FinCoT: Grounding Chain-of-Thought in Expert Financial Reasoning

Natapong Nitarach, Warit Sirichotedumrong, Panop Pitchayarthorn,
Pittawat Taveekitworachai, Potsawee Manakul, Kunat Pipatanakul

SCB 10X, SCBX Group

Abstract

This paper presents FinCoT, a structured chain-of-thought (CoT) prompting framework that embeds domain-specific expert financial reasoning blueprints to guide large language models' behaviors. We identify three main prompting styles in financial NLP (FinNLP): (1) standard prompting (zero-shot), (2) unstructured CoT (free-form reasoning), and (3) structured CoT (with explicitly structured reasoning steps). Prior work has mainly focused on the first two, while structured CoT remains underexplored and lacks domain expertise incorporation. Therefore, we evaluate all three prompting approaches across ten CFA-style financial domains and introduce FinCoT as the first structured finance-specific prompting approach incorporating blueprints from domain experts. FinCoT improves the accuracy of a general-purpose model, Qwen3-8B-Base, from 63.2% to 80.5%, and boosts Fin-R1 (7B), a finance-specific model, from 65.7% to 75.7%, while reducing output length by up to 8.9× and 1.16× compared to structured CoT methods, respectively. We find that FinCoT proves most effective for models lacking financial post-training. Our findings show that FinCoT does not only improve performance and reduce inference costs but also yields more interpretable and expert-aligned reasoning traces.

1 Introduction

Financial decision–making, such as stochastic modeling, risk assessment, portfolio optimization, and algorithmic trading (Markowitz, 1952; Black and Scholes, 1973a; Rockafellar and Uryasev, 2000; Avellaneda and Stoikov, 2008), demands precise mathematics and domain-specific reasoning (Lewkowycz et al., 2022; Wen and Zhang, 2025). Recent advances in large foundation models for finance, such as the multimodal FinTral (Bhatia et al., 2024) and language-centric Fin-R1 (Liu et al., 2025), demonstrate progress but still face challenges in interpretability and domain alignment (Nie et al., 2024; Arya.ai, 2024; Lee et al., 2025). Accordingly, these shortcomings motivate stricter control over a model's intermediate reasoning path, which we explore via prompt design.

Prompting guides LLM reasoning without extra training. Methods such as Chain-of-Thought (Wei et al., 2023), Code Prompting (Hu et al., 2023), Plan-and-Solve (Wang et al., 2023), and Self-Reflection (Renze and Guven, 2024) encourage stepwise thinking but remain domain-agnostic. In finance, this can lead to omissions in valuation checks or confusion between basis points and percentages. Yet real-world analysis follows well-defined workflows—valuation, discounting, portfolio attribution—that depend on explicit intermediate structure. Embedding such structure in the prompt helps the model verify units, formulas, and boundary conditions, improving interpretability and alignment with expert practice.

We design **FinCoT**, a zero-shot prompt that injects expert financial workflows-encoded as Mermaid blueprints-into a structured CoT template, yielding auditable reasoning without fine-tuning. Across ten CFA domains, FinCoT significantly boosts accuracy (most in quantitative areas) and produces shorter, clearer outputs than standard or structured CoT prompts. This paper offers three main contributions:

- We provide a comprehensive investigation and the first taxonomy of financial prompting–covering standard prompting, unstructured CoT, and structured CoT/FinCoT–clarifying how each paradigm addresses domain-specific reasoning requirements.
- We release nine blueprint templates—conceptual diagrams modeled after the Unified Modeling Language (UML) (Engels et al., 2000) and rendered in Mermaid syntax–that LLMs can parse as plain–text hints to drive structured reasoning both

within FinCoT and in broader domain-specific prompting scenarios.

- On 1.032k CFA–style questions across ten financial domains and four open-source LLMs, FinCoT shows notable gains–up to +17.3 pp in accuracy ($p < 0.001$)–particularly on pretrained models and quantitatively structured tasks. While less effective on instruction-tuned or niche domains, FinCoT consistently reduces verbosity (∼8× fewer tokens) and improves reasoning trace clarity under a three–point interpretability rubric.

2 Background and Related Work

2.1 Prompt Engineering

Standard Prompting (SP): Refers to the baseline technique of simply providing a natural language instruction to an LLM, without providing any intermediate 'thinking' steps, demonstrations, or explicit reasoning cues-i.e., a zero-shot setup. While the GPT-3 paper (Brown et al., 2020) popularized few-shot prompting via exemplars, more recent work formalizes and benchmarks zero-shot prompting as a distinct paradigm (Wei et al., 2022). Our implementation follows the formulation shown in Appendix Listing 1 (ZS) in (Callanan et al., 2024), and is used to represent the standard prompting baseline.

Unstructured Chain-of-Thought (UST-CoT): A general-purpose reasoning strategy using free-form CoT to establish a baseline for unconstrained prompting. These include:
- **Chain-of-Thought (CoT)** (Wei et al., 2023): Decompose reasoning into intermediate steps, encouraging the model to deliberately and systematically 'think' before finalizing an answer.
- **Code Prompting** (Hu et al., 2023): Translates problems into executable code, allowing the model to simulate logic or perform precise computations. In other words, LLMs are elicited to reason explicitly and transparently through code.
- **Plan-and-Solve** (Wang et al., 2023): Separates planning from solving, where the model first outlines a high-level plan, then executes the reasoning based on that plan.

In addition to these, other prompting techniques have emerged, such as Tree of Thoughts (ToT) (Yao et al., 2023), which explores multiple reasoning paths via tree-structured search; Graph of Thoughts (GoT) (Besta et al., 2024), which frames reasoning as a graph with LLM-generated nodes and edges for flexible information flow. These methods enhance expressiveness for general tasks; they are not tailored for finance, which requires mathematical precision and domain-specific constraints. We adopt the template from Appendix Listing 2 CoT (Callanan et al., 2024) as the default prompt for this baseline.

Structured Chain-of-Thought (ST-CoT): ST-CoT augments a prompt with tags such as `<thinking>` and `<output>` that break the model's response into explicit, modular stages, promoting incremental, easily replaceable reasoning blocks. This tag-driven format has already appeared in open-source trials.[1] Figure 1 visually contrasts ST-CoT with SP, UST-CoT, and FinCoT.

FinCoT (§3) inherits ST-CoT's structure but injects domain-specific Mermaid blueprints to ground each step in expert workflows. Unlike Universal Self-Adaptive Prompting (Wan et al., 2023), which derives few-shot exemplars from LLM memory, FinCoT encodes human-crafted financial reasoning, favoring interpretability and control. The three categories Standard Prompting (direct instruction), Unstructured CoT (free-form reasoning), and Structured CoT (tag-driven with optional expert hints) offer a unified lens for classifying financial prompting.

2.2 LLMs in Domain-Specific Financial Reasoning

Existing approaches to adapting large foundation models for financial reasoning currently fall into three broad paradigms:

Prompting-based: Methods use few-shot prompts with CFA-style queries, as in "Can GPT Models be Financial Analysts?" (Callanan et al., 2024), which evaluate ChatGPT and GPT-4 on mock CFA exams.

Fine-tuning-based: Methods adapt models via supervised fine-tuning with curated QA/classification datasets (Ma et al., 2023; Harsha et al., 2025) or continued pretraining on domain-specific corpora (Yang et al., 2020; Lee et al., 2024; Bhatia et al., 2024; Ke et al., 2025; Liu et al., 2025). While effective, these approaches rely on labeled data and general language modeling objectives, lacking structuring of financial reasoning, thus limiting interpretability and alignment with expert logic.

[1] https://gist.github.com/davidmezzetti/1fac6bd406857431f1cdc74545bdfba9

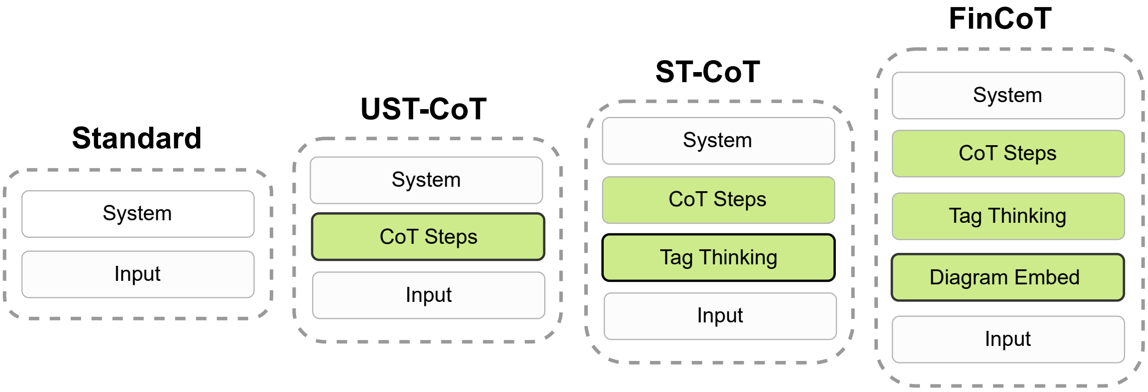
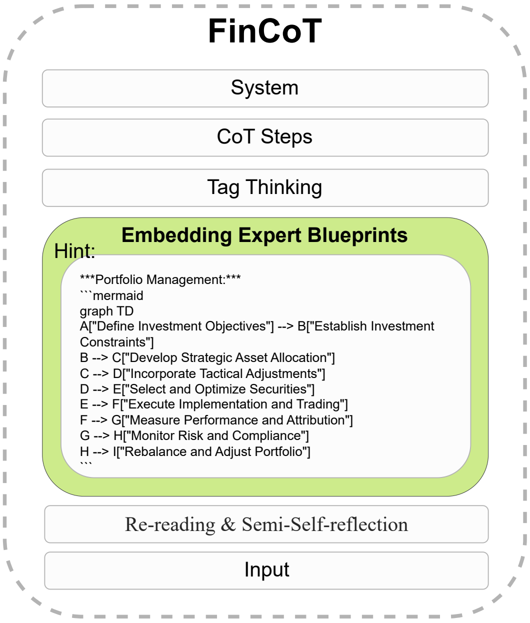


Figure 1: Taxonomy of prompting strategies by reasoning structure: SP, UST-CoT, ST-CoT, and FinCoT. Green blocks indicate added reasoning control–CoT steps, tagged thoughts, and expert diagrams (Diagram Embed)–showing the evolution toward more interpretable, domain-aligned prompts.


FinCoT
System
CoT Steps
Tag Thinking
Embedding Expert Blueprints
Hint:
Re-reading & Semi-Self-reflection
Input

~~~
***Portfolio Management:***
```mermaid
graph TD
A["Define Investment Objectives"] --> B["Establish Investment
Constraints"]
B --> C["Develop Strategic Asset Allocation"]
C --> D["Incorporate Tactical Adjustments"]
D --> E["Select and Optimize Securities"]
E --> F["Execute Implementation and Trading"]
F --> G["Measure Performance and Attribution"]
G --> H["Monitor Risk and Compliance"]
H --> I["Rebalance and Adjust Portfolio"]
```
~~~

Figure 2: Overview of FinCoT, integrates structured, expert-guided reasoning layers Mermaid diagrams, plan-and-solve scaffolding, and reflective validation to improve performance in financial tasks.

Despite advances, prior work has largely overlooked structured reasoning grounded in real-world workflows. We introduce a domain-based prompting framework designed to reflect step-by-step expert logic and evaluate it on CFA-style exam tasks.

3 FinCoT: Augmenting CoT with Structured Financial Expertise

We introduce FinCoT (Financial Chain-of-Thought), a structured prompting framework that enhances LLM reasoning in specialized financial domains. Building upon ST-CoT approaches, FinCoT explicitly embeds expert-derived problem-solving methodologies directly into prompts, guiding LLMs to follow domain-specific reasoning pathways without requiring model fine-tuning. Figure 2 illustrates the FinCoT architecture, which integrates expert-guided reasoning layers and reflective validation to improve performance in financial tasks.

3.1 The FinCoT Prompt Framework

The FinCoT prompting framework integrates the following key components and logical steps:

1. **System:** A single, top-level message that frames the task (e.g., “You are a CFA candidate; treat the following as a finance question”).
2. **Guided Step-by-Step Execution:** The prompt reserves two tag blocks <thinking> for intermediate reasoning and <output> for the final answer-thereby enforcing a structured chain-of-thought (ST-CoT) in one turn.
3. **Expert Reasoning Blueprint (via Mermaid Diagram (Sveidqvist and contributors, 2025):** A domain-specific, embedded expert blueprint with Mermaid diagram concerning the generation of diagrams (see §4), serve as a "Hint" within the context of the prompt. This blueprint explicitly outlines the recommended step-by-step problem-solving strategy for the given financial domain and is curated through a systematic process detailed in section 3.2 to ensure consistency and domain alignment.
4. **Re-Reading & Semi Self-Reflection:** Inside the <output> tags, the model briefly checks

its reasoning for consistency before committing the final answer. We call this "semi-reflection" because we drop the separate <reflection> block-avoiding per-step scoring and self-bias noted by Xu et al. (2024) yet still include a short self-check within <output>.

3.2 Embedding Expert Blueprints

The creation of effective expert reasoning blueprints involves a systematic multistage process designed to transform a wide range of financial knowledge into structured and actionable LLM diagrams.

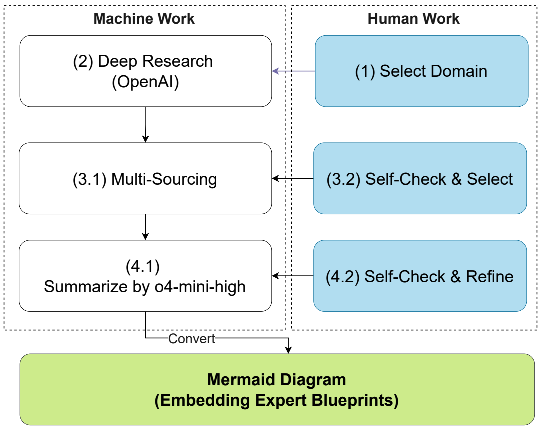

Machine Work
Human Work
(2) Deep Research (OpenAI)
(1) Select Domain
(3.1) Multi-Sourcing
(3.2) Self-Check & Select
(4.1) Summarize by o4-mini-high
(4.2) Self-Check & Refine
Convert
Mermaid Diagram (Embedding Expert Blueprints)

Figure 3: Pipeline for curating financial expert reasoning. Each stage systematically transforms raw financial knowledge into structured Mermaid diagrams for FinCoT prompting.

Curation Pipeline for Expert Reasoning: To construct expert blueprints for each financial domain, we implement a hybrid pipeline combining machine assistance and human judgment, as illustrated in Figure 3. The process includes the following stages:

1. **Scope Definition and Knowledge Aggregation:** The target CFA domain is scoped (e.g., Economics focusing on supply and demand), and relevant expert strategies are aggregated, using Deep Research[2], from diverse authoritative sources. Outputs are validated by human-in-the-loop reviewers with financial knowledge (e.g., finance graduates or CFA charterholders) to ensure conceptual accuracy and domain alignment.
2. **Validation and Synthesis:** We cross-reference and synthesize the aggregated knowledge to ensure accuracy, identify core principles and filter redundancies.
3. **Iterative Refinement into Structured Workflows:** The synthesized expert knowledge is iteratively transformed into logical step-by-step reasoning workflows. This refinement process focuses on ensuring the coherence, correctness, and clarity of the resulting problem-solving strategies for each financial domain.
4. **Mermaid Diagram Generation:** This refined workflow is translated into a Mermaid diagram (Bari et al., 2024) using its text-based syntax. We selected Mermaid due to its LLM prompt compatibility and clear visual guidance aligning with the FinCoT prompt. The diagrams are constructed based on the source content validated and synthesized first in 2, and then applied to each financial domain, with the entire collection organized and described in Appendix A as reference blueprints[3].
5. **Prompt Integration:** The text-based Mermaid blueprint is embedded as "Hint" within the FinCoT prompt template (Appendix B.2), directly guiding the LLM's reasoning.

4 Experimental Setup

Model Configurations and Inference Parameters: We selected the Qwen family of models due to their strong baseline performance in zero-shot financial reasoning. In preliminary evaluations, **Qwen2.5-7B-Instruct** achieved 69.7% accuracy on standard task prompts, substantially outperforming **Llama3.1-8B-Instruct** (46.3%), motivating its use as our primary model family. To evaluate the impact of both instruction tuning and domain-specific adaptation, we compare two model groups.

(A) General-purpose foundation models

- **Qwen2.5-7B (pretrained model)** vs. **Qwen2.5-7B-Instruct**: to assess the effect of instruction tuning on a strong general-purpose foundation model.
- **Qwen3-8B-Base (pretrained model)** vs. **Qwen3-8B**, **Qwen3-8B (Thinker)**: to isolate the impact of ST-CoT prompting within the same architecture.

[2] An AI agent for retrieving/synthesizing knowledge from public sources such as CFA notes, academic texts, and industry guides.

[3] The resulting expert blueprints are reviewed for conceptual consistency and practical correctness (but not guaranteed precision) by CFA Level III charterholders.

- **Gemma-3-12B-IT (Text-only)**: an instruction-tuned model recently released, comparable in size to Qwen3-8B. It achieved 52.81% accuracy on the Flare-CFA benchmark, outperforming Llama3.1-8B-Instruct, and serves as a competitive baseline.

(B) Financial-specific reasoning models

- **Fin-R1 (7B)**: adapted from Qwen2.5-7B-Instruct using supervised and reinforcement learning on a financial dataset distilled from DeepSeek-R1 (Liu et al., 2025).
- **DianJin-R1-7B**: fine-tuned from Qwen2.5-7B-Instruct using CFLUE, FinQA, and CCC, with GRPO to improve domain-specific reasoning (Zhu et al., 2025).
- **Fin-o1-8B**: built on Qwen3-8B and trained on the FinCoT[4] corpus using SFT and GRPO, setting a strong benchmark in financial reasoning (Qian et al., 2025).

All experiments used a maximum sequence length of 16.384k tokens. Following best practices for decoding stability (Du et al., 2025), we set the generation temperature to 0.2 to encourage focused and consistent outputs under evaluation conditions.

Prompting Strategies Compared: Our study evaluates the effectiveness of FinCoT against three baseline prompting strategies: SP, UST-CoT, and ST-CoT, which were detailed in Section 2. For clarity in this section:

- **SP:** The model receives only the target question.
- **UST-CoT:** In addition to the question, the model is given a generic cue to reason step-by-step.
- **ST-CoT:** This strategy employs structural tags (e.g., `<thinking>`, `<output>`) to guide the model in generating an organized step-by-step reasoning trace for the target question.
- **FinCoT:** A zero-shot prompting method that integrates expert domain templates (excluding the Ethics domain). Each prompt includes a Mermaid diagram as a "Hint" to guide structured financial reasoning via relevant domain insights.

While all methods operate in a zero-shot setting, FinCoT uniquely injects expert-guided structure through diagrams. Recent studies suggest that even large reasoning models may struggle with instruction-following when overloaded with reasoning cues (Li et al., 2025; Fu et al., 2025; Jang et al., 2025; Yao et al., 2025), though this remains underexplored in financial contexts. We thus evaluate whether CoT-style prompts (UST-CoT, ST-CoT, FinCoT) enhance instruction-following compared to SP.

Evaluation Benchmark: To assess financial reasoning, we use 1.032k multiple-choice questions from the CFA-Easy subset of FinEval (also referred to as Flare-CFA), originally introduced by Ke et al. (2025). This curated set reflects the rigor of CFA exams and enables evaluation across both theoretical and practical domains. Each question is categorized into one of ten CFA domains using GPT-4o with a dedicated classification prompt (see Appendix B.3), and Figure C shows the resulting domain distribution.

Evaluation Metrics: We report **accuracy** as the metric, defined as the percentage of questions where the model's prediction matches the ground truth. To assess response efficiency, we also report the **average output length** (in tokens) across questions. For statistical significance, we use a paired bootstrap test (Efron and Tibshirani, 1994) with $B{=}10k$ resamples over binary correctness scores, reporting the mean difference (Δ), 95% confidence interval, and p-value. Additionally, we measure the proportion of financial domains where a method improves accuracy by at least 1% over SP and compute the average domain-wise accuracy gain.

5 Results and Discussions

5.1 Financial Reasoning Performance

Baseline Performance: Table 1 reports zero-shot accuracies for four prompting strategies (SP, UST-CoT, ST-CoT, FinCoT) across our model suite. Under the basic SP prompt, the instruction-tuned **Qwen3-8B (Thinker)** attains the highest accuracy among general-purpose models (88.18%), while the financial model **Fin-o1-8B** leads its group at 79.65%. These strong baselines clearly highlight the effectiveness of instruction tuning and domain specificity.

Pretrained Models: On **Qwen2.5-7B**, FinCoT (All Blueprints) yields a +7.95 pp improvement

[4] The FinCoT dataset is constructed by TheFinAI and publicly available at https://huggingface.co/datasets/TheFinAI/FinCoT. It combines financial QA datasets such as FinQA, ConvFinQA, TATQA, DocMath-Eval, Econ-Logic, BizBench-QA, and DocFinQA, with GPT-4o-generated reasoning traces to enhance structured financial question answering. Note that this dataset is not derived from our prompting approach.

Prompt	Accuracy (%)								
	General models						Financial models		
	Qwen2.5-7B	Qwen2.5-7B Instruct	Qwen3-8B Base	Qwen3-8B	Qwen3-8B (Thinker)	Gemma-3-12B IT	Fin-R1 7B	DianJin-R1 7B	Fin-o1 8B
SP	54.07	69.67	63.18	74.42	88.18	52.81	65.70	78.39	**79.65**
UST-CoT	67.83 (↑13.76)	**75.68*** (↑6.01)	72.58 (↑9.40)	**82.36*** (↑7.94)	**89.05*** (↑0.87)	**77.81*** (↑25.00)	75.19 (↑9.49)	67.73 (↓10.66)	79.36 (↓0.29)
ST-CoT	**70.35*** (↑16.28)	74.52 (↑4.85)	78.49 (↑15.31)	81.01 (↑6.59)	88.18	76.74 (↑23.93)	74.32 (↑8.62)	68.80 (↓9.59)	78.39 (↓1.26)
FinCoT	62.02 (↑7.95)	74.22 (↑4.55)	**80.52*** (↑17.34)	81.10 (↑6.68)	87.21 (↓0.97)	75.58 (↑22.77)	**75.78*** (↑10.08)	**79.75*** (↑1.36)	77.23 (↓2.42)
Domain-wise performance of FinCoT									
Economics	69.09 (↑15.02)	73.26 (↑3.59)	79.26 (↑16.08)	79.55 (↑5.13)	86.92 (↓1.26)	74.61 (↑21.80)	73.45 (↑7.75)	55.52 (↓22.87)	78.00 (↓1.65)
FixedIncome	68.12 (↑14.05)	73.35 (↑3.68)	78.88 (↑15.70)	80.81 (↑6.39)	87.21 (↓0.97)	76.45 (↑23.64)	74.22 (↑8.52)	66.86 (↓11.53)	76.74 (↓2.91)
Quant.Meth.	68.02 (↑13.95)	**75.19** (↑5.52)	80.14 (↑16.96)	80.91 (↑6.49)	**87.79** (↓0.39)	75.68 (↑22.87)	74.90 (↑9.20)	65.79 (↓12.60)	77.42 (↓2.23)
EquityInvest.	69.09 (↑15.02)	74.22 (↑4.55)	79.26 (↑16.08)	80.52 (↑6.10)	86.72 (↓1.46)	76.45 (↑23.64)	74.42 (↑8.72)	62.31 (↓16.08)	78.68 (↓0.97)
Port.Manage.	67.54 (↑13.47)	74.13 (↑4.46)	**80.72** (↑17.54)	80.91 (↑6.49)	86.92 (↓1.26)	77.03 (↑24.22)	75.00 (↑9.30)	62.02 (↓16.37)	76.55 (↓3.10)
Derivatives	68.90 (↑14.83)	73.64 (↑3.97)	79.84 (↑16.66)	80.81 (↑6.39)	87.21 (↓0.97)	**77.23** (↑24.42)	**76.16** (↑10.46)	**71.80** (↓6.59)	**79.94** (↑0.29)
Fin. Reporting	**69.28** (↑15.21)	73.84 (↑4.17)	79.07 (↑15.89)	**81.10** (↑6.68)	87.02 (↓1.16)	75.87 (↑23.06)	72.87 (↑7.17)	62.69 (↓15.70)	77.23 (↓2.42)
Alter.Invest.	68.99 (↑14.92)	74.90 (↑5.23)	78.97 (↑15.79)	79.94 (↑5.52)	87.50 (↓0.68)	76.94 (↑24.13)	74.90 (↑9.20)	56.98 (↓21.41)	78.88 (↓0.77)
Corp.Issuers	68.31 (↑14.24)	74.32 (↑4.65)	79.26 (↑16.08)	79.36 (↑4.94)	87.02 (↓1.16)	**77.23** (↑24.42)	75.58 (↑9.88)	60.08 (↓18.31)	79.07 (↓0.58)

Table 1: Comparison of accuracy (%) of prompting techniques. 'FinCoT' simultaneously applies expert reasoning blueprints from all CFA domains, while each '(DomainName)' (e.g., 'Economics') row applies domain-specific blueprints individually. (↑/↓) Denote accuracy improvement or decline relative to the SP baseline, colored green for (↑) and red for (↓). **Bold** values highlight the best-performing prompt variant for each model. (*) Indicates that the accuracy improvement among the model-level prompt variants is statistically significant ($p < 0.05$) based on paired bootstrap testing; domain-specific rows are not tested for significance.

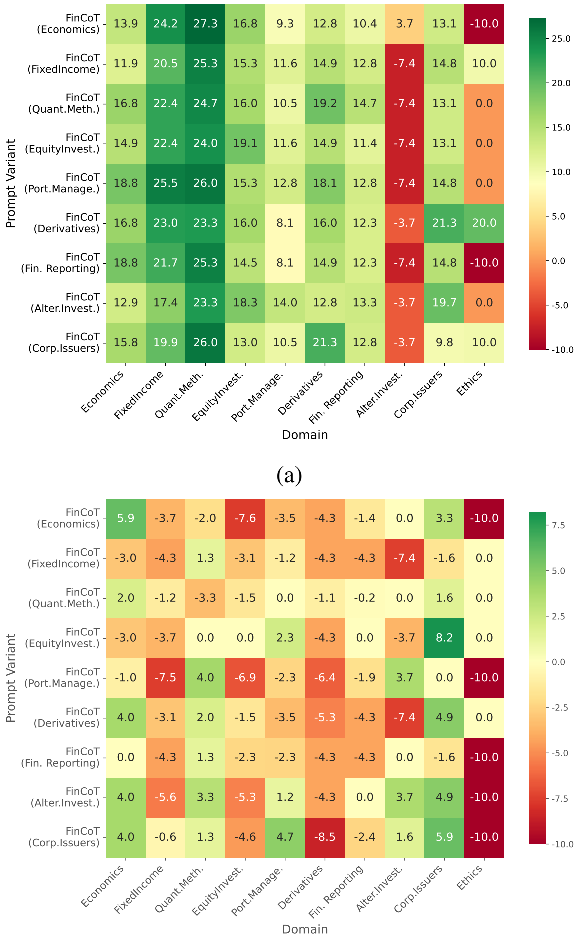
over SP (95% CI [6.30, 9.59], $p < 0.001$), while UST-CoT and ST-CoT also exceed the baseline by +13.76 pp and +16.28 pp. When FinCoT is applied using a single-domain blueprint (e.g., Financial Reporting), the gains increase substantially to +15.21 pp. Similarly, on **Qwen3-8B-Base**, FinCoT delivers the strongest overall boost (+17.35 pp, 95% CI [15.02, 19.77], $p < 0.001$). These findings further underscore the importance of structured domain knowledge, particularly for models lacking instruction or domain alignment.

Instruction Models: We evaluate three prominent instruction-tuned variants. On **Qwen2.5-7B-Instruct**, FinCoT improves accuracy by +4.55 pp over SP, compared to -1.46 pp with UST-CoT and -0.3 pp with ST-CoT. On **Qwen3-8B (Thinker)**, FinCoT yields a slight drop (–0.96 pp), while ST-CoT shows no change and UST-CoT yields a modest +0.87 pp. **Gemma-3-12B-IT**, a strong instruction-tuned baseline (52.81% SP), benefits substantially from all strategies: +25.00 pp (UST-CoT), +23.93 pp (ST-CoT), and +22.77 pp (FinCoT). Notably, domain-specific FinCoT prompts (e.g., Derivatives, Corporate Issuers) provide even larger boosts (+24.42 pp), indicating that blueprint reasoning complements instruction tuning by addressing specialized financial gaps.

Financial-Specific Models: FinCoT also helps specialized models like **Fin-R1**, confirming that blueprint prompting provides complementary gains beyond fine-tuning. However, for models with strong built-in reasoning, such as **DianJin-R1-7B** and **Fin-o1-8B**, FinCoT offers limited improvement or slight degradation—likely due to conflicts between external scaffolds and internal reasoning routines. These outcomes suggest diminishing returns for CoT prompting when domain alignment and reasoning are already deeply encoded.

Overall, FinCoT is most impactful for models lacking prior task-specific adaptation. By grounding reasoning in structured financial workflows, it bridges key gaps in zero-shot settings without requiring additional tuning. This pattern highlights a trade-off between model internalization and prompt-time controllability. *Future work could explore hybrid strategies that adapt prompting depth based on model alignment.*

Cross-domain behavior of FinCoT: This section examines how pretrained and finance-specific models respond to structured prompting with FinCoT. Each domain-specific blueprint is applied across all CFA domains to evaluate its transferability, and accuracy differences relative to SP are measured. Figure 4a and 4b visualize results



(a)

(b)

Figure 4: Accuracy improvements (%) of each FinCoT domain-specific prompt compared to Standard Prompting (SP). Subfigure (a) shows results on Qwen3-8B-Base (pretrained), while (b) shows results on Fin-o1-8B (finance-specific).

for Qwen3-8B-Base (pretrained) and Fin-o1-8B (finance-specific), while Table 1 provides a comprehensive summary of overall model-level accuracy. On Qwen3-8B-Base, FinCoT generally improves performance, though gains are not universal. Prompts from quantitative domains such as *Derivatives*, *Portfolio Management*, and *Corporate Issuers* yield average gains exceeding +13 pp. The blueprint structure provides inductive guidance that enhances decomposition, formula selection, and financial term alignment. In several domains, Qwen3-8B-Base with FinCoT matches or surpasses the SP baseline of Fin-o1-8B, despite the absence of any task-specific training.

On Fin-o1-8B, gains from FinCoT are more modest, typically within the +1–4 pp range. Minor declines appear in some domains (e.g., Fixed Income, Equity Investments), suggesting that additional scaffolding may interfere with optimized internal reasoning acquired during fine-tuning. Structured prompts may over-specify solutions or reduce instruction-following flexibility.

These findings highlight FinCoT's complementary role. For pretrained models, FinCoT acts as a lightweight yet effective augmentation layer at inference time, reducing the performance gap with fine-tuned models. For already fine-tuned models, careful prompt selection or adaptation may be necessary to preserve existing reasoning strengths without introducing conflict. A broader breakdown of FinCoT performance across additional models is provided in Appendix F, where radar plots illustrate domain-wise patterns and complement the main analysis.

5.2 Efficiency Analysis

Effective deployment of foundation models in financial settings requires balancing verbosity and accuracy while emphasizing efficiency. FinCoT offers a prompt-based alternative to fine-tuned models (Fin-o1, DianJin-R1, Fin-R1), demonstrating similar accuracy as seen in Fig. 5, with token lengths detailed in Appendix D.2 (Tab. 3). Our analysis concentrates on output tokens since prompt encoding occurs once with parallel self-attention $O(n_{\text{in}}^2)$ (rapid) (Vaswani et al., 2023), while decoding involves $O(n_{\text{out}})$ sequential steps with per-step KV-cache updates (high memory usage) (Gu et al., 2018; Shazeer, 2019). Comprehensive input and output token data are provided in Appendix D.1(Tab. 2), and recent decoding accelerations such as speculative sampling (Chen et al., 2023) and FlashAttention (Dao et al., 2022) further underscore output as the primary latency driver.

Output Token Length vs. Accuracy: In deployments, prompting strategies that achieve high accuracy with minimal output length are desirable.

For general-purpose models such as Qwen3-8B-Base and Qwen3-8B (Thinker), FinCoT reduces output length while preserving or improving accuracy. On Qwen3-8B-Base, FinCoT improves accuracy from 78.49% (ST-CoT) to 80.52% (+2.03 pp) while reducing average output from 3.42k to 0.38k tokens, an 8.9× compression. On Qwen3-8B (Thinker), FinCoT maintains 88.18% accuracy with tokens dropping from 1.35k to 1.23k ($\sim$1.1×). Results show that FinCoT's structured blueprints enable more concise, focused reasoning in general-purpose models. Among financial-

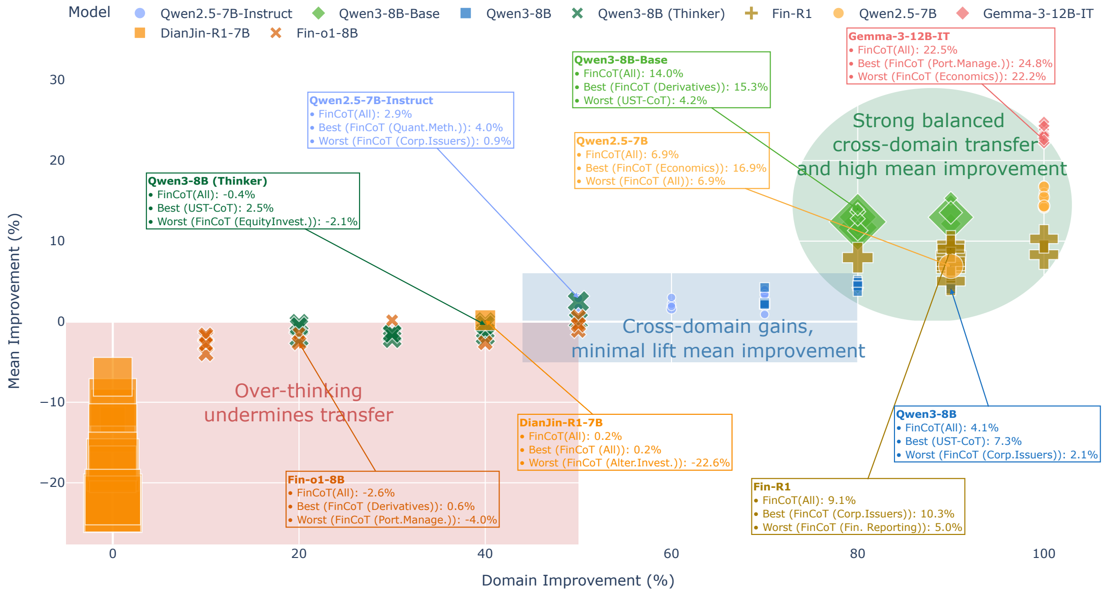


Figure 5: Comparison of prompt strategies across financial domains. Each point represents a domain-method pair, with position indicating accuracy improvement and domain coverage. Circle size encodes generated tokens.

specific models such as DianJin-R1-7B and Fin-o1-8B, FinCoT shows minimal or negative improvement and little benefit in output compression, consistent with fine-tuned models internalizing domain-specific reasoning. Fin-o1-8B suggests that excessive prompt scaffolding may interfere with latent reasoning, reducing effectiveness and leading to overthinking that undermines transfer.

Three behavioral zones emerge from these results. First, models like Gemma-3-12B-IT and Qwen3-8B-Base show high mean improvement and strong cross-domain transfer, reflecting effective generalization. Second, models such as Qwen2.5-7B-Instruct and Qwen3-8B display noticeable cross-domain transfer with lower mean improvement, suggesting limited benefit. Third, models like Fin-o1-8B and Qwen3-8B (Thinker) exhibit low cross-domain adaptability and minimal performance lift, indicating that overly detailed prompting may conflict with internal reasoning.

These findings underscore the importance of aligning the prompting strategy with a model's pretraining or fine-tuning to optimize performance and efficiency. *With limited supervision, FinCoT provides a transparent, cost-effective alternative for enhancing financial reasoning.* For a price–sensitivity analysis of token cost, see Appendix D.2.1 and Fig. 7.

6 Conclusion

We presented **FinCoT**, a zero-shot prompting framework that embeds expert-curated Mermaid diagrams within structured chain-of-thought scaffolds. By grounding reasoning in domain logic, FinCoT bridges human financial workflows with LLM outputs, without model fine-tuning. While broadly applicable, FinCoT yields strong gains for general models (e.g., Qwen3-8B-Base), but more modest or negative effects for instruction or finance-specific models (e.g., Qwen-Thinker, Fin-o1, DianJin-R1), where added structure may interfere with learned reasoning.

Relative to SP, FinCoT improves Qwen3-8B-Base by +17.33 pp and Fin-R1 by +10.08 pp ($p < 0.001$), outperforming some fine-tuned models. In contrast, instruction-tuned models like Qwen3-8B (Thinker) sometimes favor UST-CoT. Cross-domain results show blueprints from quantitative fields transfer best (+27.3 pp on Qwen3-8B-Base). FinCoT also reduces token output by up to $8\times$, offering interpretable, efficient prompting for regulated financial applications. Ultimately, FinCoT suggests that with meticulous prompt design, even general-purpose LLMs can approach the reasoning quality of fine-tuned financial experts in complex decision-making tasks.

Limitations

Our evaluation highlights several limitations. (i) Efficiency gains come mainly from reduced output tokens, but larger inputs from structured templates add cost; overall, FinCoT is still competitive, especially in long-form reasoning. (ii) Domain routing via pre-classification risks template mismatch despite safeguards; adaptive selection methods are needed. (iii) Improvements are uneven, with domains like Alternative Investments and Ethics limited by small samples (∼10–30); larger, balanced benchmarks are required. (iv) Blueprint creation requires expert effort (∼2 hours/domain), and current evaluations are multiple-choice with rubric-based interpretability.

Overall, FinCoT offers structured, auditable reasoning rather than replacing fine-tuned models. Its blueprint methodology and plug-and-play usability demonstrate prompt-level supervision as lightweight knowledge distillation with potential for law, medicine, and engineering.

Acknowledgments

We extend our gratitude to the Hatari Team for generously providing additional GPU resources, which significantly enhanced our computational capacity and enabled the successful execution of large-scale experiments. We also thank Tuntai Boriboonthana (CFA Charterholder) from InnovestX for his expert contributions in verifying blueprint fidelity and conducting random audits to ensure domain classification consistency in GPT-4o outputs. Their support and insights were invaluable to the development and evaluation of this work.

References

300Hours. 2025a. CFA Level 1 Economics Cheat Sheet.

300Hours. 2025b. CFA Level 1 Equity Investments: Our Cheat Sheet.

AnalystPrep. 2025. Economics CFA Level 1 Essential Review Summary.

Arya.ai. 2024. 5 Best Large Language Models (LLMs) for Financial Analysis.

Marco Avellaneda and Sasha Stoikov. 2008. High-Frequency Trading in a Limit Order Book. In *Proceedings of the 12th International Conference on Quantitative Finance*, pages 1–9. Cornell University. Preprint: SSRN 1085964.

Daniele De Bari, Giacomo Garaccione, Riccardo Coppola, Marco Torchiano, and Luca Ardito. 2024. Evaluating large language models in exercises of uml class diagram modeling. In *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '24)*, pages 393–399. ACM.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. Fintral: A family of gpt-4 level multimodal financial large language models. *Preprint*, arXiv:2402.10986.

Fischer Black and Myron Scholes. 1973a. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81(3):637–654.

Fischer Black and Myron Scholes. 1973b. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81(3):637–654.

Bloomberg. 2025. Bloomberg US Aggregate Bond Index.

Financial Stability Board. 2025. OTC Derivatives Market Reforms. Technical report, Financial Stability Board.

Zvi Bodie, Alex Kane, and Alan J. Marcus. 2017. *Investments*, 11th edition. McGraw-Hill Education, New York, NY.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

CAIA Association. 2025. Alternative Investment Management.

Ethan Callanan, Amarachi Mbakwe, Antony Papadimitriou, Yulong Pei, Mathieu Sibue, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2024. Can GPT models be Financial Analysts? An Evaluation of ChatGPT and GPT-4 on mock CFA Exams. In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 23–32, Jeju, South Korea. -.

CFA Institute. 2024. *CFA Program Curriculum 2025: Level II, Volume 2 – Economics*. Wiley. Includes coverage of classical, neoclassical, and endogenous growth models.

CFA Institute. 2025a. CFA Program Curriculum: Alternative Investments.

CFA Institute. 2025b. CFA Program Curriculum: Equity Investments.

CFA Institute. 2025c. CFA Program Curriculum: Equity Investments & Fixed Income.

CFA Institute. 2025d. CFA Program Curriculum: Financial Reporting and Analysis.

CFA Institute. 2025e. CFA Program Curriculum: Fixed Income (Levels I & II).

CFA Institute. 2025f. CFA Program Curriculum: Level II – Derivatives.

CFA Institute. 2025g. CFA Program Curriculum: Portfolio Management.

CFA Institute. 2025h. *CFA Program Curriculum: Quantitative Methods*. CFA Institute, Charlottesville, VA. Levels I & II; covers TVM, probability, hypothesis testing, regression, portfolio stats.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating large language model decoding with speculative sampling. *Preprint*, arXiv:2302.01318.

Aswath Damodaran. 2012. *Investment Valuation: Tools and Techniques for Determining the Value of Any Asset*, 3rd edition. Wiley, Hoboken, NJ.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Preprint*, arXiv:2205.14135.

Ilia Dichev. 2017. On the Balance Sheet-Based Model of Financial Distress Prediction. *The Accounting Review*, 92(4):1125–1152.

Weihua Du, Yiming Yang, and Sean Welleck. 2025. Optimizing Temperature for Language Models with Multi-Sample Inference. *Preprint*, arXiv:2502.05234.

Efficient Learning. 2025. Program Overview: CFA Economics.

Bradley Efron and Robert J Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC press.

Gregor Engels, Reiko Heckel, and Stefan Sauer. 2000. UML – a universal modeling language? In *Proceedings of the 21st International Conference on Application and Theory of Petri Nets (ICATPN 2000)*, volume 1825 of *Lecture Notes in Computer Science*, pages 24–38. Springer, Heidelberg.

Frank J. Fabozzi. 2012. *Bond Markets, Analysis, and Strategies*, 8th edition. Pearson/Prentice Hall.

Tingchen Fu, Jiawei Gu, Yafu Li, Xiaoye Qu, and Yu Cheng. 2025. Scaling reasoning, losing control: Evaluating instruction following in large reasoning models. *Preprint*, arXiv:2505.14810.

Brian Gordon. 2020a. CFA Exam Level 1 Economics Lecture. `https://www.youtube.com/watch?v=SvqKJnN4Tbo`.

Brian Gordon. 2020b. CFA Level I: Equity Investments Preview. `https://www.youtube.com/watch?v=SvqKJnN4Tbo`.

Brian Gordon. 2020c. CFA Level I: Equity Investments Revision. `https://www.youtube.com/watch?v=SvqKJnN4Tbp`.

Richard C. Grinold and Ronald N. Kahn. 2000. *Active Portfolio Management*. McGraw-Hill.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. *Preprint*, arXiv:1711.02281.

Chetan Harsha, Karmvir Singh Phogat, Sridhar Dasaratha, Sai Akhil Puranam, and Shashishekar Ramakrishna. 2025. Synthetic data generation using large language models for financial question answering. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 76–95, Abu Dhabi, UAE. Association for Computational Linguistics.

Yi Hu, Haotong Yang, Zhouchen Lin, and Muhan Zhang. 2023. Code Prompting: a Neural Symbolic Method for Complex Reasoning in Large Language Models. *Preprint*, arXiv:2305.18507.

John C. Hull. 2017. *Options, Futures, and Other Derivatives*, 10th edition. Pearson.

Investopedia. 2025. Aggregate Demand and Supply (AD–AS) Model.

Investopedia. 2025. Asset-Backed Security (ABS).

Investopedia. 2025a. Balance of Payments.

Investopedia. 2025b. Business Cycle: Definition & 4 Phases.

Investopedia. 2025a. Consumer Price Index (CPI).

Investopedia. 2025b. Convexity in Bonds: Definition, Meaning, and Examples.

Investopedia. 2025c. Coupon Rate Definition.

Investopedia. 2025d. Credit Spread Definition.

Investopedia. 2025e. Duration Definition and Its Use in Fixed Income Investing.

Investopedia. 2025. Exchange Rate Definition.

Investopedia. 2025a. Financial Statements: List of Types and How to Read Them.

Investopedia. 2025b. Fixed-Income Security Definition, Types, and Examples.

Investopedia. 2025a. Fundamental Analysis: Principles, Types, and How to Use It.

Investopedia. 2025b. Government Regulations: Do They Help Businesses?

Investopedia. 2025c. Gross Domestic Product (GDP) Formula.

Investopedia. 2025d. How to Read a Financial Analysis Report.

Investopedia. 2025. Interest Rate Definition & Impact.

Investopedia. 2025a. Law of Supply and Demand.

Investopedia. 2025b. Monetary Policy Definition.

Investopedia. 2025c. Phillips Curve: Trade-Off Between Inflation and Unemployment.

Investopedia. 2025d. Porter's Five Forces Definition.

Investopedia. 2025e. Price Elasticity of Demand.

Investopedia. 2025f. Simple vs. Compound Interest: Definition and Formulas. Explains TVM basics and formula examples.

Investopedia. 2025g. The Rule of 72: What It Is and How to Use It in Investing. Provides quick doubling-time approximation.

Investopedia. 2025a. When to Rebalance a Bond Portfolio.

Investopedia. 2025b. Yield Curve Definition & Types.

iPassFinanceExams. 2025. CFA Economics Study Tips.

Raymond James. 2025. Fixed Income Strategies.

Doohyuk Jang, Yoonjeon Kim, Chanjae Park, Hyun Ryu, and Eunho Yang. 2025. Reasoning model is stubborn: Diagnosing instruction overriding in reasoning models. *Preprint*, arXiv:2505.17225.

Robert A. Jarrow and Stuart M. Turnbull. 1995. Pricing Derivatives on Financial Securities Subject to Credit Risk. *Journal of Finance*, 50(1):53–85.

Robert A. Jarrow and Stuart M. Turnbull. 1996. *Derivative Securities*. South-Western College Publishing.

Kaplan Schweser. 2025. Level I CFA Economics Study Tips.

Zixuan Ke, Yifei Ming, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Demystifying Domain-adaptive Post-training for Financial LLMs. *Preprint*, arXiv:2501.04961.

Jean Lee, Nicholas Stevens, and Soyeon Caren Han. 2025. Large language models in finance (finllms). *Neural Computing and Applications*.

Sangmin Lee, Suzie Oh, Saeran Park, Guijin Son, and Pilsung Kang. 2024. FINALE : Finance domain instruction-tuning dataset with high-quality rationales via chain-of-thought prompting. In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 89–106, Jeju, South Korea. -.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. *Preprint*, arXiv:2206.14858.

Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. 2025. When thinking fails: The pitfalls of reasoning for instruction-following in llms. *Preprint*, arXiv:2505.11423.

Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, Chao Li, Sheng Xu, Dezhi Chen, Yun Chen, Zuo Bai, and Liwen Zhang. 2025. Fin-R1: A Large Language Model for Financial Reasoning through Reinforcement Learning. *Preprint*, arXiv:2503.16252.

Xuezhi Ma, Yan Zhou, Yining Wang, and 1 others. 2023. Financialqa: A reasoning benchmark for financial question answering. *arXiv preprint arXiv:2302.07304*.

Harry Markowitz. 1952. *Portfolio Selection*, volume 7 of *Journal of Finance Classics*. Wiley, New York.

Andrew Metrick and Ayako Yasuda. 2010. *Private Equity and Venture Capital*. Wiley.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. *arXiv preprint arXiv:2406.11903*.

Krishna G. Palepu, Paul M. Healy, Sue Wright, Michael Bradbury, and Jeff Coulton. 2013. *Business Analysis and Valuation: Using Financial Statements*. Cengage.

Stephen H. Penman. 2012. *Financial Statement Analysis and Security Valuation*, 5th edition. McGraw-Hill Education.

Jerald E. Pinto, Elaine Henry, Thomas R. Robinson, and John D. Stowe. 2015. *Equity Asset Valuation*. CFA Institute Investment Series. Wiley, Hoboken, NJ.

Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Han Yi, Yilun Zhao, Jimin Huang, Qianqian Xie, and Jian yun Nie. 2025. Fino1: On the transferability of reasoning-enhanced llms and reinforcement learning to finance. *Preprint*, arXiv:2502.08127.

Matthew Renze and Erhan Guven. 2024. The Benefits of a Concise Chain of Thought on Problem-Solving in Large Language Models. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, page 476–483. IEEE.

R. Tyrrell Rockafellar and Stanislav Uryasev. 2000. Optimization of Conditional Value-at-Risk. Technical Report IFOR Technical Report TR00-27, University of Florida.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *Preprint*, arXiv:1911.02150.

Clyde P. Stickney, Paul R. Brown, and James M. Wahlen. 2007. *Financial Reporting, Financial Statement Analysis, and Valuation: A Strategic Perspective*. Prentice Hall.

Knut Sveidqvist and contributors. 2025. Mermaid: A javascript-based diagramming and charting tool. GitHub repository.

Dominique Tavella and Curt Randall. 2000. *Pricing Financial Instruments: The Finite Difference Method*. Wiley.

Bruce Tuckman and Angel Serrat. 2011. *Fixed Income Securities: Tools for Today's Markets*, 3rd edition. Wiley.

UWorld Finance. 2025a. CFA® Economics: Syllabus & Sample Questions.

UWorld Finance. 2025b. CFA® Finance Study Resources.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Xingchen Wan, Ruoxi Sun, Hootan Nakhost, Hanjun Dai, Julian Eisenschlos, Sercan Arik, and Tomas Pfister. 2023. Universal self-adaptive prompting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7437–7462, Singapore. Association for Computational Linguistics.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Preprint*, arXiv:2201.11903.

Bo Wen and Xin Zhang. 2025. Enhancing reasoning to adapt large language models for domain-specific applications. *Preprint*, arXiv:2502.04384.

Wikipedia contributors. 2025. Time Value of Money. Overview of present and future value concepts.

Jeffrey M. Wooldridge. 2013. *Introductory Econometrics: A Modern Approach*, 5th edition. Cengage Learning, Boston, MA.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. 2025. Are reasoning models more prone to hallucination? *Preprint*, arXiv:2505.23646.

Jie Zhu, Qian Chen, Huaixia Dou, Junhui Li, Lifan Guo, Feng Chen, and Chi Zhang. 2025. Dianjin-r1: Evaluating and enhancing financial reasoning in large language models. *Preprint*, arXiv:2504.15716.

A Expert Reasoning Blueprints

Economics

~~~
***Economics:***
```mermaid
graph TD;
  A[Step 1: Question Breakdown] -->|Extract key terms| A1{Identify Topic}
  A1 -->|Micro: Supply & Demand, Market Structures| A2
  A1 -->|Macro: GDP, Growth, Policy, Trade| A3
  A1 -->|Currency & Regulation| A4

  A2 --> B1[Identify model: Elasticity, Cost Curves, Shutdown Points]
  A3 --> B2[Map to AD-AS, Business Cycles, Growth Theories]
  A4 --> B3[Assess Exchange Rates, Trade, Capital Flows, Regulation]

  B1 -->|Check for formula or concept?| C{Numerical or Conceptual}
  B2 --> C
  B3 --> C

  C -->|Numerical| D1[Extract data, apply formulas, check assumptions]
  C -->|Conceptual| D2[Analyze cause-effect, policy impact]

  D1 --> E[Step 4: Solution Development]
  D2 --> E
  E -->|Construct structured response| E1(Core insight + economic rationale)
  E -->|Consider alternative scenarios| E2(Assess different possibilities)

  E1 --> F[Step 5: Answer Validation]
  E2 --> F
  F -->|Check logic, principles, and assumptions| F1(Verify consistency)
  F1 -->|Ensure completeness & clarity| F2(Confirm answer structure)
```
~~~

Explanation: Step-1: Question Breakdown (A) – Extract key terms by parsing the question to see whether it focuses on microeconomics, macroeconomics, or currency/regulation topics (300Hours, 2025a; UWorld Finance, 2025a).

Step-2: Identify Topic (A1) – Microeconomics (A2): Focus on supply & demand mechanisms and market structures such as perfect competition, monopoly, oligopoly, and monopolistic competition (Investopedia, 2025a; 300Hours, 2025a). – Macroeconomics (A3): Consider aggregate demand–aggregate supply analysis, phases of the business cycle (expansion, peak, contraction, trough), and growth models (Solow, endogenous growth) (Investopedia, 2025b; CFA Institute, 2024). – Currency & Regulation (A4): Examine exchange-rate regimes (floating vs. pegged), trade balances, capital-flow impacts, and relevant government policies (Investopedia, 2025,b).

Step-3: Model Selection or Strategy Mapping (B1–B3) – Micro Models (B1): Choose elasticity calculations and cost-curve analysis (marginal/average cost, shutdown point) for supply–demand or firm-behavior questions (Investopedia, 2025e). – Macro Frameworks (B2): Apply AD–AS curves, Phillips-curve trade-offs, or business-cycle indicators to frame policy or growth analysis (Investopedia, 2025,c). – FX & Regulation (B3): Use exchange-rate determination models, balance-of-payments analysis, or regulatory impact frameworks for currency/trade questions (Investopedia, 2025a).

Step-4: Determine Numerical vs. Conceptual Approach (C) – Numerical (D1): Gather the relevant data (prices, quantities, rates), apply formulae (e.g., Elasticity $= \frac{\%\Delta Q_n}{\%\Delta P}$, GDP $= C + I + G + (X - M)$), and verify assumptions (Investopedia, 2025c). – Conceptual (D2): Construct a narrative explaining cause–effect relationships (e.g., how a monetary-policy change shifts AD or how trade barriers affect capital flows) (Investopedia, 2025b).

Step-5: Solution Development (E) – Structured Response (E1): State the core economic insight first, then provide the step-by-step rationale linking theory to the question context (Kaplan Schweser, 2025). – Alternative Scenarios (E2): Where relevant, outline best-case, base-case, and worst-case scenarios or show how supply–demand curves shift under different assumptions (iPassFinanceExams, 2025).

Step-6: Answer Validation (F) – Verify Consistency (F1): Check that numerical answers satisfy boundary conditions (e.g., correct sign on elasticity, GDP component sums). – Confirm Clarity (F2): Ensure your explanation is complete, logically ordered, and clearly communicates both the result and its limitations (UWorld Finance, 2025b).

Source: 300Hours CFA Level 1 Economics Cheat Sheet (300Hours, 2025a), UWorld Finance's CFA® Economics: Syllabus & Sample Questions (UWorld Finance, 2025a), Kaplan Schweser's Level I Economics tips (Kaplan Schweser, 2025), Efficient Learning's CFA Economics overview (Efficient Learning, 2025), iPass Finance Exams' study guide (iPassFinanceExams, 2025), AnalystPrep's essential review summary (AnalystPrep, 2025), and key Investopedia articles on supply and demand (Investopedia, 2025a), GDP (Investopedia, 2025c), and business cycles (Investopedia, 2025b) as well as Prof. Brian Gordon's CFA Exam Level 1 Economics video (Gordon, 2020a).

Fixed Income

~~~
***Fixed Income:***
```mermaid
graph TD
    A[Purpose and Scope] --> B3[Analyze Macro Conditions]
    B --> C[Assess Bond Features]
    C --> D[Risk and Yield Analysis]
    D --> E[Develop Recommendations]
    E --> F[Review Performance]

    %% Notes and detailed steps
    A --> |Set objectives| B
    B --> |Review interest rates and inflation| C
    C --> |Focus on duration, spread| D
    D --> |Assess scenarios| E
```
~~~

Explanation: Step-1: Purpose and Scope – Define the investment objective–income generation, capital preservation, hedging, or total return and establish portfolio constraints and benchmarks such as target yield, duration limits, credit quality floors, or sector allocation guidelines (Investopedia, 2025b).

Step-2: Analyze Macro Conditions – Examine current and forecast interest rate paths, since rising rates erode bond prices and falling rates support them (Investopedia, 2025); monitor inflation indicators (CPI, PPI) to gauge real yield trends (Investopedia, 2025a); and assess yield-curve shapes (normal, inverted, flat) for economic turning points and yield-curve trade opportunities (Investopedia, 2025b).

Step-3: Assess Bond Features – Identify bond type government, corporate, municipal, structured products (ABS/MBS) and note any embedded options (callable, putable, convertible) (Investopedia, 2025); review coupon structure (fixed vs. floating), payment frequency, and maturity to understand cash-flow timing and reinvestment risk (Investopedia, 2025c).

Step-4. Risk and Yield Analysis – Calculate duration to estimate price sensitivity to yield changes (Investopedia, 2025e) and convexity for non-linear price effects (Investopedia, 2025b); analyze credit spreads over benchmarks to gauge default and liquidity risk (Investopedia, 2025d); and stress-test the portfolio under parallel shifts, steepeners, and flatteners to assess P&L impacts.

Step-5: Develop Recommendations – Formulate strategies such as adjusting overall duration (shorten if rates are likely to rise), implementing barbell or laddered maturity structures, or choosing bullet portfolios to manage reinvestment and rate risk (James, 2025).

Step-6: Review Performance – Track total returns (price changes plus coupon income) against benchmarks like the Bloomberg US Aggregate Bond Index (Bloomberg, 2025); perform attribution analysis to decompose yield carry, curve roll-down, and spread effects; and revisit assumptions and rebalance when market conditions or issuer fundamentals change (Investopedia, 2025a).

Source: CFA Program Curriculum for Fixed Income (CFA Institute, 2025e), Fabozzi's Bond Markets, Analysis, and Strategies (Fabozzi, 2012), Tuckman & Serrat's Fixed Income Securities (Tuckman and Serrat, 2011), Jarrow & Turnbull's credit-risk derivatives pricing (Jarrow and Turnbull, 1995), Basel Committee papers on credit risk, Investopedia articles on fixed-income concepts (Investopedia, 2025b,b,e), Reuters coverage of convexity risk, and Dichev's balance sheet model for distress prediction (Dichev, 2017).

Quantitative Methods

~~~
***Quantitative Methods:***
```mermaid
graph TD
    A["Articulating Purpose and Context"] --> B["Collecting Input Data"]
    B --> C["Processing and Cleaning Data"]
    C --> D["Selecting Quantitative Models and Tools"]
    D --> E["Estimating Parameters and Testing Hypotheses"]
    E --> F["Interpreting Results and Communicating Findings"]
    F --> G["Monitoring and Model Reassessment"]
```
~~~

Explanation: Step-1: Articulating Purpose and Context (A) Define the research question time value of money calculations, probability distributions, hypothesis testing, regression analysis, or portfolio statistics–and establish the CFA application context (market efficiency, risk estimation, cash-flow forecasting) (CFA Institute, 2025h).

Step-2: Collecting Input Data (B) Gather historical returns, economic indicators, financial statements, and market data from reputable sources; ensure relevance by matching data to the chosen objective (e.g., interest rates for TVM, volatility for risk models) (Investopedia, 2025f).

Step-3: Processing and Cleaning Data (C) Perform data quality checks and remove outliers, handle missing values, confirm consistency–and apply transformations (normalization, log-transforms) before analysis (Wooldridge, 2013).

Step-4: Selecting Quantitative Models and Tools (D) Choose appropriate models–ARIMA for time series, linear/multivariate regression, probability distributions, or Monte Carlo simulation–and leverage CFA-recommended software or spreadsheet tools (Damodaran, 2012; Investopedia, 2025g).

Step-5: Estimating Parameters and Testing Hypotheses (E) Estimate model parameters via regression or maximum likelihood; conduct t-tests, F-tests, or chi-square tests to validate assumptions and results, with Level II emphasis on multivariate regression and sensitivity analysis (Wooldridge, 2013).

Step-6: Interpreting Results and Communicating Findings (F) Translate coefficients, p-values, and confidence intervals into actionable investment insights; prepare clear visual aids (charts, tables) to support recommendations (Bodie et al., 2017).

Step-7: Monitoring and Model Reassessment (G) Track out-of-sample performance against benchmarks; update models as new data arrive, reassess assumptions, and recalibrate parameters to maintain relevance (Wikipedia contributors, 2025).

Source: CFA Program Curriculum: Quantitative Methods (CFA Institute, 2025h), Wooldridge's Introductory Econometrics (Wooldridge, 2013), Damodaran's Investment Valuation (Damodaran, 2012), and Bodie, Kane, & Marcus's Investments (Bodie et al., 2017).

Equity Investments

~~~
***Equity Investing:***
```mermaid
graph TD
    A[Objective Setting] --> B[Market and Sector Insights]
    B --> C[Industry Competitive Analysis]
    C --> D[Company Review]
    D --> E[Valuation and Risks]
    E --> F[Investment Decision]

    %% Step-specific highlights
    B --> |Look at growth patterns| C
    C --> |Evaluate competitors' positions| D
    D --> |Check financial health| E
    E --> |Combine insights into strategy| F
```
~~~

Explanation Step-1: Objective Setting (A) Define your investment objectives–capital appreciation, dividend income, or total return in line with your risk tolerance and investment horizon; consider constraints such as liquidity needs, tax implications, regulatory requirements, and any specific mandates (CFA Institute, 2025b).

Step-2: Market and Sector Insights (B) Assess macro indicators (GDP growth, interest rates, inflation) to gauge the overall market environment and identify sectors poised for growth or decline based on economic trends, technological shifts, and consumer behavior (Investopedia, 2025a).

Step-3: Industry Competitive Analysis (C) Apply Porter's Five Forces to evaluate industry attractiveness–competitive rivalry, threat of new entrants, bargaining power of suppliers and buyers, and substitute threats–and assess each firm's market share and competitive moat (Investopedia, 2025d).

Step-4: Company Review (D) Examine financial statements (income statement, balance sheet, cash flows) to measure profitability, liquidity, and stability; evaluate management's track record and strategic vision; and review corporate governance structures to ensure alignment with shareholder interests (Investopedia, 2025d; Bodie et al., 2017).

Step-5: Valuation and Risks (E) Use valuation methods–Discounted Cash Flow (DCF), Price-to-Earnings (P/E) ratios, Dividend Discount Models (DDM)–to estimate intrinsic value; identify key risks such as market volatility, operational challenges, regulatory changes, and competitive threats (Pinto et al., 2015).

Step-6: Investment Decision (F) Formulate your Buy, Hold, or Sell recommendation based on the above analyses and determine how the position fits within the broader portfolio–considering diversification, correlation, and overall risk–return objectives (300Hours, 2025b).

Source: CFA Program Curriculum's Equity Investments module (CFA Institute, 2025b), Investopedia's guides on fundamental analysis (Investopedia, 2025a), Porter's Five Forces stock analysis (Investopedia, 2025d), and reading financial reports (Investopedia, 2025d), 300Hours' CFA Level 1 Equity Cheat Sheet (300Hours, 2025b), Bodie, Kane & Marcus's Investments (Bodie et al., 2017), Pinto et al.'s Equity Asset Valuation (Pinto et al., 2015), and CFA Level I Equity video lectures by Prof. Brian Gordon (Gordon, 2020b,c).

Portfolio Management

~~~
***Portfolio Management:***
```mermaid
graph TD
    A["Define Investment Objectives"] --> B["Establish Investment Constraints"]
    B --> C["Develop Strategic Asset Allocation"]
    C --> D["Incorporate Tactical Adjustments"]
    D --> E["Select and Optimize Securities"]
    E --> F["Execute Implementation and Trading"]
    F --> G["Measure Performance and Attribution"]
    G --> H["Monitor Risk and Compliance"]
    H --> I["Rebalance and Adjust Portfolio"]
```
~~~

Explanation: Step-1: Define Investment Objectives – Clarify whether the portfolio is aimed at capital growth, income generation, or a balanced mix. Specify expected returns, risk tolerance, and liquidity needs. This step forms the foundation for aligning investment strategy with client mandates (CFA Institute, 2025g).

Step-2: Establish Investment Constraints – Define legal, regulatory, tax, and unique client considerations such as ESG preferences or geographic limits. These constraints ensure feasibility and compliance of portfolio design (CFA Institute, 2025g).

Step-3: Develop Strategic Asset Allocation – Allocate across major asset classes (equities, fixed income, alternatives, cash) based on expected returns and risk tolerance. Use models from Modern Portfolio Theory and CAPM to inform allocation (Markowitz, 1952; Bodie et al., 2017).

Step-4: Incorporate Tactical Adjustments – Introduce short-term adjustments to the strategic allocation based on market outlook or economic indicators. These shifts aim to enhance returns through asset or sector rotation (Grinold and Kahn, 2000).

Step-5: Select and Optimize Securities – Apply quantitative screens and qualitative research to choose securities. Use optimization techniques such as mean-variance optimization or the Black-Litterman model to maximize risk-adjusted returns (Bodie et al., 2017; Grinold and Kahn, 2000).

Step-6: Execute Implementation and Trading – Implement trade strategies that minimize costs and slippage, considering market impact and liquidity. Align execution with strategic intentions (CFA Institute, 2025g).

Step-7: Measure Performance and Attribution – Track performance using return metrics, Sharpe ratio, alpha, and beta. Perform attribution to evaluate decisions across asset allocation, sector, and security selection (Grinold and Kahn, 2000).

Step-8: Monitor Risk and Compliance – Use tools like Value-at-Risk (VaR), stress testing, and tracking error to monitor portfolio risk. Ensure compliance with constraints and regulations (CFA Institute, 2025g).

Step-9: Rebalance and Adjust Portfolio – Periodically adjust the portfolio to maintain alignment with the strategic asset allocation as market conditions evolve.

Source: CFA Program Curriculum's Portfolio Management module (CFA Institute, 2025g), Bodie, Kane & Marcus's Investments for portfolio theory and risk-return optimization (Bodie et al., 2017), Grinold & Kahn's Active Portfolio Management for advanced attribution and optimization techniques (Grinold and Kahn, 2000), and Markowitz's seminal Portfolio Selection on diversification and risk-adjusted returns (Markowitz, 1952).

Derivatives

~~~~~
***Derivatives:***
```mermaid
graph TD
    A[Define Objective and Context] --> B[Identify Derivative Instrument]
    B --> C[Understand Contract Specifications]
    C --> D[Gather Market Data]
    D --> E[Apply Valuation Models]
    E --> F[Assess Risks: Market, Counterparty, etc.]
    F --> G[Construct Payoff Diagrams or Strategies]
    G --> H[Interpret Results and Make Recommendations]
    H --> I[Review, Monitor, and Adjust Strategies]

    %% Example labels or notes (optional)
    A --> |Hedging, speculation, arbitrage| B
    C --> |Features like notional amount, expiration| D
    D --> |Market prices, volatility, risk-free rates| E
    F --> |Sensitivity to Greeks: Delta, Gamma, Vega, etc.| G
    H --> |Adjust based on changing market conditions| I
````
~~~~~

Explanation: Step-1: Define Objective and Context – Clarify the purpose of using derivatives: hedging, speculation, or arbitrage. Identify relevant constraints, such as regulatory limitations or portfolio mandates (CFA Institute, 2025f; Hull, 2017).

Step-2: Identify Derivative Instrument – Choose the appropriate derivative: options, futures, forwards, swaps, or structured/exotic products (Jarrow and Turnbull, 1996).

Step-3: Understand Contract Specifications – Review contract parameters, including the underlying asset, strike price, expiration, settlement method (physical or cash), and style (European, American) (CFA Institute, 2025f).

Step-4: Gather Market Data – Collect input variables such as spot price, volatility, risk-free rate, dividends, and term structure of interest rates (Hull, 2017).

Step-5: Apply Valuation Models – Apply pricing frameworks suited to the derivative:

- Black-Scholes model for European options (Black and Scholes, 1973b).
- Binomial Tree for path-dependent or American-style options (Hull, 2017).
- Cost-of-carry model for futures and forwards (Jarrow and Turnbull, 1996).
- Finite-difference methods for complex derivatives (Tavella and Randall, 2000).

Step-6: Assess Risks – Use Greeks (Delta, Gamma, Vega, Theta, Rho) to evaluate sensitivity to market factors. Consider counterparty and credit risk in OTC markets (Hull, 2017; Board, 2025).

Step-7: Construct Payoff Diagrams or Strategies – Visualize outcomes using payoff graphs. Design strategies such as straddles, collars, or protective puts based on desired exposure (Hull, 2017).

Step-8: Interpret Results and Make Recommendations – Translate model output into actionable insights: confirm hedge effectiveness, profit potential, or risk exposure.

Step-9: Review, Monitor, and Adjust Strategies – Continuously monitor derivative positions in light of market conditions, risk metrics, and investment objectives (Board, 2025).

Source: Based on Hull's comprehensive treatment of markets and pricing models (Hull, 2017), the CFA Institute Level II Derivatives readings (CFA Institute, 2025f), Black & Scholes's seminal option pricing model (Black and Scholes, 1973b), Jarrow & Turnbull's practical engineering perspective (Jarrow and Turnbull, 1996), Tavella & Randall's numerical finite-difference techniques (Tavella and Randall, 2000), and the Basel Committee's OTC derivatives reforms for regulatory context (Board, 2025).

Financial Reporting

~~~
***Financial Reporting:**
```mermaid
graph TD
A[Articulating Purpose and Context] --> B[Collecting Input Data]
    B --> C[Processing Data]
    C --> D[Analyzing and Interpreting Processed Data]
    D --> E[Developing and Communicating Conclusions]
    E --> F[Doing Follow-Up]

    A --> |Defines goals, tools, and audience| B
    B --> |Gather data on economy and industry| C
    C --> |Use tools like ratios and charts| D
    D --> |Interpret data for conclusions| E
    F --> |Periodic review and iteration| A
```
~~~

Explanation: Step-1: Articulating Purpose and Context
Define the objectives of the analysis–such as assessing profitability, liquidity, or solvency. Identify stakeholders (e.g., investors, creditors, management) and tailor the analysis to their needs. Set the framework, including accounting standards (IFRS or US GAAP) and the time horizon (CFA Institute, 2025d).

Step-2: Collecting Input Data
Gather primary financial statements: income statement, balance sheet, and cash flow statement. Supplement this with industry benchmarks and macroeconomic data. Ensure the quality, accuracy, and completeness of all collected data (Investopedia, 2025a).

Step-3: Processing Data
Standardize data for comparability by adjusting for non-recurring items or differences in accounting policies. Compute financial ratios such as ROE, current ratio, and debt-to-equity. Use visualizations (e.g., charts, graphs) to uncover trends and patterns (Stickney et al., 2007).

Step-4: Analyzing and Interpreting Processed Data
Assess financial health by interpreting computed ratios. Benchmark against peer companies and industry averages. Identify strengths and weaknesses to determine strategic implications (Palepu et al., 2013).

Step-5: Developing and Communicating Conclusions
Summarize findings in a clear, concise report. Offer actionable recommendations–e.g., restructuring debt or improving efficiency. Tailor communication style and depth to fit the audience, whether board members, analysts, or external investors.

Step-6: Doing Follow-Up
Monitor outcomes of implemented actions and assess whether financial targets are met. Update the analysis regularly with new data and refine recommendations. Incorporate feedback to improve future analysis cycles.

Source: CFA Program Curriculum's Financial Reporting and Analysis readings covering ratio analysis, cash flow analysis, and IFRS/GAAP standards (CFA Institute, 2025d) alongside Investopedia's overview of financial statement components (Investopedia, 2025a), Paul R. Brown's strategic perspective on statement analysis and valuation (Stickney et al., 2007), and Palepu & Healy's MBA-level treatment of business analysis and valuation using financial statements (Palepu et al., 2013).

Alternative Investments

~~~
***Alternative Investments:***
```mermaid
graph TD
    A["Define Investment Objectives and Mandate"] --> B["Identify Alternative Asset Classes"]
    B --> C["Conduct Manager and Strategy Due Diligence"]
    C --> D["Perform Valuation and Pricing Analysis"]
    D --> E["Assess Risk and Liquidity"]
    E --> F["Allocate Alternatives in Portfolio"]
    F --> G["Monitor Performance and Rebalance"]
~~~
```

Explanation: Step-1: Define Investment Objectives and Mandate – Clarify the purpose of including alternative investments–whether for diversification, higher return potential, or hedging against market volatility. Define constraints such as time horizon, liquidity needs, regulatory frameworks, and risk tolerance (CFA Institute, 2025a).

Step-2: Identify Alternative Asset Classes – Explore the universe of alternatives, including hedge funds, private equity, real estate, infrastructure, commodities, and venture capital. Assess how each class contributes to portfolio diversification via low correlation to traditional assets (Bodie et al., 2017; CAIA Association, 2025).

Step-3: Conduct Manager and Strategy Due Diligence – Evaluate managers based on their track record, investment philosophy, risk management, and operational quality. Understand the specific strategies (e.g., long/short, event-driven, global macro) and their alignment with investment mandates (CAIA Association, 2025; Metrick and Yasuda, 2010).

Step-4: Perform Valuation and Pricing Analysis – Address the unique valuation challenges of illiquid assets. Use models like discounted cash flow (DCF) or mark-to-model, and apply appropriate liquidity or opacity discounts. Compare performance with custom or market benchmarks (Metrick and Yasuda, 2010).

Step-5: Assess Risk and Liquidity – Identify key risks including market, manager, and operational risks. Analyze downside risk and tail event exposure. Evaluate liquidity risks, such as lock-up periods and redemption windows, that may affect rebalancing ability (CFA Institute, 2025a).

Step-6: Allocate Alternatives in Portfolio – Determine appropriate weighting of alternative assets, guided by expected return, volatility, and correlation with traditional investments. Make strategic allocation decisions with room for tactical adjustments based on market conditions (Bodie et al., 2017).

Step-7: Monitor Performance and Rebalance – Track returns over time, evaluate them against relevant benchmarks, and assess if performance remains consistent with expectations. Rebalance periodically to ensure alignment with objectives, risk profile, and current market landscape (CAIA Association, 2025).

Source: CFA Program Curriculum's Alternative Investments readings covering hedge funds, private equity, real assets, and due diligence frameworks (CFA Institute, 2025a)–together with Metrick & Yasuda's deep dive into private equity and venture capital (Metrick and Yasuda, 2010), CAIA Association's comprehensive CAIA-level materials on hedge funds, real estate, commodities, and other alternatives (CAIA Association, 2025), and Bodie, Kane & Marcus's chapters on alternative asset classes and portfolio integration in *Investments* (Bodie et al., 2017).

Corporate Issuers

```
***Corporate Issuer Analysis:***
```mermaid
graph TD
    A["Corporate Issuer Overview"] --> B["Industry Classification"]
    B --> C["Sector Trends and Competitive Landscape"]
    A --> D["Financial Statement Analysis"]
    D --> E["Profitability, Liquidity, Leverage"]
    A --> F["Credit Risk Assessment"]
    F --> G["Rating Agencies and Default Probabilities"]
    A --> H["Capital Structure and Issuance History"]
    H --> I["Bond Issuances and Debt Maturities"]
    A --> J["Corporate Governance and Management"]
    J --> K["Board Quality and Managerial Competence"]
    A --> L["Valuation and Investment Analysis"]
    L --> M["DCF, Relative Valuation, Multiples"]
```
```
```

Explanation: Step-1: Corporate Issuer Overview – Begin with a high-level understanding of the firm's business model, market positioning, and strategic objectives. This foundational context is essential for both equity and fixed income analysis (CFA Institute, 2025c).

Step-2: Industry Classification and Sector Trends – Classify the firm by sector or sub-sector (e.g., financials, consumer discretionary) and evaluate the competitive landscape. Analyze market trends, industry growth prospects, and systemic risks. This industry context shapes performance expectations and relative valuation (Penman, 2012).

Step-3: Financial Statement Analysis and Key Metrics – Analyze income statement, balance sheet, and cash flow data. Focus on metrics like revenue growth, operating margin, return on equity, and leverage. This step reveals the firm's financial health and operational efficiency (Penman, 2012; CFA Institute, 2025c).

Step-4: Credit Risk Assessment and Rating Measures – Evaluate creditworthiness through agency ratings (e.g., S&P, Moody's), credit spreads, and financial ratios. Analyze the probability of default and credit cycle indicators. This step is vital for bondholders and fixed income portfolio managers (Fabozzi, 2012).

Step-5: Capital Structure, Issuance History, and Debt Profile – Examine the firm's financing structure, including the mix of debt vs. equity, historical issuance patterns, and maturity schedules. This informs views on solvency and refinancing risks (Fabozzi, 2012).

Step-6: Corporate Governance and Leadership Quality – Assess governance practices such as board independence, shareholder rights, and disclosure quality. Evaluate the management team's execution track record and alignment with shareholder interests (CFA Institute, 2025c).

Step-7: Valuation and Investment Analysis – Use valuation models like DCF, P/E, or EV/EBITDA to derive intrinsic value. Develop an investment thesis based on fundamental insights. These valuation techniques are central to both equity and credit investing (Penman, 2012).

Source: CFA Program Curriculum's Equity Investments and Fixed Income readings–which cover firm analysis, industry evaluation, and credit assessment frameworks (CFA Institute, 2025c)–along with Penman's Financial Statement Analysis and Security Valuation for accounting-to-valuation linkages (Penman, 2012), and Fabozzi's Bond Markets, Analysis, and Strategies for credit risk and corporate debt issuance insights (Fabozzi, 2012).

B Prompt Template

B.1 Structured Chain-of-Thought (ST-CoT)

ST-CoT for CFA Exam

```
You are a CFA (chartered financial analyst) taking a test to evaluate your knowledge of finance. You think step-by-step approach
to answer queries.

Follow these steps:
1. Think through the problem step by step within the <thinking> tags.
2. Provide your final, concise answer within the <output> tags.

The <thinking> sections are for your internal reasoning process only.
Do not include any part of the final answer in these sections.
The actual response to the query must be entirely contained within the <output> tags.

### Response Format:
<thinking>
[Reasoning through options A, B, and C to understand and solve the problem.]
</thinking>
<output>
"answer": [Final your answer (A , B , or C )]
</output>
```

B.2 FinCoT

FinCoT for CFA Exam

```
You are taking a test for the Chartered Financial Analyst (CFA) program designed to evaluate your knowledge of different topics in
finance. You think step-by-step approach with reflection to answer queries.

Follow these steps:
1. Think through the problem step by step reflect and verify while reasoning within the <thinking> tags.
2. Please and put the answer your final, concise answer within the <output> tags.

The <thinking> sections are for your internal reasoning process only.
Do not include any part of the final answer in these sections.
The actual response to the query must be entirely contained within the <output> tags.

Hint:{THOUGHT.get("embedding_expert_blueprints_[i]")}

### Response Format:
<thinking>
[Think step by step and respond with your thinking and the correct answer (A, B, or C ), considering the specific sector.]
</thinking>

<output>
"sector": [The sector being addressed],
"question": [The financial question],
"answer": [Reflect and verify the final answer (A, B, or C)]
</output>
```

B.3 Classify Domain

Classify Domain CFA Exam

```
SYSTEM_INSTRUCTION = """You are a CFA expert. Categorize the given CFA question into exactly one
of these categories:

    Ethical and Professional Standards
    - Code of Ethics, Standards of Professional Conduct, professional integrity
    - Professional responsibilities, ethical decision-making, client interests
    Category code: Ethics

    Quantitative Methods
    - Statistical analysis, probability theory, hypothesis testing
    - Time value of money, financial mathematics, regression analysis
    Category code: Quant.Meth.

    Economic Analysis and Market Forces
    - Microeconomics: supply, demand, market structures
    - Macroeconomics: GDP, inflation, monetary policy, economic cycles
    Category code: Economics

    Financial Reporting and Analysis
    - Financial statements, accounting standards, ratio analysis
    - Balance sheets, income statements, cash flow analysis
    Category code: Fin.Reporting

    Corporate Finance and Issuers
    - Capital structure, dividend policy, corporate governance
    - Mergers & acquisitions, capital budgeting, risk management
    Category code: Corp.Issuers

    Equity Investments
    - Stock valuation, equity markets, company analysis
    - Market efficiency, equity portfolio management
    Category code: EquityInvest.

    Fixed Income Investments
    - Bond markets, yield curves, duration analysis
    - Credit analysis, fixed income portfolio management
    Category code: FixedIncome

    Derivative Instruments
    - Options, futures, forwards, swaps
    - Hedging strategies, derivative pricing, risk management
    Category code: Derivatives

    Alternative Investments
    - Real estate, private equity, hedge funds
    - Commodities, structured products, crypto assets
    Category code: Alter.Invest.

    Portfolio Management
    - Asset allocation, portfolio construction, rebalancing
    - Risk management, performance measurement, client objectives
    Category code: Port.Manage.

Respond with only the single most appropriate category code, nothing else. For example: Ethics,
Port.Manage., etc.
"""
```

C Domain Distribution

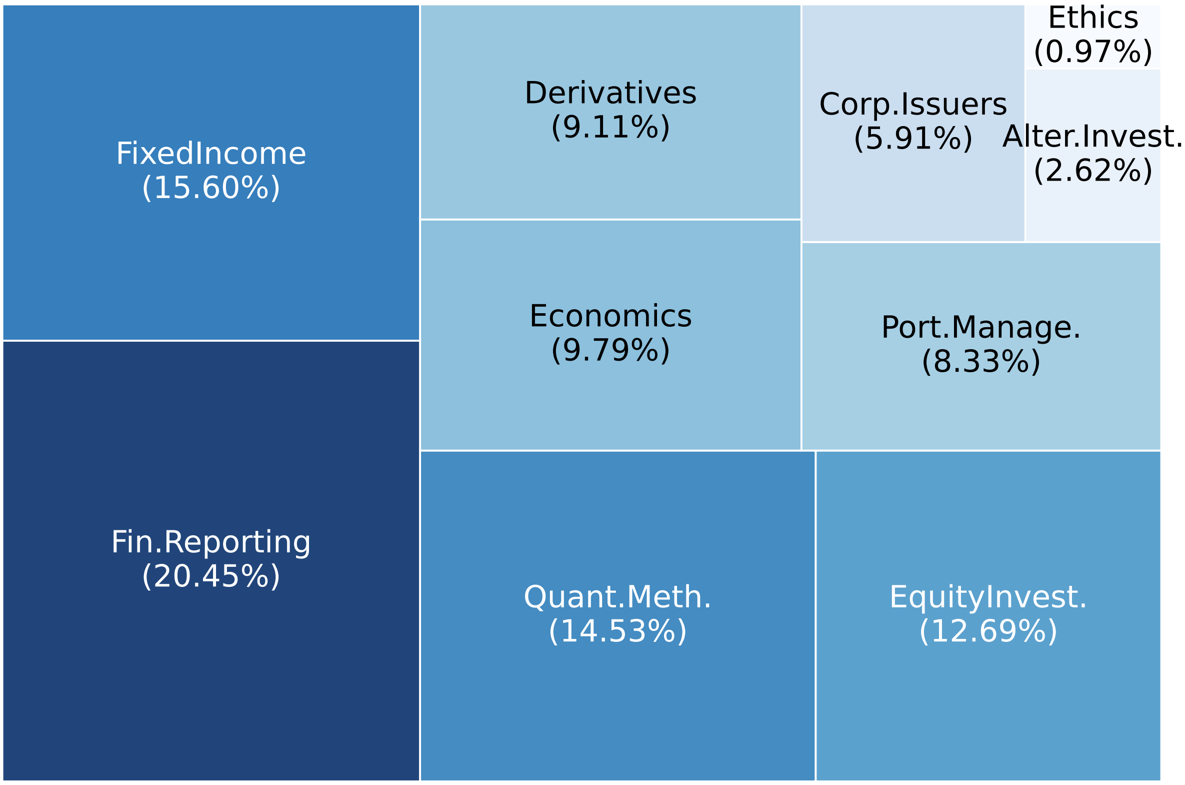


Figure 6: GPT-4o classified the benchmark domain distribution of CFA. A random sample of 100 items was manually audited by a financial expert to validate domain labels.

D Average Input and Output Tokens

D.1 Average Input Tokens

Prompt	**Average Input Tokens (k)**								
	Qwen2.5-7B	Qwen2.5-7B Instruct	Qwen3-8B Base	Qwen3-8B	Gemma-3-12B IT	Qwen3-8B (Thinker)	Fin-R1	DianJin-R1 7B	Fin-o1-8B
SP	**0.07***	**0.07***	**0.07***	**0.07***	**0.07***	**0.07***	**0.07***	**0.07***	**0.07***
UST-CoT	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09
ST-CoT	0.18	0.18	0.18	0.18	0.19	0.18	0.18	0.18	0.18
FinCoT (All Blueprints)	1.75	1.75	1.75	1.75	1.78	1.75	1.75	1.75	1.75
Domain-wise performance of FinCoT									
FinCoT (Economics)	0.55	0.55	0.55	0.55	0.56	0.55	0.55	0.55	0.55
FinCoT (FixedIncome)	0.34	0.34	0.34	0.34	0.36	0.34	0.34	0.34	0.34
FinCoT (Quant.Meth.)	0.33	0.33	0.33	0.33	0.34	0.33	0.33	0.33	0.33
FinCoT (EquityInvest.)	0.34	0.34	0.34	0.34	0.36	0.34	0.34	0.34	0.34
FinCoT (Port.Manage.)	0.33	0.33	0.33	0.33	0.35	0.33	0.33	0.33	0.33
FinCoT (Derivatives)	0.39	0.39	0.39	0.39	0.44	0.39	0.39	0.39	0.39
FinCoT (Fin. Reporting)	0.37	0.37	0.37	0.37	0.38	0.37	0.37	0.37	0.37
FinCoT (Alter.Invest.)	0.32	0.32	0.32	0.32	0.34	0.32	0.32	0.32	0.32
FinCoT (Corp.Issuers)	0.39	0.39	0.39	0.39	0.41	0.39	0.39	0.39	0.39

Table 2: Comparison of prompting techniques: average input token length (k) across models. **Bold** values highlight the prompt variant that uses the least tokens for each model.

D.2 Average Output Tokens

Prompt	**Average Output Tokens (k)**								
	Qwen2.5-7B	Qwen2.5-7B Instruct	Qwen3-8B Base	Qwen3-8B	Gemma-3-12B IT	Qwen3-8B (Thinker)	Fin-R1	DianJin-R1 7B	Fin-o1-8B
SP	0.45	**0.05***	0.89	0.32	**0.27***	1.52	0.88	2.18	**0.46***
UST-CoT	0.48	0.28	**0.31***	0.46	0.39	1.50	**0.58***	2.28	0.53
ST-CoT	**0.39***	0.22	3.42	**0.25***	0.31	1.35	2.22	7.20	0.58
FinCoT (All Blueprints)	2.22	0.29	0.38	0.36	0.32	**1.23***	1.92	**1.60***	0.79
Domain-wise performance of FinCoT									
FinCoT (Economics)	0.36	0.38	0.99	0.39	0.38	1.25	**2.01**	12.65	0.76
FinCoT (FixedIncome)	0.42	0.27	4.55	0.30	**0.32**	1.24	2.31	**8.31**	0.81
FinCoT (Quant.Meth.)	0.48	0.27	3.07	**0.31**	0.35	1.22	2.17	8.60	0.80
FinCoT (EquityInvest.)	**0.32**	0.31	7.18	0.37	0.34	1.19	2.16	10.07	0.78
FinCoT (Port.Manage.)	0.38	**0.26**	0.56	0.30	0.33	1.20	2.14	9.46	0.79
FinCoT (Derivatives)	0.36	0.30	**0.42**	0.39	0.34	1.24	2.05	5.54	0.81
FinCoT (Fin. Reporting)	0.46	0.28	0.93	0.33	0.34	1.19	2.13	8.76	**0.73**
FinCoT (Alter.Invest.)	0.47	**0.26**	0.50	0.38	0.34	1.23	2.16	11.53	0.77
FinCoT (Corp.Issuers)	0.52	**0.26**	1.18	0.32	0.33	**1.16**	2.08	11.37	0.82

Table 3: Comparison of prompting techniques: average output token length (k) across models. **Bold** values highlight the prompt variant that uses the least tokens for each model. (*) Indicates that the change in average output token count among the model-level prompt variants is statistically significant ($p < 0.05$) based on paired bootstrap testing; domain-specific rows are not tested for significance.

D.2.1 Efficiency of Input and Output Cost in Simulation

This appendix reports a cost–efficiency analysis under realistic output–input price ratios. Let I and O denote the average input and output tokens for a (prompt, model) pair. For a price ratio r,

$$\mathrm{Cost}(r) = I + r\,O, \qquad \mathrm{Efficiency}(r) = \frac{\mathrm{Cost}_{\mathrm{baseline}}(r)}{\mathrm{Cost}_{\mathrm{prompt}}(r)} = \frac{I_{\mathrm{base}} + r\,O_{\mathrm{base}}}{I_{\mathrm{prompt}} + r\,O_{\mathrm{prompt}}}.$$

Units and normalization. We measure cost in "input-token dollars": the effective input price is 1, and $r = \mathrm{price}_{\mathrm{out}}/\mathrm{price}_{\mathrm{in,eff}}$ carries the output premium. This rescaling makes Efficiency dimensionless and invariant to any common price factor.
Break-even and sensitivity. For a candidate prompt p vs. baseline b, the break-even ratio solving $\mathrm{Cost}_p(r) = \mathrm{Cost}_b(r)$ is

$$r^\star = \frac{I_p - I_b}{O_b - O_p} \qquad (O_p \neq O_b).$$

If $O_p < O_b$, then $\dfrac{d}{dr}\mathrm{Efficiency}(r) = \dfrac{O_b I_p - O_p I_b}{(I_p + rO_p)^2} > 0$: the candidate improves as r increases; if $O_p > O_b$, the trend reverses. When $O_p = O_b$, ranking depends only on inputs (I_b vs. I_p) and is independent of r.
Caching and effective input price. With prompt caching,

$$p_{\mathrm{in,eff}}(K) = p_{\mathrm{read}} + \frac{p_{\mathrm{write}}}{K}, \qquad r(K) = \frac{\mathrm{price}_{\mathrm{out}}}{p_{\mathrm{in,eff}}(K)},$$

where K is the number of reuses. Hence $r(K)$ increases monotonically in K and approaches $\mathrm{price}_{\mathrm{out}}/p_{\mathrm{read}}$ as $K \to \infty$.
Price instantiation (grid for plots). From public price points we use $r \in \{5, 6.9, 8, 14.29, 22.22, 40, 44.44, 50, 80\}$: *GPT-5*[5] input \$1.25/MTok (cached \$0.125/MTok), output \$10/MTok ⇒ r=10/1.25=8, r_{cached}=10/0.125=80; *Claude Opus 4.1*[6] input \$15/MTok, output \$75/MTok; caching write \$18.75/MTok, read \$1.50/MTok ⇒ r=5 (no cache), 6.9 (K=2), 14.29 (K=5), 22.22 (K=10), 40 (K=50), 44.44 (K=100), and the read-only limit 50 ($K \to \infty$). We display r on a log scale because the grid spans an order of magnitude (5–80).
Worked example (illustrative). Baseline $(I_b, O_b) = (100, 300)$; candidate $(I_p, O_p) = (250, 150)$. At $r = 8$ (GPT-5 no cache): $\mathrm{Cost}_b = 100 + 8 \cdot 300 = 2500$, $\mathrm{Cost}_p = 250 + 8 \cdot 150 = 1450$, so Efficiency ≈ 1.72. Break-even $r^\star = (250 - 100)/(300 - 150) = 1$; the candidate dominates for $r > 1$.
Note (scope). All models evaluated in this appendix are *open-source*. The curves simulate dollar costs by pairing the measured (I, O) token counts from these models with *provider API prices* (GPT-5 and Claude Opus 4.1)—prices are used for *simulation only*; no paid API runs were executed for these experiments.

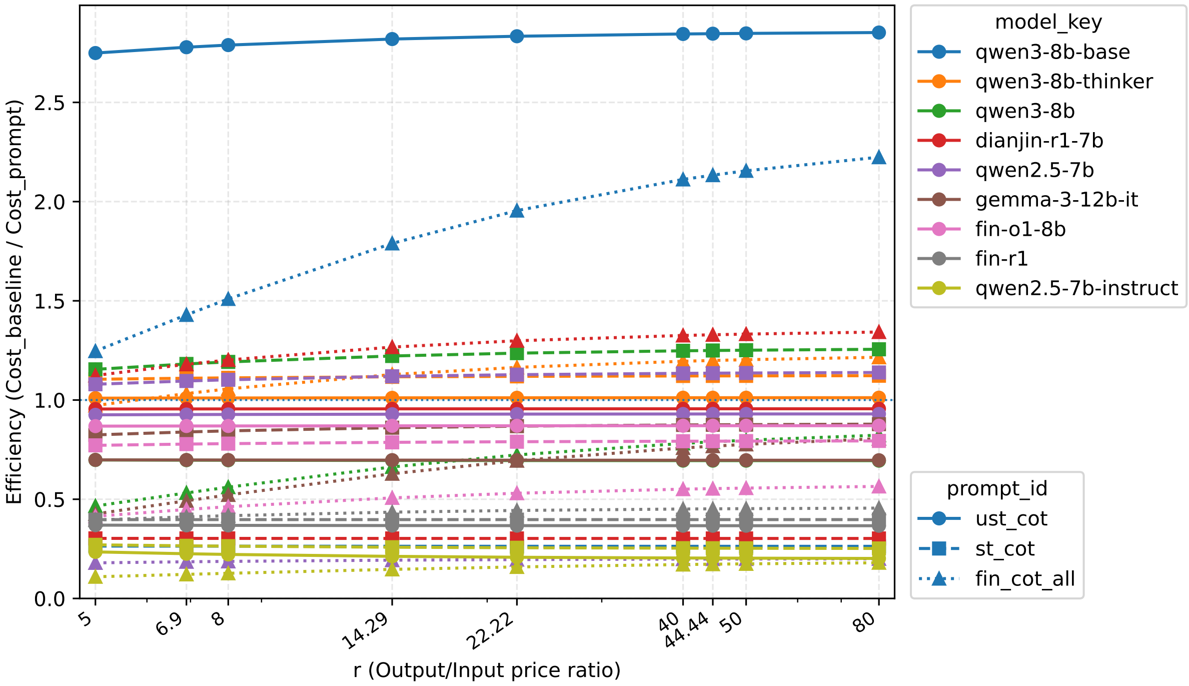


Figure 7: **Cost–efficiency vs. price ratio** r for UST-CoT, ST-CoT, and FinCoT-All across models. Efficiency is $\mathrm{Cost}_{\mathrm{baseline}}/\mathrm{Cost}_{\mathrm{prompt}}$ with $\mathrm{Cost} = I + rO$; values > 1 indicate lower cost than the baseline. r values use provider prices for GPT-5 and Claude Opus 4.1 as described above. *Notation:* MTok = million tokens; USD per MTok.

[5] OpenAI API pricing: https://openai.com/api/pricing/.

[6] Anthropic pricing: https://www.anthropic.com/pricing#api.

E Significance Testing

E.1 Accuracy

Model	Baseline	Comparison	Δ (pp)	95% CI (pp)	p-value	Significant
Qwen2.5-7B	SP	UST-CoT	-13.76	[-15.89, -11.63]	0.0000	✓
	SP	ST-CoT	-16.27	[-18.60, -14.05]	0.0000	✓
	SP	FinCoT (All)	-7.94	[-9.59, -6.30]	0.0000	✓
	UST-CoT	ST-CoT	-2.52	[-3.49, -1.65]	0.0000	✓
	UST-CoT	FinCoT (All)	5.81	[4.46, 7.27]	0.0000	✓
	ST-CoT	FinCoT (All)	8.33	[6.69, 9.98]	0.0000	✓
Qwen2.5-7B-Instruct	SP	UST-CoT	6.00	[4.65, 7.56]	0.0000	✓
	SP	ST-CoT	4.84	[3.59, 6.20]	0.0000	✓
	SP	FinCoT (All)	4.55	[3.29, 5.91]	0.0000	✓
	UST-CoT	ST-CoT	-1.16	[-1.84, -0.58]	0.0000	✓
	UST-CoT	FinCoT (All)	-1.45	[-2.23, -0.78]	0.0000	✓
	ST-CoT	FinCoT (All)	-0.29	[-0.68, 0.00]	0.1024	–
Qwen3-8B-Base	SP	UST-CoT	-9.40	[-11.24, -7.66]	0.0000	✓
	SP	ST-CoT	-15.31	[-17.54,-13.18]	0.0000	✓
	SP	FinCoT (All)	-17.35	[-19.77,-15.02]	0.0000	✓
	UST-CoT	ST-CoT	-5.91	[-7.36, -4.55]	0.0000	✓
	UST-CoT	FinCoT (All)	-7.96	[-9.59, -6.30]	0.0000	✓
	ST-CoT	FinCoT (All)	-2.04	[-3.00, -1.26]	0.0000	✓
Qwen3-8B	SP	UST-CoT	7.95	[6.30, 9.69]	0.0000	✓
	SP	ST-CoT	6.60	[5.14, 8.14]	0.0000	✓
	SP	FinCoT (All)	6.70	[5.23, 8.24]	0.0000	✓
	UST-CoT	ST-CoT	-1.35	[-2.13, -0.68]	0.0000	✓
	UST-CoT	FinCoT (All)	-1.26	[-1.94, -0.68]	0.0000	✓
	ST-CoT	FinCoT (All)	0.10	[0.00, 0.29]	0.7370	–
Qwen3-8B (Thinker)	SP	UST-CoT	-0.87	[-1.45, -0.39]	0.0002	✓
	SP	ST-CoT	0.00	[0.00, 0.00]	2.0000	–
	SP	FinCoT (All)	0.96	[0.39, 1.55]	0.0002	✓
	UST-CoT	ST-CoT	0.87	[0.39, 1.45]	0.0002	✓
	UST-CoT	FinCoT (All)	1.83	[1.07, 2.71]	0.0000	✓
	ST-CoT	FinCoT (All)	0.96	[0.39, 1.55]	0.0002	✓
Gemma-3-12B-IT	SP	UST-CoT	-24.999	[-27.71, -22.38]	0.0000	✓
	SP	ST-CoT	-23.934	[-26.55, -21.32]	0.0000	✓
	SP	FinCoT (All)	-22.765	[-25.39, -20.16]	0.0000	✓
	UST-CoT	ST-CoT	1.065	[0.48, 1.74]	0.0002	✓
	UST-CoT	FinCoT (All)	2.234	[1.36, 3.20]	0.0000	✓
	ST-CoT	FinCoT (All)	1.169	[0.58, 1.84]	0.0000	✓
Fin-R1	SP	UST-CoT	-9.49	[-11.34, -7.75]	0.0000	✓
	SP	ST-CoT	-8.62	[-10.37, -6.88]	0.0000	✓
	SP	FinCoT (All)	-10.07	[-11.92,-8.24]	0.0000	✓
	UST-CoT	ST-CoT	0.87	[0.39, 1.45]	0.0002	✓
	UST-CoT	FinCoT (All)	-0.58	[-1.07, -0.19]	0.0042	✓
	ST-CoT	FinCoT (All)	-1.45	[-2.23, -0.78]	0.0000	✓
Dianjin-R1-7B	SP	UST-CoT	10.662	[8.82, 12.60]	0.0000	✓
	SP	ST-CoT	9.594	[7.75, 11.43]	0.0000	✓
	SP	FinCoT (All)	-1.361	[-2.13, -0.68]	0.0000	✓
	UST-CoT	ST-CoT	-1.068	[-1.74, -0.48]	0.0000	✓
	UST-CoT	FinCoT (All)	-12.023	[-14.05, -10.08]	0.0000	✓
	ST-CoT	FinCoT (All)	-10.956	[-12.89, -9.01]	0.0000	✓
Fino1-8B	SP	UST-CoT	0.294	[0.00, 0.68]	0.0984	–
	SP	ST-CoT	1.264	[0.58, 1.94]	0.0000	✓
	SP	FinCoT (All)	2.429	[1.55, 3.39]	0.0000	✓
	UST-CoT	ST-CoT	0.970	[0.39, 1.55]	0.0000	✓
	UST-CoT	FinCoT (All)	2.134	[1.26, 3.00]	0.0000	✓
	ST-CoT	FinCoT (All)	1.165	[0.58, 1.84]	0.0000	✓

Table 4: Paired bootstrap significance testing ($B = 10{,}000$ samples) for accuracy differences across prompt strategies. Δ indicates average accuracy difference (in percentage points), with 95% confidence intervals (CI) and p-values. A result is considered statistically significant if $p < 0.05$.

E.2 Average Output Tokens

Model	**Baseline**	**Comparison**	Δ (k)	95% CI (k)	p-value	Significant
Qwen2.5-7B	SP	UST-CoT	-0.09867	[-0.09867, -0.09867]	0.0000	✓
	SP	ST-CoT	0.02273	[0.02273, 0.02273]	0.0000	✓
	SP	FinCoT (All)	-0.11555	[-0.11555, -0.11555]	0.0000	✓
	UST-CoT	ST-CoT	0.12140	[0.12140, 0.12140]	0.0000	✓
	UST-CoT	FinCoT (All)	-0.01688	[-0.01688, -0.01688]	0.0000	✓
	ST-CoT	FinCoT (All)	-0.13828	[-0.13828, -0.13828]	0.0000	✓
Qwen2.5-7B-Instruct	SP	UST-CoT	-0.227	[-0.227, -0.227]	0.0000	✓
	SP	ST-CoT	-0.169	[-0.169, -0.169]	0.0000	✓
	SP	FinCoT (All)	-0.241	[-0.241, -0.241]	0.0000	✓
	UST-CoT	ST-CoT	0.058	[0.058, 0.058]	0.0000	✓
	UST-CoT	FinCoT (All)	-0.014	[-0.014, -0.014]	0.0000	✓
	ST-CoT	FinCoT (All)	-0.072	[-0.072, -0.072]	0.0000	✓
Qwen3-8B-Base	SP	UST-CoT	-0.584	[-0.584, -0.584]	0.0000	✓
	SP	ST-CoT	2.522	[2.522, 2.522]	0.0000	✓
	SP	FinCoT (All)	-0.516	[-0.516, -0.516]	0.0000	✓
	UST-CoT	ST-CoT	3.106	[3.106, 3.106]	0.0000	✓
	UST-CoT	FinCoT (All)	0.068	[0.068, 0.068]	0.0000	✓
	ST-CoT	FinCoT (All)	-3.038	[-3.038, -3.038]	0.0000	✓
Qwen3-8B	SP	UST-CoT	0.141	[0.141, 0.141]	0.0000	✓
	SP	ST-CoT	-0.067	[-0.067, -0.067]	0.0000	✓
	SP	FinCoT (All)	0.048	[0.048, 0.048]	0.0000	✓
	UST-CoT	ST-CoT	-0.208	[-0.208, -0.208]	0.0000	✓
	UST-CoT	FinCoT (All)	-0.093	[-0.093, -0.093]	0.0000	✓
	ST-CoT	FinCoT (All)	0.115	[0.115, 0.115]	0.0000	✓
Qwen3-8B (Thinker)	SP	UST-CoT	-0.018	[-0.018, -0.018]	0.0000	✓
	SP	ST-CoT	-1.271	[-1.271, -1.271]	0.0000	✓
	SP	FinCoT (All)	-0.168	[-0.168, -0.168]	0.0000	✓
	UST-CoT	ST-CoT	-1.253	[-1.253, -1.253]	0.0000	✓
	UST-CoT	FinCoT (All)	-0.150	[-0.150, -0.150]	0.0000	✓
	ST-CoT	FinCoT (All)	1.103	[1.103, 1.103]	0.0000	✓
Gemma-3-12B-IT	SP	UST-CoT	0.11985	[0.11985, 0.11985]	0.0000	✓
	SP	ST-CoT	0.03661	[0.03661, 0.03661]	0.0000	✓
	SP	FinCoT (All)	0.04523	[0.04523, 0.04523]	0.0000	✓
	UST-CoT	ST-CoT	-0.08324	[-0.08324, -0.08324]	0.0000	✓
	UST-CoT	FinCoT (All)	-0.07462	[-0.07462, -0.07462]	0.0000	✓
	ST-CoT	FinCoT (All)	0.00862	[0.00862, 0.00862]	0.0000	✓
Fin-R1	SP	UST-CoT	1.526	[1.526, 1.526]	0.0000	✓
	SP	ST-CoT	1.338	[1.338, 1.338]	0.0000	✓
	SP	FinCoT (All)	1.035	[1.035, 1.035]	0.0000	✓
	UST-CoT	ST-CoT	-0.188	[-0.188, -0.188]	0.0000	✓
	UST-CoT	FinCoT (All)	-0.491	[-0.491, -0.491]	0.0000	✓
	ST-CoT	FinCoT (All)	-0.303	[-0.303, -0.303]	0.0000	✓
Dianjin-R1-7B	SP	UST-CoT	0.10023	[0.10023, 0.10023]	0.0000	✓
	SP	ST-CoT	5.02159	[5.02159, 5.02159]	0.0000	✓
	SP	FinCoT (All)	-0.57669	[-0.57669, -0.57669]	0.0000	✓
	UST-CoT	ST-CoT	4.92136	[4.92136, 4.92136]	0.0000	✓
	UST-CoT	FinCoT (All)	-0.67692	[-0.67692, -0.67692]	0.0000	✓
	ST-CoT	FinCoT (All)	-5.59828	[-5.59828, -5.59828]	0.0000	✓
Fino1-8B	SP	UST-CoT	0.06788	[0.06788, 0.06788]	0.0000	✓
	SP	ST-CoT	0.11765	[0.11765, 0.11765]	0.0000	✓
	SP	FinCoT (All)	0.33273	[0.33273, 0.33273]	0.0000	✓
	UST-CoT	ST-CoT	0.04977	[0.04977, 0.04977]	0.0000	✓
	UST-CoT	FinCoT (All)	0.26485	[0.26485, 0.26485]	0.0000	✓
	ST-CoT	FinCoT (All)	0.21508	[0.21508, 0.21508]	0.0000	✓

Table 5: Paired bootstrap significance testing ($B = 10{,}000$ samples) for average output token differences across prompt strategies. Δ indicates mean difference in output length (in thousands of tokens), with 95% confidence intervals and p-values. A result is significant if $p < 0.05$.

F Radar Behavior Accuracy

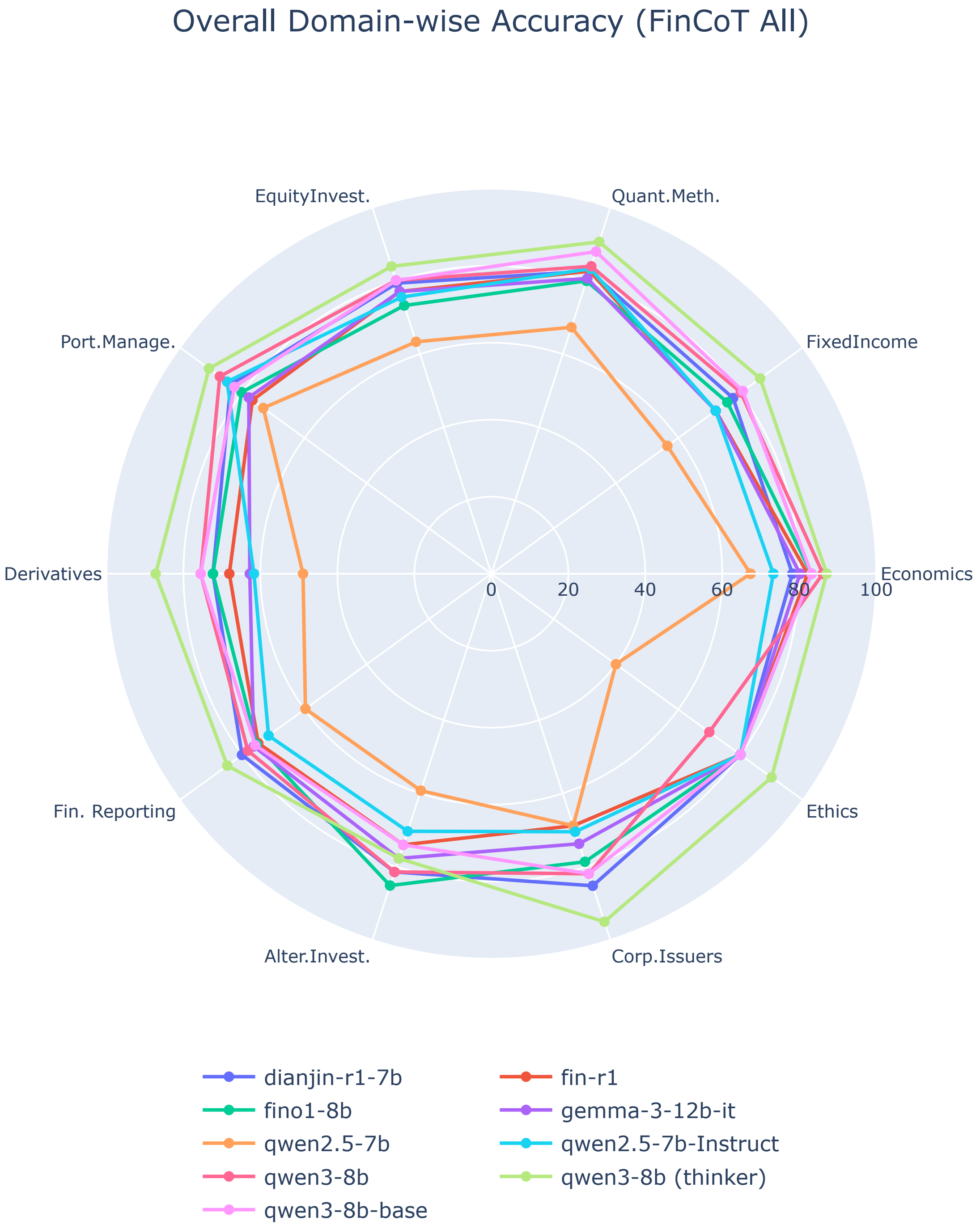
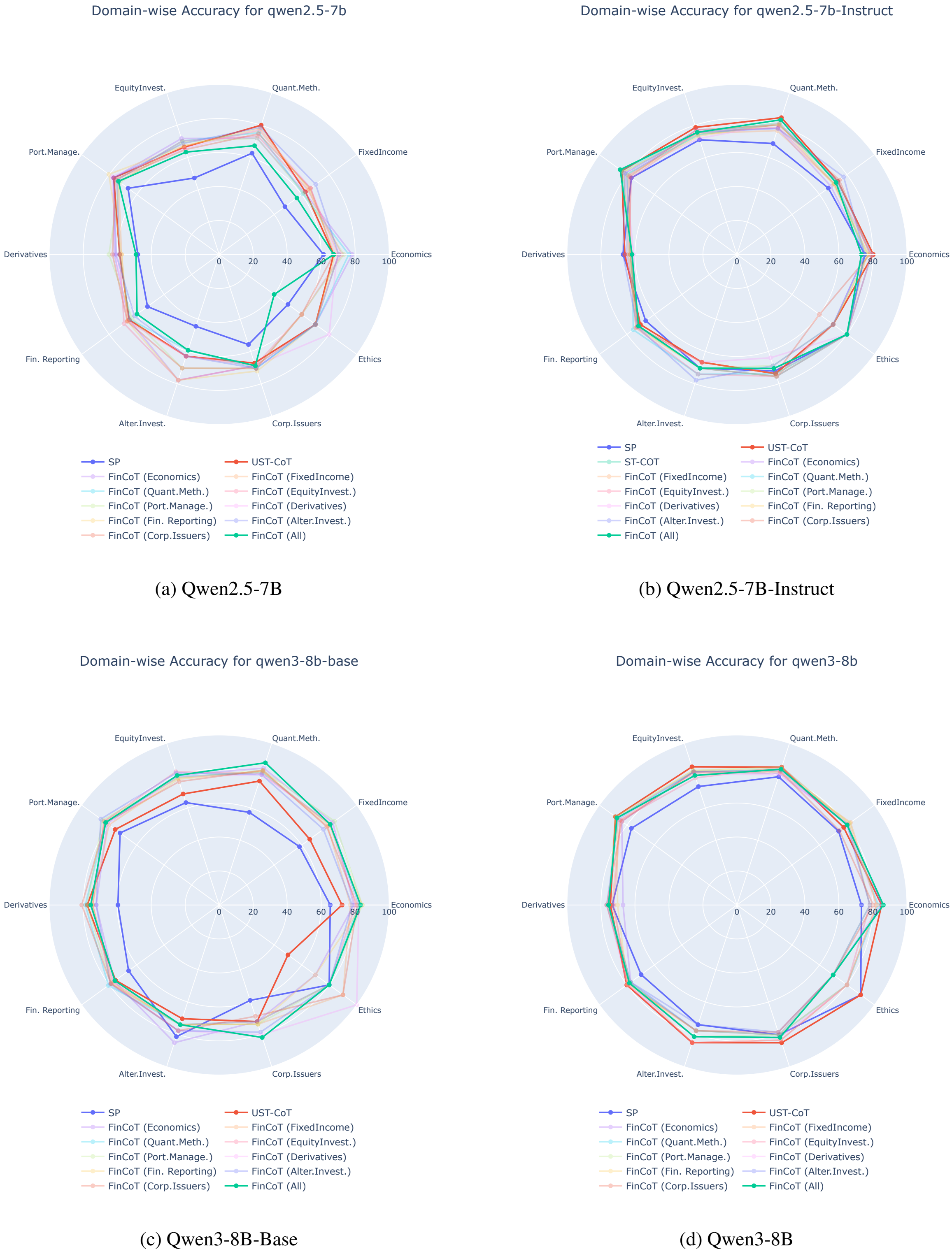


Figure 8: Overall FinCoT behaviour accuracy.



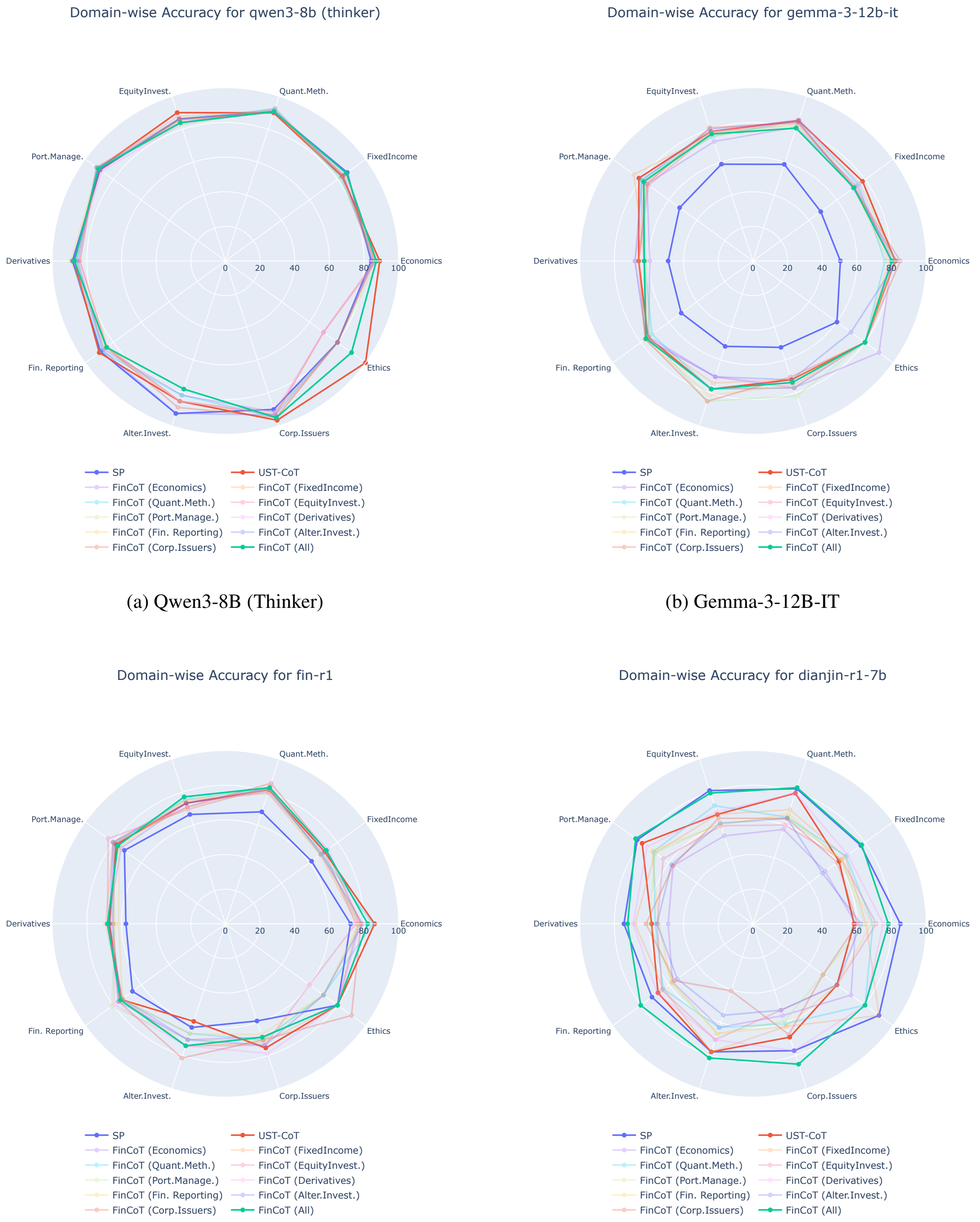
(a) Qwen2.5-7B

(b) Qwen2.5-7B-Instruct

(c) Qwen3-8B-Base

(d) Qwen3-8B

Figure 9: Radar charts for each model variant.



(a) Qwen3-8B (Thinker)

(b) Gemma-3-12B-IT

(c) Fin-R1

(d) Dianjin-R1-7B

Figure 10: Radar charts for each model variant (charts 5–8).

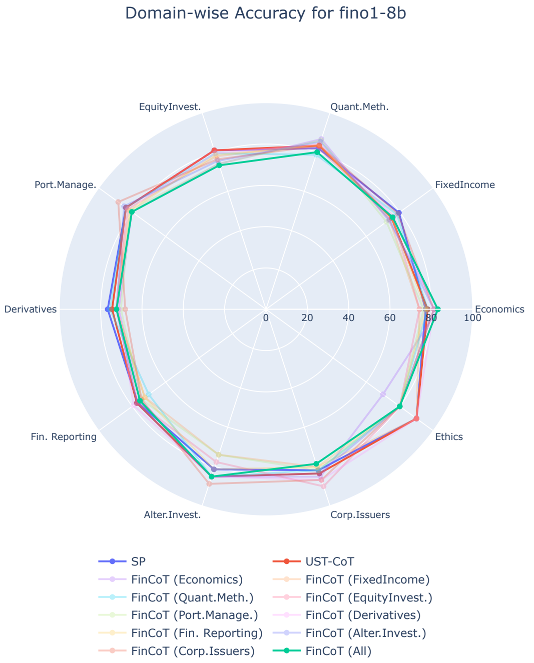


Figure 11: Radar charts for each model variant (charts 9).