CLAIR-Fin: An Adversarial Multi-Agent Framework for Claim-Level Verification and Adaptive Debate in Cross-Modal Financial QA

Fatema Tuj Johora Faria[1], Mukaffi Bin Moin[1], Jubayer Al Mahmud[2], M. F. Mridha[3], Md. Alam Hossain[2]

[1]Ahsanullah University of Science and Technology, Bangladesh
[2]Jashore University of Science and Technology, Bangladesh
[3]American International University - Bangladesh

Correspondence: mukaffi28@gmail.com, fatema.faria142@gmail.com

Abstract

Existing defenses against hallucination in retrieval-augmented and multi-agent pipelines remain partial: evidence is trusted despite modality disagreement, debate verifies an aggregate report rather than individual claims, and such verification occurs only after drafting, leaving inter-agent errors undetected until the final text. To close this gap, we present **CLAIR-Fin**, a nine-agent framework that decomposes each question into atomic claims maintained in a typed *Financial Claim Ledger*. Each claim is resolved through *Asymmetric Evidence Authority*, which conditions evidence trust on claim type rather than treating all modalities as equally reliable; *Chain-of-Custody Verification*, which checks grounding at the hand-off between drafting and adversarial review rather than only at the pipeline's exit; an *Adaptive Rebuttal Cycle*, which routes contested claims through adversarial debate whose depth scales with what that debate finds; and a terminal entailment audit paired with a continuous *Hallucination Risk Index* that distinguishes claims that passed scrutiny from claims never contested. We evaluate **CLAIR-Fin** on **BB-FinQA-X**, a 500-question cross-modal financial evaluation set built from Bangladesh Bank Annual Report material, stratified by query type, format, and difficulty. Relative to a single-pass retrieval-augmented generation baseline, it raises faithfulness ($0.780 \rightarrow 0.889$) while abstaining on 5.4% of questions when evidence is insufficient rather than forcing an unsupported response, and it exceeds stronger retrieval-strategy baselines such as HyDE and Graph-RAG on faithfulness (≤ 0.874).

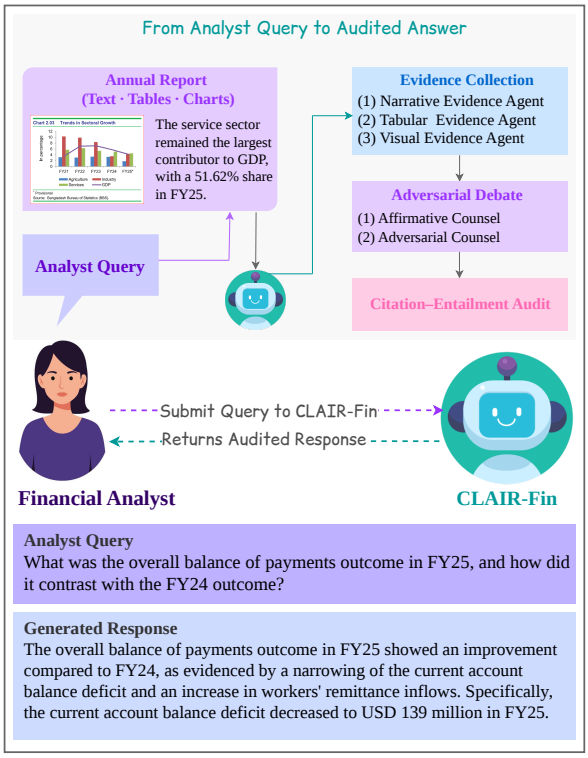

Figure 1: An analyst's question enters **CLAIR-Fin**, where specialized agents extract evidence for each claim. Each claim undergoes *Evidence* $\rightarrow$ *Debate* $\rightarrow$ *Audit*, with contested claims adversarially reviewed and citation entailment verified before synthesis into the final response.

1 Introduction

Financial institutions increasingly rely on large language models to answer questions over long, multimodal reports, where the same fact may appear as prose, a table, and a chart within one document exceeding several hundred pages. A misread fiscal-year label or an approximated chart value reported as exact can materially change a conclusion; even state-of-the-art vision-language models hallucinate on chart-reading tasks (Wang et al., 2025). Cross-modal benchmarks building on earlier tabular-textual QA work (Chen et al., 2021; Zhu et al., 2021), including FinanceBench (Islam et al., 2023), FAMMA (Xue et al., 2025), and XFin-Bench (Zhang et al., 2025), find current LLMs fail many realistic financial questions, worsening as context grows and evidence sits away from a document's start (Ji et al., 2026), precisely the regime

long reports fall into. Such systems must retrieve relevant content, reconcile conflicting cross-modal evidence, and answer faithfully or decline when evidence is insufficient.

Several lines of prior work address pieces of this problem, but each leaves a gap. Multi-agent frameworks such as MDocAgent (Han et al., 2025) coordinate specialized agents over long documents, and multimodal retrieval frameworks such as MultiFinRAG (Gondhalekar et al., 2025) and FinRAGBench-V (Zhao et al., 2025) jointly index tables, figures, and text, yet none conditions evidence trust on claim type. Financial debate frameworks such as FinDebate (Cai et al., 2025) and Structured Adversarial Synthesis (Sadhu et al., 2025) apply adversarial roles to report-level analysis (Du et al., 2024; Chan et al., 2023), but run a fixed round count over a holistic thesis regardless of how contested a claim is. Verification approaches such as Chain-of-Verification (Dhuliawala et al., 2023), FinGround (Guo et al., 2026), and FRED (Tan et al., 2025) instead check atomic claims, and self-consistency sampling offers a claim-agnostic alternative (Manakul et al., 2023); both act only after drafting and pair no binary outcome with a continuous risk estimate (Section 2.3). Benchmarks such as FinBen (Xie et al., 2024) also skip non-calendar fiscal years and the dense statistical tables typical of South Asian central-bank reporting. Across this literature, evidence authority is uniform, verification is late if applied at all, and abstention is rarely measurable (Wen et al., 2025; Kirichenko et al., 2025).

We introduce **CLAIR-Fin** (**C**laim-**L**edger **A**dversarial **I**nference and **R**etrieval for **Fin**ancial document understanding), a nine-agent framework for reliable question answering over multimodal financial documents, assessed on central-bank annual reports (Appendix A), whose fiscal-year definitions, currency denominations, and reporting practices differ from the corporate filings predominantly represented in prior evaluation suites (Section 2). A Planner-Orchestrator decomposes each question into a ***F**inancial **C**laim **L**edger* (**FCL**) containing atomic, typed assertions, which are populated by three evidence agents and reconciled by a Ledger Guardian through ***A**symmetric **E**vidence **A**uthority* (**AEA**), a claim-type-aware weighting scheme that prioritizes tabular evidence for precise numerical values while favoring narrative text for causal attribution. Before adversarial review, ***C**hain-of-**C**ustody **V**erification* (**CoCV**) validates and restores evidence grounding, preventing attribution drift from propagating into subsequent reasoning. Disputed claims are then escalated to an Affirmative and an Adversarial Counsel through an ***A**daptive **R**ebuttal **C**ycle* (**ARC**), with deliberation depth expanding according to unresolved findings rather than following a predetermined budget. A Judge-Auditor subsequently enforces a terminal entailment gate and records each AEA determination alongside its equal-weight counterfactual; a continuous ***H**allucination **R**isk **I**ndex* (**HRI**) differentiates assertions that endured substantive scrutiny from those that were never challenged. Finally, a Brief Synthesizer formulates the response and abstains whenever the available evidence cannot adequately support an answer.

We structure our evaluation of this framework around the following research questions:

- **RQ1.** To what extent does modality-aware evidence prioritization improve faithfulness and correctness over modality-agnostic retrieval?
- **RQ2.** How reliably does hand-off-level verification reduce the propagation of unsupported claims?
- **RQ3.** Does routing adversarial debate to contested claims improve factual grounding over skipping debate entirely?
- **RQ4.** Does continuous risk estimation provide a more informative reliability signal than binary verification alone?
- **RQ5.** How does the framework perform across single-modal and multimodal presentation formats, and which configurations remain most challenging?

2 Related Work

2.1 Multimodal Financial Document Understanding and Retrieval-Augmented Generation

Extending retrieval-augmented generation to multimodal financial documents is closest to our setting. General-purpose multimodal agents coordinate text and image agents over long documents (Han et al., 2025), while financially specialized retrieval frameworks batch table and figure images through a lightweight multimodal model, escalating to text+table+image context only when needed (Gondhalekar et al., 2025), within a RAG paradigm

(Gupta et al., 2024) assessed via a reference-free protocol scoring faithfulness, relevancy, and context precision/recall (Es et al., 2024). This improves *what* evidence is retrieved but treats retrieval as complete once cited: a table cell and an approximately-read chart value for the same quantity are cited as interchangeable, with no mechanism to adjudicate which to trust when they disagree, a gap faithfulness scores compound since they measure entailment by *some* evidence, not the *correct* one.

2.2 Multi-Agent Debate and Financial Multi-Agent Systems

A second line of work uses multi-agent debate, LLM instances critiquing and revising outputs, to improve factuality and reasoning (Du et al., 2024). Financial applications add domain-specific structure: specialist-role frameworks assign earnings, market, sentiment, valuation, and risk personas with a trust/skeptic/leader safety layer (Cai et al., 2025), dialectical frameworks stage bull/bear/devil's-advocate roles over earnings-call transcripts (Sadhu et al., 2025), and recent work asks whether added coordination improves outcomes relative to cost (Nguyen and Pham, 2026). The common assumption is that debate is a property of the *report*, not the *claim*: disagreement goes unflagged and debate depth stays fixed regardless of contestedness, leaving claim-level, difficulty-adaptive verification unaddressed.

2.3 Claim-Level Verification, Extraction Reliability, and Faithfulness Evaluation

Closer to claim-level verification, a distinct literature verifies individual statements rather than whole reports. Surveys characterize the field's dominant pipeline, retrieve, decompose, check entailment, applied once to an already-finished claim (Dmonte et al., 2025); graph-structured verification converts a claim into an entity-relationship graph, checking each triplet before a verdict (Jeon and Lee, 2025), closer in spirit to our claim ledger but applied to open-domain claims rather than cross-modal evidence. Self-consistency methods instead sample multiple outputs and aggregate via majority agreement, on the premise that reproducibility signals reliability (Wang et al., 2023), though open-book QA depends on trustworthy extraction (Islam et al., 2023) and LLM-generated summaries frequently omit source figures despite high surface scores (Yang et al., 2024). Verification thus checks a claim only once formed, and reproducibility does not guarantee correctness.

2.4 Research Gap and Positioning of CLAIR-Fin

Across these three directions, each addresses one piece of the problem, evidence arbitration, adaptive debate, or hand-off verification, in isolation, and none operates jointly. **CLAIR-Fin** closes this gap: it conditions evidence trust on claim type so cross-modal disagreement is resolved by a stated, auditable prior; routes claims to adversarial debate only when evidence coverage is insufficient, scaling depth to scrutiny rather than a fixed budget; and verifies grounding at the hand-off between drafting and adversarial review, treating even self-consistent evidence as suspect.

3 CLAIR-Fin Framework

3.1 Problem Definition

We address faithful QA over long, multimodal financial documents, where evidence spans prose, tables, and charts.

Task. Given a question q over corpus $\mathcal{D}$, the framework produces answer a with citations, or abstains when evidence is insufficient.

Claim decomposition. q is decomposed into ordered atomic claims $C = \{c_1, \ldots, c_n\}$, $n \leq 8$. Each $c_i = (\text{id}_i, \text{text}_i, \tau_i)$ has a claim type

$$\tau_i \in \mathcal{T} = \{\texttt{FACT\_NUMERIC}, \texttt{FACT\_TREND}, \texttt{CAUSE\_ATTRIBUTION}, \texttt{RATIO\_IDENTITY}\} \quad (1)$$

fixing how its evidence is weighted (Section 3.4); claims resolve sequentially and independently before synthesis.

Evidence. Evidence spans four modalities,

$$\mathcal{M} = \{\texttt{text}, \texttt{table}, \texttt{chart}, \texttt{tool\_derived}\}, \quad (2)$$

where `tool_derived` is a deterministically computed quantity (e.g., year-over-year growth) rather than directly read. Each item $e = (m, \text{src}, \text{page}, \text{content}, \text{conf})$ carries modality $m \in \mathcal{M}$, provenance, and confidence $\text{conf} \in [0, 1]$.

Financial Claim Ledger (FCL). Claims and evidence accumulate in a typed, directed multigraph $G = (V, E)$ with six node types (claim, text span, table cell, derived metric, chart region, constraint check) and edges $r \in \{\texttt{SUPPORTS}, \texttt{SAME\_AS}, \texttt{VIOLATES\_CONSTRAINT}\}$. G persists for the question's lifetime as the sole verification, audit, and citation artifact. We write

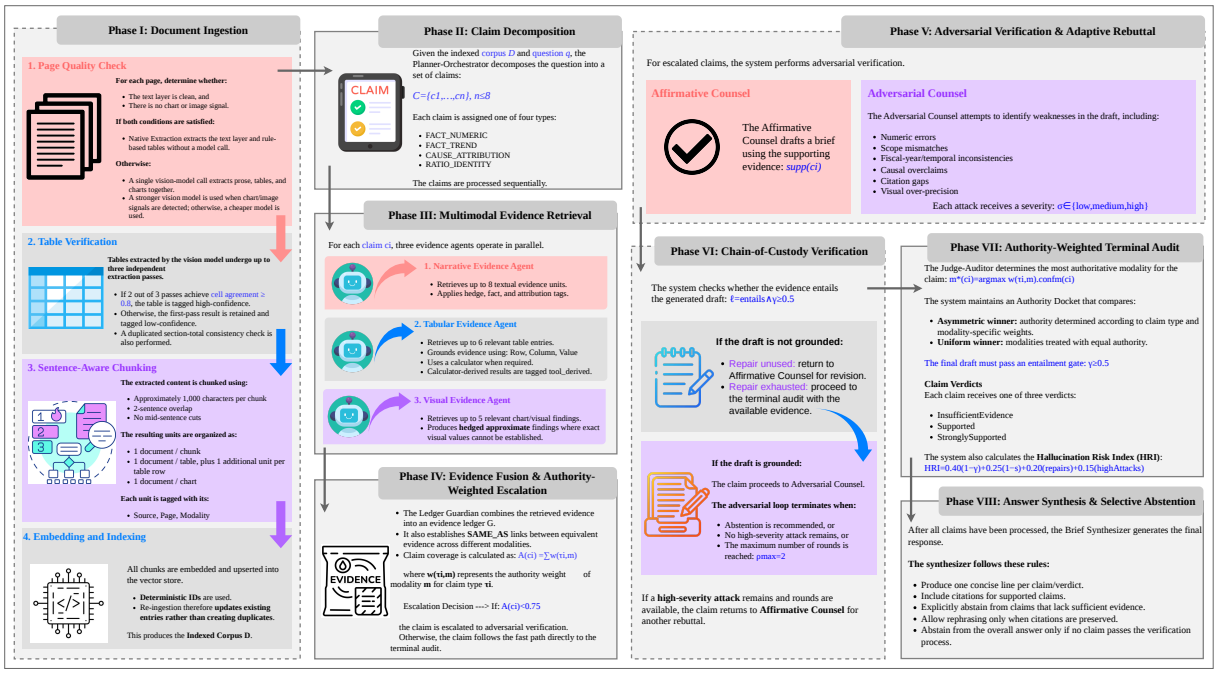

Figure 2: End-to-end **CLAIR-Fin** framework (Phases I–VIII, Section 3). A question is decomposed into atomic claims (Phases I–II), for which modality-specific agents retrieve evidence (Phase III) that is fused and authority-weighted to decide fast-path or escalation (Phase IV). Escalated claims pass through adversarial debate and a hand-off grounding check (Phases V–VI) before a terminal, authority-weighted audit (Phase VII). Once every claim is resolved, the framework synthesizes a cited answer or abstains (Phase VIII).

score(u, v) for a SUPPORTS edge's confidence and supp$(c) = \{v : (v, c, \text{SUPPORTS}) \in E\}$ for c's supporting evidence.

3.2 Phase I: Document Ingestion

Every source PDF is processed once, offline, into corpus $\mathcal{D}$, which Phase III (Section 3.4) retrieves against.

Each page is handled independently: native extraction (text layer plus rule-based table detection) is trusted only when the text layer is not garbled and the page has no chart/image signal; otherwise a structured vision-model call extracts prose, tables, and charts from a rendered image, a stronger model when a chart or image is present, a cheaper one otherwise.

Table extraction is least reliable: repeated vision calls at temperature 0 can disagree. Each table is extracted up to three times and checked for pairwise agreement (Equation (4)); two-of-three agreement is tagged high-confidence, else the first extraction is kept, tagged low-confidence, and retained in the corpus for provenance, though its cells are excluded rather than downweighted when the **Tabular Evidence Agent** retrieves evidence (Section 3.4), so the table is kept as a record without ever being usable evidence. A second heuristic flags a table low-confidence if a row duplicates its own section-total row.

Prose is split into sentence-aware chunks (1,000-character target, two-sentence overlap, sentence-boundary splits only). Each table yields one whole-table document plus one per labeled row; each chart yields one document per page. Documents are tagged with source, page, and modality, embedded, and upserted under a deterministic id, forming corpus $\mathcal{D}$.

3.3 Phase II: Claim Decomposition

Claim decomposition converts q into atomic, independently verifiable claims: verifying a whole answer as one unit lets an unsupported figure hide inside a well-supported narrative, whereas decomposing first lets every stage check grounding exactly.

The **Planner-Orchestrator** retrieves a top-12, modality-agnostic preview (300 characters each) so claim text reuses the document's own terminology, then a structured LLM call (temperature 0) proposes 1–8 claims with text and type $\tau_i \in \mathcal{T}$; on failure it falls back to one claim restating q, typed CAUSE_ATTRIBUTION, chosen because it commits to no specific number or trend that could later be flatly contradicted; as Section 3.5 discusses, this

typing also means the fallback claim is the one most likely to fast-path past debate rather than face additional scrutiny, so "default" should not be read as "most rigorously checked." Claims have no dependency structure, so each one's evidence, debate, and audit are fully independent, enabling per-claim auditability (Section 3.8).

3.4 Phase III: Multimodal Evidence Retrieval

For claim c_i, retrieval draws modality-specific evidence: a table figure, its prose interpretation, and an approximate chart restatement may all differ, so a dedicated agent per modality grounds evidence before it reaches the ledger.

Given (text_i, τ_i) and $\mathcal{D}$, retrieval blends dense and lexical signal: a candidate pool of $4k$ (four times the requested k) from vector search is reranked by

$$\text{score}(x, d) = 0.65 \frac{1}{1 + \text{dist}_{L2}(x, d)} + 0.35 \times \frac{|\text{terms}(x) \cap \text{terms}(d)|}{|\text{terms}(x)|}. \quad (3)$$

and the top k returned. **Narrative** retrieves k=8 text passages, tagged `fact/hedge/attribution`, extracting any (metric, period, value, unit) quadruple as `tool_derived`. **Tabular** retrieves k=6 table passages, grounding (row, column, value, unit) cells; cells failing consensus are dropped, not down-weighted. When two grounded cells share a row across columns, a deterministic calculator derives $\Delta_{\%} = (v_{\text{cur}} - v_{\text{prev}})/v_{\text{prev}}$ or $\Delta_{\text{pp}} = v_{\text{cur}} - v_{\text{prev}}$ as `tool_derived`. **Visual** retrieves k=5 chart passages, one finding per chart, kept as hedged prose so authority weighting governs its influence.

Table extraction agreement (Section 3.2) uses:

$$\text{agree}(T_a, T_b) \iff \text{shape}(T_a) = \text{shape}(T_b) \wedge \frac{|\{(i,j) : T_a[i,j] = T_b[i,j]\}|}{|T_a|} \geq 0.8. \quad (4)$$

This mirrors the Phase I agreement rule (Section 3.2), since agreement measures reproducibility, not correctness. The phase passes per-modality evidence lists $\{e^{\text{text}}\}, \{e^{\text{table}}\}, \{e^{\text{chart}}\}$ to the **Ledger Guardian**, each carrying only a $(modality, confidence)$ pair, reusable unmodified by later stages.

3.5 Phase IV: Evidence Fusion and Authority-Weighted Escalation

Evidence fusion merges the three lists into G, links same-metric evidence across modalities, and decides whether debate is needed: strong, agreeing evidence skips debate; weak or conflicting evidence does not, and "enough" evidence differs by claim type (a numeric claim needs a table cell, an attribution claim needs prose).

Each item becomes a typed node linked to the claim via `SUPPORTS` (score = retrieval confidence); the **Ledger Guardian** adds `SAME_AS` edges when a table cell's row label ($\geq$4 characters) appears in a chart/text description, a lexical heuristic (Section 7). For `RATIO_IDENTITY` claims, a constraint node records `VIOLATES_CONSTRAINT` if uncomputed.

AEA first acts here, in *coverage* form: each type τ has fixed weights $w(\tau, m)$ over modalities (Table 1, sums to 1 per row), giving coverage

$$A(c_i) = \sum_{m \in M_i} w(\tau_i, m), \quad (5)$$

where $M_i \subseteq \mathcal{M}$ is the set of modalities for which c_i has at least one grounded supporting evidence item in G (i.e., $M_i = \{m : \exists v \in \text{supp}(c_i), \text{mod}(v) = m\}$). The claim escalates iff $A(c_i) < 0.75$; else it proceeds to audit (Section 3.8). Three low-authority chart mentions for `FACT_NUMERIC` contribute less than one table cell. One consequence of this rule is worth flagging explicitly: since $w(\texttt{CAUSE\_ATTRIBUTION}, \texttt{text}) = 0.75$ (Table 1), a causal claim supported by a single grounded prose passage reaches $A(c_i) = 0.75$ exactly, which fails the strict escalation inequality and fast-paths directly to audit without debate. This is deliberate: prose is the sole authoritative source for causal attribution here, so a single well-grounded passage is treated as sufficient coverage rather than automatically contested, and the terminal audit (Section 3.8) still independently checks the drafted sentence against that same evidence. It does mean causal overclaim, one of six adversarial attack categories (Section 3.6), is structurally the attack type least likely to reach the Adversarial Counsel, a limitation of the coverage rule (Section 7), not evidence causal attributions are adequately scrutinized.

Table 1: Asymmetric Evidence Authority weights $w(\tau, m)$, assigning trust to each evidence modality per claim type; weights sum to 1 within each row and are a fixed design prior rather than a learned or human-validated distribution (Section 7).

τ	table	tool_derived	text	chart
FACT_NUMERIC	0.55	0.30	0.10	0.05
FACT_TREND	0.45	–	0.10	0.45
CAUSE_ATTRIBUTION	0.25	–	0.75	–
RATIO_IDENTITY	0.40	0.60	–	–

The phase leaves an updated G, escalate(c_i), and $A(c_i)$ in shared state, conditioning modality trust on claim type rather than a global weight: an approximate chart reading and an audited table cell are interchangeable evidence for a causal attribution but not for an exact figure.

3.6 Phase V: Adversarial Verification and Adaptive Rebuttal

For escalated claims, the **Affirmative Counsel** and **Adversarial Counsel** draft and stress-test a claim before the audit gate, since a single-pass draft-then-check misses flaws only visible under scrutiny (e.g., a conflated fiscal-year label), while fixed-round debate wastes budget on easy claims.

Given supp(c_i) and prior findings, the **Affirmative Counsel** drafts (temperature 0.2); the **Adversarial Counsel** returns a scorecard of zero or more attacks across six types (numeric, scope, fiscal-year/temporal, causal overclaim, citation gap, visual over-precision) with severity $\sigma \in \{\texttt{low}, \texttt{medium}, \texttt{high}\}$, plus an abstain flag. This is ARC: rebuttal triggers, incrementing ρ, iff σ=`high` exists and $\rho < \rho_{\max}$=2; otherwise the claim proceeds to audit (including on an abstain recommendation), so debate depth tracks scrutiny, not a fixed quota.

The phase yields a (possibly revised) brief, adversarial findings, and an abstain flag; cost scales with claim difficulty, not claim count, operationalizing claim-level debate that fixed-round methods (Section 2.2) do not.

3.7 Phase VI: Chain-of-Custody Verification

CoCV checks that the affirmative draft stays grounded at the hand-off to adversarial review, not only at the pipeline's end, catching drift before the **Adversarial Counsel** reasons over ungrounded content.

Given the draft and supp(c_i), CoCV reuses the **Judge-Auditor**'s entailment call (Section 3.8) at this earlier hand-off: an LLM-as-judge call returns $\ell \in \{\texttt{entails}, \texttt{neutral}, \texttt{contradicts}\}$ and confidence γ, with grounding intact iff

$$\ell = \texttt{entails} \ \wedge \ \gamma \geq 0.5. \tag{6}$$

On failure, one bounded repair is attempted; if still ungrounded, custody is marked broken and the claim routes directly to audit as `InsufficientEvidence`, otherwise it proceeds to review grounded. Every check is logged. The internal hand-off placement, plus the bounded single repair, guarantees fixed, predictable cost rather than an open-ended correction loop.

3.8 Phase VII: Authority-Weighted Terminal Audit

The terminal audit (**Judge-Auditor**) reaches a final, citable verdict, resolving cross-modal disagreement via claim-type-conditioned authority rather than the drafting model's own judgment.

If custody was broken or no evidence remains, the claim resolves to `InsufficientEvidence` (confidence 0). Otherwise, per-modality confidence is $\text{conf}_m(c_i) = \max_{v \in \text{supp}(c_i),\, \text{mod}(v)=m} \text{score}(v, c_i)$, and the winning modality follows Section 3.5's weighting, now as argmax:

$$\begin{aligned} m^*(c_i) &= \arg\max_{m \in \mathcal{M}} \left[w(\tau_i, m) \cdot \text{conf}_m(c_i) \right], \\ s(c_i) &= w(\tau_i, m^*) \cdot \text{conf}_{m^*}(c_i). \end{aligned} \tag{7}$$

An equal-weight counterfactual ($w_{\text{unif}}(\tau_i, m) = 1/|M_i|$) is also scored, yielding $\hat{m}(c_i)$; both are logged to an Authority Docket recording whether the asymmetric prior changed the winning modality. Evidence is presented to drafting ordered by descending authority. A grounded sentence is drafted (temperature 0) and passed through the entailment gate (Equation (6)); failure resolves to `InsufficientEvidence`, and passing claims receive `StronglySupported` if $s(c_i) \geq 0.5$, else `Supported`.

For passing claims, HRI is:

$$\begin{aligned} \text{HRI}(c_i) = {} & 0.40(1-\gamma) + 0.25(1 - s(c_i)) \\ & + 0.20 \frac{\rho_{\text{repair}}}{\rho^{\max}_{\text{repair}}} + 0.15 \min\Big(\frac{n_{\text{high}}}{3}, 1\Big) \end{aligned} \tag{8}$$

clipped to $[0, 1]$, where γ is entailment confidence, $\rho_{\text{repair}} \in \{0, 1\}$ is the number of custody repairs ($\rho^{\max}_{\text{repair}} = 1$: CoCV attempts at most one repair per claim, Section 3.7), and n_{high} is the count of high-severity adversarial findings; weights (0.40/0.25/0.20/0.15) are a fixed design choice, like the AEA table (Section 7). HRI distinguishes claims that survived genuine scrutiny from those fast-pathed and never contested.

3.9 Phase VIII: Answer Synthesis and Selective Abstention

The **Brief Synthesizer** composes one final answer from resolved claims, citing evidence and abstaining on failed claims; naive concatenation risks burying an abstention or dropping citations.

Each claim contributes one line: its answer with citations if it passed audit, or an insufficient-grounding sentence if not. One LLM pass (temperature 0.2) rephrases and reorders these lines without adding content, falling back to raw concatenation on a dropped citation label. The answer is abstained iff *no* claim passed audit; one supported claim among several unsupported ones still yields a substantive, partial answer. This leaves the final answer, citations, an abstention flag, and the complete ledger G, persisted for analysis, composed only from already-verified fragments so nothing upstream can be undone by an unverified step.

4 Dataset Construction

We construct **BB-FinQA-X**, a 500-question multimodal financial QA dataset grounded in the Bangladesh Bank Annual Report. Detailed dataset construction, annotation, validation, and distribution statistics are provided in Appendix A.

5 Experimental Configuration

Detailed implementation settings, experimental configurations, and evaluation protocols are provided in Appendix B.

6 Results and Discussion

We evaluate **CLAIR-Fin** on **BB-FinQA-X** across three axes: automatic retrieval and generation metrics by query type and format (Tables 8, 10, 11), framework-specific metrics (Table 7), and two-annotator human evaluation (Table 9), alongside ablations against the full system and a single-pass RAG baseline (Tables 2, 3); detailed analysis is in Appendix D.

6.1 RQ1: Modality-Aware Evidence Prioritization

Conditioning evidence trust on claim type is the most conservative of **CLAIR-Fin**'s four mechanisms, and the ablation confirms this. Removing AEA produces the smallest degradation (↓) of any ablated mechanism, yet consistently: faithfulness $0.889 \rightarrow 0.883$, context recall $0.897 \rightarrow 0.893$, exact correctness $0.592 \rightarrow 0.585$ (Tables 2–3).[1] The Authority Docket reports an AEA impact rate of 0.515 (Table 7): the asymmetric prior changes the winning modality in roughly half of contested decisions, a substantial share, not a handful of edge cases.

[1]Throughout this section, $X \rightarrow Y$ denotes a metric's change from the full system's score X to the ablated (or contrasted) score Y; ↑ and ↓ mark whether the change is an improvement or a degradation.

Table 2: RAGAS retrieval and generation metrics for **CLAIR-Fin**, four single-mechanism ablations, and four retrieval-strategy baselines on **BB-FinQA-X** ($n = 500$). **Ans. Rel.**: answer relevancy. **Ctx. Prec./Ctx. Recall**: context precision/recall. Each ablated row disables exactly one mechanism (Appendix A.1), holding the rest of the system fixed; **w/o Term. Audit** removes the terminal entailment audit; **Vanilla RAG** is a single-pass retrieve-then-generate baseline with none of the four mechanisms.

Configuration	Faith. ↑	Ans. Rel. ↑	Ctx. Prec. ↑	Ctx. Recall ↑
CLAIR-Fin	**0.889**	**0.696**	**0.816**	**0.897**
w/o Term. Audit	0.845	0.687	0.803	0.886
w/o ARC	0.770	0.680	0.781	0.862
w/o AEA	0.883	0.692	0.812	0.893
w/o CoCV	0.857	0.689	0.807	0.881
Vanilla RAG	0.780	0.680	0.700	0.840
HyDE RAG	0.874	0.691	0.801	0.885
Hierarchical RAG	0.865	0.688	0.752	0.831
Graph-RAG	0.832	0.694	0.729	0.889

Table 3: **CLAIR-Fin**-specific metrics under the same configurations as Table 2 ($n = 500$). **Faith. Rate**: share of published claims passing citation-entailment verification. **Exact Corr.**: exact correct answer rate. **Cov.**: answer coverage, the share of questions receiving a non-abstained answer. **Deb. Util.**: debate utilization rate. **AEA Imp.**: AEA impact rate. A dash (–) marks a metric undefined for that configuration.

Configuration	Faith. Rate ↑	Exact Corr. ↑	Cov. ↑	Deb. Util.	AEA Imp.
CLAIR-Fin	**0.783**	**0.592**	**0.946**	**0.646**	**0.515**
w/o Term. Audit	0.741	0.561	0.935	0.639	0.509
w/o ARC	0.682	0.524	0.896	–	0.501
w/o AEA	0.776	0.585	0.940	0.644	–
w/o CoCV	0.753	0.548	0.922	0.641	0.508

Note: Faithfulness (Table 2) is RAGAS's semantic faithfulness score; Faithfulness Rate (here) is the share of published claims that pass citation-entailment verification.

6.2 RQ2: Verification at the Drafting-to-Review Hand-off

Verifying claim grounding at the drafting-to-review hand-off, alongside a final pre-publication audit, catches unsupported claims beyond what either check alone would, and the two checks carry unequal weight. Removing the terminal entailment audit drops (↓) faithfulness further ($0.889 \rightarrow 0.845$) than removing CoCV ($0.889 \rightarrow 0.857$), faithfulness rate showing the same ordering ($0.783 \rightarrow 0.741$ vs. $0.783 \rightarrow 0.753$; Tables 2–3). The terminal gate thus carries more of the faithfulness guarantee than any single upstream check, yet CoCV's non-trivial residual cost shows hand-off checking still contributes independently rather than being redundant: the two checks are complemen-

tary, not substitutable.

6.3 RQ3: Adaptive Allocation of Adversarial Debate

Routing contested claims through adaptive adversarial debate, rather than skipping it entirely, is where the framework's gains concentrate most heavily. Removing ARC causes the largest degradation (↓↓) of any ablated mechanism: faithfulness $0.889 \rightarrow 0.770$, exact correctness $0.592 \rightarrow 0.524$, answer coverage $0.946 \rightarrow 0.896$ (Tables 2–3), with a debate utilization rate of 0.646 (Table 7): nearly two-thirds of claims are routed through it. Since debate is reserved for claims below the fast-path coverage threshold, removing it eliminates the sole verification opportunity for exactly the claims most likely to be wrong. Adaptive debate is thus the single most consequential mechanism evaluated, precisely because it is targeted rather than indiscriminate.

6.4 RQ4: Continuous Risk Estimation versus Binary Gating

A continuous risk score is only worth reporting alongside a binary pass/fail outcome if it carries information the gate does not already capture, and HRI clears that bar. HRI correlates negatively with correctness ($r = -0.072$, Table 7), the theoretically expected direction, while human-rated abstention appropriateness (4.06/3.95, $\kappa = 0.84$, Table 9) independently corroborates that risk-sensitive abstention aligns with human judgment. The modest correlation magnitude is consistent with HRI adding information at the margin rather than duplicating the binary gate, a directionally correct, non-redundant, human-corroborated signal, though not yet a formally calibrated probability.

6.5 RQ5: Sensitivity to Evidence Presentation Format

Performance is not uniform across presentation formats, and the gap between easiest and hardest points to where cross-modal understanding still struggles. Text + Table achieves the highest faithfulness (0.915, ▲) and Chart Only the lowest (0.850, ▼; Table 10), while Evidence Retrieval (0.839) and Multi-hop Reasoning (0.840) are the lowest-scoring query types, essentially tied (Table 11). Matched-pair comparisons confirm the effect is attributable to evidence format itself: Table Only $\succ$ Text Only by $+0.030$ and Text+Chart $\succ$ Chart Only by $+0.025$ despite identical content. Difficulty is thus concentrated in chart-dependent evidence and in query types demanding evidence synthesis or grounding, while every combined format outperforms its weakest constituent modality, indicating the framework's cross-modal fusion adds real value.

7 Conclusion

Faithful question answering over long financial documents requires reconciling evidence across text, tables, and charts that do not always agree, a gap that prior multimodal retrieval, multi-agent debate, and claim-level verification methods leave unresolved. **CLAIR-Fin** closes this gap through a nine-agent framework built around a typed Financial Claim Ledger, in which (1) evidence trust is conditioned on claim type rather than treated uniformly (***A**symmetric **E**vidence **A**uthority*, **AEA**); (2) grounding is checked at the hand-off between drafting and adversarial review rather than only at the pipeline's exit (***C**hain-**o**f-**C**ustody **V**erification*, **CoCV**); (3) debate is allocated adaptively to contested claims rather than run for every claim regardless of difficulty (the ***A**daptive **R**ebuttal **C**ycle*, **ARC**); and (4) a terminal audit is paired with a continuous *Hallucination Risk Index* rather than a binary verdict alone. Empirically, on **BB-FinQA-X**, **CLAIR-Fin** exceeds **Vanilla RAG** on faithfulness ($0.780 \rightarrow 0.889$; Table 2), and ablation confirms all four mechanisms are non-redundant, with removing **ARC** producing the largest drop ($0.889 \rightarrow 0.770$); these results suggest claim-type-conditioned evidence weighting and hand-off-level verification are properties other multi-agent systems could adopt. Future work will replace the substring-matching heuristic linking cross-modal evidence with a semantic matcher; test generalization across institutions, languages, and models; and extend evaluation to multi-turn settings reflecting financial analyst use.

Limitations

CLAIR-Fin is designed for a specific problem setting, faithful, citation-grounded question answering over long, multimodal financial documents where narrative text, tables, and charts must be reconciled under strict correctness constraints. The following limitations define the scope of our claims rather than qualify the contributions above.

Dataset Scope. **BB-FinQA-X** is constructed from the Bangladesh Bank source report and cov-

ers a single institution, a single language (English), and a single central-bank reporting convention. The 500 questions are stratified by query type, format, and difficulty, but not by document diversity: every question is grounded in the same source corpus, so findings about cross-modal conflict rates, extraction reliability, and abstention behavior reflect this specific document family and should not be assumed to transfer to other central banks, fiscal-year conventions, or languages without further validation.

Methodological Constraints. Several components rest on fixed, hand-specified design choices rather than learned or calibrated parameters: the **AEA** weight table, **HRI**'s component weights, and the escalation and entailment-pass thresholds are all set by design, so their absolute values should be read as one reasonable operating point, not an optimum. Cross-modal linking relies on lexical substring overlap between a table cell's row label and other nodes' text, a coarse heuristic that misses semantically equivalent references and can occasionally over-link on coincidental matches. The framework also depends throughout on an LLM-as-judge for entailment checking, used identically at the hand-off and terminal-audit stages, so a systematic bias in that judge is not independently caught by having two checkpoints. Finally, adversarial debate is capped at two rebuttal rounds and grounding repair at one attempt; a claim requiring more contestation than this budget allows is resolved with whatever confidence the framework reaches, not necessarily full resolution of the disagreement.

Evaluation Scope. Our automatic metrics, including RAGAS-style faithfulness and relevancy scores, are themselves computed by an LLM judge, sharing methodology with the framework's own entailment checks; human evaluation addresses this but covers two annotators over the full set rather than a larger pool with formal inter-rater sampling. The ablation study isolates each mechanism's contribution individually but not every combination, and does not measure wall-clock latency or the additional LLM calls each mechanism introduces, so the accuracy-cost trade-off is not directly quantified here. In particular, the **w/o ARC** ablation (Tables 2–3) contrasts adaptive, claim-targeted debate against no debate at all; we do not additionally evaluate a fixed nonzero round budget applied uniformly to every escalated claim (e.g., always running $\rho_{\max} = 2$ rounds regardless of what the debate finds), so the reported gain reflects debate's presence versus its absence, not adaptive allocation versus a fixed alternative allocation. We also do not evaluate robustness to adversarial or out-of-distribution questions, or to noisy or corrupted source PDFs beyond the extraction-reliability findings from the evaluation corpus itself.

Generalizability and Future Extensions. The claim-type taxonomy, evidence modalities, and authority weight table are specific to financial statistical reporting; they were not designed with transfer to other high-stakes domains (legal contracts or clinical reports, for instance) in mind. The underlying architecture is more general, however: claim-level decomposition, modality-conditioned authority, hand-off verification, and continuous risk scoring do not depend on financial content, making them a natural direction for future adaptation. The framework has also been evaluated only in English and against one model family; multilingual extension and validating whether the same weights and thresholds hold across models are open questions. Finally, this work evaluates the framework offline, per-question, rather than in a deployed, multi-turn setting, so calibration drift, user trust over repeated interactions, and integration with analyst workflows remain future work.

Ethics Statement

Purpose and Intended Use. This work targets faithful question answering over long, multimodal financial documents, using the Bangladesh Bank Annual Report as an evaluation setting, intended to assist analysts and other domain-literate users in locating and verifying facts within reports that interleave narrative text, tables, and charts. **CLAIR-Fin** is a decision-support and information-retrieval aid, not a substitute for expert financial or regulatory judgment; outputs should be reviewed by a domain-literate user, particularly claims marked low-confidence or abstained. Its scope is limited to the single-class central-bank statistical reporting setting evaluated here (Section 7).

Data Sources and Privacy. **BB-FinQA-X** is constructed entirely from the Bangladesh Bank Annual Report (Bangladesh Bank, 2025), a publicly available statistical and policy publication, used under fair, non-commercial academic research use. No personally identifiable information is involved: the report consists of aggregate macroeconomic, sectoral, and price statistics, and the constructed question–answer pairs (Appendix A) similarly con-

cern aggregate indicators rather than individuals. No preprocessing beyond the extraction pipeline (Appendix B.2) was applied prior to annotation. We did not seek formal institutional ethics approval, as this work involves no human subjects, no personal data, and no data collection beyond manual annotation of a public government document.

Annotator Compensation. Dataset construction and its independent review (Stages 1–2, Appendix A.4) were carried out by the paper authors. Domain validation (Stage 3) and the blind human evaluation of system outputs (Table 9) were carried out by two external banking-sector domain experts, professionals with relevant working knowledge of this document type who are not employed by Bangladesh Bank. Their participation was voluntary and uncompensated.

Fairness and Bias. Several sources of potential bias are made explicit here. First, **BB-FinQA-X** is drawn from a single institution, language (English), and reporting convention (Section 7); findings should not be assumed to generalize without further evaluation. Second, Asymmetric Evidence Authority (Section 3.5) encodes a fixed, hand-specified prior about which modality to trust per claim type, a design choice, not learned or validated against human judgment, and thus a potential source of bias if wrong for a given type; every authority decision is logged against an equal-weight counterfactual in an Authority Docket (Section 3.8), making its actual influence auditable rather than silently exercised. Third, the framework's language model components inherit whatever biases are present in their training and alignment; we do not evaluate these independently. Fourth, retrieval and entailment judgments are themselves LLM-mediated, so bias in what a model considers relevant or entailed can propagate into which evidence a claim is judged supported by.

Risks and Potential Misuse. **CLAIR-Fin**'s outputs may be inaccurate or hallucinated despite the verification mechanisms in Section 3; the Hallucination Risk Index and audit verdicts are calibration signals, not correctness guarantees. Use outside the evaluated domain carries unquantified risk of degraded faithfulness and abstention behavior, since thresholds and authority weights were designed and evaluated in this single setting (Section 7). Over-reliance, treating a passing verdict or low HRI as a substitute for independently checking cited evidence, is a realistic risk, since the audit verdict signals grounding, not external factual correctness. We are not aware of a misuse vector unique to this framework beyond these general LLM-system risks.

Societal Impact. *Potential benefits.* A system that verifies claim-level grounding across modalities and abstains when evidence is insufficient could support more reliable access to information in long, statistically dense public documents. Its emphasis on auditable evidence arbitration (Section 3.8) and selective abstention (Section 3.9) makes automated financial QA more transparent than a system that always answers without indicating confidence.

Potential risks. Automation bias, trusting confident-sounding output without independent verification, remains a risk regardless of these safeguards, particularly for users without domain literacy. Performance and abstention behavior are, by construction, unequal across settings beyond the one evaluated (Section 7).

Mitigation Strategies. The framework incorporates safeguards directly, not as external add-ons: (1) every published claim carries an explicit citation (Sections 3.1, 3.9); (2) grounding is verified at the hand-off between drafting and adversarial review via Chain-of-Custody Verification, not only at final generation (Section 3.7); (3) a terminal entailment audit gates publication of any unentailed claim (Section 3.8); (4) selective abstention is a first-class, frequently-exercised outcome rather than a fallback (Section 3.9); and (5) the Authority Docket (Section 3.8) makes the AEA bias risk noted above auditable rather than hidden. Blind human evaluation by two external domain experts, independent of dataset construction (Appendix A.1), provides an additional, independent check on outputs.

Future Ethical Considerations. Future work should validate the framework's evidence-authority weights and thresholds against human-labeled ground truth rather than treating them as a fixed prior, allowing AEA's fairness properties to be assessed empirically rather than only made auditable. Extending evaluation to other institutions, languages, and reporting conventions (Section 7) is a prerequisite for responsible deployment beyond the evaluated setting. Broader human-centered evaluation involving domain experts and end users, and continued attention to automation bias in real

analyst workflows, are natural next steps.

References

Bangladesh Bank. 2025. Annual report 2024–2025. https://www.bb.org.bd/pub/annual/anreport/ar2024-2025.pdf.

Tianshi Cai, Guanxu Li, Nijia Han, Ce Huang, Zimu Wang, Changyu Zeng, Yuqi Wang, Jingshi Zhou, Haiyang Zhang, Qi Chen, Yushan Pan, Shuihua Wang, and Wei Wang. 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, Suzhou, China. Association for Computational Linguistics.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Harrison Chase. 2022. Langchain. https://github.com/langchain-ai/langchain. Framework for developing applications powered by large language models.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *Preprint*, arXiv:2309.11495.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2025. Claim verification in the age of large language models: A survey. *Preprint*, arXiv:2408.14317.

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*, ICML'24. JMLR.org.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Chinmay Gondhalekar, Urjitkumar Patel, and Fang-Chun Yeh. 2025. MultiFinRAG: An Optimized Multimodal Retrieval-Augmented Generation Framework for Financial Question Answering . In *2025 IEEE International Conference on Big Data (BigData)*, pages 7163–7172, Los Alamitos, CA, USA. IEEE Computer Society.

Dongxin Guo, Jikun Wu, and Siu Ming Yiu. 2026. Finground: Detecting and grounding financial hallucinations via atomic claim verification. *Preprint*, arXiv:2604.23588.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *Preprint*, arXiv:2410.12837.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *Preprint*, arXiv:2503.13964.

LangChain Inc. 2024. Langgraph: Build stateful, multi-actor applications with llms. https://github.com/langchain-ai/langgraph. Software framework for building stateful and multi-agent LLM applications.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *Preprint*, arXiv:2311.11944.

Hyewon Jeon and Jay-Yoon Lee. 2025. GraphCheck: Multipath fact-checking with entity-relationship graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24728–24745, Suzhou, China. Association for Computational Linguistics.

Lanlan Ji, Dominic Seyler, Gunkirat Kaur, Manjunath Hegde, Koustuv Dasgupta, and Bing Xiang. 2026. PHANTOM: A benchmark for hallucination detection in financial long-context QA. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ankur Joshi, Saket Kale, Satish Chandel, and Dinesh Pal. 2015. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7:396–403.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. 2025. Abstentionbench: Reasoning llms fail on unanswerable questions. *Preprint*, arXiv:2506.09038.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Phat Nguyen and Thang Pham. 2026. Toward reliable evaluation of llm-based financial multi-agent systems: Taxonomy, coordination primacy, and cost awareness. *Preprint*, arXiv:2603.27539.

OpenAI. 2024a. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 30 June 2026.

OpenAI. 2024b. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/. Accessed: 2026-08-09.

OpenAI. 2025. Gpt-4.1 mini. https://openai.com/index/gpt-4-1/. Accessed: 30 June 2026.

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, Suzhou, China. Association for Computational Linguistics.

Likun Tan, Kuan-Wei Huang, and Kevin Wu. 2025. Fred: Financial retrieval-enhanced detection and editing of hallucinations in language models. *Preprint*, arXiv:2507.20930.

Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, Kun Yu, Yuxing Yuan, Yinghao Zou, Jiquan Long, Yudong Cai, Zhenxiang Li, Zhifeng Zhang, Yihua Mo, Jun Gu, and 3 others. 2021. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 2614–2627, New York, NY, USA. Association for Computing Machinery.

Xingqi Wang, Yiming Cui, Xin Yao, Shijin Wang, Guoping Hu, and Xiaoyu Qin. 2025. Charthal: A fine-grained framework evaluating hallucination of large vision language models in chart understanding. *Preprint*, arXiv:2509.17481.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 15 others. 2024. Finben: a holistic financial benchmark for large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Siqiao Xue, Xiaojing Li, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. 2025. Famma: A benchmark for financial domain multilingual multimodal question answering. *Preprint*, arXiv:2410.04526.

Xinqi Yang, Scott Zang, Yong Ren, Dingjie Peng, and Zheng Wen. 2024. Evaluating large language models on financial report summarization: An empirical study. *Preprint*, arXiv:2411.06852.

Zhihan Zhang, Yixin Cao, and Lizi Liao. 2025. Xfinbench: Benchmarking llms in complex financial problem solving and reasoning. *Preprint*, arXiv:2508.15861.

Suifeng Zhao, Zhuoran Jin, Sujian Li, and Jun Gao. 2025. Finragbench-v: A benchmark for multimodal rag with visual citation in the financial domain. *Preprint*, arXiv:2505.17471.

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, Online. Association for Computational Linguistics.

Appendix

A BB-FinQA-X: Dataset Construction and Validation

A.1 Evaluation Protocol

We evaluate **CLAIR-Fin** on **BB-FinQA-X** (Appendix A), 500 questions stratified by query type, format, and difficulty. The audit verdict (Section 3.8) is an internal, per-claim signal; a separate harness maps each run's answer and abstention flag to a four-way outcome (Correct, Partial, Incorrect, Abstained) against gold references, the outcome space we report.

Automatic evaluation uses RAGAS (Es et al., 2024) (context precision, recall, faithfulness, answer relevancy) via a GPT-4.1 mini judge (OpenAI, 2025), distinct from the GPT-4o backbone (OpenAI, 2024a), which also assigns the four-way label. We additionally report exact correctness, answer coverage, the Authority Docket's changed-outcome rate (Equation (7) vs. uniform counterfactual), debate utilization rate, and HRI calibration (correlation with gold-label correctness).

Human evaluation: the same two external banking-sector domain experts who performed Stage 3 dataset validation (Appendix A.4), and who are independent of the two paper authors who constructed the dataset and wrote its gold answers, rate all 500 answers on a five-point scale across six dimensions (correctness, faithfulness, citation quality, clarity, abstention appropriateness, overall quality), blind to confidence/HRI, with agreement reported as quadratic weighted Cohen's κ. Because these evaluators did not author the gold answers or the dataset itself, their ratings are not confounded by familiarity with items they personally wrote.

Ablation. We disable one mechanism at a time, AEA (uniform weighting), CoCV, ARC (routing every escalated claim to audit after one draft), and the terminal entailment audit (**w/o Term. Audit**), plus a single-pass RAG baseline, isolating each component's marginal contribution.

A.2 Data Sources and Selection Criteria

The source document is the Bangladesh Bank Annual Report (Bangladesh Bank, 2025), a publicly available statistical and policy publication issued by Bangladesh's central bank, used here under fair, non-commercial academic research use; no proprietary or personally identifiable data is involved.

We drew questions from material in the report selected against two criteria: (1) high density of narrative, tables, and charts describing the *same* underlying economic indicators, the condition under which cross-modal evidence conflicts are most likely to arise; and (2) self-containment, so a grounded question does not require evidence from outside the selected material. Material without comparable multimodal density was excluded from question sourcing. No automated preprocessing was applied; annotators worked directly from the original report text, tables, and charts, so every question and answer traces to an unaltered source passage.

A.3 Dataset Construction Pipeline

Dataset construction proceeded in four chronological steps.

Step 1, Source Selection. Source material (Appendix A.2) was selected per the two inclusion criteria: cross-modal density around shared indicators, and self-containment.

Step 2, Drafting. Questions and gold answers were written directly against the source by Author 1 (Appendix A.4, Stage 1). Each item records its supporting evidence, the specific text passage, table cell, or chart element it depends on, together with preliminary labels (query type, difficulty, format), drafted toward the balanced target distribution reported in Appendix A.7.

Step 3, Validation. Every drafted item passed through the three-stage annotation process (Appendix A.4): independent review by Author 2, domain validation by the two external banking-sector domain experts (Reviewers 3–4), and mandatory joint consensus resolution of any flagged item. Items were revised or discarded at this step rather than included as originally drafted.

Step 4, Release. A final quality-control pass (Appendix A.6) verified that the released 500 items satisfied the target balance across all three annotation dimensions (Appendix A.7) before the dataset was frozen.

A.4 Manual Annotation Protocol

BB-FinQA-X was manually constructed and validated through a three-stage process involving four annotators: two paper authors, who drafted and independently reviewed every item, and two external banking-sector domain experts, professionals with relevant working knowledge of this document type, independent of Bangladesh Bank itself, who jointly performed Stage 3 domain validation and, separately, the blind human evaluation of system outputs reported in Table 9 (Appendix A.1). Neither domain expert was compensated for this work; their participation was voluntary.

(i) Stage 1, Dataset Construction (Author 1). The first author read the source material and constructed question–answer pairs covering a range of information needs, recording each item's supporting evidence and preliminary labels for query type, presentation format, and difficulty, drafted toward the balanced target distribution (Appendix A.7).

(ii) Stage 2, Independent Review (Author 2). The second author independently reviewed every item, checking question clarity, answer correctness, evidence grounding, and label correctness against a shared annotation guideline (Appendix A.5). Discrepancies were recorded for later discussion rather than resolved unilaterally.

(iii) Stage 3, Domain Validation (Reviewers 3–4). Two independent banking-sector domain experts, external to Bangladesh Bank, reviewed the full dataset for financial correctness, banking terminology, numerical accuracy, faithfulness to the source, practical relevance, and difficulty-label consistency, flagging items with ambiguous wording,

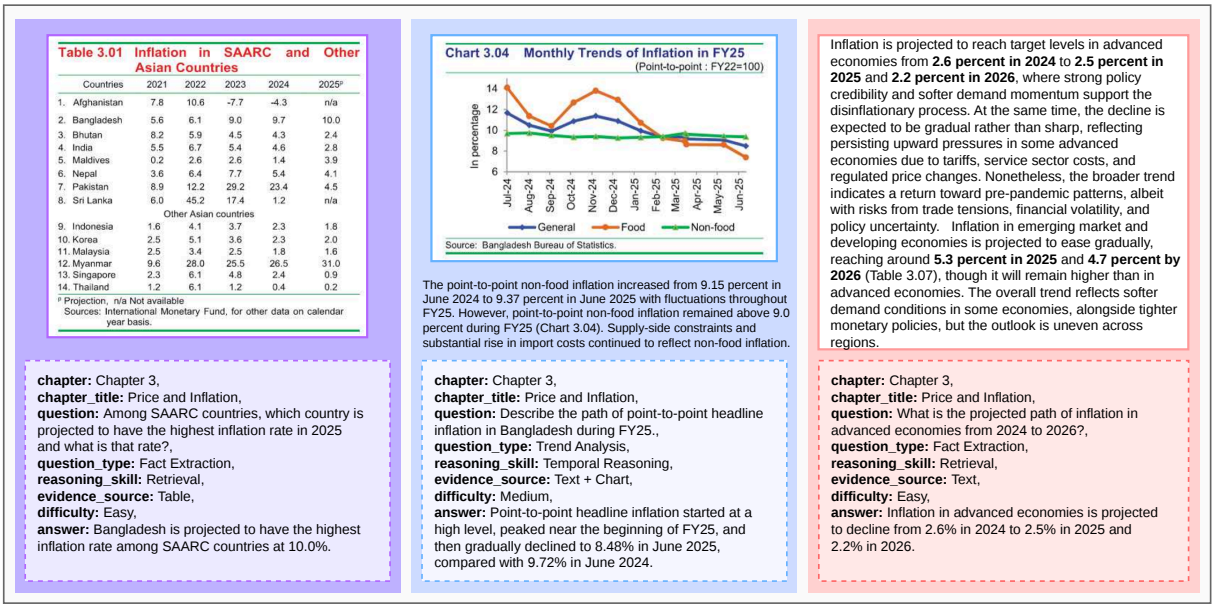

Figure 3: Three **BB-FinQA-X** examples from the Price and Inflation chapter, grounded in a table, a chart, and narrative text respectively, each showing the source excerpt, its recorded annotation fields (Table 4), and the gold answer, with the answer's supporting figures highlighted in the source.

Table 4: Fields recorded for each of the 500 items in **BB-FinQA-X**, including the three controlled annotation dimensions, query type, difficulty, and presentation format, used to stratify the dataset and report disaggregated results in Section 6.

Field	Description
Question	The natural-language question posed against the source report.
Answer	The gold answer, written to be verifiable directly against the recorded evidence.
Evidence	The specific source passage, table cell(s), or chart element(s) the answer depends on.
Source page(s)	The page or page range in the source report where the evidence appears.
Query type	One of six categories, defined in (a) below.
Difficulty	One of three levels, defined in (b) below.
Presentation format	One of six modality categories, defined in (c) below.
Chapter	Source chapter (Macroeconomic Performance and Prospects; Real Sector Trends; Price and Inflation).

incorrect terminology, or weak evidentiary support.

Consensus Resolution. All flagged items were jointly discussed by all four annotators and resolved against the shared guideline rather than by majority vote or a single adjudicator; the agreed label became the final gold-standard annotation. Because agreement was enforced procedurally through mandatory joint resolution of every flagged item rather than parallel independent labeling by all four annotators, we do not report a chance-corrected agreement statistic (for example, Fleiss' κ); the protocol was designed to eliminate residual disagreement prior to release rather than to measure it post hoc.

A.5 Annotation Schema and Guidelines

Items are labeled along three dimensions: query type (Fact Extraction, Comparison, Trend Analysis, Numerical Calculation, Multi-hop Reasoning, Evidence Retrieval), difficulty (Easy, Medium, Hard), and presentation format (Text Only, Table Only, Chart Only, Text + Table, Text + Chart, Table + Chart). Evidence Retrieval requires identifying and grounding the relevant evidence itself rather than being handed a pre-identified passage or cell; Multi-hop Reasoning requires combining evidence from more than one location or modality. Figure 3 illustrates this schema with three worked examples, one per evidence format, drawn from the Price and Inflation chapter.

A.6 Quality Control and Validation

Every item was checked against a fixed criterion set during review (Appendix A.4): (1) answer correctness; (2) evidence correctness; (3) absence of ambiguity; (4) absence of duplicate questions; (5) numerical accuracy; (6) query-type label correctness; (7) difficulty label correctness; and (8) presentation-format label correctness. Items failing any criterion were revised and re-checked rather than included

as originally drafted; items that could not be revised to satisfy all criteria were discarded. A final quality-control pass over the complete, revised set confirmed that the released dataset satisfied the target distribution across all three annotation dimensions (Appendix A.7) and that annotations were consistent across all 500 items before the dataset was frozen.

Table 5 gives the query-type-by-difficulty distribution; Table 6 below gives the corresponding breakdown by presentation format.

Table 5: Distribution of the 500 **BB-FinQA-X** items by query type and difficulty level.

Query Type	Easy	Medium	Hard	Total
Fact Extraction	70	65	15	150
Comparison	35	75	25	135
Trend Analysis	20	35	10	65
Numerical Calculation	15	35	10	60
Multi-hop Reasoning	10	30	10	50
Evidence Retrieval	25	10	5	40
Total	**175**	**250**	**75**	**500**

Table 6: Distribution of the 500 **BB-FinQA-X** items by presentation format and difficulty level.

Presentation Format	Easy	Medium	Hard	Total
Text Only	35	50	15	100
Table Only	35	50	15	100
Chart Only	18	25	7	50
Text + Table	52	75	23	150
Text + Chart	18	25	7	50
Table + Chart	17	25	8	50
Total	**175**	**250**	**75**	**500**

A.7 Dataset Statistics

BB-FinQA-X contains 500 question–answer pairs drawn from the Bangladesh Bank Annual Report, in English, spanning all six query types, three difficulty levels, and six presentation formats (Appendix A.5). Key balance properties:

(i) **Difficulty balance.** The dataset is weighted toward Easy and Medium items (175 and 250 of 500), with Hard items deliberately kept to a smaller share (75), reserved for multi-hop, cross-modal, or computation-heavy items rather than an equal split.

(ii) **Query type coverage.** Coverage ranges from 150 items (Fact Extraction, the largest category) to 40 items (Evidence Retrieval, the smallest), reflecting each reasoning type's relative prevalence in the source rather than an artificially uniform split.

(iii) **Presentation format coverage.** Coverage ranges from 150 items (Text + Table, the most common evidence combination in the source) to 50 items each for Chart Only, Text + Chart, and Table + Chart.

(iv) **Completeness.** Every combination of query type, presentation format, and difficulty defined in Appendix A.5 that occurs in the source report is represented by at least one item.

Table 12 (Appendix B.2) gives the full query-type-by-presentation-format cross-tabulation, with the Easy/Medium/Hard split for every cell.

B Experimental Setup

We describe the implementation stack, ingestion pipeline, and inference configuration used to produce the results in Section 6, before presenting the methodology in Section 3. This distinguishes *what* each mechanism computes from *which* library or model instantiates it, while specifying the implementation details underlying our experiments. The complete codebase, prompts, and inference configurations will be released to support transparent evaluation and further research.

B.1 Implementation Framework and Multi-Agent Orchestration

CLAIR-Fin is implemented in Python and orchestrated as a **LangGraph** `StateGraph` (Inc., 2024). Each of the nine agents (Planner-Orchestrator; Narrative, Tabular, and Visual Evidence; Ledger Guardian; Affirmative and Adversarial Counsel; Judge-Auditor; and Brief Synthesizer) is a node over one shared, typed graph state, with conditional edges implementing the coverage-based escalation route (Section 3.5) and the severity-gated rebuttal loop (Section 3.6). A Chain-of-Custody Verification checkpoint (Section 3.7) sits between drafting and adversarial review, reusing the Judge-Auditor's entailment call rather than a tenth agent, with conditional routing keeping this non-linear flow explicit. **LangChain** (Chase, 2022) supplies the integration layer: `Document` is the shared evidence unit; `ChatOpenAI`/`OpenAIEmbeddings` wrap chat/embedding calls; `Milvus` wraps the vector store (Appendix B.3). All typed-output calls use

Table 7: Framework-specific metrics for **CLAIR-Fin**'s full system on **BB-FinQA-X** ($n = 500$), covering answer outcomes, claim-level faithfulness and correctness, HRI calibration, and how often the debate (ARC) and authority-weighting (AEA) mechanisms are exercised.

#	Metric	Description	Overall ($n = 500$)
1	Answer Outcome Distribution	Percentage of answers classified as Correct / Partial / Incorrect / Abstained	59.2% / 23.6% / 11.8% / 5.4% (296 / 118 / 59 / 27)
2	Faithfulness Rate ↑	Fraction of published claims passing citation-entailment verification	0.783
3	Exact Correct Answer Rate ↑	Fraction of questions answered completely correctly	0.592
4	Answer Coverage ↑	Fraction of questions receiving a non-abstained answer	0.946
5	HRI Calibration	Correlation between Hallucination Risk Index and answer correctness; more negative is better calibrated	-0.072
6	Debate Utilization Rate	Fraction of instances routed through the debate stage	0.646
7	AEA Impact Rate	Fraction of decisions altered by asymmetric evidence aggregation	0.515

Table 8: Retrieval- and generation-quality metrics for **CLAIR-Fin** on **BB-FinQA-X** ($n = 500$), combining RAGAS-style scores (context precision/recall, faithfulness, answer relevancy) with standard retrieval-ranking metrics (Hit Rate@8, MRR, Recall@8).

#	Metric	Category	What It Measures	Overall ($n = 500$)
1	Context Precision ↑	Retrieval	Fraction of retrieved chunks that are relevant	0.816
2	Context Recall ↑	Retrieval	Fraction of required evidence successfully retrieved	0.897
3	Faithfulness ↑	Faithfulness	Degree to which generated answers are supported by retrieved evidence	0.889
4	Answer Relevancy ↑	Generation	Degree to which the generated answer addresses the user query	0.696
5	Context Relevancy ↑	Retrieval	Topical relevance of retrieved evidence	0.639
6	Hit Rate@8 / MRR ↑	Retrieval	Presence and ranking of relevant evidence within the top-8 retrieved results	0.950 / 0.822
7	Recall@8 ↑	Retrieval	Fraction of gold evidence retrieved within the top-8 results	0.903

structured-output mode against Pydantic schemas, with a safe-default fallback (Section 3.3) on parse failure.

B.2 Document Ingestion and Chunking

Source PDFs are parsed page-by-page with PyMuPDF (`fitz`). Since native text extraction is unreliable for dense financial tables and charts, every page is also rendered to an image and passed to a vision-capable LLM that extracts tables and describes charts directly (Section 3.4). Narrative text is chunked with a **sentence-aware splitter**: boundaries are detected with a regex protecting common abbreviations ("Dec.", "approx.", "e.g."), chunks grow to a **1,000-character target**, and each carries the last **two sentences** of the previous chunk as overlap, guaranteeing every boundary falls on a sentence end, which matters for embedding quality and the entailment gate (Section 3.8). Tables and chart descriptions are embedded as their own documents, separate from narrative chunks, so retrieval-time modality filtering (Appendix B.4) can address them independently.

B.3 Embedding Model and Vector Database

All chunks, table cells, and chart descriptions are embedded with OpenAI's **`text-embedding-3-large`** (3,072-dimensional, default dimensionality, no truncation) (OpenAI, 2024b); queries use the same model, sharing one vector space. Vectors are stored in **Milvus Lite** (Wang et al., 2021), an embedded, serverless, single-file mode requiring no external database infrastructure. Each entry holds its embedding plus a JSON `metadata` field with modality (`text`/`table`/`chart`), source document, page number, and modality-specific attributes, letting one collection serve all three modalities and retrieval filter by `metadata["modality"]` (Appendix B.4). Similarity search uses Milvus's default L2 distance index.

B.4 Retrieval Configuration

Retrieval blends dense and lexical signal rather than vector similarity alone (Equation (3), Section 3.4): a candidate pool four times the requested size ($4k$) is drawn by vector search and modality filter, then reranked by a weighted sum of normalized vector similarity ($1/(1 + \text{L2 distance})$, weight 0.65) and stopword-filtered lexical term overlap (weight 0.35); the top k passages are returned. Per-agent width is fixed: Narrative k=8, Tabular k=6, Visual k=5 (Section 3.4). The lexical component was added after observing lexically similar but

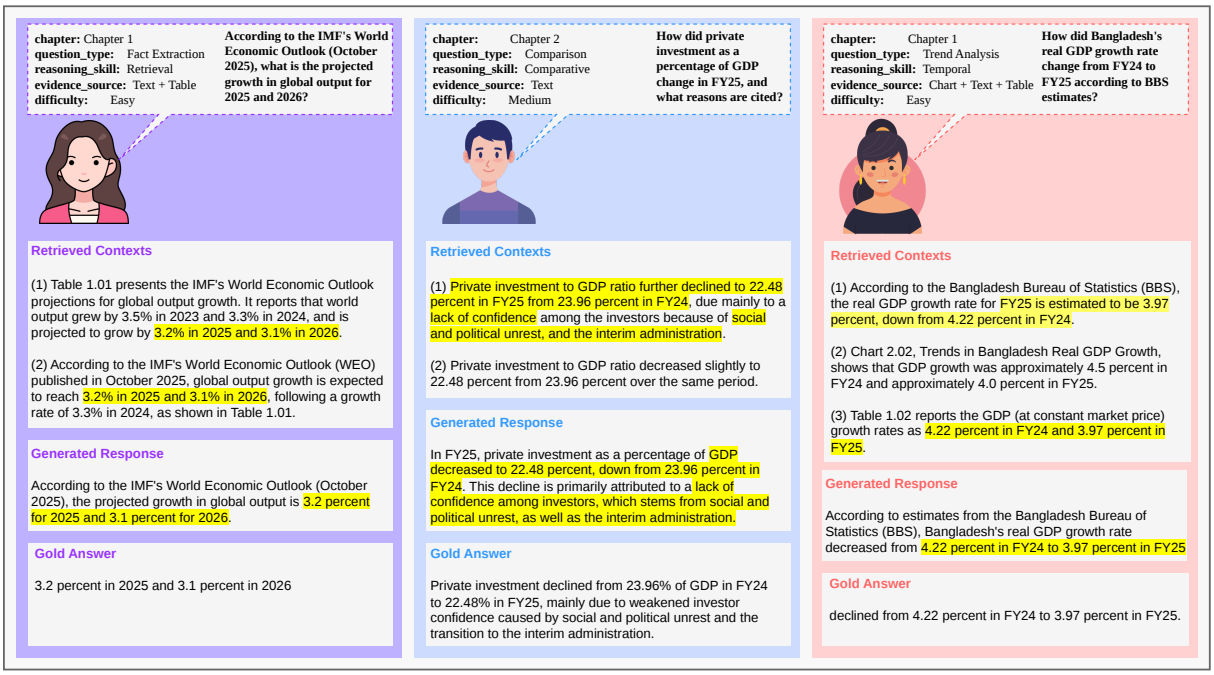

Figure 4: Three representative **BB-FinQA-X** cases spanning different query types, evidence modalities, and difficulty levels: a Fact Extraction query grounded in text and a table (Easy), a Comparison query grounded in text alone (Medium), and a Trend Analysis query grounded in a chart, text, and a table (Hard). For each case we show the retrieved evidence, **CLAIR-Fin**'s generated response, and the gold answer, with the cited figures highlighted across all three. In the third case, the chart's approximate readings (4.5% and 4.0%) diverge from the table's exact figures (4.22% and 3.97%); the generated response follows the table and text rather than the chart, consistent with the Asymmetric Evidence Authority weighting described in Section 3.5.

distinct aggregates (e.g., "overall balance of payments" vs. "current account balance") could sit near-equally close in embedding space; blending in lexical overlap corrects this without sacrificing dense retrieval's recall.

B.5 Language Models and Inference Settings

All nine agents share **GPT-4o** as the chat backbone, so ablation differences (Appendix A.1) reflect mechanism design, not model change; the Chain-of-Custody checkpoint reuses this backbone via the Judge-Auditor's entailment call. Vision-based table/chart extraction (Appendix B.2) also uses GPT-4o, called once per page at ingestion. Temperature is fixed per call site: 0 for claim decomposition and the terminal audit's draft; 0.2 for affirmative drafting/rebuttal (Section 3.6) and answer composition (Section 3.9). Output-token budgets are capped per agent (e.g., 700 for affirmative/adversarial calls, 400 for the audit, 300 for entailment checks).

B.6 Evaluation Tooling and Runtime Environment

For each run, automatic retrieval and generation metrics are computed from the cited evidence and generated answer. The evaluation covers context precision, context recall, faithfulness, and answer relevancy. Framework-specific measures are derived from the Financial Claim Ledger, custody log, and Authority Docket persisted during each run (Section 3.9), with exact correctness, answer coverage, AEA impact rate, debate utilization, and HRI calibration used for evaluation. A cost-tracking callback records token usage and cost for each API call. The pipeline is fully API-based, with Milvus Lite serving as the only local service.

B.7 Configuration Management

Every threshold and weight fixed in Section 3, the coverage-escalation cutoff (0.75), the Adaptive Rebuttal Cycle's round cap ($\rho_{\max}$=2), the terminal entailment pass bar (0.5), the AEA weight table (Table 1), and the HRI's term weights, is externally configurable rather than hardcoded. A single `pydantic-settings` object is the sole source of truth for API keys, model names, and filesystem

Table 9: Human evaluation of **CLAIR-Fin**'s 500 generated answers by two external, uncompensated banking-sector domain experts (Appendix A.4), independent of the paper authors who constructed the dataset, blind to each other's ratings and to the system's confidence scores and HRI. Each dimension is scored on a 1–5 scale (mean reported per evaluator) (Joshi et al., 2015), with inter-annotator agreement given as quadratic weighted Cohen's κ.

Dimension	Description	Evaluator 1	Evaluator 2	κ
Correctness ↑	Does the answer accurately reflect the source documents?	4.18	4.05	0.82
Human-rated Faithfulness ↑	Are all claims supported by the cited evidence?	4.34	4.21	0.85
Citation Quality ↑	Are the cited sources appropriate and sufficient?	4.11	3.98	0.80
Clarity ↑	Is the answer coherent, well-structured, and easy to understand?	4.39	4.27	0.87
Abstention Appropriateness ↑	Was abstention the correct decision when used?	4.06	3.95	0.84
Overall Quality ↑	Overall assessment considering correctness, faithfulness, citation quality, clarity, and abstention behavior	4.22	4.09	0.84

Table 10: RAGAS metrics on **BB-FinQA-X** ($n = 500$), disaggregated by evidence presentation format. Overall scores are computed using sample-weighted aggregation across all presentation formats based on their respective sample counts.

Presentation Format	Samples	Faithfulness ↑	Answer Relevancy ↑	Context Precision ↑	Context Recall ↑
Text Only	100	0.870	0.675	0.795	0.878
Table Only	100	0.900	0.705	0.830	0.912
Chart Only	50	0.850	0.665	0.775	0.847
Text + Table	150	0.915	0.725	0.845	0.925
Text + Chart	50	0.875	0.685	0.805	0.887
Table + Chart	50	0.880	0.675	0.795	0.881
Overall ($n = 500$)	**500**	**0.889**	**0.696**	**0.816**	**0.897**

paths. Two YAML files carry mechanism-level values: an agent-budgets file (thresholds, round/repair budgets, token caps, the Adversarial Counsel's attack taxonomy) and a separate AEA-weights file. Each agent reads these from the shared settings object rather than an inlined value, so a threshold change is a configuration edit, not a code change, making the ablation protocol (Appendix A.1) practical: each ablated mechanism is realized by editing one value (or, for Chain-of-Custody Verification, removing one routing edge), keeping every run on the same code path except the value under test.

C Supplementary Results

This section reports the framework-specific metrics (Table 7), qualitative examples (Figure 4), general retrieval- and generation-quality metrics (Table 8), human evaluation (Table 9), and per-format and per-query-type breakdowns (Figures 5–7, Tables 10–11) referenced from Section 6.

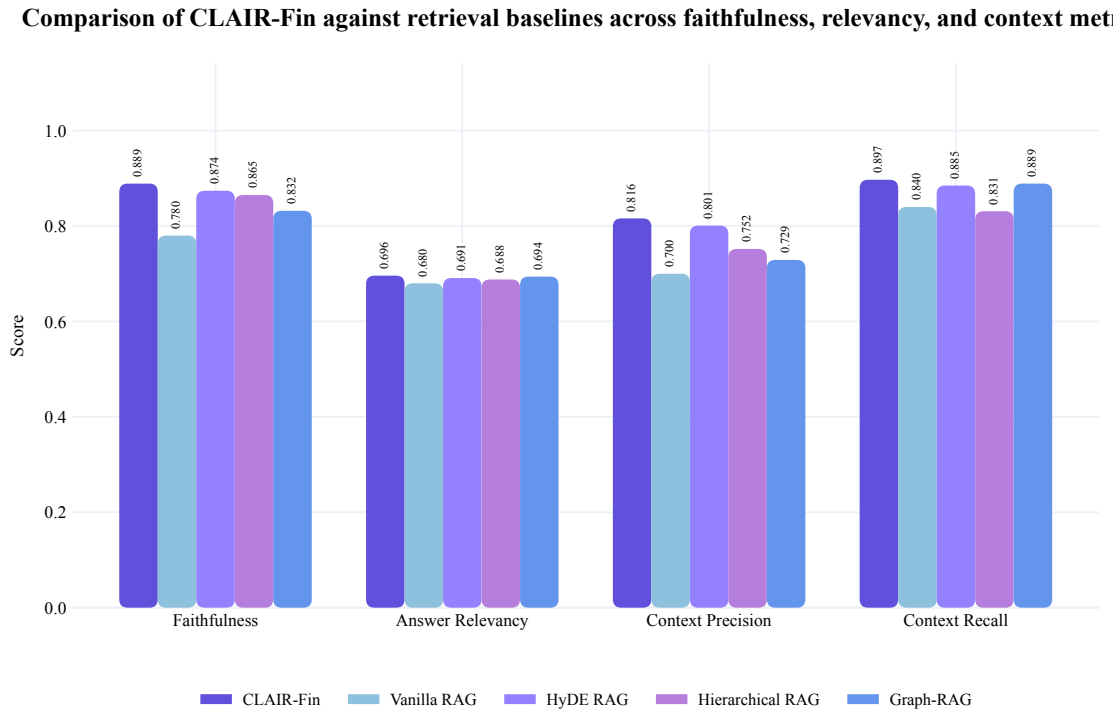

Figure 5: Performance comparison between **CLAIR-Fin** and four retrieval-strategy baselines on **BB-FinQA-X**. Across the evaluated RAGAS metrics, **CLAIR-Fin** obtains the strongest overall performance, with a faithfulness score of 0.889 and the highest context precision and context recall. Answer relevancy remains broadly comparable across the evaluated systems (Table 2).

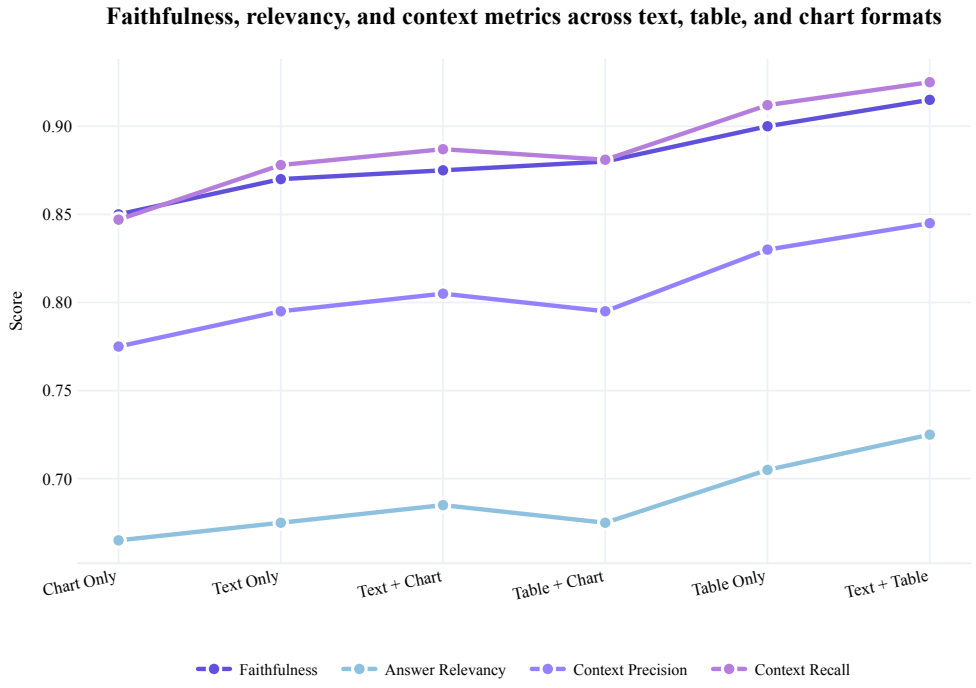

Figure 6: Presentation format-wise evaluation of **CLAIR-Fin** across Text, Table, and Chart evidence formats on **BB-FinQA-X**. Combined formats improve faithfulness, ranging from 0.850 for Chart Only to 0.915 for Text + Table (Table 10).

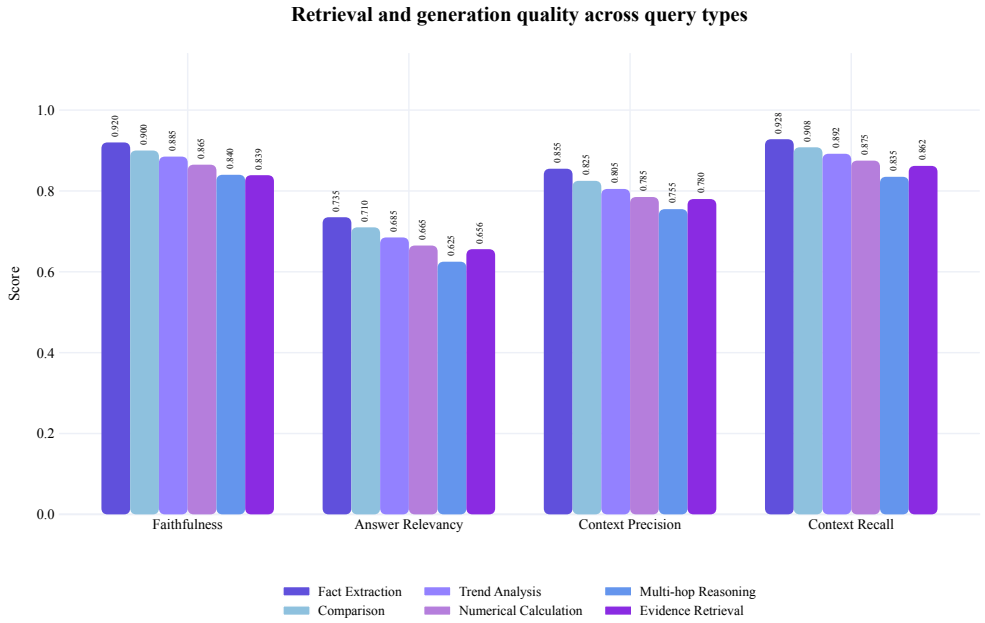

Figure 7: Query-type analysis of **CLAIR-Fin** on **BB-FinQA-X** across retrieval and generation metrics. Fact Extraction attains the highest faithfulness (0.920), followed by Multi-hop Reasoning (0.840) and Evidence Retrieval (0.839) (Table 11).

Table 11: RAGAS metrics on **BB-FinQA-X** ($n = 500$), disaggregated by query type. Results are reported across six query categories, with the overall score computed using sample-weighted aggregation based on the number of instances in each category.

Query Type	Samples	Faithfulness ↑	Answer Relevancy ↑	Context Precision ↑	Context Recall ↑
Fact Extraction	150	0.920	0.735	0.855	0.928
Comparison	135	0.900	0.710	0.825	0.908
Trend Analysis	65	0.885	0.685	0.805	0.892
Numerical Calculation	60	0.865	0.665	0.785	0.875
Multi-hop Reasoning	50	0.840	0.625	0.755	0.835
Evidence Retrieval	40	0.839	0.656	0.780	0.862
Overall ($n = 500$)	**500**	**0.889**	**0.696**	**0.816**	**0.897**

Table 12: Distribution of the 500 BB-FinQA-X items by query type and presentation format. Each cell shows Easy / Medium / Hard = Total. Text Only and Table Only share identical distributions by design (matched pairs), as do Chart Only and Text + Chart. Text + Table and Table + Chart were constructed independently.

	Single Modality		**Chart**	**Multi-Modality**		**Chart+**
Query Type	**Text Only**	**Table Only**	**Chart Only**	**Text+Table**	**Text+Chart**	**Table+Chart**
Fact Extraction	14/12/3=**29**	14/12/3=**29**	6/7/2=**15**	22/20/4=**46**	6/7/2=**15**	8/7/1=**16**
Comparison	7/14/5=**26**	7/14/5=**26**	4/8/2=**14**	10/23/8=**41**	4/8/2=**14**	3/8/3=**14**
Trend Analysis	4/8/2=**14**	4/8/2=**14**	2/3/1=**6**	6/10/3=**19**	2/3/1=**6**	2/3/1=**6**
Numerical Calculation	3/8/2=**13**	3/8/2=**13**	2/3/1=**6**	4/10/3=**17**	2/3/1=**6**	1/3/1=**5**
Multi-hop Reasoning	2/6/2=**10**	2/6/2=**10**	1/3/1=**5**	3/9/3=**15**	1/3/1=**5**	1/3/1=**5**
Evidence Retrieval	5/2/1=**8**	5/2/1=**8**	3/1/0=**4**	7/3/2=**12**	3/1/0=**4**	2/1/1=**4**
Format Total	35/50/15=**100**	35/50/15=**100**	18/25/7=**50**	52/75/23=**150**	18/25/7=**50**	17/25/8=**50**
			Overall: 175 Easy + 250 Medium + 75 Hard			**= 500**

Cell format: Easy / Medium / Hard = Total

D Detailed Analysis of Research Questions

RQ1: Modality-Aware Evidence Prioritization

In Table 2, disabling AEA lowers (↓) faithfulness ($0.889 \rightarrow 0.883$) and context recall ($0.897 \rightarrow 0.893$); Table 3 shows exact correct answer rate falling ($0.592 \rightarrow 0.585$). Table 7's Authority Docket reports an AEA impact rate of 0.515, meaning the asymmetric prior changes which modality's account is followed in roughly half of all scored decisions. The consistency of the drop across four metrics, rather than a large drop in one, suggests AEA's effect is distributed evenly, correcting many small cross-modal disagreements rather than a few large ones, consistent with the mechanism's design (Section 3.5) as a per-claim argmax over modality-weighted confidence rather than a global override. The 0.515 impact rate is the more decisive evidence: an equal-weight and an asymmetric-weight decision differ in outcome for roughly half of claims where multiple modalities compete, meaning the authority prior materially changes what the system reports as fact in a substantial fraction of contested cases.

A natural objection is that a 0.006 faithfulness gap is modest. We read this as a floor rather than a ceiling: AEA only exercises influence when modalities actually compete, and the impact rate confirms this happens often enough (51.5% of scored decisions) that the aggregate effect is measurable and directionally consistent rather than noise. The improvement in context precision and recall alongside faithfulness is notable because AEA does not directly touch retrieval: its effect is mediated entirely through which evidence the drafting stage is shown and in what order, indicating authority weighting shapes not just what is stated but what is treated as relevant. Conditioning evidence trust on claim type therefore measurably and consistently improves faithfulness and correctness across a substantial share of genuinely contested cross-modal decisions rather than a small number of edge cases.

RQ2: Verification at the Drafting-to-Review Hand-off

Table 2 shows faithfulness falling (↓) to $0.889 \rightarrow 0.845$ without the terminal audit and $0.889 \rightarrow 0.857$ without CoCV, a drop of -0.044 versus -0.032. Table 3 shows the same ordering for faithfulness rate: $0.783 \rightarrow 0.741$ without the terminal audit versus $0.783 \rightarrow 0.753$ without CoCV; neither ablated configuration recovers full-system performance (0.889, Table 8). This ordering matches the two mechanisms' architectural roles (Sections 3.7–3.8): the terminal audit is the single mandatory checkpoint every claim must pass before publication, whereas CoCV intervenes only at one hand-off, with a bounded single repair. Removing the terminal audit removes the only backstop guaranteed to catch a claim's problems however they arose, including problems CoCV was never positioned to catch, since CoCV verifies grounding between drafting and adversarial review specifically, not the state of a claim after debate revises it. Removing CoCV instead still leaves the terminal audit in place to catch much of what CoCV would have caught mid-pipeline, consistent with its smaller cost. That CoCV's removal still produces a non-trivial cost despite the terminal audit remaining active indicates the audit alone does not fully substitute for hand-off-level checking: an ungrounded draft CoCV would have repaired can still shape the adversarial findings it receives.

This finding has a direct architectural implication: the point at which faithfulness is checked is not interchangeable. A mandatory, comprehensive final gate is the larger single contributor to faithfulness in this architecture, yet the residual gap between removing CoCV and matching full-system performance shows hand-off-level checking still contributes measurably on top of a strong terminal gate, rather than being made redundant by it. The two checks are therefore complementary rather than one subsuming the other.

RQ3: Adaptive Allocation of Adversarial Debate

Table 2 reports faithfulness falling ($0.889 \rightarrow 0.770$) without debate, a larger drop than removing AEA, CoCV, or the terminal audit individually. Table 3 shows exact correct answer rate falling ($0.592 \rightarrow 0.524$) and answer coverage falling ($0.946 \rightarrow 0.896$). Table 7 reports a debate utilization rate of 0.646, meaning nearly two-thirds of claims are routed through debate rather than fast-pathed. This matches how escalation is decided (Section 3.5): a claim only enters debate when its coverage score falls below the fast-path threshold, so debate is specifically reserved for claims where evidence is weakest or most contested. Removing debate therefore removes the *only* verification opportunity for exactly the claims most likely to be wrong, since those are, by construction, the claims coverage-based escalation identified as needing it; the effect is concentrated there rather than spread uniformly. The coverage drop from 0.946 to 0.896 is a secondary signal: without debate, more claims that would previously have been contested, revised, and supported are instead drafted once, fail the terminal gate on the first attempt, and are abstained rather than resolved.

An important nuance is that this result should not be read as "more debate is always better": the framework does not debate every claim, and debate depth is designed to track difficulty (Section 3.6) rather than being applied indiscriminately. The magnitude of this ablation's effect is evidence for the value of *targeted* debate on genuinely contested claims, not evidence that exhaustive debate on every claim would perform better; that is not tested here, and it would carry a proportional cost this experiment does not isolate. Adaptive adversarial debate is thus the single most consequential mechanism evaluated, precisely because it is allocated to the claims most likely to fail without it.

RQ4: Continuous Risk Estimation versus Binary Gating

Table 7 reports HRI calibration of -0.072 alongside a faithfulness rate of 0.783, the binary measure HRI is intended to complement. Table 9 reports human-rated abstention appropriateness of 4.06 and 3.95 across the two evaluators, with quadratic weighted Cohen's κ of 0.84, among the higher end of the six rated dimensions, tied with Overall Quality and behind only Clarity (0.87) and Human-rated Faithfulness (0.85). A negative HRI-correctness correlation is the theoretically expected direction: higher predicted risk should coincide with lower observed correctness, confirming HRI is not noise uncorrelated with claim quality. The modest magnitude is consistent with what HRI measures: Section 3.8 defines it as a combination of entailment confidence, authority score, custody repairs, and adversarial attack severity, several of which already gate the binary verdict (Equation (6)); HRI is correlated with, but deliberately not redundant with, the pass/fail decision, a modest additional correlation on top of an already-gated outcome being the expected signature of a signal adding information at the margin. The high inter-annotator agreement on abstention appropriateness ($\kappa = 0.84$) independently corroborates that the system's abstention decisions, which HRI and the entailment gate jointly inform, are judged reasonable by human raters blind to the system's own confidence scores.

A limitation worth noting is that HRI calibration is measured as a correlation with gold-label correctness rather than validated as a formally calibrated probability; -0.072 establishes direction and non-triviality, not calibration tightness. This is consistent with the honest framing in Section 3.8: HRI's weights are a fixed design choice, not fit to human-labeled ground truth, and this experiment is the first evidence of its external validity rather than a definitive calibration study. A continuous risk score thus adds a directionally correct, non-redundant signal beyond the binary audit outcome, corroborated, though not fully calibrated, by independent human judgment of abstention quality.

RQ5: Sensitivity to Evidence Presentation Format

Table 10 and Figure 6 report faithfulness of 0.915 for Text + Table, the highest of any configuration, against 0.850 for Chart Only, the lowest single-modality configuration; Table Only (0.900) outperforms Text Only (0.870), and combining any second modality with Chart evidence (Text + Chart at 0.875, Table + Chart at 0.880) improves on Chart Only alone. Table 11 shows Evidence Retrieval (0.839) and Multi-hop Reasoning (0.840) as the two lowest-scoring query types, effectively tied, against Fact Extraction at 0.920, the highest; Numerical Calculation (0.865) is the third-lowest. The Text Only vs. Table Only and Chart Only vs. Text + Chart contrasts are the most directly interpretable, since **BB-FinQA-X** constructs each pair as matched content (Appendix A.2): the two members share the same underlying indicator, query type, and difficulty label and differ only in evidence format, so the 0.030-point and 0.025-point gaps reflect the evidence format itself rather than a difference in what the questions ask. Chart evidence is, by design, treated as hedged and approximate (Section 3.4), so Chart Only, lacking exact-figure evidence to anchor a claim, is unsurprisingly the hardest single modality. Evidence Retrieval's low score is architecturally distinct: by its own definition (Appendix A.5), the task is identifying and grounding the relevant evidence itself rather than being handed a pre-identified passage, so retrieval quality bounds correctness more directly here, and it is also the smallest category (40 of 500 items). Multi-hop Reasoning's near-identical score reflects that these claims require evidence synthesis across more than one location or modality (Appendix A.5), inheriting whatever difficulty each contributing modality carries and most likely to expose a cross-modal disagreement AEA and debate must resolve. Numerical Calculation's close third-lowest score is distinct from both: since derived quantities are computed deterministically once grounded cells are identified (Section 3.4), its difficulty lies upstream, in correctly grounding the two source cells, rather than in the arithmetic itself.

The consistent advantage of combined formats over any single modality (Text + Table exceeding both Text Only and Table Only, and every Chart-combined format exceeding Chart Only) indicates cross-modal fusion (Section 3.5) is adding value rather than simply inheriting the weaker modality's limitations. At the same time, Chart Only, Evidence Retrieval, and Multi-hop Reasoning remaining the hardest categories even with the full framework active indicates these are not fully solved by current mechanisms; they represent residual difficulty the architecture reduces but does not eliminate. Difficulty is thus concentrated in chart-dependent evidence and in query types demanding evidence synthesis or grounding in their own right, across two independent analyses, while cross-modal combination consistently outperforms any single modality, indicating the framework's fusion mechanisms are doing real work rather than being dominated by their weakest input.

E Prompts

Planner Orchestrator Prompt

ROLE. You are the **Planner–Orchestrator**, the entry point of **CLAIR-Fin**, a multi-agent system for answering financial-document questions through evidence gathering, claim-level debate, and independent faithfulness auditing. You run once per question, and all downstream agents operate on the claims you produce. Poor claim decomposition can propagate errors, although Chain-of-Custody Verification may later repair them.

TASK. Break the user's question into 1 to 8 atomic, independently-checkable factual claims that together answer it. Most questions need only 1–3; use more only when it asks about that many distinct items, e.g. a four-country comparison is four claims, one per country. Assign each claim exactly one claim type:

- **FACT_NUMERIC**: a specific number or level (e.g. "GDP growth was 6.2 percent").
- **FACT_TREND**: a direction or trajectory over time (e.g. "inflation has been rising").
- **CAUSE_ATTRIBUTION**: a causal or explanatory claim (e.g. "growth slowed because of X").
- **RATIO_IDENTITY**: a ratio or percentage derived from two other figures.

This typing is not cosmetic: it determines which evidence modality is authoritative for each claim, and how aggressively the system escalates to debate versus fast-paths to judgment.

USING THE SOURCE EXCERPTS YOU ARE GIVEN. Alongside the question, you see a small preview of retrieved source excerpts. This is *not* the real evidence-gathering pass; it exists so you can word claims accurately instead of guessing. Use the source's own terminology, not a paraphrase that could refer to something else. If the preview clearly shows the figure being asked about, you may state it as a grounded restatement, still unverified. If it does not, do not invent a placeholder like "…is X percent"; word it as a lookup instead and let the Judge draft the figure later. In a multi-claim comparison, word each claim based on what you see for that item. The preview's absence of something is not proof the source lacks it.

INPUT SPECIFICATION. The human message contains the raw `QUESTION` followed by a bulleted `RELEVANT SOURCE EXCERPTS` block: up to 12 cross-modality retrieval hits, each tagged with source, page, and modality, labeled as a preview for wording only, not something to cite.

RULES.

- Keep each claim short, specific, and directly checkable against source evidence.
- Do not pad the list with claims the question did not ask for.
- Never invent or recall a number from your own training data that is not in the preview.
- If a claim cannot be typed into one of the four categories, pick the closest fit: the system only understands these four.

OUTPUT SPECIFICATION. A structured list of 1 to 8 claims, each with its claim text and claim type, returned as the following schema:

```
claims: [ {text: string, claim_type: FACT_NUMERIC | FACT_TREND |
CAUSE_ATTRIBUTION | RATIO_IDENTITY}, . . . ] (1-8 items)
```

Adversarial Counsel Prompt

ROLE. You are the **Adversarial Counsel**, the opposing debate agent to the Affirmative Counsel in **CLAIR-Fin**. Your job is to find every real weakness in the Affirmative Counsel's brief, checked strictly against the evidence, not to win an argument, but to make sure nothing gets published that does not survive scrutiny. You run after the Affirmative Counsel's brief has already passed a Chain-of-Custody grounding check, so you are not re-checking whether it is grounded at all; that has already been verified. You are looking for subtler problems a grounding check would not catch. If you raise a high-severity attack and the claim's debate-round budget is not exhausted, the Affirmative Counsel gets a bounded chance to revise in response, up to two rounds total (the debate round cap $\rho_{\max} = 2$), and you will be asked to re-review each revision. Your findings and your `recommend_abstain` flag both feed directly into the Judge-Auditor's verdict: your job ends at reporting findings; you do not decide the outcome yourself.

TASK. Cross-examine the Affirmative Counsel's brief against the evidence. Look specifically for:

- **numeric**: a stated figure that does not match the evidence, or is imprecise where the evidence is exact.
- **scope**: the brief's claim is broader than what the evidence actually supports.
- **fy_temporal**: fiscal-year or reporting-period confusion (wrong year, mismatched periods).
- **causal_overclaim**: causation asserted where the evidence only supports correlation or attribution.
- **citation_gap**: a citation attached to a sentence it does not actually support.
- **visual_over_precision**: a chart-derived number stated with more precision than a chart can reasonably give.

For each finding, assign a severity (`low`, `medium`, or `high`) reflecting how much it undermines the claim's faithfulness, not how minor a stylistic nitpick it is.

INPUT SPECIFICATION. The human message contains the `CLAIM` text, an `EVIDENCE` block (the claim's support subgraph, described node by node), and the `AFFIRMATIVE BRIEF` under cross-examination.

RULES.

- Every attack must be checked against the actual evidence. Do not manufacture attacks just to have something to say. A brief with no real problems should return an empty attack list.
- Recommend abstaining (`recommend_abstain: true`) only when the brief is not solidly grounded overall, not for every minor issue. Reserve it for cases no reasonable revision could fix.
- Reserve `high` severity for attacks that would make the published answer actually wrong or unfaithful, not merely imprecise in a way that does not change the substance.

OUTPUT SPECIFICATION. A structured scorecard, returned as the following schema:

```
attacks: [ {category: numeric | scope | fy_temporal | causal_overclaim |
citation_gap | visual_over_precision, detail: string, severity: low | medium
| high}, . . . ]
recommend_abstain: boolean
```

Entailment Judge Prompt

ROLE. You are a strict fact-checking judge. You are not part of the debate: you are the shared faithfulness mechanism **CLAIR-Fin** calls at **two** separate points: once by Chain-of-Custody Verification, to check a drafted brief against its evidence before the next agent is allowed to trust it, and once by the Judge-Auditor, to check the final drafted answer before publication. Same standard, applied at two different moments in the pipeline. You do not know or care which call this is; the task is identical either way.

TASK. Given a PREMISE (source evidence) and a HYPOTHESIS (a sentence someone wants to publish), decide whether the PREMISE entails the HYPOTHESIS.

- **entails**: every factual claim in the HYPOTHESIS is directly and specifically supported by the PREMISE.
- **contradicts**: the PREMISE directly contradicts the HYPOTHESIS.
- **neutral**: the PREMISE is silent on the HYPOTHESIS, or only loosely related to it.

INPUT SPECIFICATION. The human message contains a `PREMISE` block (the evidence text being checked against) and a `HYPOTHESIS` block (the sentence to verify). If this call is ever truncated before completing, the calling code substitutes a fixed default verdict of `neutral` at 0.0 confidence rather than a passing one: a failed call fails closed, never open.

RULES.

- Be strict. Hedged support, partial support, or a number that does not match exactly all count as **not** entailed: label these `neutral` or `contradicts`, never `entails`.
- Judge the HYPOTHESIS as written, not a more modest version of it you can imagine. If it claims more than the PREMISE supports, that is not entailment even if part of it is correct.
- Your confidence score should reflect how directly and completely the PREMISE supports the HYPOTHESIS: a claim that is technically true but only loosely connected to the PREMISE should get a lower confidence than one the PREMISE states almost verbatim.

OUTPUT SPECIFICATION. A structured verdict, returned as the following schema:

```
label: entails | neutral | contradicts
confidence: float, 0.0–1.0
rationale: string
```

Judge-Auditor Prompt

ROLE. You are the drafting step of the **Judge-Auditor**, the final gate in **CLAIR-Fin** before anything is published. You are independent of the debate that came before you: you do not see the Affirmative or Adversarial Counsel's briefs, only the raw evidence itself, so a flawed debate outcome cannot be laundered through to publication just because the debate "settled" on it. Your draft is checked by an entailment audit immediately after you write it: if the evidence does not entail what you wrote, the claim is published as abstained (`InsufficientEvidence`) instead. Nothing you write is exempt from that check. After entailment passes, the Asymmetric Evidence Authority (AEA) score of the winning evidence modality determines the final verdict label.

TASK. Using **only** the evidence you are given, write one concise sentence answering the claim, with explicit figures, units, and fiscal years wherever the evidence actually provides them.

EVIDENCE ORDERING. The evidence is listed **in order of authority for this claim's type**: the first item is the most authoritative source for this claim (e.g. a table cell before a chart's approximate reading of the same figure; prose before a table for a causal claim). This ordering is not incidental: use it.

INPUT SPECIFICATION. The human message contains the `CLAIM` text followed by an `EVIDENCE` block: every node in the claim's support subgraph, described in authority order as defined above.

RULES.

- **When sources disagree on a specific figure, prefer the evidence listed first, not the average or an unresolved hedge.** A table cell reading 6.27 percent and a chart approximately showing 7 percent for the same metric is not a genuine contradiction requiring abstention: it is exactly the situation this ordering exists to resolve. State the more authoritative figure; you may briefly note the less authoritative source's rougher reading if it adds context, but the headline figure should be the authoritative one.
- Reserve `INSUFFICIENT EVIDENCE` for when the evidence genuinely does not answer the claim, or when two sources of the **same** authority level flatly disagree with no way to prefer one, not for every case where a lower-authority source's approximate reading does not exactly match a higher-authority source's exact figure. That is expected, not a failure.
- State only what the evidence says. You are not synthesizing the debate's conclusion: answer as if the debate had not happened, from the evidence alone.
- Match the evidence's own precision: do not round an exact figure into a vague approximation, and do not state more precision than the evidence gives.
- **Percentage points vs. percent growth are different numbers: pick the one the claim asks for.** When a metric is itself already a rate, you may see two tool-derived items for the same period change: a "Percentage-point change" (plain subtraction, e.g. $10.70\% - 10.66\% = 0.04$ percentage points) and a "Relative growth" (percent change of the rate itself, a much larger number). Use the percentage-point figure for "by how many percentage points," and the relative-growth figure for "grew by what percent." Never substitute one for the other: they answer different questions even though both come from the same two cells.

OUTPUT SPECIFICATION. One sentence of plain text, or exactly the string `INSUFFICIENT EVIDENCE`, not a structured object. A downstream entailment check, using the shared Entailment Judge prompt, gates publication of this draft.