Prompt Engineering for Finance Tasks: What Changes, What Doesn't

Gabriel Colón
COGS 150 — UC San Diego • June 3, 2025

Abstract

Large language models (LLMs) can ace structured finance tasks, but do prompt techniques change *how* they reason or only *how* they sound? I ran a controlled experiment using GPT-4o on three practical finance task types: calculating a stock trade's profit or loss, comparing risk between investments, and proposing a simple diversified portfolio. I varied five prompt strategies: Role Prompting, Chain-of-Thought (CoT), Few-Shot examples, Explicit Instruction, and Self-Consistency. Across 75 trials GPT-4o achieved 100% task accuracy. The techniques did not move correctness, but they changed the *shape* of the explanations and the surface style of the outputs. CoT and Self-Consistency produced the clearest step-by-step rationales, while Role Prompting mainly shifted tone. I also observed a formatting effect: about 19% of all responses included emojis, concentrated in CoT and Role prompts, especially for the portfolio task. These results fit prior work that CoT improves performance in reasoning-heavy settings yet can also produce plausible but unfaithful narratives, which matters in regulated domains like finance. I close with a plan for a follow-up study on harder, ambiguous finance problems, new rubric-based scoring, and checks for explanation faithfulness.

Introduction

Prompt engineering promises leverage: better instructions, better results. In practice, teams want to know whether a technique changes correctness, speed, or just presentation. Finance is a good sandbox because many tasks are crisp and easy to grade. The core question I asked was:

> **How do different prompt techniques affect LLM performance and reasoning on stock-market investment tasks?**

I tested five common techniques and compared outputs on accuracy and qualitative qualities like clarity, structure, and tone.

Related Work

Chain-of-Thought (CoT) prompting elicits stepwise reasoning and can boost performance on math and logic problems. Self-Consistency samples multiple reasoning paths and votes on the answer to improve reliability on complex reasoning tasks. Methods that interleave reasoning with tool use, like ReAct, help on tasks that need retrieval or external actions. Tree-of-Thought extends CoT by exploring solution branches and pruning them like a search tree. Zero-shot CoT shows that even “Let’s think step by step” can unlock latent reasoning skills.

At the same time, explanation *faithfulness* is an open problem. CoT rationales can be fluent yet unfaithful to the model’s actual decision process, which is risky if users infer trust from a tidy narrative. Recent surveys catalog broader prompting families, including least-to-most decomposition and iterative self-refinement, which I recommend for future, harder tasks.

Tasks and Prompt Techniques

I used three finance tasks that map to common assistant use-cases and have clear grading rules:

- **Stock Trade P&L**
  Compute the profit or loss given buy price, sell price, and shares. Graded exact numeric correctness.

- **Risk Assessment**
  Choose the safer option based on basic risk–return principles, for example, a government bond vs. a penny stock. Graded correct/incorrect vs. an expert key.

- **Portfolio Diversification**
  Suggest a simple balanced allocation, graded against the 60/40 reference.

Prompt techniques tested were: Role Prompting, Chain-of-Thought, Few-Shot, Explicit Instruction, and Self-Consistency.

Experimental Setup

The experiment used a single model variant (ChatGPT-4o) to control for model differences. The design was 3 tasks × 5 prompt strategies × 5 trials each, for a total of 75 outputs. Grading was based on numeric correctness for the P&L task, matching the expert key for the risk assessment, and alignment to a simple diversified baseline for the portfolio task. I also recorded whether the output used emojis as a light-weight marker for surface-level stylistic change.

A Google Sheet was used for data capture; trial-level examples are available in my working notes.

Results

Accuracy: Across all 75 trials, the model answered correctly on every task instance. This ceiling effect meant differences between prompt types did not appear in accuracy for these structured tasks.

Explanation structure and tone: CoT and Self-Consistency produced the most organized, stepwise explanations. Role Prompting mainly shifted tone — for example, sounding like “a stock broker” — without adding much reasoning depth. These findings align with prior research showing prompting often affects *communication style* more than correctness in deterministic settings.

Formatting and emoji usage: Out of 75 responses, 14 included emojis. Usage was concentrated in CoT and Role prompts, especially for the portfolio task:

- Role: 4 of 25
- CoT: 8 of 25
- Self-Consistency: 2 of 25
- Few-Shot and Explicit: 0 of 25 each

The total emoji rate was about 18.7%. Portfolio tasks, being more open-ended, invited more stylistic flourishes.

Discussion

The ceiling effect here suggests that on crisp, well-specified tasks, GPT-4o’s correctness is unaffected by prompt style. What changes is the *presentation* — clarity, structure, and tone — which can be critical in finance contexts where trust, comprehension, and regulatory compliance matter.

CoT and Self-Consistency make outputs appear rigorous, but explanation faithfulness remains a concern. A tidy reasoning chain is not necessarily proof of actual reasoning; in regulated domains, this distinction matters. The stylistic effects observed, such as emoji usage, show that

prompt templates can also serve as UX controls, influencing tone and format in predictable ways.

Limitations

The tasks were intentionally simple and fully specified, so the study underestimates how much prompt choice might matter on ambiguous, noisy, or multi-objective problems. There was no human-rated preference testing, meaning I measured correctness, not perceived helpfulness or trust. Only one model variant was tested.

Future Work

A follow-up study will target harder finance scenarios — such as market forecasting with incomplete data and multi-criteria decision making — and introduce richer prompting strategies like least-to-most decomposition, self-refinement loops, and tool-assisted reasoning. New metrics will include time-to-answer, token usage, and rubric-based explanation clarity, along with adversarial tests for faithfulness. Human evaluation will measure perceived clarity, trust, and satisfaction across different prompt styles.

Methods Detail

This was a within-task comparison of five prompt types. For each task, I ran five trials with each technique, randomizing numeric values and wording. The prompts included:

- **Role Prompting**: “You are a professional stock broker…” followed by task-specific instruction.
- **Chain-of-Thought**: “Think through the calculation step by step before giving the final answer.”
- **Few-Shot**: Two worked examples followed by a new query.
- **Explicit Instruction**: Output constraints like “Return only the final number, no symbols.”
- **Self-Consistency**: Multiple reasoning paths with majority voting.

Observation coding included noting whether an answer used emojis and where they appeared.

Ethical and Practical Considerations

No prompts or outputs constituted financial advice; all were for research purposes. In future versions, CoT outputs will be paired with faithfulness audits to ensure that presented steps are essential to the final answer. For accessibility, style controls like emojis and visual scaffolding could be used deliberately to support novice users without undermining professionalism.

Conclusion

In this controlled test on three structured finance tasks, GPT-4o reached a correctness ceiling unaffected by prompt type. The differences were in explanation clarity and stylistic elements, which have real implications for trust and UX in finance applications. CoT and Self-Consistency offered the clearest reasoning trails, while Role Prompting primarily altered tone. The next step is to move beyond correctness into evaluating the faithfulness, clarity, and user perception of AI explanations in more complex, ambiguous domains.