E2A-Bench: Benchmarking Evidence-to-Action Reliability in Financial Chart Reasoning

Anonymous ACL submission

Abstract

Can financial vision–language models (VLMs) turn chart evidence into reliable action recommendations? Existing hallucination evaluations are mostly claim-centric: they assess whether generated statements are supported, but not whether evidence remains traceable through rationale, confidence, and final action. We introduce E2A-Bench, a 969-query evidence-to-action benchmark for financial chart reasoning, constructed from 323 HS300 constituents under three input modalities with deterministic OHLCV-derived evidence anchors. E2A-Bench evaluates four reliability links through UCR, RCI, ECI, and NDR, a coverage-aware chain score for decision reliability. Evaluating 20 VLMs reveals three failures hidden by scalar hallucination scores: the lowest-UCR model ranks near the bottom by NDR due to only 6.4% directional coverage; oracle-aided verification reduces unsupported claims but can collapse actionable coverage; and financial fine-tuning amplifies the BUY:SELL ratio by 4.21–4.68× across strict base–fine-tuned pairs. These results show that financial VLM evaluation should trace the full evidence-to-action chain rather than optimize a single hallucination score.

1 Introduction

In financial chart reasoning, vision–language models (VLMs) are expected to identify technical signals such as candlestick patterns, moving-average crossovers, and volume changes (Huang et al., 2025; Lee et al., 2025; Ding et al., 2024). They may further issue an action recommendation among BUY, SELL, HOLD, and UNCERTAIN, together with a confidence estimate. Reliability in this setting is therefore not reducible to factual correctness at the text level. A decision-reliable model must ground its recommendation in verifiable chart evidence without systematically avoiding directional decisions when sufficient evidence exists. Otherwise, a fabricated "golden cross," volume breakout, or trend reversal may become the stated rationale for a BUY recommendation, turning a local hallucination into a decision-level risk.

However, existing evaluations do not fully capture this requirement.Chart hallucination benchmarks are mostly claim-centric, assessing whether local statements are supported by visual or structured evidence, whereas financial multimodal benchmarks are typically task-centric, reporting aggregate scores or final-answer accuracy (Wang et al., 2024; Xue et al., 2024; Liu et al., 2025b; Shu et al., 2025). Both perspectives are useful, but they share an *evaluation-unit mismatch*: they evaluate claim correctness or task outcomes rather than *evidence-backed action*This mismatch yields two blind spots: *coverage blindness*, where hallucination scores can favor models that avoid directional recommendations, and *chain blindness*, where evaluations do not verify whether evidence remains consistent from chart-derived facts to claims, rationales, confidence estimates, and final actions (Geifman and El-Yaniv, 2017; Wang et al., 2025; Ravichander et al., 2025).Consequently, a model may appear reliable by producing few unsupported claims while offering little actionable value, or by issuing plausible rationales whose actions or confidence estimates are not supported by the underlying evidence.

We frame this issue as a **coverage–grounding frontier**, as shown in Fig. 1: financial VLM reliability should be measured not by unsupported-claim reduction alone, but by whether a model improves evidence grounding while preserving actionable coverage (Geifman and El-Yaniv, 2017; Liu et al., 2025a). A method that lowers hallucination by suppressing BUY or SELL recommendations does not necessarily improve decision reliability. A meaningful improvement should move models toward the favorable region of this frontier, where they act when evidence is sufficient and ground

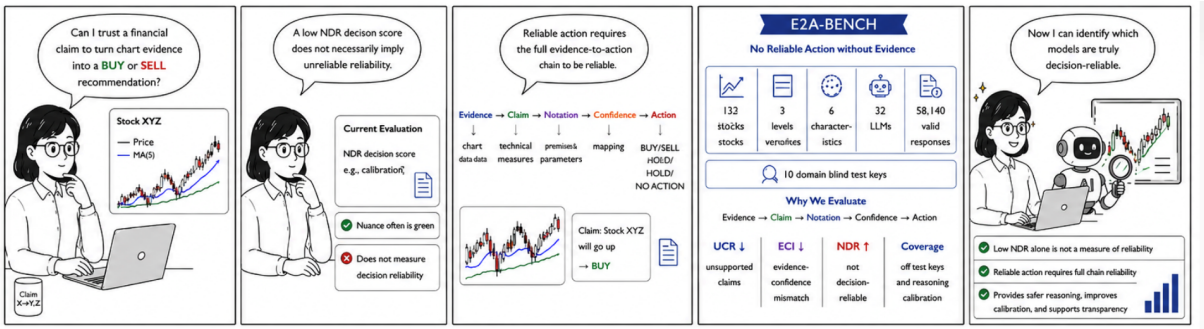


Figure 1: The coverage–grounding frontier in financial VLM evaluation. Existing evaluations are often coverage-blind and chain-blind: they measure whether claims are supported, but not whether models preserve actionable coverage or maintain full-chain reliability. `E2A-Bench` addresses both by evaluating evidence-to-action reliability through grounding, coherence, calibration, and net decision reliability.

their actions in verifiable evidence, coherent rationales, and calibrated confidence (Geng et al., 2024; Liu et al., 2025a). This framing also clarifies the scope of our evaluation: we do not assess whether models predict market returns, but whether their chart-based recommendations are supported by a verifiable evidence chain.

To operationalize this frontier, we introduce `E2A-Bench`, a 969-query evidence-to-action benchmark over 323 HS300 constituents and three input modalities: text-only, chart-only, and multimodal. Each query is built from a 30-trading-day OHLCV window and paired with ten deterministic technical-indicator evidence anchors, yielding 3,230 ground-truth facts for claim verification and evidence-strength estimation. Models produce a technical analysis, rationale, action, and confidence score. `E2A-Bench` evaluates the resulting evidence-to-action chain with UCR for unsupported claims, RCI for reasoning–action consistency violations, ECI for overconfident actions under weak evidence, and NDR for evidence-gated, coverage-aware net decision reliability.

We instantiate `E2A-Bench` by evaluating 20 VLMs spanning proprietary, open-weight, and financially adapted systems. The experiments reveal three patterns. First, scalar hallucination scores can misrank decision reliability: Gemma-4-E2B-It has the lowest overall unsupported-claim rate (39.9%), but only 6.4% directional coverage and NDR $= 5.1\%$, whereas Qwen2.5-VL-3B-Instruct achieves the highest NDR (32.1%) with 82.5% coverage. Second, in chart-only intervention analysis, oracle-aided verification can sharply reduce unsupported claims but collapse actionable coverage, yielding negative mean ΔNDR. Third, financial fine-tuning redistributes risk: strict base–fine-tuned pairs show inconsistent NDR changes but consistent BUY:SELL amplification of 4.21–4.68$\times$. Together, these results motivate evaluating financial VLMs along the coverage–grounding frontier rather than by scalar hallucination scores alone.

Our contributions are threefold.

- We introduce `E2A-Bench`, a 969-query benchmark for coverage-aware evidence-to-action reliability in financial chart reasoning.
- We propose a chain-level protocol for grounding, reasoning–action consistency, evidence–confidence calibration, and directional coverage.
- We evaluate 20 VLMs and uncover three failures: coverage-driven rank reversal, verification-induced coverage collapse, and BUY-side amplification after financial fine-tuning.

2 Related Work

Financial chart reasoning with VLMs. Recent VLMs have substantially improved chart, document, and structured-image reasoning. General-purpose model families such as Qwen2.5-VL (Bai et al., 2025) and Qwen3-VL (Qwen Team, 2025) strengthen visual parsing, table/chart understanding, and multimodal reasoning, while chart-oriented methods such as MatCha (Liu et al., 2023b), DePlot (Liu et al., 2023a) and ChartLLaMA (Han et al., 2023) improve chart comprehension through chart derendering, plot-to-table conversion, chart-specific pretraining, or instruction tuning. Finance-oriented systems such as FinLLaVA (Xie et al., 2024) and PyFi (Zhang et al., 2025) further adapt VLMs to financial tables, charts, and image-based financial reasoning. These advances make financial chart reasoning increasingly feasible, but they do not establish whether

Benchmark	Scale	#Inp.	Main evaluation target	Ev.	Act.	Cov.
CharXiv	≈2K charts	1	Realistic chart QA	✗	✗	✗
ChartHal	≈2K claims	1	Hallucination detection	✓	✗	✗
VisFinEval	≈1K Q&A	1	Financial chart QA	✗	✗	✗
FinChart-Bench	≈5K Q&A	1	Financial chart QA	✗	✗	✗
E2A-Bench (ours)	**≈1K Q&A**	**3**	**Evidence-to-action reliability**	✓	✓	✓

Table 1: Comparison with representative chart and financial multimodal benchmarks. "#Inp." denotes the number of supported input types among text-only, chart-only, and multimodal inputs. "Ev." indicates explicit evidence-grounding evaluation; "Act." indicates decision-oriented action evaluation; and "Cov." indicates coverage-aware scoring.

generated recommendations are grounded in verifiable chart evidence, coherent with the stated rationale, calibrated in confidence, and actionable under sufficient evidence.

Benchmarking hallucination in financial chart reasoning. Existing benchmarks provide important foundations for evaluating chart and financial VLMs. ChartQA and CharXiv evaluate chart question answering and realistic chart reasoning, while ChartHal targets fine-grained hallucination in chart understanding (Masry et al., 2022; Wang et al., 2024, 2025). Financial benchmarks such as FAMMA, VisFinEval, and FinChart-Bench extend multimodal evaluation to finance-specific QA, business scenarios, and real-world financial charts (Xue et al., 2024; Liu et al., 2025b; Shu et al., 2025). However, these evaluations are mostly claim-centric or task-centric: they assess whether local statements are supported, or whether final answers are correct. E2A-Bench is complementary: it shifts the evaluation unit to coverage-aware evidence-to-action reliability, testing whether chart evidence remains traceable through claims, rationales, confidence estimates, and final actions.

3 E2A-Bench: Evidence-to-Action

To evaluate financial VLMs along the coverage–grounding frontier, we introduce E2A-Bench, an evidence-to-action benchmark that tests whether chart-based recommendations are grounded, coherent, calibrated, and actionable.

3.1 Design Principles

E2A-Bench is guided by three principles.

Evidence verifiability. Since financial charts do not uniquely determine a correct trading action, we use deterministic OHLCV-derived facts as evidence anchors rather than future returns or human investment judgments.

Decision orientation. The evaluation target is a recommendation with its rationale and confidence estimate, rather than a descriptive chart summary.

Coverage-aware chain integrity. Reliable action requires evidence to remain consistent through claims, rationale, confidence, and final action, while avoiding artificial reliability gains from systematic non-action.

3.2 Task Formulation

For each stock s and evaluation date d_0, E2A-Bench uses the preceding 30 trading days of OHLCV data, where OHLCV denotes open, high, low, close, and volume. Let D_{s,d_0} denote this market window and let $\mathcal{T} = \{\text{text}, \text{chart}, \text{multi}\}$ denote the supported input types. Each input is constructed by a modality-specific formatter:

$$x_{s,d_0}^{(\tau)} = \phi_\tau(D_{s,d_0}), \qquad \tau \in \mathcal{T}, \tag{1}$$

where text serializes the OHLCV records, chart renders the candlestick chart with moving averages and volume bars, and multi provides both.

Given $x_{s,d_0}^{(\tau)}$, the evaluated VLM M produces

$$r_{s,d_0}^{(\tau)} = M\left(x_{s,d_0}^{(\tau)}\right) = (y, \rho, a, c), \tag{2}$$

where y is the technical analysis, ρ is the decision rationale, $a \in \{\text{BUY}, \text{SELL}, \text{HOLD}, \text{UNCERTAIN}\}$ is the action, and $c \in \{1, \ldots, 5\}$ is the confidence score. The task does not assign a ground-truth trading action; instead, it exposes the response components needed to evaluate evidence-backed recommendations.

3.3 Evidence Anchor Construction

Financial charts do not provide a unique ground-truth trading action. We therefore construct reference evidence at the chart-fact level. For each market window D_{s,d_0}, E2A-Bench applies a deterministic fact function:

$$F_{s,d_0} = g(D_{s,d_0}), \tag{3}$$

where F_{s,d_0} denotes OHLCV-derived technical facts used as evidence anchors.

As summarized in Table 2, the anchors cover common technical-analysis dimensions, including moving averages, price–volume behavior, trend,

Fact Key	Compact Definition
MA alignment	Ordering of MA5, MA10, and MA20 on the last day
MA cross	Golden/death cross of MA5 and MA10 in the last 5 days
Volume change	Current volume relative to the 20-day mean
Price breakout	Close price relative to the 20-day high–low band
Trend direction	Linear-regression slope of 20-day close prices
Volatility	Annualized volatility of recent returns
Recent 5-day change	Five-day cumulative price movement
Vol–price divergence	Directional relation between recent price and volume changes
Price position	Close percentile in the 20-day high–low range
Short-term momentum	Three-day price acceleration pattern

Table 2: Compact summary of the ten deterministic OHLCV-derived evidence anchors used in E2A-Bench.

volatility, range position, and short-term momentum. Each anchor is computed directly from open, high, low, close, and volume records using fixed rules. These anchors serve two purposes: they provide reference facts for verifying model-generated technical claims, and they define evidence strength for confidence calibration. We additionally audit 200 randomly sampled chart–label pairs and observe no inconsistency between the rendered charts and their OHLCV-derived labels.

Importantly, F_{s,d_0} is not a trading-action label. A BUY or SELL recommendation is not evaluated by future return, but by whether the model's stated claims, rationale, and confidence are consistent with verifiable chart evidence. This design separates evidence-to-action reliability from market prediction while keeping the evaluation reproducible.

3.4 Evaluation Protocol

Overview. E2A-Bench evaluates each response as an evidence-to-action chain rather than as isolated generated text. As illustrated in Fig. 2, the protocol parses a model response, verifies its claims against deterministic evidence anchors, and scores whether evidence remains reliable through the decision process. Given evidence anchors F_{s,d_0} and a response $r = (y, \rho, a, c)$, the protocol checks four

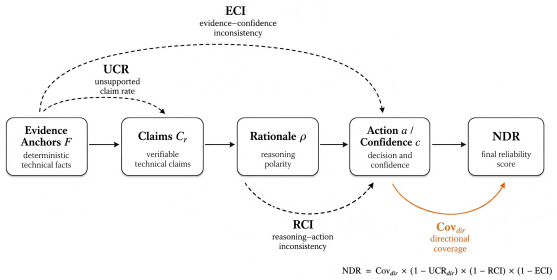


Figure 2: Evidence-to-action scoring protocol in E2A-Bench. A model response is parsed into claims, rationale polarity, action, and confidence; claims are verified against deterministic OHLCV evidence anchors; and the response is scored along grounding, reasoning–action consistency, evidence–confidence calibration, and directional coverage.

reliability links:

$$\begin{aligned} F_{s,d_0} \to \mathcal{C}_r &: \text{evidence grounding}, \\ \rho \to a &: \text{reasoning–action consistency}, \\ F_{s,d_0} \to (a, c) &: \text{evidence–confidence calibration}, \\ a \in \mathcal{A}_{\text{dir}} &: \text{directional coverage}. \end{aligned} \tag{4}$$

The first three links define local diagnostic errors, while directional coverage accounts for whether the model preserves actionable outputs. Together, they instantiate the coverage–grounding frontier: a model is reliable only when it acts under sufficient evidence and keeps its action grounded, coherent, and calibrated.

Response parsing. Before scoring, each response is normalized into the variables used by the diagnostic protocol:

$$\begin{aligned} \Pi(r) &= (\mathcal{C}_r, \sigma_r, a, c), \\ \mathcal{C}_r &= \text{Claims}(y, \rho), \qquad \sigma_r = \sigma(\rho). \end{aligned} \tag{5}$$

where $\mathcal{C}_r$ is the set of verifiable technical claims extracted from the generated analysis and rationale, with each claim mapped to an evidence anchor. The variable $\sigma_r \in \{\text{bullish}, \text{bearish}, \text{neutral}, \text{unknown}\}$ denotes the coarse polarity of the rationale. The parsed action a and confidence score c are used for action-level reliability scoring. This parsing step exposes response components for evidence-grounded evaluation; it does not judge trading profitability.

Local diagnostic errors. Using the parsed tuple $\Pi(r)$ and the matched evidence anchors $F = F_{s,d_0}$, we define three response-level diagnostics for the first three links of the evidence-to-action chain. Let $\mathbf{1}[\cdot]$ denote the indicator function and let $\mathcal{A}_{\text{dir}} = \{\text{BUY}, \text{SELL}\}$.

Unsupported Claim Rate (UCR) measures whether the model's verifiable technical claims are supported by the evidence anchors. Let $\mathrm{Supp}(q, F)$ indicate whether claim $q \in \mathcal{C}_r$ is supported by its corresponding anchor. We define

$$\mathrm{UCR}(r; F) = \frac{\sum_{q \in \mathcal{C}_r} \mathbf{1}[\neg \mathrm{Supp}(q, F)]}{\max(|\mathcal{C}_r|, 1)}. \quad (6)$$

When no verifiable claim is extracted, this convention sets $\mathrm{UCR} = 0$; such cases are not treated as reliable decisions, but are handled by the coverage term.

Reasoning–Action Consistency Inconsistency (RCI) measures whether the final action contradicts the polarity of the stated rationale:

$$\mathrm{RCI}(r) = \mathbf{1}[(\sigma_r, a) \in \{(\mathrm{bullish}, \mathrm{SELL}), (\mathrm{bearish}, \mathrm{BUY})\}]. \quad (7)$$

Evidence–Confidence Inconsistency (ECI) measures high-confidence directional action under weak evidence. Let $S(F)$ denote the number of unambiguous directional or extreme signals in the evidence anchors, as specified in App. A. We define

$$\mathrm{ECI}(r; F) = \mathbf{1}[S(F) = 0 \wedge a \in \mathcal{A}_{\mathrm{dir}} \wedge c \geq 4]. \quad (8)$$

Together, UCR, RCI, and ECI diagnose grounding failure, reasoning–action inconsistency, and evidence–confidence miscalibration, respectively.

Coverage-aware reliability. The local diagnostics evaluate the quality of directional decisions, but they do not account for whether the model acts only under sufficient chart evidence. For a response r with evidence anchors F_r, let $S(F_r)$ denote the number of unambiguous directional or extreme signals. We define the evidence gate

$$g_\theta(F_r) = \mathbf{1}[S(F_r) \geq \theta], \quad (9)$$

where θ is the minimum evidence strength required for a directional action to count toward reliable coverage. For a response set $\mathcal{R}$ under a fixed model, modality, and method, the evidence-gated directional coverage is

$$\mathrm{Cov}_{\mathrm{dir},\theta} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \mathbf{1}[a_r \in \mathcal{A}_{\mathrm{dir}} \wedge g_\theta(F_r) = 1]. \quad (10)$$

We compute $\mathrm{UCR}_{\mathrm{dir}}$ by averaging response-level UCR over directional responses. By convention, RCI and ECI also denote directional-conditioned averages unless otherwise specified. We then define thresholded *Net Decision Reliability* as

$$\mathrm{NDR}_\theta = \underbrace{\mathrm{Cov}_{\mathrm{dir},\theta}}_{\text{coverage}} \cdot \underbrace{(1 - \mathrm{UCR}_{\mathrm{dir}})\,(1 - \mathrm{RCI})\,(1 - \mathrm{ECI})}_{\text{chain reliability}}. \quad (11)$$

Statistic	Value
Stocks	323 HS300 constituents
Input modalities	text-only, chart-only, multimodal
Benchmark queries	969
Evidence anchors	3,230 total facts
Input token length, mean ± std.	text-only: 1628.8 ± 61.2; chart-only: 1027.0 ± 0.0; multimodal: 2466.8 ± 61.2

Table 3: Benchmark-level statistics of `E2A-Bench`.

The main paper uses $\theta = 1$, written as NDR for brevity, so that a directional action contributes to reliable coverage only when at least one strong chart signal is present. We analyze the sensitivity of NDR_θ to alternative evidence thresholds in App. C.2. The multiplicative form reflects a necessary-condition view of reliable action: a decision is not reliable if it is absent, unsupported, inconsistent with its rationale, or overconfident under weak evidence. It should not be interpreted as assuming statistical independence among the factors, nor as measuring realized trading returns.

Reporting convention. We report $\mathrm{UCR}_{\mathrm{all}}$ as the average unsupported-claim rate over all responses, which is the closest analogue to a conventional hallucination score. We also report $\mathrm{UCR}_{\mathrm{dir}}$ to measure grounding quality when the model commits to a directional action. In contrast, RCI and ECI are reported on directional responses by default, since their operational meaning concerns action-level consistency and calibration. Non-directional outputs such as HOLD and UNCERTAIN reduce $\mathrm{Cov}_{\mathrm{dir},\theta}$ rather than mechanically lowering consistency or calibration errors. If no directional response is produced, we set $\mathrm{NDR}_\theta = 0$ and mark directional-conditioned diagnostics as not applicable. This convention separates evidence-backed action from conservative non-action, which scalar hallucination rates can conflate.

3.5 Validation and Benchmark Statistics

Scorer validation. `E2A-Bench` uses deterministic OHLCV-derived evidence anchors, so human annotation is not used to define trading-action ground truth. We instead use professional annotations from anonymous industry practitioners to validate the automatic scorer. Annotators judge raw model responses using plain-language questions aligned with UCR, RCI, and ECI. The annotations show substantial inter-annotator agreement, and the scorer matches professional consensus la-

Model	Size	Cov. ↑	UCR_{all} ↓	UCR_{dir} ↓	RCI ↓	ECI ↓	NDR ↑
Proprietary							
Gemini-3.1-Pro	–	**59.0**	70.0	67.5	5.2	7.7	**16.8**
Gemini-3.1-Flash	–	37.9	58.1	53.8	2.7	6.3	16.0
Claude-Sonnet-4.6	–	55.1	71.5	69.4	3.6	1.9	16.0
GPT-5.5	–	25.9	61.2	47.6	**2.0**	4.0	12.8
GPT-5.4	–	15.2	65.4	54.4	2.7	4.1	6.5
GPT-5.4-mini	–	8.3	**55.9**	**46.5**	2.5	**1.2**	4.3
Open-Weight							
Qwen2.5-VL-3B-Instruct[a]	3B	**82.5**	53.1	55.0	10.9	2.9	**32.1**
Qwen2.5-VL-7B-Instruct[b]	7B	53.9	47.9	48.5	4.4	5.7	25.0
Gemma-4-31B-It	31B	61.5	62.8	59.3	4.4	6.0	22.5
Qwen3.5-4B	4B	52.7	59.1	54.6	12.3	3.7	20.2
Qwen3.5-9B	9B	64.2	65.0	63.6	14.5	5.1	19.0
Qwen3.5-2B	2B	54.0	67.1	61.3	8.4	2.7	18.6
Qwen3-VL-4B-Thinking[c]	4B	71.9	75.2	74.1	6.6	8.8	15.9
Qwen3.5-27B	27B	48.1	70.7	69.3	13.1	1.9	12.6
Gemma-4-E4B-It	4B	20.7	60.2	52.9	4.5	**0.0**	9.4
Gemma-4-E2B-It	2B	6.4	**39.9**	**14.5**	**1.6**	4.8	5.1
Financial Fine-Tuned							
PyFi-QwenVL-3B-COT-47K[a]	3B	62.4	**40.5**	**41.8**	26.8	**5.0**	**25.3**
PyFi-QwenVL-7B-COT-47K[b]	7B	49.5	50.7	54.1	10.6	12.5	17.8
Amsi-fin-o1[c]	4B	**81.3**	69.2	73.7	**5.3**	14.5	17.3
FinLLaVA	7B	36.9	50.7	47.2	15.9	15.1	13.9

Table 4: Main trimodal leaderboard on `E2A-Bench`. All values are percentages averaged over the three input modalities. Cov. denotes directional coverage, and NDR denotes coverage-aware net decision reliability. Within each group, **bold** and underlined entries mark the best and second-best values. Superscripts indicate strict base–fine-tuned pairs: [a]Qwen2.5-VL-3B, [b]Qwen2.5-VL-7B, and [c]Qwen3-VL-4B-Thinking.

bels with high reliability. Detailed protocols and agreement statistics are reported in App. A.5.

Benchmark statistics. Table 3 reports the reusable benchmark structure: stock universe, input modalities, benchmark queries, evidence anchors, and input-token lengths. These statistics characterize the scale and input context of `E2A-Bench`; additional distributional details are provided in App. A.3.

4 Experiments and Findings

4.1 Setup

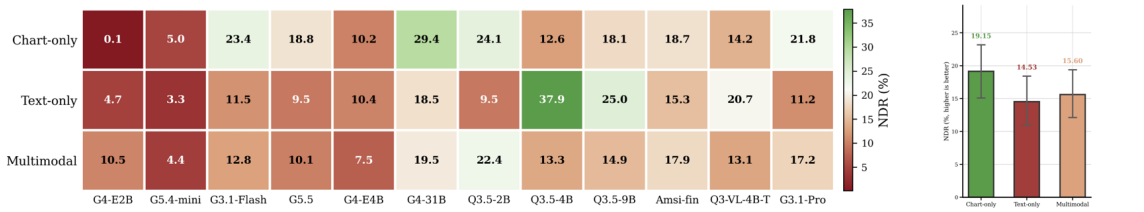
Models. We evaluate 20 VLMs across three families: proprietary systems, open-weight models, and financial fine-tunes. The proprietary set includes Gemini-3.1-Pro/Flash (Google DeepMind, 2026a, 2025), GPT-5.5/5.4/5.4-mini (OpenAI, 2026a,b), and Claude-Sonnet-4.6 (Anthropic, 2026). The open-weight set includes Qwen2.5-VL-3B/7B-Instruct (Bai et al., 2025), Qwen3-VL-4B-Thinking (Qwen Team, 2025), Qwen3.5-2B/4B/9B/27B (Qwen Team, 2026), and Gemma-4-E2B/E4B/31B-It (Google DeepMind, 2026b). The financial set includes Amsi-fin-o1 (AITRADER, 2025), FinLLaVA (Xie et al., 2024), and PyFi-QwenVL-3B/7B-COT-47K (Zhang et al., 2025). Details are reported in App. C.

4.2 Overall Performance

Table 4 reports the main trimodal leaderboard across 20 VLMs. Overall, reliability is better characterized by the coverage–grounding frontier than by a scalar hallucination score.

Low hallucination does not imply reliable action. Qwen2.5-VL-3B-Instruct achieves the highest NDR (32.1%) with high directional coverage (82.5%), whereas Gemma-4-E2B-It has the lowest UCR_{all} (39.9%) but only 6.4% coverage and NDR = 5.1%. This result shows that low hallucination can reflect conservative non-action rather than evidence-backed decision reliability.

Model families show distinct reliability profiles. Open-weight models achieve the strongest overall reliability, with Qwen2.5-VL-3B/7B ranking first and second by NDR. Proprietary models are comparatively conservative: the best proprietary model, Gemini-3.1-Pro, reaches only 16.8% NDR despite



(a) Baseline NDR heatmap across representative models and input modalities. (b) Mean baseline NDR by modality.

Figure 3: Input modality effects on baseline NDR. *(a)* Model-wise NDR across text-only, chart-only, and multimodal inputs. *(b)* Mean NDR by modality with 95% bootstrap confidence intervals. Modality effects are model-dependent; chart-only input achieves the highest average baseline NDR.

moderate coverage. Financially adapted models are competitive but uneven; PyFi-QwenVL-3B reaches 25.3% NDR, while other fine-tuned models are limited by reasoning–action inconsistency, calibration error, or grounding loss.

Reliable action requires balancing coverage and chain quality. The strongest models do not minimize one diagnostic error in isolation; they preserve actionable coverage while keeping grounding, reasoning–action consistency, and calibration errors controlled. Financially adapted models illustrate this trade-off: PyFi-QwenVL-3B-COT-47K reaches a competitive NDR (25.3%), but suffers from a high reasoning–action inconsistency rate, whereas Amsi-fin-o1 maintains high coverage (81.3%) but is limited by weaker grounding and calibration. Among proprietary models, Gemini-3.1-Pro obtains the highest NDR (16.8%), yet remains below the best open-weight and financially adapted systems under our protocol. Overall, the leaderboard supports the central premise of `E2A-Bench`: decision reliability should be evaluated by movement along the coverage–grounding frontier, not by unsupported-claim rate alone.

4.3 Diagnostic Analyses

4.3.1 Input Modality Effects

Input modality affects NDR in a model-dependent manner. As shown in Fig. 3, the best input form differs by model: Qwen3.5-4B peaks with text-only input (37.9%), while Gemma-4-31B and Qwen3.5-2B peak with chart-only input (29.4% and 24.1%). On average, chart-only input gives the strongest baseline NDR (19.15%), outperforming text-only (14.53%) and multimodal input (15.60%). Therefore, direct chart access can improve evidence-to-action reliability, whereas multimodal input does not consistently yield better decisions, motivating modality-specific reporting.

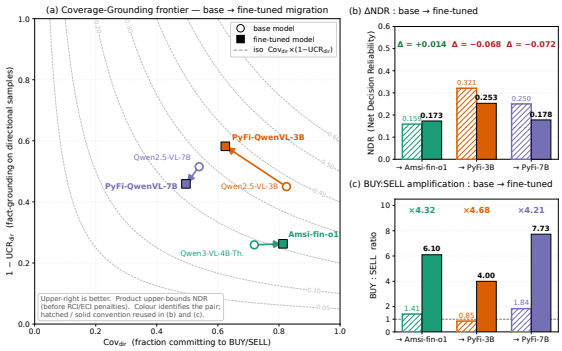


Figure 4: **Risk redistribution under financial fine-tuning.** *(a)* Base-to-fine-tuned migration on the coverage–grounding plane, where the dashed curves show iso-$\mathrm{Cov}_{\mathrm{dir}}(1-\mathrm{UCR}_{\mathrm{dir}})$ contours. *(b)* Fine-tuning produces inconsistent NDR changes across pairs. *(c)* All pairs show strong BUY:SELL amplification, indicating a consistent shift toward BUY actions despite heterogeneous reliability changes.

4.3.2 Fine-Tuning Redistributes Risk

Fig. 4 compares strict base–fine-tuned pairs by their coverage–grounding migration, NDR change, and BUY:SELL shift.

Reliability gains are base-dependent. Financial fine-tuning moves different models in different directions on the frontier. Amsi-fin-o1 improves NDR modestly (+0.014), mainly through higher directional coverage with nearly unchanged grounding. PyFi-QwenVL-3B improves grounding but loses coverage, while PyFi-QwenVL-7B moves downward and leftward; both decrease NDR (−0.068 and −0.072). Thus, financial fine-tuning does not yield a uniform reliability gain, but redistributes risk across coverage, grounding, coherence, and calibration.

BUY-side amplification is consistent. Despite heterogeneous NDR changes, all three fine-tuned models shift strongly toward BUY actions. The BUY:SELL ratio increases from 1.41 to 6.10, from 0.85 to 4.00, and from 1.84 to 7.73, corresponding to 4.21–4.68× amplification. This stable action-distribution shift suggests that financial

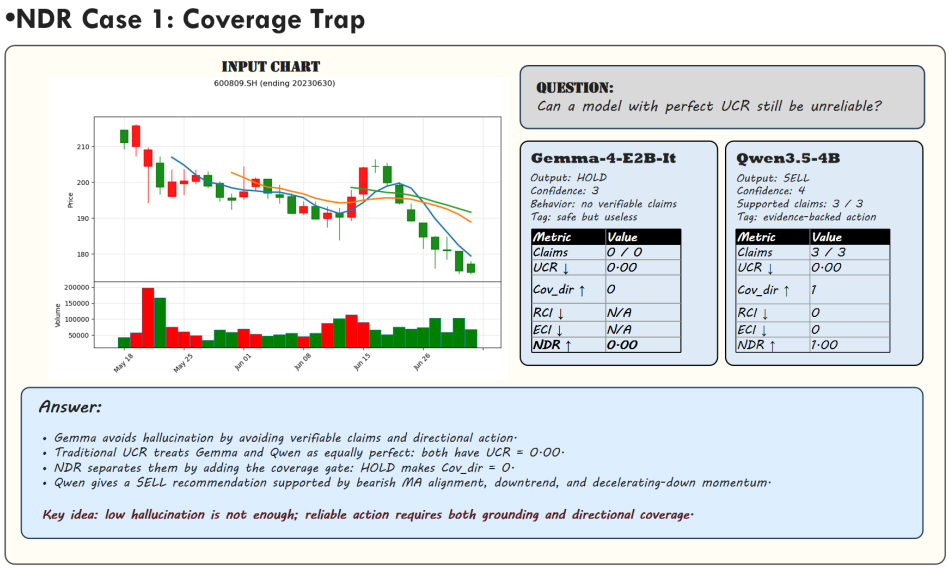


Figure 5: Coverage trap in claim-level hallucination evaluation. Gemma-4-E2B-It avoids verifiable technical claims and outputs a non-directional HOLD decision, yielding zero observed UCR but no actionable coverage. In contrast, Qwen3.5-4B issues a directional SELL recommendation whose claims are supported by the evidence anchors. This case illustrates why claim-level hallucination scores alone can overestimate conservative non-action, while NDR distinguishes non-action from evidence-backed decision reliability.

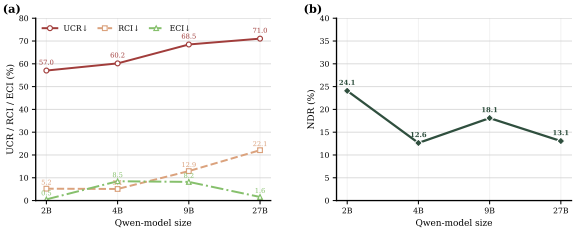


Figure 6: Scaling behavior within Qwen3.5 under chart-only inputs. *(a)* Local diagnostics (UCR, RCI, ECI) show non-monotonic and partially conflicting trends across model sizes. *(b)* NDR also varies non-monotonically, peaking at 2B.

adaptation should be audited not only for aggregate reliability, but also for directional bias.

4.3.3 Scaling Is Not Monotonic

Larger Qwen3.5 models do not consistently improve chain-level reliability. We analyze chart-only inputs, a common setting for financial chart reasoning, while holding the model family fixed. As shown in Fig. 6(a), scaling from 2B to 27B does not monotonically improve the local diagnostics. UCR increases from 57.0 to 71.0, suggesting weaker factual grounding at larger scales, while RCI and ECI follow non-monotonic and partially conflicting trends. Thus, larger models are not uniformly better at preserving grounding, reasoning–action consistency, and confidence calibration.

NDR follows a separate scaling pattern. Fig. 6(b) shows that NDR peaks at 2B (24.1), drops at 4B (12.6), partially recovers at 9B (18.1), and declines again at 27B (13.1). This trajectory differs from any single local diagnostic and indicates that net decision reliability is shaped by the interaction between coverage and chain quality. The result reinforces a central point of E2A-Bench: parameter scale alone is not a reliable proxy for evidence-to-action reliability.

4.3.4 Qualitative Illustration

Fig. 5 illustrates the coverage trap behind the rank reversal. Gemma-4-E2B-It obtains zero observed UCR by avoiding verifiable claims and issuing HOLD, whereas Qwen3.5-4B gives an evidence-backed SELL recommendation. This contrast shows why NDR must account for directional coverage rather than hallucination avoidance alone.

5 Conclusion

We introduced E2A-Bench, a coverage-aware evidence-to-action benchmark for financial VLM reliability. By tracing chart evidence through claims, rationale, confidence, and action, E2A-Bench reveals failures hidden by scalar hallucination scores. Our results show that low hallucination can reflect non-action, verification can collapse coverage, and financial fine-tuning can amplify BUY-side bias. These findings suggest that financial VLM evaluation should account for the coverage–grounding frontier, rather than relying on hallucination reduction alone.

Limitations

E2A-Bench is built on HS300 constituents at a fixed evaluation date, so it does not yet cover different market regimes, asset classes, or cross-market trading conventions. This controlled snapshot improves comparability across models, but future versions should extend the benchmark to bull, bear, and sideways regimes, as well as to broader equity markets and other financial instruments. Such extensions would test whether evidence-to-action reliability remains stable under changing volatility, liquidity, and macro-market conditions.

E2A-Bench relies on deterministic OHLCV-derived evidence anchors and pattern-based claim detectors to ensure reproducible and auditable evaluation. This design avoids subjective trading-action labels, but it may miss rare paraphrases, nuanced technical claims, or domain-specific reasoning patterns not covered by the current detector. Future work can improve coverage by expanding multilingual claim patterns, incorporating more robust semantic parsers, and adding professional adjudication for ambiguous cases, while preserving the core principle of grounding evaluation in verifiable evidence.

References

AITRADER. 2025. Amsi-fin-o1 MLX 8-bit: Quantized financial vision-language model for apple silicon. Hugging Face model card. https://huggingface.co/AITRADER/Amsi-fin-o1-MLX-8bit. Accessed: 2026-05-25.

Anthropic. 2026. Introducing Claude Sonnet 4.6. Model release. https://www.anthropic.com/news/claude-sonnet-4-6. Accessed: 2026-05-25.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. *CoRR*, abs/2502.13923.

Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. 2024. Large language model agent in financial trading: A survey. *arXiv preprint arXiv:2408.06361*.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6577–6595. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 Flash model card. Model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf. Accessed: 2026-05-25.

Google DeepMind. 2026a. Gemini 3.1 Pro model card. Model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-05-25.

Google DeepMind. 2026b. Gemma 4 model overview. Technical documentation. https://ai.google.dev/gemma/docs/core. Accessed: 2026-05-25.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal LLM for chart understanding and generation. *CoRR*, abs/2311.16483.

Kung-Hsiang Huang, Hou Pong Chan, Yi R. Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2025. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Transactions on Knowledge and Data Engineering*.

Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. 2025. Large language models in finance (FinLLMs). *Neural Computing and Applications*, 37(30):24853–24867.

Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023a. Deplot: One-shot visual language reasoning by plot-to-table translation. In *ACL (Findings)*, Findings of ACL, pages 10381–10399. Association for Computational Linguistics.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. 2023b. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. In *ACL (1)*, pages 12756–12770. Association for Computational Linguistics.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025a. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Zhaowei Liu, Xin Guo, Haotian Xia, Lingfeng Zeng, Fangqi Lou, Jinyi Niu, Mengping Li, Qi Qi, Jiahuan Li, Wei Zhang, Yinglong Wang, Weige Cai, Weining Shen, and Liwen Zhang. 2025b. Visfineval: A scenario-driven chinese multimodal benchmark for holistic financial understanding. In *EMNLP*, pages 24088–24146. Association for Computational Linguistics.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL (Findings)*, Findings of ACL, pages 2263–2279. Association for Computational Linguistics.

OpenAI. 2026a. Introducing GPT-5.4. Model release. https://openai.com/index/introducing-gpt-5-4/. Accessed: 2026-05-25.

OpenAI. 2026b. Introducing GPT-5.5. Model release. https://openai.com/index/introducing-gpt-5-5/. Accessed: 2026-05-25.

Qwen Team. 2025. Qwen3-VL technical report. *CoRR*, abs/2511.21631.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. Official blog. https://qwen.ai/blog?id=qwen3.5. Accessed: 2026-05-25.

Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. 2025. Halogen: Fantastic LLM hallucinations and where to find them. In *ACL (1)*, pages 1402–1425. Association for Computational Linguistics.

Dong Shu, Haoyang Yuan, Yuchen Wang, Yanguang Liu, Huopu Zhang, Haiyan Zhao, and Mengnan Du. 2025. Finchart-bench: Benchmarking financial chart comprehension in vision-language models. *CoRR*, abs/2507.14823.

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. *CoRR*, abs/2509.17481.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. In *NeurIPS*.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L. Gwet. 2013. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13:61.

Qianqian Xie, Dong Li, Mengxi Xiao, Zihao Jiang, Ruoyu Xiang, Xiao Zhang, Zhengyu Chen, Yueru He, Weiguang Han, Yuzhe Yang, Shunian Chen, Yifei Zhang, Lihang Shen, Daniel S. Kim, Zhiwei Liu, Zheheng Luo, Yangyang Yu, Yupeng Cao, Zhiyang Deng, and 20 others. 2024. Open-FinLLMs: Open multimodal large language models for financial applications. *CoRR*, abs/2408.11878.

Siqiao Xue, Tingting Chen, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. 2024. FAMMA: A benchmark for financial domain multilingual multimodal question answering. *CoRR*, abs/2410.04526.

Yuqun Zhang, Yuxuan Zhao, and Sijia Chen. 2025. PyFi: Toward pyramid-like financial image understanding for VLMs via adversarial agents. *CoRR*, abs/2512.14735.

Appendix

A Detailed Benchmark Specification

This appendix expands the compact evidence-anchor summary in Table 2. We specify the label spaces, deterministic computation rules, and strong-signal indicators used for evidence-confidence evaluation. All anchors are computed from the 30-trading-day OHLCV window and serve as evidence labels rather than trading-action labels.

A.1 Execution Records and Re-run Consistency

The core benchmark contains 969 unique stock–modality queries. Additional archived response records may arise from chart-only intervention analyses or paired baseline executions. These records are used for experimental comparison and consistency checking, but are not counted as additional benchmark queries.

For consistency checking, we compare aggregate metrics with and without paired chart-only baseline executions. Deduplication changes UCR, RCI, and ECI by at most 0.4 percentage points and changes NDR by at most 0.005; model rankings remain unchanged. Thus, the storage convention does not affect the substantive conclusions.

A.2 Strong-Signal Rules for Evidence–Confidence Evaluation

ECI penalizes high-confidence directional actions when the chart evidence is weak. To estimate evidence strength, `E2A-Bench` computes a strong-signal count $S(F)$ from the evidence anchors. As specified in Table 5, only directional or extreme labels are counted as strong signals. Descriptive anchors may still be used for claim verification, but do not by themselves justify high-confidence directional action.

Excluded anchors. Not all evidence anchors in Table 7 contribute to $S(F)$. MA alignment, volatility, and recent 5-day change are excluded from the strong-signal set because they are descriptive or slowly varying signals rather than standalone directional triggers. MA alignment may overlap with trend direction, volatility is not directional by itself, and recent 5-day change is treated as contextual evidence. All three anchors remain available for claim verification under UCR.

Signal Type	Qualifying Labels
MA cross	`golden_cross`, `death_cross`
Price breakout	`breakout_high`, `breakout_low`
Volume anomaly	`heavy_volume`, `light_volume`
Strong trend	`uptrend` or `downtrend` with $\lvert\text{slope}\rvert > 0.30\%$/day
Vol–price divergence	`bullish_divergence`, `bearish_divergence`
Extreme price position	`high_zone`, `low_zone`
Momentum acceleration	`accelerating_up`, `accelerating_down`

Table 5: Evidence-anchor labels counted as strong signals for $S(F)$. A sample has weak evidence when $S(F) = 0$; ECI fires when such a sample receives a high-confidence directional action.

Quantity	Value
Stocks	323
Input conditions	3
Benchmark queries	969
Evaluated models	20
Completed response cells	19,380
Completion rate	100%
Max scoring slots / query	31
Scoring-slot capacity	30,039
Realized verifiable judgments	112,369

Table 6: Core trimodal benchmark and scoring inventory. Chart-only intervention settings are treated as experimental conditions and are not counted as additional core benchmark queries.

A.3 Benchmark Inventory and Output Distribution

Table 6 summarizes the core trimodal inventory of `E2A-Bench`. The core benchmark contains 323 stocks under three input modalities, yielding 969 benchmark queries and 19,380 completed response cells across 20 models. Each query exposes up to 31 scoring slots: at most 27 verifiable claim atoms plus RCI, ECI, action, and confidence.

Fig. 7 shows the final-action distribution over evaluated model outputs. Non-directional responses dominate: HOLD and UNCERTAIN account for 63.6% of parsed actions. This motivates coverage-aware scoring, since low unsupported-claim rates can coincide with limited directional action.

A.4 Full Specification of Evidence Anchors

Table 7 lists the ten OHLCV-derived evidence anchors used by `E2A-Bench` for claim verification and evidence-strength estimation.

#	Fact Key	Dimension	Inputs	Computation Rule	Output Labels	ECI Signal
1	MA alignment	MA stacking	Close	Compare MA5, MA10, and MA20 on the last day.	`bullish`, `bearish`, `mixed`	No
2	MA cross	Short-term crossover	Close	Detect golden/death cross of MA5 and MA10 within the last 5 days.	`golden_cross`, `death_cross`, `no_cross`	Yes
3	Volume change	Volume anomaly	Volume	Compare current volume with the 20-day mean; heavy if $> 1.5\times$, light if $< 0.7\times$.	`heavy_volume`, `light_volume`, `normal_volume`	Yes
4	Price breakout	Range breakout	Close, High, Low	Compare the close price with the 20-day high–low band.	`breakout_high`, `breakout_low`, `within_range`	Yes
5	Trend direction	Price trend	Close	Fit a linear trend to 20-day close prices; use 0.15%/day as the trend-label threshold.	`uptrend`, `downtrend`, `sideways`	Yes†
6	Volatility	Return variability	Close	Compute annualized standard deviation of recent returns.	`high_volatility`, `medium_volatility`, `low_volatility`	No
7	Recent 5-day change	Short-window return	Close	Compute 5-day cumulative price change; significant if $\lvert\Delta\rvert > 3\%$.	`significant_rise`, `significant_drop`, `flat`	No
8	Vol–price divergence	Price-volume relation	Close, Volume	Compare 10-day price change with recent volume change.	`bullish_divergence`, `bearish_divergence`, `volume_price_sync_up`, `volume_price_sync_down`, `no_divergence`	Yes
9	Price position	Range position	Close, High, Low	Compute the close percentile in the 20-day high–low range; high if ≥ 0.80, low if ≤ 0.20.	`high_zone`, `mid_zone`, `low_zone`	Yes
10	Short-term momentum	Momentum acceleration	Close	Compare recent daily change with short-window average change.	`accelerating_up`, `decelerating_up`, `flat_momentum`, `decelerating_down`, `accelerating_down`	Yes

Table 7: Full specification of the ten deterministic OHLCV-derived evidence anchors. "Strong" indicates whether the anchor contributes to the strong-signal count $S(F)$ used by ECI. †Trend direction uses 0.15%/day for label assignment and a stricter 0.30%/day threshold for strong-signal counting.

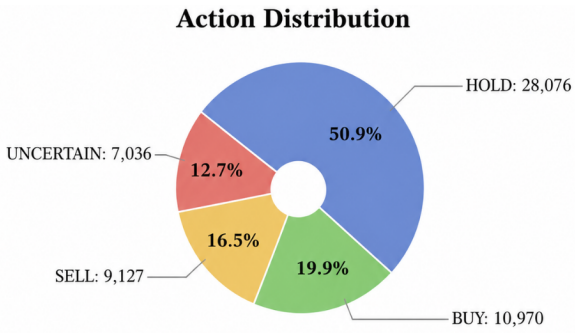


Figure 7: **Final-action distribution of evaluated model outputs.** HOLD and UNCERTAIN account for 63.6% of parsed actions, motivating coverage-aware evaluation beyond claim-level hallucination.

Diagnostic	AC1	Precision	Recall	F1
UCR	0.71	91.3%	88.7%	90.0%
RCI	0.74	87.6%	84.2%	85.9%
ECI	0.82	93.1%	90.5%	91.8%

Table 8: Human validation of the automatic scorer. AC1 measures inter-annotator agreement among four anonymous industry annotators; precision, recall, and F1 are computed against human-consensus labels.

A.5 Human Validation of the Automatic Scorer

Annotation protocol. We validate the automatic scorer using professional annotations from four anonymous industry practitioners with finance-related experience. The annotators participated under written agreements and were informed of the research purpose, intended data usage, annotation procedure, and confidentiality requirements. Human annotation is used only to validate the response-level scorer; it is not used to define ground-truth trading actions.

Annotators were shown the rendered candlestick chart and the raw model response. They were not shown model identities, test-time methods, automatic labels, or metric names. Each response was judged using three plain-language questions: whether the analysis contains unsupported chart claims, whether the reasoning contradicts the final recommendation, and whether the expressed confidence is appropriate given the chart evidence. These questions correspond to UCR, RCI, and ECI, respectively.

Agreement and scorer performance. We compute inter-annotator agreement on the overlapping subset using Gwet's AC1. We use Gwet's AC1 rather than Cohen's κ because some error types

have low positive-class prevalence, where κ can be sensitive to marginal distributions (Wongpakaran et al., 2013). The four anonymous industry annotators show substantial agreement across all three diagnostics, with AC1 scores of 0.71 for UCR, 0.74 for RCI, and 0.82 for ECI. We then evaluate the automatic scorer against human-consensus labels. As shown in Table 8, all three diagnostics exceed 85% F1, indicating that the scorer is a reliable proxy for large-scale evidence-to-action evaluation.

The scorer performs strongest on ECI, where the signal-count rule is comparatively deterministic. RCI has the lowest recall, reflecting cases with implicit concessive reasoning or weak polarity cues.

A.6 Prompt Template and Language Setting

The raw prompts and model responses in `E2A-Bench` are in Simplified Chinese, matching the A-share market setting and the HS300 stock universe. For readability, paper examples are translated into English, while all metric computation is performed on the original model outputs. The framework is language-flexible and can be adapted by replacing the prompt template, response parser, and claim-pattern detectors. Fig. 8 shows an English translation of the base inference prompt used in the standard financial chart reasoning task.

A.7 License

We will release the code, evaluation scripts, more prompt templates, processed benchmark metadata, and rendered chart images upon acceptance. The released artifacts will be distributed under the MIT License. The chart images in `E2A-Bench` are rendered from publicly available OHLCV market data and do not contain personal or user-generated information. Raw market data redistribution, if restricted by the original data provider, will be handled by releasing data-construction scripts and source specifications rather than redistributing restricted raw records.

B Scoring Implementation Details

Indicator notation. We use $\mathbf{1}[\cdot]$ to denote the indicator function, which returns 1 if the condition is true and 0 otherwise.

Claim support for UCR. Each extracted technical claim is mapped to one deterministic evidence anchor. A claim is counted as unsupported if it contradicts the matched anchor or asserts a technical signal that is absent from the anchor. When no verifiable claim is extracted, response-level UCR is set to zero by convention; such responses are not treated as reliable actions, but are penalized through the coverage term in NDR.

Rationale polarity for RCI. For RCI, the rationale is mapped to a coarse polarity in $\{\text{bullish}, \text{bearish}, \text{neutral}, \text{unknown}\}$. RCI fires only when a directional rationale explicitly contradicts the final directional action, namely bullish reasoning with SELL or bearish reasoning with BUY. Neutral or unknown rationales, as well as non-directional actions, are not counted as reasoning–action consistency violations.

Evidence strength for ECI. For ECI, evidence strength is computed from the strong-signal count $S(F)$ defined in App. A.2. Weak evidence corresponds to $S(F) = 0$, and high confidence corresponds to $c \geq 4$ on the five-point confidence scale. Thus, ECI fires when a model issues a high-confidence directional action under no strong veri-

Base inference prompt

You are a professional A-share technical analyst. Please perform technical analysis based on the provided stock data and give a trading recommendation.

Important: Only describe information that can be directly observed from the chart or data. Do not list signals that cannot be verified.

Please strictly follow the output format below:

[Technical Analysis]
Use 3–5 sentences to describe the recent technical trend of the stock. Only include observations that can be directly read from the chart or data. Do not use a fixed template or mechanically enumerate signals.

[Rationale]
Based on the technical analysis above, provide the reasoning behind your overall judgment.

[Trading Recommendation]
Action: BUY / SELL / HOLD (choose exactly one)
Confidence: 1–5 (1 = very uncertain, 5 = very certain)

Please answer in Simplified Chinese.

Figure 8: English translation of the base inference prompt. The original prompt is written in Simplified Chinese and used for model inference; examples in the paper are translated for readability.

Threshold	Diagnostic interpretation
$\theta = 0$	No evidence gate. Any directional action can contribute to coverage, regardless of strong-signal count.
$\theta = 1$	Main-paper setting. A directional action contributes only when at least one strong chart signal is present.
$\theta > 1$	Stricter evidence gate. Coverage decreases, and retained actions correspond to denser chart evidence.
$\theta \to \infty$	Conservative endpoint. Coverage approaches zero for all models. The rate of decay reflects how often a model issues directional actions under increasingly dense evidence.

Table 9: Operational interpretation of the evidence threshold θ in NDR_θ. The threshold changes only the aggregation rule over existing response logs.

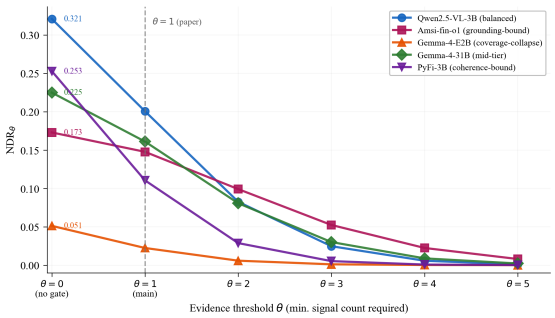


Figure 9: **NDR sensitivity to evidence threshold θ.** *(a)* NDR_θ curves for five representative models as θ increases from 0 (no evidence gate) to 5 (strict gate). The dashed vertical line marks $\theta = 1$, the main-paper setting. Coverage-collapse models (Gemma-4-E2B, orange) start low at $\theta = 0$ and decay quickly; grounding-bound models (Amsi, red) start high but decay faster than balanced models (Qwen2.5-VL-3B, blue) because their chain reliability is lower. *(b)* Rank stability table: top-5 and bottom-3 models at $\theta = 1$, with their ranks at $\theta = 0$ and $\theta = 3$. Rank order is largely stable for high-NDR models, confirming that the main-paper ranking is robust to threshold choice.

fied chart signal.

C Experimental Details and Additional Results

C.1 Model Access and Generation Settings

All evaluations are run once under deterministic decoding whenever supported. For open-weight and financially adapted models, we use vLLM on 8 NVIDIA L20 GPUs with temperature 0. For API-based proprietary models, we set temperature to 0 when the option is available; otherwise, we use the most deterministic setting exposed by the provider.

For reproducibility, we report the proprietary model identifiers used in our experiments:

- Gemini-3.1-Flash: `gemini-3.1-flash-lite-preview`
- Gemini-3.1-Pro: `gemini-3.1-pro-preview`
- GPT-5.5: `gpt-5.5`
- GPT-5.4: `gpt-5.4`
- GPT-5.4-mini: `gpt-5.4-mini`
- Claude-Sonnet-4.6: `claude-sonnet-4-6`

C.2 Thresholded NDR Analysis

The main paper defines thresholded net decision reliability in Eq. (11) and reports NDR as shorthand for NDR_1. Here, we analyze how NDR_θ changes when the evidence threshold θ varies from 0 to 5. This analysis does not require rerunning any model. All model actions, confidence scores, extracted claims, and evidence anchors are already stored in the response logs; changing θ only changes the aggregation rule in Eqs. (9) to (11). Following the main definition, θ controls the thresholded coverage term, while the local chain-reliability terms remain the directional-conditioned diagnostics used in the main evaluation.

Threshold semantics. Table 9 summarizes the role of θ. The ungated case $\theta = 0$ measures how reliable a model appears when any directional action can contribute to coverage. The main setting $\theta = 1$ introduces a minimal evidence requirement, filtering out directional actions made under no strong chart signal. Larger thresholds impose progressively stricter evidence requirements, which helps diagnose whether a model's apparent reliability depends on sparse-evidence decisions. In the limit $\theta \to \infty$, coverage vanishes for all models; before that endpoint, the decay pattern reveals how much reliable action each model can sustain under increasingly dense evidence.

Empirical interpretation. Fig. 9 provides a diagnostic view of why models obtain different reliability profiles. Models that already start low at $\theta = 0$ are primarily limited by directional coverage or local chain quality, rather than by the evidence threshold itself. Models that decay sharply as θ increases rely more heavily on directional actions made under sparse evidence. By contrast, models that remain comparatively high under mild thresholds preserve more reliable actions when a minimal evidence requirement is imposed. This analysis supports the $\theta = 1$ main setting: it removes actions

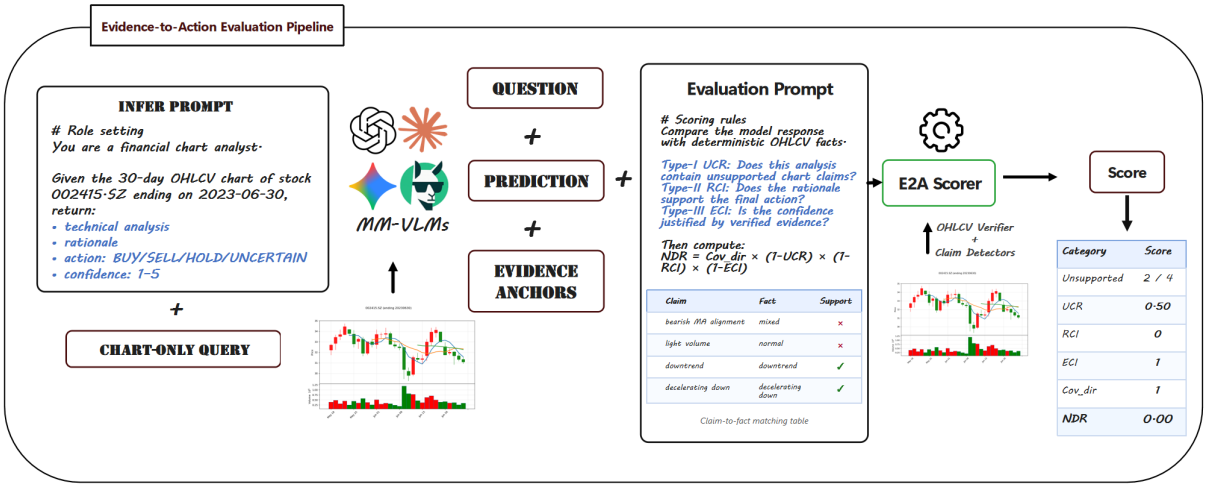


Figure 10: Example of evidence-to-action scoring. Given a chart-only query, the VLM generates a technical analysis, rationale, action, and confidence score. The scorer parses the response into claims, rationale polarity, action, and confidence, then verifies extracted claims against deterministic OHLCV evidence anchors. In this example, the model issues a high-confidence SELL recommendation with coherent reasoning, but two claims are unsupported and the confidence is not justified by sufficiently strong verified evidence. The response therefore receives $\mathrm{UCR} = 0.50$, $\mathrm{RCI} = 0$, $\mathrm{ECI} = 1$, and $\mathrm{NDR} = 0$, illustrating that reliable action requires grounding, coherence, calibration, and directional coverage simultaneously.

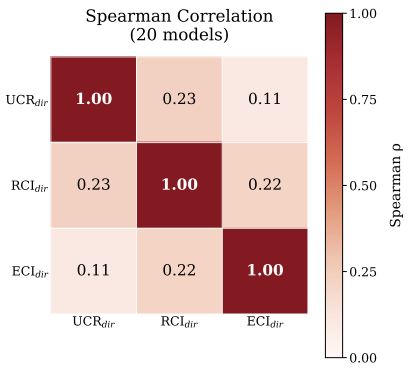


Figure 11: Local diagnostics capture non-redundant failure modes. The heatmap reports pairwise Spearman correlations among $\mathrm{UCR}_{\mathrm{dir}}$, RCI, and ECI across 20 models. All off-diagonal correlations are small ($\rho \leq 0.23$), indicating that factual grounding, reasoning–action consistency, and evidence–confidence calibration measure largely distinct aspects of evidence-to-action reliability.

unsupported by any strong chart signal while avoiding the severe coverage collapse induced by overly strict thresholds. Thus, NDR_θ is best viewed as a robustness and diagnostic analysis of Eq. (11), not as a separate leaderboard family.

C.3 Non-Redundancy of Local Diagnostics

To examine whether the three local diagnostics capture distinct failure modes, we compute pairwise Spearman correlations among $\mathrm{UCR}_{\mathrm{dir}}$, RCI, and ECI across the 20 evaluated models. As shown in Fig. 11, all off-diagonal correlations are small: $\rho = 0.23$ for $\mathrm{UCR}_{\mathrm{dir}}$–RCI, $\rho = 0.11$ for $\mathrm{UCR}_{\mathrm{dir}}$–ECI, and $\rho = 0.22$ for RCI–ECI. This low correlation indicates that factual grounding, reasoning–action consistency, and evidence–confidence calibration are not interchangeable measurements of the same error. A model can ground its technical claims while still contradicting its own rationale, or remain coherent while being overconfident under weak evidence. These results support the chain-level design of `E2A-Bench`: a single hallucination score would conflate distinct reliability links, whereas UCR, RCI, and ECI provide complementary diagnostics for identifying the binding failure mode of each model.

C.4 Case Study: Evidence-to-Action Scoring

Fig. 10 illustrates the full scoring pipeline on a single chart-only response. The example complements the formal protocol in Sec. 3.4 by showing how generated claims are extracted, matched against deterministic OHLCV anchors, and converted into local diagnostics and final NDR. Although the model issues a directional SELL recommendation with coherent reasoning, unsupported claims and overconfident evidence use break the chain, yielding zero NDR.

C.5 Contrastive Case: Claim-Level Factuality vs. Evidence-Backed Action

Table 10 provides a detailed evidence audit for a single chart-only input. Both models receive the same input: stock 600519.SH (Kweichow Moutai), using the 30-day OHLCV window ending on 2023-06-30. The ground-truth fact labels are computed deterministically from the OHLCV data.

This case illustrates why UCR and NDR answer different evaluation questions. UCR measures whether generated claims are unsupported, so a conservative response can appear reliable by making few verifiable claims and avoiding directional commitment. NDR instead evaluates whether the model can produce an evidence-backed action. Qwen2.5-VL-3B-Instruct makes one imperfect claim, but its SELL recommendation remains directionally covered, coherent with the bearish rationale, and calibrated under multiple verified signals. Thus, NDR separates conservative non-action from actionable evidence-backed reliability.

C.6 Qualitative Case: Imperfect Claims with Reliable Action

Fig. 12 complements the main-text coverage-trap example by showing that NDR is not a perfection-based hallucination score. Qwen3.5-4B makes one unsupported claim, but its SELL recommendation remains directionally covered, coherent with the bearish rationale, and supported by multiple verified evidence anchors. In contrast, Gemma-4-E2B-It again avoids verifiable claims and outputs HOLD, obtaining zero observed UCR but zero NDR due to the absence of directional coverage. This case illustrates that NDR penalizes conservative non-action while still recognizing actionable recommendations whose evidence-to-action chain remains largely intact.

Item	**Content**
Stock / date	600519.SH (Kweichow Moutai), 2023-06-30
Input modality	Chart-only candlestick PNG with MA5/MA10/MA20 and volume
Ground-truth OHLCV fact labels	
MA cross	Death cross: MA5 crossed below MA10 on day -3
5-day price change	-4.2%; significant drop threshold: $> 3\%$
Volume ratio	$1.6\times$ 20-day mean; heavy-volume threshold: $> 1.5\times$
Price position	22nd percentile of the 20-day high–low band
Trend direction	Downtrend; slope $= -0.19\%$/day
Signal count	4 unambiguous bearish signals; $\theta = 1$ evidence gate passed
Case A: Gemma-4-E2B-It	
Response (abridged)	"The MA5 appears to have crossed below the MA10 ... volume patterns are ambiguous ... I cannot determine the future trend with sufficient confidence."
Action / Confidence	HOLD / 2
Claim audit	Death cross: ✓ supported; volume ambiguity: not counted as a verifiable unsupported claim; price position: not claimed
UCR (response)	0.0%; no unsupported claims
Coverage gate	**Fail**; action $\notin \mathcal{A}_{\text{dir}}$
NDR contribution	0; coverage gate blocks downstream reliability contribution
Case B: Qwen2.5-VL-3B-Instruct	
Response (abridged)	"The MA5 has crossed below the MA10 ... 5-day change is -4.2% ... volume $1.8\times$ the 20-day average ... price at 22nd percentile ..."
Action / Confidence	SELL / 4
Claim audit	Death cross: ✓; 5-day change: ✓; volume ratio: ✗ stated $1.8\times$, actual $1.6\times$; price position: ✓
UCR (response)	25.0%; 1 of 4 audited claims unsupported
Coverage gate	**Pass**; action $\in \mathcal{A}_{\text{dir}}$ and signal count ≥ 1
RCI gate	**Pass**; bearish rationale with SELL action
ECI gate	**Pass**; confidence 4/5 under multiple verified signals
NDR contribution	Positive; coverage, coherence, and calibration gates passed
Aggregate decomposition over 323 stocks	
Gemma NDR	$6.4\% \times 85.5\% \times 98.4\% \times 95.2\% = 5.1\%$
Qwen NDR	$82.5\% \times 45.0\% \times 89.1\% \times 97.1\% = 32.1\%$

Table 10: Evidence audit for a single-input contrast between claim-level factuality and evidence-backed action. Gemma-4-E2B-It obtains zero observed UCR by avoiding unsupported claims and directional action, whereas Qwen2.5-VL-3B-Instruct has non-zero UCR but passes the coverage, coherence, and calibration gates.

•NDR Case 2: Imperfect Claims, Still Reliable

INPUT CHART

601628.SH (ending 20230630)

QUESTION:
Can a model with non-zero UCR still receive high NDR?

Gemma-4-E2B-It

Output: HOLD
Confidence: 3
Behavior: no claim atoms
Tag: abstention-like safety

Metric	Value
Claims	0 / 0
UCR ↓	0·00
Cov_dir ↑	0
RCI ↓	N/A
ECI ↓	N/A
NDR ↑	0·00

Qwen3.5-4B

Output: SELL
Confidence: 4
Supported claims: 3 / 4
Signals: 4 verified

Metric	Value
Claims	3 / 4
UCR ↓	0·25
Cov_dir ↑	1
RCI ↓	0
ECI ↓	0
NDR ↑	0·75

Answer:

- *Qwen makes one unsupported claim, so UCR is not perfect·*
- *However, its SELL action remains directionally covered and confidence is supported by four verified signals·*
- *Because RCI = 0 and ECI = 0, the chain does not break; NDR only discounts by the unsupported-claim rate·*
- *Gemma remains UCR-perfect but receives NDR = 0 because it never provides a directional action·*

Key idea: NDR is coverage-aware, not hallucination-only; it rewards evidence-backed action even when grounding is not flawless·

Figure 12: **Imperfect claims can still yield reliable action.** Qwen3.5-4B makes one unsupported claim, resulting in non-zero UCR, but its SELL action remains directionally covered, coherent with the bearish rationale, and supported by multiple verified evidence signals. This example shows that NDR is not a perfection-based hallucination score; it evaluates whether an actionable recommendation remains sufficiently evidence-backed across the full chain.