoring the overall context. In contrast, Domain*o*1s-finance comprehensively considers both Positive and Negative facts to draw conclusions. As shown in Figure 10, Qwen-2.5-Instruct starts making reasoning errors after generating "However," incorrectly classifying the user as a non-consumer, leading to an incorrect result. In comparison, Domain*o*1s-legal avoids errors through structured reasoning paths and tree search.

D Answer Length and Inference Time

In this section, we present the reasoning chain length and inference time of Domain*o*1s and baselines in generating answers for stock investment recommendations and legal reasoning QA tasks. The reasoning chain length is measured by the average number of words rather than tokens in the responses to ensure fair comparison across different models. As shown in Table 8 and Table 9, o1-like models (OpenO1-Llama, OpenO1-Qwen, and our Domain*o*1s) have longer reasoning chains than other baselines, among which our Domain*o*1s and Domain-CoT have the longest reasoning chains. Although Domain*o*1s exhibits longer inference time compared to the baselines, this is attributed to its generation of longer and higher-quality reasoning chains and the search for optimal reasoning paths, ultimately leading to superior accuracy metrics.

E PROOF-Score Generation

Figure 11 is the prompt template for instructing GPT-4o to generate PROOF-Scores.

F Construction of Sub-datasets

We select 8 data points from each of the 55 stocks in the stock investment recommendation dataset. Each data point contains more than 10 positive tweets and 10 negative tweets. We construct five groups of tweets, with each group containing 10 tweets but different ratios of positive to negative tweets: 1:9 (1 positive, 9 negative), 3:7 (3 positive, 7 negative), 5:5 (5 positive, 5 negative), 7:3 (7 positive, 3 negative), and 9:1 (9 positive, 1 negative). Figure 12 shows a construction example and model responses. The example has a Positive:Negative ratio of 7:3. Tweets with yellow background are positive, while those with blue background are negative. Finance-Llama-3's response only considers the positive tweets, completely ignoring the negative ones. In contrast, Domain*o*1s-finance considers both positive and negative tweets to arrive at the correct answer.

We extract 500 questions from the legal reasoning QA dataset. Each question contains key conditions necessary for answering the question. We remove these key conditions from the questions, making them impossible to answer. Figure 13 shows a construction example.

Question

For a given set of facts, estimate their overall impact on AAPL's stock price movement. Answer with Positive/ Negative. Here are the facts:

2022-09-29

- Bank of America downgrades Apple ($AAPL) from "Buy" to "Neutral."
- Rosenblatt raises $AAPL price target to $189 from $160 and upgrades to Buy.
- Apple's stock plummets, down by over 5% on the day.
- Apple's market cap falls, contributing to the S&P 500 staying positive while Apple trades negative.
- Apple's CEO warned about potential demand concerns.
- Apple's procurement chief fired after viral TikTok incident.
- Apple's stock hits a new low of the day.
- Apple's share of the market reaches a record low.
- Apple reportedly working on a feature to track iPhone users without GPS.
- Apple downgraded to "sell" by Jefferies.
- Apple given a "strong buy" rating by JPMorgan.
- Apple reportedly planning to make electric vehicles.
- Apple's stock options trading actively, with significant put volume.
- Apple's stock included in a 10% decline pattern of FAANG stocks.
- Apple's stock is a focus in options trading with notable activity.
- Apple targeted in discussions about potential short squeeze.

2022-09-30

- Apple stock entered bear market territory on September 30, 2022.
- Apple is down 22% year-to-date; meta down 60%, amzn down 34%.
- Apple's stock price closed at $141.9 on September 30, 2022.
- Apple fired its long-serving procurement head due to a TikTok incident.
- Apple's stock declined 4.91% on September 30, 2022.
- Apple raised its stock price target to $189 by Rosenblatt Securities.
- Apple's Korea offices were raided by antitrust officials.
- Large trades occurred in Apple: $1 billion plus.
- Apple option alert: Jan 20 $140 calls bid at $12.5.
- Apple's 14 Pro and 14 Max demand reports were mentioned.
- Apple's stock price broke a ten-dollar threshold.
- Apple's stock is part of the Dow's performance drag.
- Apple's stock failed to protect key support levels.
- Apple's stock is down 3% from previous closing price.
- Apple's stock faced multiple simultaneous negative factors.
- Apple's stock showed algos-driven patterns.
- Apple's price target for next week was highlighted as a top analyst estimate.

2022-10-01

- Relative Strength scan identifies AAPL as strong stock for trading.
- AAPL reported among top stocks by mention on trading day of October 1, 2022.
- Jim Cramer associated with AAPL's stock performance issues.
- AAPL ranked sixth in year-to-date stock performance list.
- Retail traders' selling week noted, including AAPL.
- AAPL targeted in live day-trading room offers.
- A wave 5 throw-over pattern identified for AAPL.
- Top analyst price targets discussed for AAPL for the following week.
- AAPL stock showed a significant price drop, entering bear market territory.
- AAPL's head of procurement fired due to TikTok incident.
- Apple reportedly reducing new iPhone production plans.
- AAPL mentioned in discussions around most overvalued stocks.
- Rosenblatt raised AAPL's stock price target on strong demand for devices.
- AAPL included in options trade highlights with significant percentage increase.

2022-10-02

- Trade Ideas' Test Drive registration is open for trading tools.
- Last week, $AAPL had an 8.13% weekly decline since March 2020.
- $AAPL was downgraded by Bank of America.
- $AAPL's last large trade was below key technical levels.
- $AAPL production cuts were reported alongside $NIKE inventory builds.
- $AAPL's stock and options saw significant positive gains in options trades.
- $AAPL is one of the last major tech stocks moving into Stage 4.
- $AAPL hit extreme oversold conditions for the first time in months.
- $AAPL's last large trade was on September 23, 2022.
- $AAPL's 2022 YTD stock performance noted.
- Free trials and stock options alerts available on Discord.
- $AAPL included in discussions of potential stock market bottom scenarios.

2022-10-03

- Bank of America's downgrade led Big Tech stocks, with Apple ($AAPL) experiencing a significant decline.
- Apple ($AAPL) saw a 8% drop into quarter end, unique among US Big Tech names.
- Apple's September App Store net revenue fell 5% year-over-year.
- TSMC is raising prices for all clients, including Apple.
- Apple ($AAPL) is dealing with a potential short squeeze as shares approached $140.
- Apple ($AAPL) option activity included multiple calls and puts with high open interest.
- Apple ($AAPL) hit new 12-week lows as of October 3, 2022.
- Italian antitrust fine for Apple ($AAPL) and Amazon was cancelled on appeal.
- Apple ($AAPL) calls were active going into the week of October 3, 2022.
- Apple ($AAPL) included in various trading watches and options lists.
- Apple ($AAPL) hit price targets and saw analyst price adjustments.

Ground Truth: Positive

Qwen-2.5-Instruct Answer

The overall impact of the given facts on AAPL's stock price movement is **predominantly negative**. The facts indicate a series of downgrades, warnings about demand, negative market reactions, and other unfavorable factors, such as stock hitting new lows and dragging market performance. **These all had a negative impact on AAPL's stock price.**

Negative

Domaino1s-finance Answer

The analysis focuses on Apple's stock performance amid a series of operational, strategic, and market responses impacting its valuation and investor sentiment as of late September to early October 2022. The backdrop includes operational downsizes, changes in leadership, strategic projects in EV vehicles and location tracking technology, and investor reactions around big tech stock declines. **We will examine several factors to drive the analysis, including historical performance, valuation metrics, market position, strategic initiatives, investor sentiment, risks, and recent performance.**

Historically, Apple's stock has shown resilience but faced challenges this year, down 22% by mid-September, impacting overall market sentiment. The stock's identity as a key player within the tech sector has persisted despite setbacks. The inclusion in the Big Tech segment, typically volatile, weighs on its performance. **Historically, Apple has shown capability for strong comebacks, driven by product cycles and strategic positioning.** However, this period aligns during a broader tech market decline.

Apple's stock has been targeted for its high valuation issues, with fluctuating analyst price targets reflecting mixed sentiment. Analysts have mixed ratings, with Rosenblatt an upgrade, Jefferies a downgrade, and JPMorgan a strong buy, indicating uncertainty in valuation. Despite Jefferais's downgrade, Rosenblatt has raised the target price to $189, suggesting some confidence in Apple's potential. Market views see company-valuation pressured by operational challenges and external critiques, such as antitrust raids.

Apple continues to be a dominant player in the tech market, yet faces challenges in maintaining its market share, recording a historical low in market share share significance. Despite its usual market influence, the current administration raids in South Korea may dampen its prowess. Valuable name space remains due to ongoing negotiations and strategic initiatives in EVs; however, near-term signals suggest competitive pressures or stage transitions ahead.

Recently, Apple has engaged in strategic research on EVs and location tech, but facing challenges in applications due to app store declines and supply chain costs from TSMC's price increases. Innovative projects announced bolster future growth, yet current execution hurdles affect immediate competitiveness and sentiment. While Apple reduces production plans for new iPhones, strategic divestitures or consolidations within the short term remain limited.

Investor sentiment appears mixed, with notable trading activity in options reflecting apprehension amid stage-four challenges. Recent sentiment shows increased attention in bear market phases, targeting Apple's demand concerns and misgivings over big tech valuations. **Yet, persistent option gains and analyst upside estimates suggest a belief in recovery or speculative interest, providing some optimism amidst a challenging period.**

Risks include operational missteps such as procurement lead firing, antitrust impacts, and potential demand softening. Apple faces multiple simultaneous concerns from operational disruptions in regions like South Korea to reputational damage from global markets. Investors worry algorithms driving intense sell pressures and fear of a short squeeze adding complexity to risks facing Apple's stock.

Apple's recent stock trajectory features an otherworldly decline, grade-point negativity against Big Tech peers, and significant price milestones. Despite entering bear market territory and negative days, signs of recovery include jeers of bearish sentiment turned around by bullish analyst estimates and price target revisions. Overall patterns hint at technical rebound prospects, attracting trading attention.

Financial adjustments through buybacks or restructuring are not Significantly highlighted amidst strategic recalibration and current challenge stations. Stock inclusion in significant activity of Big Tech and analysis circles underscores evaluation of market position and certification reforms. Recent market stage navigation represents a complex balance for Apple's financial strategy navigating through turbulent waters.

Despite the confluence of negative operational, strategic, and market pressures, the overall impact on Apple appears tilted towards a positive outlook. Optimistic forecasts from some analysts, ongoing strategic exploration in new tech domains, and technical signs of recovery provide a supportive backdrop. Confidence in recovery potentials appears rising amidst strategic maneuvers and market sentiment shifts, yielding positive implications for Apple's stock movement. **Price Movement: Positive.**

Figure 9: Questions and answer demonstrations for Domain*o*1s-finance and base model.

Question

Here is a service agreement:
If you are a consumer, the laws of the country in which you reside will apply to any claim, cause of action, or dispute you have against us that arises out of or relates to these Terms or the Facebook Products, and you may resolve your claim in any competent court in that country that has jurisdiction over the claim. **In all other cases**, you agree that the claim must be resolved exclusively in the U.S. District Court for the Northern District of California or a state court located in San Mateo County. You also agree that you submit to the personal jurisdiction of either of these courts for the purpose of litigating any such claim, and that the laws of the State of California will govern these Terms and any claim, without regard to conflict of law provisions.
Here is a question about it: As **a Facebook user who lives in Canada**, can I sue Facebook in Canada?

Ground Truth: Yes

Qwen-2.5-Instruct Answer

If you are a consumer and reside in a country other than the United States, any claim against Facebook must be resolved in the competent court in the country where you reside. **However, for non-consumer cases**, including those from Canadian residents, the claim must be resolved exclusively in the U.S. District Court for the Northern District of California or a state court located in San Mateo County, California. **No.**

Domaino1s-legal Answer

The question is whether a Facebook user in Canada can sue Facebook in Canada under a service agreement. The agreement states that consumers can rely on the laws of their country and resolve disputes in local courts. For non-consumers, disputes must be resolved in U.S. courts. **The analysis focuses on the user's consumer status and the jurisdictional provisions.**

1. The agreement allows consumers to rely on the laws of their country of residence, so Canadian law applies to a Canadian user.
2. Consumers are given the option to resolve disputes in their home country's courts, which implies jurisdiction in Canada.
3. The U.S. court jurisdiction applies only to non-consumers, which is irrelevant if the user qualifies as a consumer.
4. Canadian consumer protection laws generally allow consumers to sue companies like Facebook in Canadian courts.

- The agreement states: **"If you are a consumer, the laws of the country in which you reside will apply" and "you may resolve your claim in any competent court in that country."**
- **The user is likely a consumer and not subject to the non-consumer jurisdiction clause.**
- Canadian courts typically uphold consumer protection laws, allowing users to sue within their jurisdiction.

A Canadian Facebook user classified as a consumer can sue Facebook in Canada. The agreement permits disputes to be resolved in local courts, and Canadian law supports this. **The answer is: Yes.**

Figure 10: Questions and answer demonstrations for Domain*o*1s-legal and base model.

Prompt for PRISM-Score evaluation

For a given question and an LLM's response, score the response from the following angles. You need to provide the three scores in array format, i.e. [, ,].

- Reasoning Completeness (RC): Evaluates the completeness of reasoning and logical coherence in the model's answer.
1 point: Complete lack of reasoning process or logical confusion, conclusions without basis, unable to support the answer.
3 points: Basic reasoning exists but with obvious logical gaps or missing key steps, overall unconvincing.
5 points: Clear and logically rigorous reasoning, but lacks some minor information or background explanation, affecting comprehensiveness.
7 points: Complete reasoning process, rigorous logic with clear expression, covers all relevant points without gaps or omissions.

- Domain Safety (DS): Measures the safety and appropriateness of model output in specific domains.
1 point: Answer contains serious safety concerns or inappropriate content, such as illegal information, high-risk advice, or serious ethical issues.
3 points: Answer has minor potential safety concerns or inappropriate wording, but overall acceptable with no obvious malicious intent.
5 points: Answer is completely safe but lacks explicit warnings or restrictions for certain potential risks or sensitive content.
7 points: Answer is impeccable in terms of safety and appropriateness, not only fully compliant with domain safety standards but also provides relevant warnings or restrictions.

- Factual Accuracy (FA): Assesses the accuracy of facts stated in the model's answer.
1 point: Answer contains multiple major factual errors, obviously contradicts the question or common sense, conclusions are false.
3 points: Facts are basically correct but contain significant minor errors or omissions affecting accuracy and credibility.
5 points: Facts are accurate, covering main points in the question, but lacks background information or some details are not fully developed.
7 points: Facts are impeccable, all statements are accurate and comprehensive, while providing necessary background information or reliable evidence.

Here is the question: {question}

Here is the response from LLM: {response}

Figure 11: Prompt template for generating PROOF-Score.

Question: For a given set of facts, estimate their overall impact on AAPL's stock price movement. Answer with Positive/Negative. Here are the facts:
In the past 5 days, there are the following facts:

Apple to start making MacBooks in Vietnam by mid-2023
Apple TV app for Android rumored
Partnership with ATT and SpaceMobile
In talks for NFL Sunday Ticket
Partnership with Goldman Sachs on banking initiative
Hit $132.5 breakout level as a bull sign
Apple is one of the most valuable companies globally

Hit new low since mid-June on December 13, 2022
JPMorgan cuts price target to $190 from $200
Approaching 52-week low as of December 23

Ground Truth: Negative.

Finance-Llama-3:
- Apple's production diversification to Vietnam shows strategic supply chain expansion.
- Apple's Android TV app development indicates new market opportunities.
- Apple's partnership with ATT and SpaceMobile suggests telecom sector growth.
- Apple's partnership with Goldman Sachs strengthens fintech presence.
- Apple maintains its position as one of the most valuable global companies.

Based on the given facts, the overall impact on AAPL's stock price movement would be **Positive**. The company's strategic expansion in manufacturing, partnerships, and market presence suggests strong growth potential.

Domaino1s-finance:
The facts present a mix of strategic initiatives and market performance indicators for Apple (AAPL). Key strategic moves include manufacturing diversification, potential market expansions, and new partnerships. However, there are also concerning market performance indicators and analyst actions. Analysis should focus on weighing these opposing factors to determine likely price impact.

The stock hitting new lows since June 2022 indicates deteriorating market performance. This suggests a weakening trend in investor confidence despite Apple's historically strong market position.

JPMorgan's price target cut from $200 to $190 suggests concerns about valuation levels. While still implying upside from current levels, the downward revision indicates diminishing confidence in near-term price appreciation potential.

Apple maintains its position as one of the most valuable companies globally, demonstrating continued market leadership and strong fundamental business position.

Multiple positive strategic moves are evident:
- Vietnam manufacturing expansion shows supply chain diversification
- Android TV app development indicates market expansion efforts
- Partnerships with ATT, SpaceMobile, and Goldman Sachs suggest business growth initiatives
- NFL Sunday Ticket negotiations demonstrate content strategy expansion

However, these are mostly longer-term initiatives with uncertain near-term impact.

Technical indicators show mixed signals:
- Positive: $132.5 breakout level as a bull sign
- Negative: Approaching 52-week lows, suggesting deteriorating investor confidence

Approaching 52-week lows and new lows since June represent significant technical weakness, indicating elevated risk levels and potential further downside.

Recent price action has been notably weak, with the stock:
- Hitting new lows since June
- Approaching 52-week lows
- Showing technical weakness despite positive strategic announcements

While strategic initiatives suggest positive long-term prospects, near-term market performance and analyst actions indicate significant headwinds.

Price Movement: **Negative**
Despite numerous positive strategic initiatives, the immediate market signals (new lows, analyst downgrades, technical weakness) suggest continued downward pressure on the stock price. While long-term prospects remain solid, near-term price movement is likely to be dominated by current market weakness and deteriorating investor confidence.

Figure 12: Example of stock sub-dataset construction.

Example 1: If I upload content to Amazon, can Amazon then allow a third party to use that content?

Example 2: Is Change of Recommendation permitted as long as the board determines that such change is required to fulfill its fiduciary obligations?

Example 3: If a court located in particular state (A) can exercise personal jurisdiction over an individual (B) named in the fact pattern. Answer with Yes/No.

Figure 13: Example of legal sub-dataset construction.

G Explainability Analysis

Table 10 demonstrates the ablation experiments of PROOF-Score without fine-tuning (w/o Fine-tune) and without solution expansion & sampling (w/o Sample). In the w/o Fine-tune experiment, we use the Qwen-2.5-Instruct model without fine-tuning on our data and prompt it to separate each step with "\n" to facilitate our solution expansion & sampling. The results indicate that the PROOF-Score of the model without fine-tuning is lower than Domain*o*1s, demonstrating that Domain*o*1s learns to generate superior-quality reasoning processes from our constructed high-quality fine-tuning datasets. Meanwhile, the PROOF-Score of the model without solution expansion & sampling is similar to Domain*o*1s, which suggests that the role of solution expansion & sampling is more reflected in improving the quality of reasoning paths to enhance model accuracy (as shown in Table 5-7). From the perspective of PROOF-Score, the difference is not easily distinguishable, as the model can output highly interpretable answers regardless of whether solution expansion & sampling is used.

Stock		Legal	
w/o Fine-tune	6.212	w/o Fine-tune	5.067
w/o Sample	6.351	w/o Sample	6.548
Domain*o*1s-finance	**6.359**	**Domain*o*1s-legal**	**6.677**

Table 10: Comparison of PROOF-Score between Domain*o*1s with w/o Fine-tune and w/o Sample.

Stock			Legal		
Model	**TIGERScore**	**Errors**	**Model**	**TIGERScore**	**Errors**
Qwen-2.5-Instruct	0.00	0.00	Qwen-2.5-Instruct	-2.41	0.74
Llama-3-Instruct	-0.50	0.50	Llama-3-Instruct	-3.23	0.81
OpenO1-Llama	0.00	0.00	OpenO1-Llama	-0.10	0.10
OpenO1-Qwen	0.00	0.00	OpenO1-Qwen	-0.13	0.13
Finance-LLM	-4.00	1.00	Open-Australian-Legal	-6.40	1.60
Finance-Chat	0.00	0.00	DISC-LawLLM	-4.00	1.00
Finance-Llama-3	-6.00	2.00	Law-LLM	-2.45	1.11
FinGPT	0.00	0.00	Law-Chat	-3.45	0.86
Llama-2-taiwan-btc	0.00	0.00	Lawma	-3.76	0.94
SEP	0.00	0.00	**Domain*o*1s-legal**	**-0.03**	**0.03**
Domain*o*1s-finance	**0.00**	**0.00**			

Table 11: Comparison of TIGERScore and error rates between Domain*o*1s and baselines on stock and legal tasks (using TIGERScore-7B). TIGERScore represents the average error score in responses (lower absolute values indicate better answer quality), while Errors show the average number of errors per response (lower values indicate better answer quality).

In addition to our proposed PROOF-Score, we evaluate Domain*o*1s on other metrics. TIGER-

Stock			Legal		
Model	**TIGERScore**	**Errors**	**Model**	**TIGERScore**	**Errors**
Qwen-2.5-Instruct	0.00	0.00	Qwen-2.5-Instruct	-0.76	0.19
Llama-3-Instruct	0.00	0.00	Llama-3-Instruct	-1.26	0.32
OpenO1-Llama	0.00	0.00	OpenO1-Llama	0.00	0.00
OpenO1-Qwen	0.00	0.00	OpenO1-Qwen	0.00	0.00
Finance-LLM	-4.00	1.00	Open-Australian-Legal	-10.80	3.00
Finance-Chat	-0.50	0.50	DISC-LawLLM	-3.40	0.80
Finance-Llama-3	-8.00	2.00	Law-LLM	-4.09	1.28
FinGPT	0.00	0.00	Law-Chat	-1.10	0.27
Llama-2-taiwan-btc	0.00	0.00	Lawma	-1.75	0.44
SEP	0.00	0.00	**Domaino1s-legal**	**0.00**	**0.00**
Domaino1s-finance	**0.00**	**0.00**			

Table 12: Comparison of TIGERScore and error rates between Domaino1s and baselines on stock and legal tasks (using TIGERScore-13B).

Score (Jiang et al., 2023) is an explainable reference-free evaluation metric based on LLaMA-2, which provides error analysis through natural language instructions and demonstrates the error analysis process. It can be used to evaluate a wide range of text-generation tasks. Table 11 and Table 12 show the evaluation results using TIGERScore-7B and TIGERScore-13B models respectively.

As shown in Table 11 and Table 12, Domaino1s achieves the highest scores (TIGERScore) and lowest error rates (Errors) in both stock and legal tasks, which indicates that Domaino1s' answers are more accurate and have more reasonable explanations compared to the baselines. Figure 14 shows examples of TIGERScore-13B's evaluation of answers from Domaino1s, Qwen-2.5-Instruct, and Law-Chat. We observe several potential limitations with TIGERScore-13B in evaluating answers. For instance, (1) TIGERScore-13B judges answers based on its knowledge, and when there are discrepancies with the evaluated model's answers (e.g., Qwen-2.5-Instruct), it considers their errors and reflects this in both TIGERScore and Errors, even when TIGERScore-13B's answers sometimes disagree with the ground truth. Additionally, (2) when the evaluated model (e.g., Law-Chat) provides only an option as the answer, TIGERScore-13B cannot assess the correctness of logic and reasoning as there are no explanations. In such cases, TIGERScore-13B can only score based on its judgment of answer correctness, which may lead to incorrect evaluations as mentioned in (1). In contrast, when the evaluated model (e.g., Domaino1s-legal) provides long comprehensive answers, TIGERScore-13B no longer strictly scores based on the model's conclusions. These results demonstrate the importance of incorporating more powerful language models as evaluators while highlighting the contrast between TIGERScore and our PROOF-Score. PROOF-Score primarily focuses on answer explainability, safety, and factuality, rather than considering answer accuracy as the sole important metric, which helps reduce hallucinations in the evaluation model during answer assessment.

Question: Here is a service agreement:
··· If you do post content or submit material, and unless we indicate otherwise, you grant Amazon a nonexclusive, royalty-free, perpetual, irrevocable, and **fully sublicensable right to use**, reproduce, modify, adapt, publish, perform, translate, create derivative works from, distribute, and display such content throughout the world in any media. ···
Here is a question about it:
If I upload content to Amazon, can Amazon then allow a third party to use that content?
Answer with Yes/No.
Ground Truth Answer: Yes.

Answer of Qwen-2.5-Instruct: **Yes.** According to the agreement, when you upload content, you grant Amazon 'fully sublicensable' rights, which means they can allow third parties to use it.
Evaluation of TIGERScore-13B: {'num_errors': 1, 'score': -4.0, 'errors': {'error_0': {'location': 'Yes. According to the agreement, when you upload content, you grant Amazon 'fully sublicensable' rights, which means they can allow third parties to use it.', 'aspect': 'Accuracy', 'explanation': "**The output is incorrect.** The correct answer is 'No' because while the agreement grants Amazon certain rights, it explicitly states that 'Amazon has the right but not the obligation to monitor and edit or remove any activity or content' and 'Amazon takes no responsibility and assumes no liability for any content posted by you or any third party.' This means Amazon cannot guarantee or authorize third party usage of content. The output should be corrected to 'No'.", 'severity': 'Major', 'score_reduction': '4'}}}

Answer of Law-Chat: **Yes.**
Evaluation of TIGERScore-13B: {'num_errors': 1, 'score': -4.0, 'errors': {'error_0': {'location': 'Yes', 'aspect': 'Accuracy', 'explanation': ""**The output is incorrect.** The correct answer is 'No' because the service agreement states that Amazon reserves the right to remove or edit such content, but does not regularly review posted content. This means that Amazon may allow a third party to use the content, but it is not guaranteed. The output should be corrected to 'No'."", 'severity': 'Major', 'score_reduction': '4'}}}

Answer of Domaino1s-legal: The service agreement allows users to upload various types of content to Amazon. In return, Amazon expects to receive extensive rights to this uploaded content. The focus of the analysis should be on whether these rights granted to Amazon include the sublicensing of content to third parties.

The service agreement grants Amazon a nonexclusive, royalty-free, perpetual, irrevocable, and fully sublicensable right to use the uploaded content. This means Amazon has the right to allow others to use the content. The term "fully sublicensable" specifically suggests that Amazon can transfer its rights to third parties. Therefore, the logical conclusion is that Amazon can indeed allow a third party to use the content if it has been uploaded under the terms of this agreement.

The relevant phrase from the agreement is: "a nonexclusive, royalty-free, perpetual, irrevocable, and fully sublicensable right." This provision clearly indicates that Amazon possesses the right to sublicense the content to others. The use of "fully sublicensable" supports the conclusion that Amazon has the flexibility to grant usage rights to any third party.

Based on the broad rights granted to Amazon including the ability to sublicense the content, it is concluded that Amazon can allow a third party to use the uploaded content. The answer to the question is, therefore, "**Yes.**"

Evaluation of TIGERScore-13B: {'num_errors': 0, 'score': -0.0, 'errors': {}}

Figure 14: Example of TIGERScore-13B's evaluation.

H Reasoning Process Analysis

In this section, we introduce a new perspective to analyze how Domaino1s's reasoning process assists user decision-making. Taking stock investment recommendation tasks in the financial domain as an example, previous research has primarily focused on prediction accuracy as the main evaluation metric (Koa et al., 2024; Xu and Cohen, 2018; Sawhney et al., 2020; Yang et al., 2022). However, as shown in Table 2, even the previously most advanced financial LLMs or LLMs extensively trained on stock data achieve prediction accuracy close to 50% (random choice) for this binary classification problem. This confirms that LLMs are still "Wall Street Neophytes" in stock prediction (Xie et al., 2023a). In reality, many domain problems in the real world often do not have unified or unique correct answers, as they typically involve dynamically changing environmental conditions, the interplay of multiple influencing factors, and user-specific preferences. When facing these problems, users can use Domaino1s's responses as references to aid their judgment by providing explainable analysis processes and recommendations. For example, Figure 15 shows part of the model's response when asking Domaino1s-finance for stock investment advice. The model's response includes both short-term and long-term analysis of stock prices and provides both short-term and long-term investment recommendations based on user investment habits.

Question: For a given set of facts, estimate their overall impact on APPL's stock price movement, and give **investment advice**.
For stock AAPL, the closing price changes over the last 30 trading days as of 2022-12-27 are : Date 2022-11-14: decreases 0.01366; Date 2022-11-15: increases 0.003418; ⋯
The monthly average of adjusted closing prices over the past 3 years is Month 2020-01: adjusted close price is 77.06512850806827; Month 2020-02: ⋯
In the past 5 days, there are the following facts:
2022-12-23 : - Apple is approaching its 52-week low. - Apple hit a new ex-dividend date for $0.1667 on December 27, 2022. ⋯

Response: Analysis looks at Apple stock (AAPL) from 2020-2022, focusing on late 2022. Main goal is to check stock trends and future direction based on price data and market factors.

AAPL grew well from 2020-2021, rising from $77 to $170. Stock did well during COVID-19. In 2022, prices started falling from $170 high.

Stock seems expensive to some traders. Lost 25.7% in 2022. Analysts are giving lower price targets due to market conditions.

•••

Advice: **WAIT to buy.** Stock likely to fall more. While Apple is strong long-term, **wait for price to stabilize above $133-135 before buying**. If **buying for very long-term**, considering **implementing a staged buying strategy rather than committing all capital at once.**

Figure 15: Example of Domain$o$1s-finance's response on stock investment recommendation. Blue bold text indicates short-term investment advice, and red bold text indicates long-term investment advice.