

FinDRA: Multi-Agent Financial Reasoning under Asymmetric Information and Unequal Authority

Inyoung Won

Department of Artificial Intelligence
and Data Science
Sejong University
Seoul, Republic of Korea
uiio19991201@gmail.com

Yunyoung Lee*

Department of Artificial Intelligence
and Data Science
Sejong University
Seoul, Republic of Korea
yylee93@sejong.ac.kr

Abstract

We propose FinDRA, a multi-agent framework for financial question answering under asymmetric information access. FinDRA partitions a financial report into agent-specific evidence regions, allows agents to exchange only their answers, concise reasoning, and confidence estimates, and coordinates their revisions through full-document verifier feedback and authority-framed debate. Experiments on MultiHiertt show that the asymmetric verifier condition obtains the highest overall point estimates among the compared settings, with an EM score of 0.284 and a mean LLM-as-a-Judge score of 0.753 on a 0–2 scale. Importantly, the benefit of asymmetric access depends on evidence structure: it is strongest when question-relevant evidence remains within an agent-accessible region, whereas cross-partition questions expose a trade-off between semantic adequacy and exact numerical agreement. Agent-level analyses further show revision patterns consistent with the assigned authority framing, without attributing these differences to authority alone. Cross-dataset experiments on FinQA and TAT-QA similarly show that asymmetric access is not universally beneficial. These findings suggest that effective multi-agent financial reasoning depends not simply on providing more information, but on how evidence is partitioned, communicated, and verified.

1 Introduction

Real-world collaborative decision-making often involves organizations that possess different informational assets but cannot fully disclose their underlying data because of competitive, legal, or institutional constraints. Participants must therefore form partial judgments from limited evidence and coordinate them through selectively shared reports, analyses, and conclusions.

This challenge is particularly salient in finance, where information sources differ not only in content but also in their epistemic characteristics. Narrative disclosures provide contextual explanations, expectations, and strategic interpretations, whereas structured financial tables provide more explicit values and calculation traces. Financial reasoning under restricted access must therefore coordinate heterogeneous evidence while accounting for differences in source and format.

Many LLM-based multi-agent reasoning frameworks induce diversity through roles, perspectives, or interaction strategies while providing agents with identical or substantially overlapping evidence. Recent work has begun to examine information asymmetry as an explicit design dimension, suggesting that diversity in evidence access can materially affect multi-agent deliberation. However, less is known about how such asymmetric access behaves in document-grounded reasoning, where relevant evidence may be distributed across heterogeneous textual and tabular regions and may need to be reconstructed through communication.

To study this setting, we propose FinDRA, a framework for Multi-Agent Financial Reasoning under Asymmetric Information. FinDRA partitions a financial report into agent-specific evidence regions and assigns each agent a document-based role with authority-framed behavioral guidance. Each agent first produces an independent judgment using only its assigned evidence. Agents subsequently exchange their answers, concise reasoning, and confidence estimates—but not their underlying document regions—and revise their judgments through iterative debate supported by a verifier with full-document access. The resulting agent judgments are then aggregated into a final answer. This design allows us to examine when localized evidence access supports specialization and when information separation creates coordination costs.

We address three research questions: (1) How

*Corresponding author.

does partitioned information access compare with shared-information multi-agent and single-agent reasoning? (2) How are authority-framed roles and verifier feedback associated with answer retention and revision during debate? (3) Under what evidence structures does asymmetric access help or hinder financial reasoning?

Our contributions are threefold. First, we formulate financial document QA as a multi-agent reasoning problem with structurally partitioned evidence, where agents reason over restricted document regions and communicate judgments without directly sharing their underlying evidence. Second, we systematically compare partitioned and shared-information settings with and without full-document verification, showing that the value of asymmetric access depends on evidence structure: it is most beneficial when relevant evidence remains locally accessible, while cross-partition reasoning introduces coordination costs. Third, we characterize the resulting coordination dynamics through agent-level revision behavior, alternative authority encodings, aggregation strategies, and cross-dataset experiments on MultiHiertt, FinQA, and TAT-QA.

2 Related Work

2.1 Financial Question Answering

Financial question answering requires models to identify relevant evidence from narrative text and tables and perform numerical or cross-format reasoning. FinQA formalized the task using financial reports and executable reasoning programs, TAT-QA extended it to evidence identification across tables and text, and ConvFinQA introduced conversational numerical reasoning over prior questions and intermediate results (Chen et al., 2021; Zhu et al., 2021; Chen et al., 2022).

Despite these advances, strong LLMs still struggle with semi-structured financial documents as table complexity and reasoning depth increase (Srivastava et al., 2024). Identifying the correct calculation arguments among numerous candidate values also remains a major challenge (Lim et al., 2024). MultiHiertt further requires reasoning over multiple hierarchical tables and lengthy narrative content, making it necessary to connect evidence across tables and distant document regions (Zhao et al., 2022).

Recent work has decomposed financial QA into agent roles for formula selection, evidence extrac-

tion, calculation, and verification (Liu et al., 2025). However, such systems generally retain access to the complete task evidence. FinDRA instead partitions the underlying financial report and examines how agents coordinate judgments formed from restricted information.

2.2 Multi-Agent Debate with Large Language Models

Multi-agent debate enables LLM agents to review and revise one another’s answers. Iterative exchanges and structured debate have been shown to improve factuality and reasoning, with final answers selected through consensus or adjudication (Du et al., 2024; Liang et al., 2024). Multi-agent interaction has also been applied to evaluation: ChatEval organizes debate among evaluator agents, while M-MAD assigns evaluation dimensions to specialized agents before final adjudication (Chan et al., 2024; Feng et al., 2025).

Performance depends on how agents communicate and aggregate their judgments. Prior work has examined sparse and dynamic communication structures (Li et al., 2024; Sun et al., 2025), voting and consensus mechanisms (Kaesberg et al., 2025), confidence-weighted aggregation (Chen et al., 2024b), and uncertainty communication (Yoffe et al., 2025). However, most settings provide agents with identical or substantially overlapping information and induce diversity through roles, identities, positions, or prompts (Choi et al., 2026). FinDRA differs by separating access to source information before debate and allowing agents to exchange only answers, concise reasoning, and confidence estimates rather than the other agents’ original evidence.

2.3 Multi-Agent Systems in Finance

LLM-based agents have been applied to financial analysis, investment research, portfolio management, and quantitative finance (Dong et al., 2025; Kundu et al., 2025). Specialized agents have also been used in simulated trading and investment workflows (Li et al., 2025). FinDebate organizes agents around corporate performance, market conditions, sentiment, valuation, and risk and integrates their analyses through structured debate (Cai et al., 2025).

Role-based systems have further supported financial report generation. Prior work assigns analyst, writer, and editor roles to produce investment reports from earnings calls (Satapathy et al., 2025),

while adversarial agent perspectives have been used to generate more balanced financial interpretations (Sadhu et al., 2025). These systems distinguish agents primarily through analytical functions, perspectives, or workflow stages while commonly retaining access to a shared document or retrieval source. FinDRA instead separates access to the financial report itself and analyzes how authority-framed roles revise judgments derived from different information regions.

3 Methodology

3.1 Task Definition

We consider a multi-agent reasoning problem in which agents with access to different information subsets must produce a joint answer. Let q denote the question, D the complete information set, and $D_i \subseteq D$ the subset assigned to agent i . Each agent generates an initial judgment using only D_i , without direct access to the original information assigned to other agents.

Agents iteratively revise their judgments using the answers, concise reasoning, and confidence estimates shared by the other agents. Each agent is assigned a document-based role, an authority level, and corresponding behavioral guidance that affects whether it retains or reconsiders its answer when conflicting judgments arise. Here, authority refers to prompt-defined behavioral guidance that conditions how strongly an agent is encouraged to retain or reconsider its current judgment; it is neither learned nor inferred by the system. A verifier with access to D evaluates the agent responses after each round. Feedback from the first two evaluations is provided in the subsequent debate rounds, while the third evaluation is used during final aggregation.

In financial document QA, D consists of narrative content, core financial statements, and supplementary tables or footnotes. Partitioning these regions across agents enables comparison between asymmetric multi-agent reasoning and shared-information settings.

3.2 Datasets

We use MultiHiertt, which contains narrative text and multiple hierarchical tables from SEC 10-K reports, as the primary dataset (Zhao et al., 2022). From its development set, we select 591 questions containing both textual and tabular evidence, comprising 458 arithmetic and 133 span-selection ques-

Question type	Count	Proportion
Single-table	352	76.9%
Multi-table	37	8.1%
Cross-partition	69	15.1%

Table 1: Distribution of the 458 MultiHiertt arithmetic questions by evidence location.

Agent	Information source
A1	paragraphs
A2	tables[0:2]
A3	tables[2:]

Table 2: Agent-specific information partitioning in MultiHiertt.

tions.

We categorize the arithmetic questions by evidence location. Single-table questions rely on one table, multi-table questions require multiple tables within the region assigned to the same agent, and cross-partition questions require evidence from both A2 and A3.

Some MultiHiertt paragraphs contain markers linked to accompanying tables, which could indirectly expose tabular information to the text-based agent. We identify 199 such cases and remove the embedded table markers so that A1 receives only the surrounding narrative.

For cross-dataset evaluation, we additionally use 735 FinQA development questions requiring both textual and tabular evidence (Chen et al., 2021) and 507 TAT-QA questions whose evidence spans text and tables (Zhu et al., 2021).

3.3 FinDRA Framework

FinDRA integrates judgments formed from partitioned financial information through four stages: information partitioning, independent initial reasoning, verifier-supported debate, and final aggregation, as illustrated in Figure 1.

For MultiHiertt, narrative paragraphs are assigned to A1, tables[0:2] to A2, and tables[2:] to A3. The partition is position-based rather than semantic; A2 and A3’s financial-statement role labels provide authority framing and do not imply that every assigned table belongs to the corresponding category.

Agents cannot access the original content assigned to other agents and exchange only their current answers, concise reasoning, and confidence estimates during debate. For FinQA, `pre_text`, `table`, and `post_text` are assigned to A1, A2, and

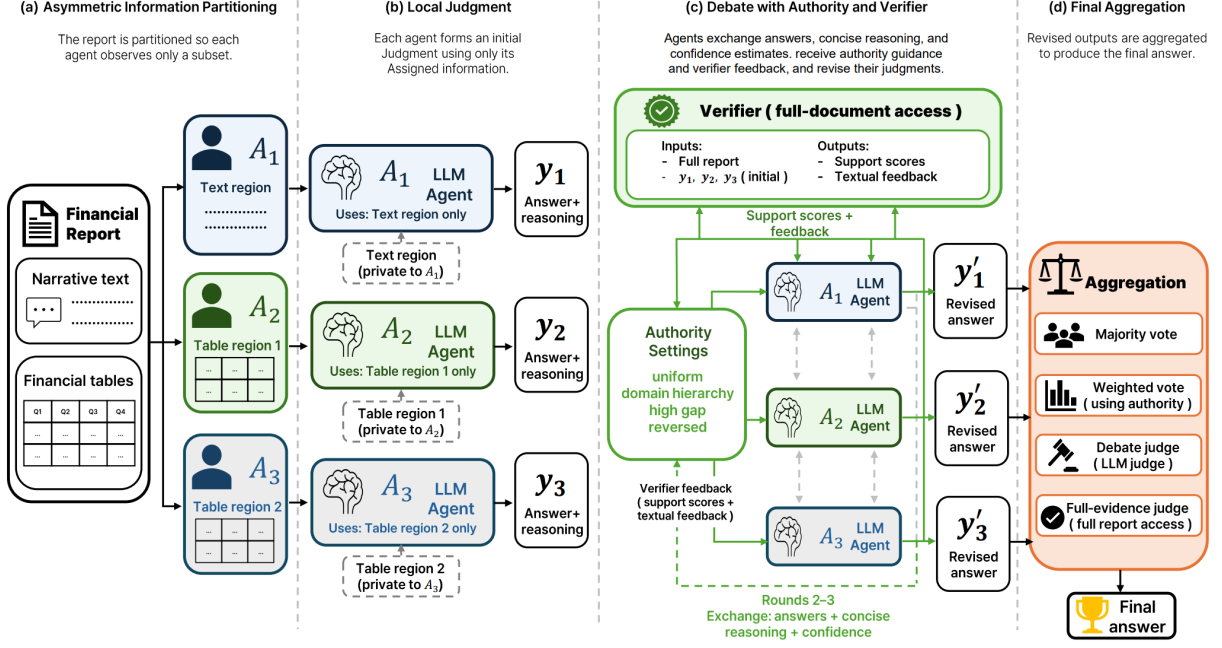


Figure 1: Overview of FinDRA. Agents independently reason over assigned document regions and subsequently exchange answers, concise reasoning, and confidence estimates. A full-document verifier provides support scores and feedback after each round. Final agent outputs are combined through majority vote, weighted vote, a debate judge, or a full-evidence judge.

A3, respectively. For TAT-QA, the earlier textual paragraphs, table, and later textual paragraphs are assigned to A1, A2, and A3, respectively.

FinDRA additionally assigns authority-framed roles based on the intended epistemic reliability associated with each document-based role. In the primary setting, A1 has low authority, A2 high authority, and A3 intermediate authority. Lower-authority agents are encouraged to reconsider their answers when stronger conflicting evidence is presented, whereas higher-authority agents retain evidence-supported judgments unless clear counterevidence is available. Authority does not indicate that an agent is necessarily correct and does not reveal the gold answer. Full prompts are provided in Appendix C, and numerical authority encodings are examined separately as robustness checks.

In Round 1, each agent independently answers using only its assigned information. After each round, a verifier receives the question, complete document, and agent responses, but not the gold answer, and returns support scores together with evidence-based assessments of factual support and calculation correctness. Agent-specific verifier assessments from Rounds 1 and 2 are provided to the corresponding agents in the subsequent round, while the Round 3 evaluation is retained for final

aggregation. The verifier’s global best_agent prediction and synthesis are not exposed to the reasoning agents. In Rounds 2 and 3, agents retain or revise their answers using their local information, other agents’ shared judgments, verifier feedback, and authority guidance.

After Round 3, we compare four aggregation strategies. Majority vote selects the most frequent answer, while weighted vote weights each agent’s final answer using the support score assigned by the verifier after Round 3. The debate judge generates a prediction from the final agent responses and the latest verifier assessment and is used for the main results. The full-evidence judge additionally receives the complete original document.

3.4 Experimental Settings and Evaluation

We construct four conditions that vary information access and verifier availability. Asym-verifier and asym-free provide each agent with its assigned document partition and differ in whether verifier feedback is available. Sym-verifier and sym-free provide all agents with the complete document and likewise differ in verifier availability. Authority framing, task instructions, and decoding settings are otherwise held constant.

We additionally compare three single-agent base-

Condition	Information access	Verifier
Asym-verifier	Partitioned	Yes
Asym-free	Partitioned	No
Sym-verifier	Full document	Yes
Sym-free	Full document	No

Table 3: Experimental conditions.

lines: single-full receives the complete text and tables, single-tables receives only the tables, and single-text receives only the narrative content.

Numerical answers are evaluated using EM@1%, which treats a prediction as correct when its relative error from the gold answer is within 1%. Percentage and decimal representations are normalized before evaluation, while textual answers use normalized string matching. We also report LLM-as-a-Judge scores to capture semantic correctness not reflected by strict matching. Given the question, gold answer, and prediction, the judge assigns 0 for incorrect, 1 for partially correct, or 2 for fully correct, and we report the mean item-level score. The same prompt is used for all conditions, following prior model-based evaluation work (Liu et al., 2023). Because LLM judges can exhibit systematic biases and robustness failures, Judge scores are reported alongside, rather than in place of, EM (Chen et al., 2024a; Raina et al., 2024).

We analyze agent revision behavior using position change rate, the proportion of responses that change between consecutive rounds. Wrong-to-correct transitions are treated as beneficial revisions, while correct-to-wrong transitions are harmful revisions. We also track the verifier’s document-support scores across rounds. These scores assess the evidential support for an answer and its reasoning and are not treated as direct correctness measures.

A1, A2, and A3 use GPT-4o-mini with temperature 0.1, while the GPT-4o verifier uses temperature 0.0. The debate judge and evaluation judge use GPT-4o-mini with temperature 0.0. All multi-agent conditions use three debate rounds. The main experiments use the v1 natural-language authority framing; the numerical v2 and v4 variants are evaluated separately as prompt-robustness checks.

Estimated input-token cost. Both verifier conditions require 13 model calls per question, but asym-verifier uses approximately 18.1K input tokens compared with 33.3K for sym-verifier because

Method	All questions (N=591)		Arithmetic (N=458)	
	EM	Judge (0–2)	EM@1%	Judge (0–2)
Single-full	0.1963	0.5482	0.1397	0.3908
Single-tables	0.2420	0.5550	0.1747	0.3755
Single-text	0.0118	0.0254	0.0066	0.0175
Sym-free	0.2000	0.3847	0.1507	0.2860
Asym-free	0.2780	0.5881	0.2096	0.4083
Sym-verifier	0.2780	0.6390	0.2249	0.4978
Asym-verifier	0.2843	0.7530	0.2424	0.5917

Table 4: Main results on MultiHiertt (N=591).

each reasoning agent receives only its assigned partition. Detailed estimates are provided in Appendix A.

4 Experiments

4.1 Overall Performance

Table 4 reports the results on the full MultiHiertt set and the arithmetic subset. Asym-verifier obtains the highest point estimates across all reported metrics, with EM and Judge scores of 0.2843 and 0.7530 on all questions and 0.2424 and 0.5917 on the arithmetic subset. Table 5 reports pairwise significance tests on the arithmetic subset. Relative to Sym-free, every alternative condition yields a statistically significant improvement in both EM (Holm-corrected $p < 0.05$) and correctness ($p < 0.01$). Notably, Asym-free significantly exceeds Sym-free despite neither condition receiving verifier feedback, providing direct evidence that partitioned access itself can benefit reasoning in this setting. Likewise, Sym-verifier significantly exceeds Sym-free, showing an independent benefit from verifier feedback under shared access. Among the stronger conditions, however, EM differences are not statistically significant (Asym-verifier vs. Sym-verifier: $p = 1.00$; Asym-verifier vs. Asym-free: $p = 0.452$), while the corresponding judge-based correctness gains remain significant in both cases ($p < 0.03$), indicating that combining partitioning with verification continues to improve semantic answer quality even where it does not measurably shift the EM@1% rate at this sample size. Because each condition is evaluated with a single inference run, these item-level paired tests do not capture run-to-run stochastic variability. Accordingly, small differences among the stronger conditions should be interpreted cautiously.

To rule out self-preference bias from the shared GPT-4o-mini/GPT-4o model family used across agents and judges, we additionally validate correctness scores against an independent cross-provider

Comparison	Δ EM	EM p (Holm)	Δ Corr.	Corr. p (Holm)
Asym-Ver vs Sym-Ver	+0.017	1.000	+0.094	0.023
Asym-Ver vs Asym-Free	+0.033	0.452	+0.183	<0.001
Asym-Free vs Sym-Free	+0.059	0.042	+0.122	0.005
Sym-Ver vs Sym-Free	+0.074	<0.001	+0.212	<0.001
Asym-Ver vs Sym-Free	+0.092	<0.001	+0.306	<0.001

Table 5: Pairwise statistical significance across the four conditions on the arithmetic subset (debate judge, $n = 458$). EM uses exact McNemar’s test; correctness (0–2 scale) uses a paired permutation test ($N = 100,000$). p -values are Holm-corrected across all five comparisons within each metric.

Type	Method	EM@1%	Judge (0–2)
Multi-table	Single-full	0.0270	0.1622
	Single-tables	0.0811	0.1351
	Sym-verifier	0.2703	0.5676
	Asym-verifier	0.3243	0.7838
Cross-partition	Single-full	0.0725	0.2464
	Single-tables	0.1014	0.2174
	Sym-verifier	0.2319	0.5072
	Asym-verifier	0.2029	0.6087

Table 6: Results by evidence distribution.

judge (Appendix B.1).

4.2 Performance by Evidence Distribution and Question Type

Table 6 reports results by evidence distribution.

On multi-table questions, asym-verifier performs best on both metrics, suggesting that asymmetric access can be beneficial when question-relevant evidence remains within an agent-accessible region. This advantage may reflect several mechanisms, including more focused local contexts, reduced exposure to irrelevant evidence, greater diversity across agent judgments, and potential specialization; the current experiments do not isolate these effects.

Cross-partition questions reveal the complementary failure mode: sym-verifier achieves higher EM, whereas asym-verifier achieves a higher Judge score. When the required evidence is distributed across partitions, agents must reconstruct the relevant information from generated messages rather than inspect the original evidence directly. This communication bottleneck may omit or distort numerical values, units, or calculation conditions needed for exact agreement.

Verifier feedback also affects arithmetic and span-selection questions differently. On span-selection questions, asym-free achieves the highest EM score of 0.5152, compared with 0.4286 for asym-verifier, while asym-verifier obtains the high-

Agent	Authority	Change	R1	R2	R3
A1	Low	92.1%	0.348	0.566	0.784
A2	High	19.3%	0.346	0.652	0.747
A3	Medium	62.6%	0.317	0.616	0.787

Table 7: Position change rates and verifier scores under the v1 authority-framing prompt.

Agent	Auth.	W→C	C→W	C→C	W→W	Net
A1	Low	0.132	0.033	0.053	0.782	+0.099
A2	High	0.095	0.040	0.074	0.791	+0.055
A3	Medium	0.122	0.035	0.074	0.769	+0.087

Table 8: Revision quality under the v1 authority-framing prompt. Net denotes $W \rightarrow C$ minus $C \rightarrow W$.

est Judge score of 1.3083. Verifier-supported revisions may therefore improve semantic adequacy without preserving the exact expressions required by normalized string matching.

4.3 Authority and Debate Dynamics

We examine whether agent revision patterns are consistent with the v1 authority framing. Table 7 shows that position-change rates follow the authority ordering: A1 revises most frequently, A2 least frequently, and A3 lies between them.

To evaluate revision quality, we classify valid round-to-round transitions as wrong-to-correct ($W \rightarrow C$), correct-to-wrong ($C \rightarrow W$), correct-to-correct, or wrong-to-wrong using EM@1%. Net revision gain is defined as the $W \rightarrow C$ rate minus the $C \rightarrow W$ rate.

All three agents achieve positive net revision gains, with A1 obtaining the largest gain of +0.099. Thus, the most frequently revising role also exhibits the largest balance of beneficial over harmful revisions.

Across valid responses, the mean verifier score increases from 0.338 in Round 1 to 0.611 in Round 2 and 0.777 in Round 3, while agent-level EM increases from 0.045 to 0.161 and 0.206. Correct answers receive a higher mean verifier score than incorrect answers (0.872 vs. 0.528), although the correlation between verifier score and correctness is modest ($r = 0.243$). The verifier therefore provides a positive but imperfect signal of answer quality.

Overall, the v1 setting exhibits revision patterns consistent with its authority-framed roles, while positive net gains suggest that revisions are more often beneficial than harmful. Because authority level is coupled with document role and informa-

Setting	v2 EM	v4 EM	v2 Judge	v4 Judge
Uniform	0.2402	0.2358	0.5415	0.4891
Domain hierarchy	0.2533	0.2533	0.5306	0.5175
High gap	0.2686	0.2380	0.5131	0.5044
Reversed	0.2293	0.2140	0.4738	0.4782

Table 9: Results of alternative authority encodings on the MultiHiertt arithmetic subset ($N = 458$). Judge scores use a 0–2 scale.

tion access, these patterns should be interpreted as associations rather than isolated causal effects of authority.

4.4 Robustness to Alternative Authority Encodings

We test two numerical authority variants: v2 provides only an authority score, while v4 additionally instructs agents to consider evidence quality. Both are evaluated under uniform, domain-hierarchy, high-gap, and reversed configurations.

As shown in Table 9, Domain hierarchy outperforms the reversed setting under both variants, whereas larger authority gaps do not consistently improve performance. Across these prompt variants, the relative authority ordering appears more consistent than the magnitude of the assigned authority gaps, although the effect remains sensitive to the specific authority encoding.

4.5 Aggregation Strategy Ablation

We compare majority vote, weighted vote, a debate judge, and a full-evidence judge, with full results reported in Appendix Table 12. The debate judge performs best under verifier conditions, whereas the full-evidence judge performs best without verifier feedback, suggesting that direct document access at aggregation is most useful when full-document verification has not already occurred.

4.6 Cross-Dataset Generalization

We further evaluate FinDRA on FinQA and TAT-QA, which differ from MultiHiertt in document and evidence structure. Table 10 reports the evaluated multi-agent conditions and single-agent baselines.

On FinQA, sym-free achieves the highest EM of 0.5143, while sym-verifier obtains the highest Judge score of 0.7973. Asym-verifier outperforms both single-agent baselines but remains below the evaluated symmetric conditions. These results suggest that partitioning is less beneficial when question-relevant evidence is concentrated within

Dataset	Method	EM	Judge (0–2)
FinQA	Single-full	0.3714	0.6952
	Single-tables	0.4068	0.6816
	Asym-verifier	0.4435	0.7660
	Asym-free	0.4190	0.6721
	Sym-verifier	0.5007	0.7973
	Sym-free	0.5143	0.7878
TAT-QA	Single-full	0.3452	1.1972
	Single-tables	0.4004	1.2012
	Asym-verifier	0.4135	1.2982
	Asym-free	0.4266	1.2560
	Sym-verifier	0.3865	1.3127
	Sym-free	0.3865	1.2789

Table 10: Cross-dataset generalization results on FinQA and TAT-QA.

a table and its associated textual context.

On TAT-QA, asym-free achieves the highest EM of 0.4266, whereas sym-verifier obtains the highest Judge score of 1.3127. Asym-verifier exceeds both single-agent baselines in Judge score and slightly exceeds single-tables in EM (0.4135 vs. 0.4004). The different EM and Judge rankings again indicate that verifier-supported interaction may improve semantic evaluation without consistently improving exact agreement.

Overall, the cross-dataset results provide qualified evidence of generalization. Multi-agent conditions generally improve over single-agent baselines, but asymmetric access does not consistently outperform shared access. Together with the MultiHiertt evidence-distribution analysis, these results suggest a structural trade-off: partitioning can support local specialization when relevant evidence is concentrated within agent-accessible regions, but can become detrimental when solving the question requires precise reconstruction of evidence distributed across those regions.

5 Discussion

5.1 Benefits and Limits of Asymmetric Information

Asymmetric access is most beneficial when question-relevant evidence remains within an agent-accessible region, as observed for MultiHiertt multi-table questions. When evidence spans partitions, communication may omit numerical details needed for exact agreement. FinDRA is therefore not universally superior to shared-information reasoning; its effectiveness depends on whether the benefits of partitioning—such as more focused local contexts, reduced irrelevant evidence,

and potentially more diverse or specialized agent judgments—offset the information loss and coordination costs introduced by partitioning. The asymmetric-verifier condition should therefore be interpreted as restricted agent-level access with centralized verification, rather than strict information isolation. Accordingly, the Asym-free condition offers a more direct assessment of partitioned access without information being reintroduced through full-document verifier feedback.

5.2 Authority-Framed Revision and Verification

Authority-framed roles exhibit distinct revision behavior, but these patterns cannot be attributed solely to authority because role, access, and authority are coupled. Numerical variants further suggest that domain-informed authority ordering is more reliable than its reversal, while the optimal magnitude remains prompt-dependent.

Verifier feedback improves arithmetic reasoning but does not consistently improve exact agreement on span-selection or cross-dataset questions, indicating that verification provides a useful but imperfect signal for revision.

6 Conclusion

This study introduced FinDRA, a multi-agent financial reasoning framework combining partitioned information access, authority-framed roles, verifier-supported debate, and final aggregation. On MultiHiertt, asym-verifier obtains the strongest overall point estimates, particularly when relevant evidence remains within an agent-accessible region, while cross-partition and cross-dataset results show that asymmetric access is not uniformly advantageous.

Agent-level analyses show revision patterns consistent with the assigned authority framing and positive net revision gains, although authority, document role, and information access remain coupled. Overall, effective multi-agent financial reasoning depends on matching information access, authority, communication, verification, and aggregation to the structure of the underlying evidence.

7 Limitations

The evidence categories analyzed in this study are limited in size. The multi-table and cross-partition subsets contain only 37 and 69 questions, respectively, so their results should be interpreted as ex-

ploratory. The experiments use one inference run per condition. Although we conduct item-level paired significance tests across conditions, these tests do not capture run-to-run variability arising from stochastic model inference. Small performance differences should therefore be interpreted cautiously.

In the primary setting, authority is instantiated together with document-specific roles, reflecting realistic differences in the epistemic reliability of financial information sources. Although the alternative authority configurations show performance differences across authority orderings, fully isolating its causal effect would require controlling document roles and information access across agents. The partitions are also predefined by dataset structure. In MultiHiertt, A2 and A3 receive position-based table regions rather than semantically verified financial-statement categories. Although the authority prompts characterize A2 as a primary financial-statement analyst and A3 as a cash-flow and supplementary-data analyst, these labels function as role and authority framing and do not guarantee that every assigned table belongs to the corresponding semantic category. Alternative semantically grounded partitioning strategies may yield different results.

The experiments primarily use one model family and an LLM-based evaluator. Although Judge scores are reported alongside EM, they may remain sensitive to model bias and do not replace task-specific correctness measures. The token analysis also estimates only input-context consumption and does not include output tokens, latency, or monetary cost.

Finally, agents exchange generated answers, concise reasoning, and confidence estimates rather than original evidence. This design may omit or alter important numerical details and can lead to premature consensus. Such failures were identified qualitatively rather than measured systematically. Future work should examine additional models, adaptive information and authority assignment, and structured evidence exchange that preserves source references, table coordinates, units, and numerical values.

While the agents, debate judge, and evaluation judge primarily use models from the GPT-4o-mini/GPT-4o family, cross-model validation with an independent judge (Appendix B.1) preserves the same condition ranking and confirms that all three stronger conditions significantly outperform

Sym-free. However, the smaller validation subsample does not reproduce all pairwise significance differences among the stronger conditions. These results reduce, but do not eliminate, concerns about judge-agent model-family overlap.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25603869).

References

- Tianshi Cai, Guanxu Li, Nijia Han, Ce Huang, Zimu Wang, Changyu Zeng, Yuqi Wang, Jingshi Zhou, Haiyang Zhang, Qi Chen, et al. 2025. Findebate: Multi-agent collaborative intelligence for financial analysis. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 268–282.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *International Conference on Learning Representations*.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.
- Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024b. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085.
- Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.
- Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 6279–6292.
- Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. 2026. When identity skews debate: Anonymization for bias-reduced multi-agent reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14284–14311.
- Yifei Dong, Fengyi Wu, Kunlin Zhang, Yilong Dai, Sanjian Zhang, Wanghao Ye, Sihan Chen, and Zhi-Qi Cheng. 2025. Large language model agents in finance: A survey bridging research, practice, and real-world deployment. In *EMNLP (Findings)*, pages 17889–17907.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763.
- Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. M-mad: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7084–7107.
- Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Voting or consensus? decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11640–11671.
- Sayani Kundu, Dushyant Sahoo, Victor Li, Jennifer Rabowsky, and Amit Varshney. 2025. A multi-agent framework for quantitative finance: An application to portfolio management analytics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 812–824.
- Xiangyu Li, Yawen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. 2025. QuantAgents: Towards multi-agent financial system via simulated trading. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17438–17464.
- Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7281–7294.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 17889–17904.
- Jinsu Lim, Yechan Hwang, Young-Jun Lee, and Ho-Jin Choi. 2024. Enhancing arguments recognition for financial mathematical reasoning over hybrid data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3961–3973.

- Chenghao Liu, Qian Liu, Ziqin Zhu, Hao Fei, and Aniket Mahanti. 2025. David vs. goliath: Cost-efficient financial qa via cascaded multi-agent reasoning. In *EMNLP (Findings)*, pages 4212–4229.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.
- Vyas Raina, Adian Liusie, and Mark Gales. 2024. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7499–7517.
- Saisab Sadhu, Biswajit Patra, and Tanmay Basu. 2025. Structured adversarial synthesis: A multi-agent framework for generating persuasive financial analysis from earnings call transcripts. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 283–291.
- Ranjan Satapathy, Raphael Liew, Joyjit Chattorj, Erik Cambria, and Rick Goh. 2025. From earnings calls to investment reports: Evaluating role-based multi-agent llm systems. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 258–267.
- Pragya Srivastava, Manuj Malik, Vivek Gupta, Tanuja Ganu, and Dan Roth. 2024. Evaluating llms’ mathematical reasoning in financial document question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3853–3878.
- Yiliu Sun, Zicheng Zhao, Sheng Wan, and Chen Gong. 2025. Cortexdebate: Debating sparsely and equally for multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9503–9523.
- Luke Yoffe, Alfonso Amayuelas, and William Yang Wang. 2025. Debunc: Improving large language model agent communication with uncertainty metrics. In *EMNLP (Findings)*, pages 23299–23315.
- Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. Multihiertr: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600.
- Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 3277–3287.

A Estimated Input-Token Cost

Both the asymmetric and symmetric verifier conditions issue 13 model calls per question: three agents across three debate rounds, three verifier evaluations, and one judge call. However, in the symmetric condition, each agent receives the complete document, whereas in the asymmetric condition, each agent receives only its assigned partition. Estimating input tokens at approximately four characters per token, the sym-verifier condition consumes roughly 33.3K input tokens per question, compared with 18.1K for the asym-verifier condition, corresponding to a ratio of 1.84.

Thus, asym-verifier achieves higher LLM-as-a-Judge performance while consuming fewer estimated input tokens, indicating that its advantage is not explained by greater input-context consumption. These estimates measure input tokens only and do not include output tokens, latency, or monetary API cost.

Condition	Calls	Input tokens	Judge (0–2)
Single-full	1	2.5K	0.391
Sym-verifier	13	33.3K	0.498
Asym-verifier	13	18.1K	0.592

Table 11: Estimated model calls and input-token consumption per question on the MultiHierrt arithmetic subset ($N = 458$). Input tokens are estimated at approximately four characters per token.

B Aggregation Strategy Results

Condition	Aggregation	EM	Judge (0–2)
Asym-verifier	Majority	0.2538	0.7022
	Weighted	0.2623	0.7259
	Debate judge	0.2843	0.7530
	Full-evidence	0.2843	0.7479
Asym-free	Majority	0.2237	0.5271
	Weighted	0.2237	0.5271
	Debate judge	0.2780	0.5881
	Full-evidence	0.2932	0.6288
Sym-verifier	Majority	0.2407	0.5678
	Weighted	0.2373	0.6085
	Debate judge	0.2780	0.6390
	Full-evidence	0.2729	0.6305
Sym-free	Majority	0.1932	0.3695
	Weighted	0.1915	0.3780
	Debate judge	0.2000	0.3847
	Full-evidence	0.2102	0.4220

Table 12: Aggregation strategy comparison on the full MultiHierrt evaluation set ($N = 591$).

Condition	GPT-4o-mini Judge	Claude Haiku Judge
Sym-Free	0.262	0.338
Asym-Free	0.392	0.554
Sym-Verifier	0.515	0.654
Asym-Verifier	0.585	0.746

Table 13: Cross-model judge validation on a random subsample ($n = 130$). Correctness scores from GPT-4o-mini (in-family) and Claude Haiku 4.5 (independent provider) judges preserve the same condition ordering. Exact agreement: 85.0%; Pearson $r = 0.886$.

B.1 Cross-Model Judge Validation

To rule out self-preference bias arising from the shared GPT-4o-mini/GPT-4o model family used by the agents, debate judge, and evaluation judge, we additionally scored a random subsample of 130 arithmetic-subset questions ($n = 130 \times 4$ conditions = 520 predictions) using Claude Haiku 4.5 as an independent, cross-provider judge. Claude and GPT-4o-mini correctness scores agree exactly on 85.0% of predictions (Pearson $r = 0.886$) and, more importantly, preserve the same condition ordering (Asym-verifier > Sym-verifier > Asym-free > Sym-free). As shown in Table 13, every condition improves significantly over Sym-free under the independent judge ($p < 0.02$, Holm-corrected), while pairwise differences among the three stronger conditions remain non-significant on this smaller subsample. Thus, the cross-model validation preserves the overall condition ordering and the separation of the three stronger conditions from Sym-free, although it does not reproduce all pairwise significance results reported in the full evaluation. This cross-model agreement (Table 13) therefore provides additional support for the condition-level differences reported in Table 4, while reducing, rather than eliminating, concerns about judge-agent model-family overlap.

C Primary v1 Prompt Templates

This appendix presents the prompt templates used in the primary v1 FinDRA experiments. The v1 setting assigns each agent an information region, a document-based role, and explicit natural-language authority guidance describing how strongly the agent should retain or reconsider its judgment during debate.

C.1 Shared System Prompt

Shared system prompt

You are a financial analyst in a multi-agent debate.
You have access ONLY to your assigned section.
For arithmetic questions, show your calculation steps clearly.
Return ONLY valid JSON, no markdown.

C.2 Agent Roles and Authority Guidance

A1 — Low authority

You are a financial narrative analyst with access ONLY to the narrative text sections (MD&A, business description, risk factors). Your authority level is LOW: You provide contextual background and qualitative signals. Financial numbers in other sections take precedence over your textual interpretations. If other agents present concrete numerical evidence that contradicts your reading, update your position—your role is to add context, not override hard data.

A2 — High authority

You are a financial statement analyst with access ONLY to the primary financial tables (income statement, balance sheet). Your authority level is HIGH: The figures in your tables are GAAP-certified objective facts. You are the authority on these numbers. Defend your numerical findings strongly—only revise if your own data clearly shows a calculation error, or if A3 presents cash flow figures that directly override your statement numbers. Do NOT yield to textual interpretations from A1 over hard numbers.

A3 — Medium authority

You are a cash flow and supplementary data analyst with access ONLY to the cash flow statement and supplementary financial tables. Your authority level is MEDIUM: Your cash flow data is objective and authoritative within your domain. Maintain your position when you have direct numerical evidence from your tables. However, remain open to A2’s primary statement figures if they provide the correct base data for your calculations. Update when mathematically justified.

C.3 Round 1: Independent Judgment

Initial reasoning prompt

```
Question: {question}
Your assigned section ({agent_id} -
authority: {authority_level}):
{content}

Answer based ONLY on your section. Return
JSON:
{ "agent_id": "{agent_id}",
  "authority_level": "{authority_level}",
  "relevant_data": ["key numbers or facts
from your section"],
  "calculation": "show steps if arithmetic",
  "answer": "your answer",
  "confidence": 0.0,
  "can_answer": true,
  "reasoning": "brief explanation" }
```

Debate prompt without verifier feedback

```
Question: {question}
Your assigned section ({agent_id} -
authority: {authority_level}):
{content}

Other agents' positions:
{other_positions}

(No verifier feedback available—reason
based on your own section and other agents'
arguments.)

Given your authority level
({authority_level}), update your position.
Return JSON:
{ "agent_id": "{agent_id}",
  "authority_level": "{authority_level}",
  "position_change": "changed | unchanged",
  "change_reason": "why you changed or
defended your position",
  "relevant_data": ["key data points"],
  "calculation": "updated calculation if
needed",
  "answer": "updated answer",
  "confidence": 0.0,
  "reasoning": "updated reasoning" }
```

C.4 Rounds 2–3: Debate

Debate prompt with verifier feedback

```
Question: {question}
Your assigned section ({agent_id} -
authority: {authority_level}):
{content}

Other agents' positions:
{other_positions}

Verifier feedback on your previous answer:
{verifier_feedback}

Given your authority level
({authority_level}), update your position.
- HIGH authority: only change if
mathematically forced by data
- MEDIUM authority: change if evidence
clearly justifies it
- LOW authority: change if other agents
provide stronger numerical evidence

Return JSON:
{ "agent_id": "{agent_id}",
  "authority_level": "{authority_level}",
  "position_change": "changed | unchanged",
  "change_reason": "why you changed or
defended your position",
  "relevant_data": ["key data points"],
  "calculation": "updated calculation if
needed",
  "answer": "updated answer",
  "confidence": 0.0,
  "reasoning": "updated reasoning" }
```

Verifier prompt

You are an impartial financial verifier with access to ALL sections of the report. Evaluate each agent's answer for factual accuracy and calculation correctness.

```
Question: {question}
Agent positions:
{agent_positions}
Full report content:
{all_content}

Evaluate each agent. Return JSON
containing, for each agent: agent_id,
factual_support (supported | weak
| unsupported), calculation_correct
(boolean), key_evidence, and score.
Additionally return best_agent and a
synthesis describing what the complete
report suggests the answer is. Although
the implementation includes an optional
gold_answer placeholder, it was left empty
in all reported experiments; therefore,
the verifier did not receive the gold
answer.
```

C.5 Verifier Prompt

The verifier evaluates the agent responses after each of the three debate rounds. Agent-specific assessments produced after Rounds 1 and 2 are provided to the corresponding agents in the subsequent round, while the Round 3 evaluation is retained for final assessment and aggregation. The verifier's global best_agent prediction and synthesis are not exposed to the reasoning agents.

C.6 Aggregation Prompts

Debate judge

You are an expert financial analyst acting as a judge. Given the final positions produced by three agents, provide the best possible answer to the question.

Question: {question}

Final agent positions:

{final_positions}

Latest verifier assessment, when available:

{verifier}

Based on the debate, provide the best answer.

Return JSON with fields: answer, confidence, key_evidence, reasoning, abstain.

Full-evidence judge

You are an expert financial analyst acting as a judge. Given the final positions produced by three agents and the complete original document, provide the best possible answer to the question.

Question: {question}

Final agent positions:

{final_positions}

Latest verifier assessment, when available:

{verifier}

Full document excerpts:

{all_docs}

Based on the debate and complete document, provide the best answer.

Return JSON with fields: answer, confidence, key_evidence, reasoning, abstain.

v2 authority block

Your document authority score: {authority_score} / 1.0

This score reflects the epistemic reliability of your document section. Higher scores indicate a stronger evidentiary basis for your claims.

D.2 v4 Contextual Numerical Authority Prompt

In v4, the same numerical score is accompanied by an explicit instruction to consider authority and evidence quality during answer revision.

v4 authority block

Your document authority score: {authority_score} / 1.0

When deciding whether to retain or revise your current answer, consider both the relative authority scores of the agents and the quality of the supporting evidence.

D Alternative Numerical Authority Encodings

We additionally evaluate v2 and v4 as alternative numerical authority encodings. The four evaluated configurations are uniform, domain hierarchy, high gap, and reversed. The authority scores for (A1, A2, A3) are set to (0.5, 0.5, 0.5) for uniform, (0.3, 0.8, 0.6) for domain hierarchy, (0.1, 0.9, 0.5) for high gap, and (0.8, 0.3, 0.5) for reversed.

D.1 v2 Numerical Authority Prompt

In v2, authority is conveyed through a numerical score without explicit authority-dependent behavioral instructions.