Phase 2 Evaluation Report Appendix

Table 1: Overall Performance Summary

This table provides the aggregate scores across all 20 queries, representing the system-wide baseline for Phase 2.

Metric	Mean Score	Interpretation
Groundedness	0.409	Moderate verbatim adherence; roughly 41% of text is direct source overlap.
Answer Relevance	0.821	High topical alignment; answers consistently address the user's intent.

Table 2: Performance by Query Category

This breakdown illustrates how the RAG system handles different levels of complexity and ambiguity.

Query Type	Count	Mean Groundedness	Mean Relevance
Direct	10	0.416	0.872
Synthesis	5	0.442	0.792
Edge Case	5	0.361	0.749

Table 3: Performance Extremes (Best and Worst)

Identifying specific successes and failures helps pinpoint where the model's retrieval or reasoning breaks down.

Query ID	Rank	Groundedness	Relevance	Primary Characteristic
Q01	Best	0.516	0.833	High-precision extraction of specific commute statistics.
Q12	2nd Best	0.484	0.731	Strong synthesis across multiple firm-level reports.
Q06	Worst	0.293	0.920	High relevance but low grounding; likely contains "narrative filler."
Q18	2nd Worst	0.308	0.806	Edge case involving mixed evidence that led to lower quote overlap.

Table 4: Detailed Evaluation Results

The full set of 20 evaluation queries and their respective scores.

Query ID	Type	Groundedness	Relevance
Q01	Direct	0.516	0.833
Q02	Direct	0.415	0.870
Q03	Direct	0.439	0.920
Q04	Direct	0.441	0.786
Q05	Direct	0.441	0.750
Q06	Direct	0.293	0.920
Q07	Direct	0.390	0.900
Q08	Direct	0.424	0.926
Q09	Direct	0.391	0.941
Q10	Direct	0.409	0.875
Q11	Synthesis	0.431	0.857
Q12	Synthesis	0.484	0.731
Q13	Synthesis	0.456	0.828
Q14	Synthesis	0.426	0.720
Q15	Synthesis	0.412	0.825
Q16	Edge Case	0.370	0.815
Q17	Edge Case	0.373	0.692
Q18	Edge Case	0.308	0.806
Q19	Edge Case	0.388	0.884
Q20	Edge Case	0.364	0.548