Phase 3: Systematic Evaluation

To evaluate prompt performance in the finance domain, I conducted a systematic, rubric-based assessment across two tasks: Information Synthesis (Task 1) and Risk Classification/Analysis (Task 2). The methodology consisted of the following components:

1. **Systematic Testing**

Each task was run using three distinct prompt frameworks, CLEAR Method, Few-Shot, and Chain-of-Thought (CoT), across three models:

- GPT-5.0
- Gemini 2.5 Flash
- Claude Sonnet 4.0

This yielded six model, prompt combinations per task, ensuring coverage of both advanced and baseline models.

2. **Rubric-Based Scoring**

Outputs were evaluated using a 4-point scale rubric across four metrics:

- Accuracy: factual correctness of financial metrics and classifications, verified against authoritative filings.
- Relevance: the extent to which the response directly addressed the specified task.
- Completeness: coverage of all required components (metrics, risks, MD&A highlights for Task 1; proper categorization and justification for Task 2).
- Domain Appropriateness: professionalism of tone, precision of financial terminology, and alignment with finance domain conventions.

The rubric ranged from 1 (Poor) to 4 (Excellent), as defined in the evaluation framework.

Metric	1 (Poor)	2 (Fair)	3 (Good)	4 (Excellent)
Accuracy	Major factual errors	Some errors, key facts missing	Mostly correct, minor errors	Completely correct, factually precise
Relevance	Off-topic, doesn’t address the task	Partially relevant	Addresses most of the tasks	Directly and fully addresses the task
Completeness	Major gaps, missing components	Covers some but not all components	Covers most required components	Covers all specified components fully
Domain Appropriateness	Uses highly informal, irrelevant, or inappropriate language; response shows	Some domain terms are present, but the language is inconsistent or includes obvious	Mostly uses correct financial terminology and tone; minor slips or slight mismatch with	Consistently professional, precise, and context-aware; language fully aligned with

	no awareness of finance domain norms.	errors/misuse; the response is partly inappropriate for the finance context.	professional standards.	finance domain conventions.

3. Blind Evaluation

To reduce evaluator bias, outputs from different prompts and models were randomized and reviewed without model labels. Scoring was performed independently before results were re-associated with model-prompt pairs.

4. Failure Case Documentation

For each task, at least one significant failure case was identified and documented per model. Failures included omission of forward-looking statements (Task 1), inconsistent risk labeling across similar contexts (Task 2), and verbosity beyond the 200-word target. Each failure was analyzed to identify whether the issue arose from prompt design limitations (e.g., ambiguity in instructions) or model capability constraints (e.g., weaker domain adaptation).

5. Comparative Analysis

After scoring, average metric performance was compared across frameworks and model categories. This enabled the identification of:

- Which prompt frameworks best supported accuracy vs. interpretability?
- Whether advanced models consistently outperformed standard models.
- Trade-offs between conciseness and reasoning depth across prompt types.