Phase 4: Synthesis and Recommendation

I. Executive Summary

Large language models (LLMs) have played a significant role across many industries. While LLMs offer considerable potential, caution is warranted when adopting them in high-stakes domains, particularly finance. The finance sector is inherently complex and requires careful handling of data and analysis. Using LLMs to summarize regulatory filings and classify risk factors presents both challenges and opportunities. This study evaluates three prompt frameworks, CLEAR, Few-Shot, and Chain-of-Thought (CoT), across GPT-5.0, Claude Sonnet 4.0, and Gemini 2.5 Flash. The performance is assessed using a four-metric rubric encompassing accuracy, relevance, completeness, and domain appropriateness, with blind scoring employed to minimize evaluation bias.

Across two phases of testing, results show that advanced models like GPT-5.0 and Claude Sonnet 4.0 consistently outperform Gemini 2.5 Flash, although performance remains highly prompt-sensitive. Few-shot prompts yielded the most stable and complete results, while CoT improved interpretability but introduced verbosity. Failure Case Documentation in both tasks is the one scored lowest overall, with omissions of forward-looking statements, numeric distortions, and inconsistent risk classifications recurring across phases. Overall, recommendations focus on prompt refinement, evidence anchoring, automated validation, and human oversight. These improvements will increase the reliability and compliance-readiness of LLM outputs in finance.

II. Methodology

Systematic Testing

Each task was executed using all three prompt frameworks across three models, producing six combinations per task. A randomized, blind evaluation approach mitigated scorer bias. Evaluations focused on both qualitative interpretation and quantitative accuracy.

Evaluation Rubric

The outputs were assessed on four dimensions:

- **Accuracy (1-4):** Correctness of extracted metrics, classifications, and alignment with source filings.
- **Relevance (1-4):** Degree to which outputs adhered to task-specific instructions.
- **Completeness (1-4):** Coverage of all required data fields or risk categories.
- **Domain Appropriateness (1-4):** Professional tone, terminology usage, and formatting aligned with industry standards.

Blind Scoring

Each output is scored without knowledge of the model or prompt type. Scores were aggregated by model and prompt framework to allow comparative analysis, reducing evaluation bias and increasing the reliability of findings.

Limitations

- **Context Truncation:** Long filings occasionally exceeded model input limits, affecting completeness.
- **Reviewer Variability:** Despite blind scoring, inter-annotator differences in interpretation could introduce minor inconsistencies.
- **Domain Coverage:** Only U.S. SEC filings were evaluated; other regulatory environments were not included.

III. Results

Phase 1 Results

Phase 1 established baseline performance on information synthesis and risk classification tasks.

- **GPT-5.0:** Delivered uniformly excellent outputs (all 4s across prompts), with summaries described as concise, accurate, and investor-ready. The CoT framework produced structured mappings that resembled professional risk registers.
- **Gemini 2.5 Flash:** Struggled under CLEAR prompts (mostly 2s), with omissions of critical numbers and risk details. Few-Shot and CoT slightly improved performance (scores of 3s), but outputs remained incomplete.
- **Claude Sonnet 4.0:** Strong performer overall, with only minor numeric discrepancies under CLEAR prompts. Few-Shot and CoT matched GPT-5.0 in accuracy and clarity, producing digestible summaries appropriate for executives.

Key Insight: Phase 1 confirmed that advanced models handle summarization tasks well, while mid-tier models like Gemini underperform without strong prompt scaffolding.

Phase 2 Results

Phase 2 introduced more complex and structured prompt formulations.

- **GPT-5.0:** Continued to perform well but revealed greater prompt sensitivity. CLEAR prompts remained rigorous but verbose; Few-Shot outputs degraded into outline-like responses with weaker completeness (score 2). CoT remained strong but less systematically organized.
- **Gemini 2.5 Flash:** Showed improvement under Few-Shot prompts (scoring 4s across most metrics, “very solid”), but CLEAR and CoT outputs remained middling (mostly 3s). The model’s reliance on Few-Shot examples highlights prompt dependency.
- **Claude Sonnet 4.0:** Emerged as the top performer. Both Few-Shot and CoT prompts scored a perfect 4 across all metrics, producing comprehensive, precise, and compliance-ready outputs. CLEAR remained strong, though occasionally less detailed.

Key Insight: Claude demonstrated the most stability across phases, with Few-Shot and CoT ensuring high-quality, reproducible results. GPT-5.0 remained powerful but inconsistent across prompt styles, while Gemini improved with Few-Shot but lagged in overall robustness.

Prompt-Level Patterns

- **CLEAR Method:** Ensured professionalism and numeric fidelity, but was frequently verbose and weaker on forward-looking nuance.
- **Few-Shot:** Most stable across models, elevating Gemini to near parity with advanced models and delivering Claude’s strongest outputs.

- **CoT:** Enhanced reasoning and interpretability, especially with Claude, but introduced verbosity and organizational drift in GPT-5.0 and Gemini.

Failure Case Documentation
Failure Case Documentation received the lowest rubric score due to recurring errors across both phases:

- **Forward-Looking Omission:** MD&A guidance, strategy, or projections are frequently omitted.
- **Numeric Distortion:** Metrics were paraphrased, rounded, or miscopied, undermining fidelity.
- Inconsistent Labeling: Similar risks are categorized differently across outputs.
- **Weak Evidence Anchoring:** Outputs lacked direct quotes or page references, reducing auditability.
- **Verbosity:** Especially in CoT outputs, exceeding word constraints.

Example: GPT-5.0 using CoT failed to capture R&D expansion guidance in an MD&A summary. The omission reduced investor insight into forward-looking strategy, highlighting prompt ambiguity in extracting forward-looking statements.

Root Causes: Ambiguous prompt design, model hallucination tendencies, and absence of automated numeric and labeling validation.

IV. Discussion

Both tasks 1 and 2 evaluations reveal important trends in model and prompt performance. GPT-5.0 remains a powerful model but is increasingly sensitive to prompt design in more structured tasks, while Gemini demonstrates meaningful improvement under Few-Shot prompts but remains inconsistent elsewhere, limiting its reliability in compliance-sensitive contexts. However, Claude Sonnet 4.0 consistently outperforms competitors by balancing accuracy, completeness, and professional tone, particularly under Few-Shot and CoT prompts.

Furthermore, prompt frameworks play a decisive role. CLEAR enforces numeric accuracy and tone but sacrifices nuance and forward-looking coverage. Few-Shot provides stability and format consistency, rescuing weaker models like Gemini. CoT enhances interpretability and reasoning but risks verbosity and inconsistency.

Moreover, Failure Case Documentation underscores ongoing vulnerabilities. Forward-looking omissions and numeric distortions present significant compliance risks in finance applications. Without explicit instructions and validation, even advanced models misrepresent material investor information.

Overall, the pattern across phases suggests that prompt engineering, evidence anchoring, and automated post-checks are as critical as model selection in ensuring reliable outputs. For finance professionals, the trade-off is clear: advanced models with well-structured Few-Shot or CoT prompts provide the strongest results, but safeguards are essential to mitigate compliance risks.

V. Recommendation

Short-Term:

- Harden prompts to explicitly extract forward-looking items and enforce numeric fidelity.
- Apply constraints on output length (≤200 words) to ensure concise, readable summaries.

Medium-Term:

- Implement evidence anchoring for all metrics and risk factors with verbatim quotes and page references.
- Automate numeric and label consistency checks to reduce human review burden.
- Incorporate adversarial examples in prompt training to cover edge cases.

Long-Term:

- Deploy human-in-the-loop review for high-materiality outputs.
- Build golden datasets of benchmark filings with ground-truth labels for continuous performance monitoring.

These measures collectively reduce compliance risk, improve interpretability, and strengthen LLM outputs for finance professionals.

VI. Conclusion

Finance demands precision, transparency, and reliability. While LLMs show promise for summarization and risk classification, consistent issues in Failure Case Documentation indicate remaining vulnerabilities. Task 1 and 2 evaluations reveal that Claude Sonnet 4.0 with Few-Shot or CoT prompts currently delivers the strongest, most reliable outputs. However, prompt refinement, automated validation, and human oversight remain essential safeguards. Implementing these measures will ensure trustworthy, compliance-ready LLM applications aligned with professional finance standards.