Prompt Engineering for Finance Tasks: What Changes, What Does Not

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

Executive Summary

- Model: GPT-4o. Design: 3 tasks × 5 prompt strategies × 5 trials = 75 outputs.
- Accuracy: 100 percent across all tasks and strategies.
- Main movement was in explanation structure and tone, not correctness.
- Chain-of-Thought and Self-Consistency produced the clearest stepwise rationales.
- Role Prompting shifted voice more than reasoning depth.
- Emoji usage appeared in 14 of 75 responses (about 18.7 percent), concentrated in CoT and Role conditions, especially on the portfolio task.

Abstract

Large language models can ace structured finance tasks, but do prompt techniques change how they reason or only how they sound? I ran a controlled experiment with 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, Few-Shot, Explicit Instruction, and Self-Consistency. Across 75 trials GPT-4o achieved 100 percent task accuracy. The techniques did not move correctness, but they changed the shape of explanations and the surface style of outputs. Chain-of-Thought and Self-Consistency produced the clearest step-by-step rationales, while Role Prompting mainly shifted tone. I also observed a formatting effect: about 19 percent of all responses included emojis, concentrated in Chain-of-Thought and Role prompts, especially for the portfolio task. These results fit prior work that Chain-of-Thought improves performance in reasoning-heavy settings yet can also produce fluent but unfaithful narratives, which matters in regulated domains like finance. I close with a plan for follow-up studies on harder, ambiguous finance problems, 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 question I asked was: How do different prompt techniques affect 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 prompting elicits stepwise reasoning and often boosts performance on math and logic problems. Self-Consistency samples multiple reasoning paths and votes on an 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 Chain-of-Thought by exploring solution branches like a search tree. Zero-shot Chain-of-Thought shows that even a simple instruction like “Let’s think step by step” can unlock latent reasoning skills. At the same time, explanation faithfulness remains an open problem. Chain-of-Thought rationales can be fluent yet unfaithful to the 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

Tasks

- Stock Trade P and L: Compute profit or loss given buy price, sell price, and shares. Graded on 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 or incorrect against an expert key.
- Portfolio Diversification: Suggest a simple balanced allocation, graded against a 60/40 reference.

Prompt Strategies

- Role Prompting
- Chain-of-Thought
- Few-Shot Examples
- Explicit Instruction
- 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 used exact numeric correctness for P and L, matching an expert key for

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

Task	Role	Chain-of-Thought	Few-Shot	Explicit Instruction	Self-Consistency
P and L	100%	100%	100%	100%	100%
Risk Assessment	100%	100%	100%	100%	100%
Portfolio Diversification	100%	100%	100%	100%	100%

Accuracy by task and prompt strategy (n=75). All conditions reached 100 percent.

Prompt Strategy	Count (of 25)	Rate
Role	4	16.0%
Chain-of-Thought	8	32.0%
Self-Consistency	2	8.0%
Few-Shot	0	0.0%
Explicit Instruction	0	0.0%

Emoji usage by prompt strategy. Total 14 of 75 responses (about 18.7 percent).

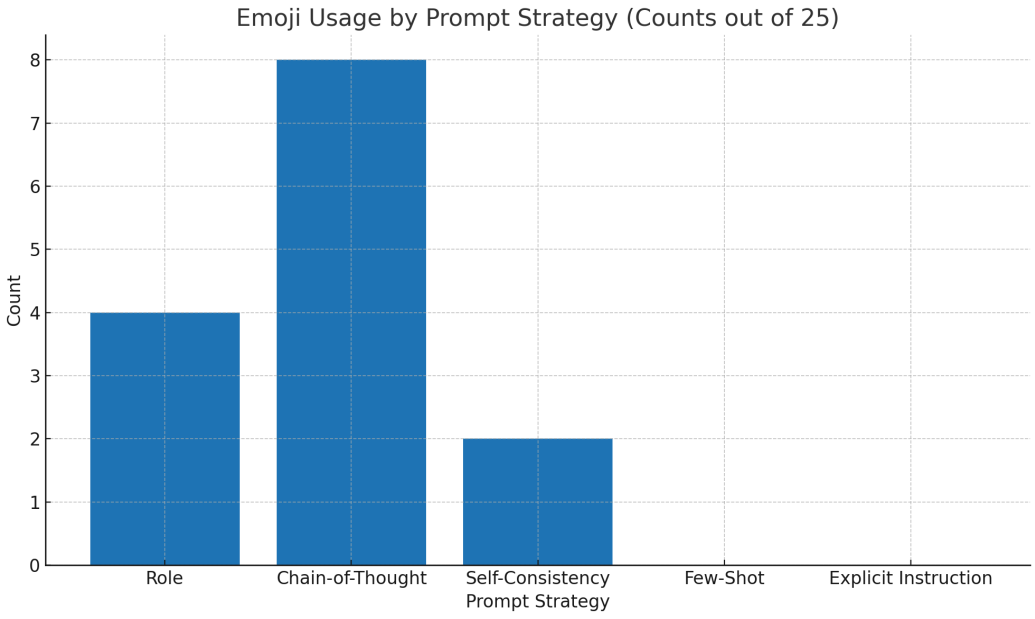

Figure 1. Emoji usage concentrated in Chain-of-Thought and Role conditions.

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: Chain-of-Thought and Self-Consistency produced the most organized, stepwise explanations. Role Prompting mainly shifted tone, such as sounding like a stock broker, without adding much reasoning depth. Formatting and emoji usage: Out of 75 responses, 14 included emojis, concentrated in Chain-of-Thought and Role prompts, especially for the portfolio task.

Discussion

The ceiling effect suggests that on crisp, well-specified tasks, correctness is largely insensitive to prompt style. What changes is the presentation, which can be critical in finance contexts where trust, comprehension, and compliance matter. Chain-of-Thought and Self-Consistency make outputs appear rigorous, but explanation faithfulness remains a concern. A tidy reasoning chain is not proof of the underlying process. The stylistic effects observed here, such as emoji usage, show that prompt templates can act as UX controls, shaping tone and formatting in predictable ways.

Limitations

- Tasks were simple and fully specified, which likely underestimates the value of prompting on ambiguous or noisy problems.
- No human-rated preference testing was included; I measured correctness, not perceived helpfulness or trust.
- Only one model variant was tested; cross-model replication is still needed.
- I did not measure latency or token usage, both relevant to practical adoption.

Future Work

- Harder scenarios: market forecasting with incomplete data, competing objectives, and noisy context.
- Richer prompting: least-to-most decomposition, self-refinement loops, and tool-assisted reasoning.
- New metrics: time to answer, token usage, rubric-based clarity, and adversarial tests for faithfulness.
- Human evaluation: measure perceived clarity, trust, and satisfaction across prompt styles.

Recommendations for Teams

- Treat prompting as a presentation and process tool on structured tasks; do not expect accuracy gains where tasks are deterministic and well-specified.

- Use Chain-of-Thought and Self-Consistency when you need clear rationales and audit trails; pair them with faithfulness checks before relying on them for compliance.
- Standardize output style through templates. If emojis or tone are undesirable, make that explicit in prompts.
- Instrument evaluation. Track incident reports, failure modes, and user feedback to refine prompts over time.

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, Chain-of-Thought outputs will be paired with faithfulness audits to test whether presented steps are essential to the final answer. For accessibility, style controls like emojis and visual scaffolding can 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. Chain-of-Thought and Self-Consistency offered the clearest reasoning trails, while Role Prompting primarily altered tone. The next step is to move beyond correctness into evaluating faithfulness, clarity, and user perception of AI explanations in more complex, ambiguous domains.