

Faithful to the Frame: Source-Framing Propagation in LLM-Agent Decision Workflows

Yi-Ting Chiu

National Tsing Hua University
chew1tim@gapp.nthu.edu.tw

Chung-Chi Chen

Human-Agent Ally Lab (HAA Lab)
National Institute of Informatics (NII), Japan
chen@nii.ac.jp

Abstract

As LLMs become reusable infrastructure for evidence access, NLP evaluation must move beyond verifying whether generated outputs are faithful to their sources. We study source-framing propagation: whether the directional effect of institutionally authored evidence is preserved or altered through LLM-mediated representations. We introduce Evidence-to-Action Propagation (E2A), an evaluation protocol that traces how fixed evidence affects reasoning, belief, and action across role-conditioned LLM receivers. In a controlled primary evaluation over 94 post-knowledge-cutoff earnings events and 24,816 LLM-agent decision traces, supplemented by targeted mechanism and scope diagnostics, raw disclosures produce a strong positive-return skew. Source-faithful canonical extraction produces no detectable change in this skew (80.1% vs. 80.0%, $p = 0.889$); no-style synthesis shifts action away from buy, while its belief and quality effects remain unresolved. Shared summaries synchronize evidence paths and rationales. A separate diagnostic shows that randomizing evidence order reduces the skew by 4.30 percentage points, although its quality effect remains unresolved. These findings reveal a faithfulness paradox: reducing hallucination risk through source-faithful representations may still preserve a source’s downstream directional influence. More broadly, the paper argues that NLP evaluation should examine not only whether models reproduce sources accurately, but how language interfaces shape the path from evidence to reasoning and action.¹

1 Introduction

LLM systems increasingly turn complex source documents into decision-facing evidence interfaces. Filings are summarized, claims are organized, risks are surfaced, and evidence is rendered into forms that downstream systems can act on. At this scale,

the evaluation problem changes. When many receivers encounter a source through the same interface, source faithfulness is not only a guard against hallucination. It can also *scale* whatever the source causes receivers to attend to, believe, and do.

Faithfulness is a central safety criterion for LLM-based summarization and evidence-grounded generation: a representation is typically considered safer when it stays close to its source and introduces no unsupported claims (Maynez et al., 2020; Fabbri et al., 2021, 2022; Laban et al., 2022; Dreyer et al., 2023). This criterion is necessary, but it treats the source as the reference object. Many sources that LLM systems summarize are not neutral records of the world. They are institutionally authored documents, such as earnings releases, legal briefs, clinical submissions, and policy reports. Such documents can be factually constrained while still shaping attention through selection, ordering, tone, comparison class, and emphasis (Li, 2008; Kothari et al., 2009; Boutron et al., 2010; Kamenica and Gentzkow, 2011; Loughran and McDonald, 2011; Davis et al., 2012).

We study source-framing propagation: whether the downstream directional pattern elicited by an institutionally authored source is preserved, attenuated, or synchronized by LLM-mediated representation regimes. Here, *source framing* refers to author-controlled presentation choices in the source, not to fabrication or hallucination. A source can be accurate and still directionally organized. Rather than exhaustively tracking linguistic mechanisms of framing, this study confronts a more consequential downstream question: how different representations change what role-conditioned LLM receivers cite, reason from, believe, and decide.

To address this, we introduce **Evidence-to-Action Propagation (E2A)**, a workflow-level evaluation protocol. Given a fixed evidence packet and a population of non-interacting LLM receivers, E2A traces five linked layers: *Evidence Use*, *Ra-*

¹  [Code & Data](#)

tionale, Belief, Action, and Quality. This shifts evaluation from asking only whether a generated artifact is supported by its source to asking how the representation changes downstream use of that source.

We instantiate E2A in earnings-event workflows. Earnings disclosures provide bounded, timestamped, institutionally authored evidence packets with observable short-window outcome guardrails (Ball and Brown, 1968; Brown and Warner, 1985; MacKinlay, 1997). Our primary matrix compares five representation regimes over a fixed evidence packet: raw disclosure, canonical no-synthesis extraction, a shared LLM summary, independent LLM summaries, and a structured evidence ledger. Additional controlled experiments isolate prompt-style sensitivity, evidence order, local lexical de-framing, and entity-prior scope. Across this integrated evaluation, the primary matrix yields 24,816 LLM-agent decision traces across four model families and six role-conditioned profiles; the targeted controls are reported as separate diagnostic matrices.

Our results reveal a *faithfulness paradox*. Raw disclosures elicit a strong positive-return skew, and source-faithful canonical extraction produces no detectable change (80.1% vs. 80.0%, $p = 0.889$), a non-detection rather than evidence of equivalence. Primary no-style synthesis shifts action away from buy, but its belief and quality effects remain unresolved; sharing a single summary reduces diversity in receivers' evidence paths and rationales. Thus, fidelity alone does not remove source-elicited directional influence, and representation choices affect distinct downstream layers.

The paradox points to evidence order as a tractable intervention target. In a separate diagnostic that holds the evidence-unit multiset fixed, randomizing evidence order reduces the skew by 4.30 percentage points. This offers a content-preserving candidate mitigation, but not a complete or cost-free solution: the skew remains, and the associated quality effect is unresolved. More broadly, **accuracy does not imply neutrality, and faithfulness does not imply safety**. The key evaluation question is therefore not only whether models distort sources, but how the representation regime changes the path from source evidence to downstream traces.

Our contributions are:

- We identify *source-framing propagation* as a

failure mode for LLM-agent workflows.

- We introduce *Evidence-to-Action Propagation* as a **five-layer** evaluation protocol for tracing the path from source evidence to downstream agent decisions under a fixed evidence packet.
- We provide a **controlled empirical study** in post-cutoff earnings workflows, showing that source-faithful extraction can carry forward a source-elicited directional skew, while synthesis, sharing, formatting, and evidence order affect distinct downstream layers.

2 Related Work

Source fidelity and artifact-level evaluation.

Faithfulness evaluation in summarization and evidence-grounded generation typically treats the source document as the reference object. A generated artifact is considered faithful when its claims are supported by the source, and unsafe when it introduces unsupported content, contradicts the source, or fails factual consistency checks (Maynez et al., 2020; Fabbri et al., 2021, 2022; Laban et al., 2022; Dreyer et al., 2023; Chen et al., 2023). This work is essential for detecting hallucination and source inconsistency, but its unit is usually the generated artifact. We retain source fidelity as a necessary condition and shift the evaluation object to the downstream trace that the artifact elicits.

Institutionally authored evidence and source framing.

Institutionally produced documents can be accurate without being neutral. They are often constrained by law, professional norms, disclosure rules, or scientific reporting standards, while still shaping attention through selection, tone, ordering, comparison class, and emphasis. Kamenica and Gentzkow (2011) formalize this logic as persuasion through accurate signals: an informed sender can affect receiver posteriors without fabricating evidence. Financial disclosure research provides a concrete empirical case: earnings-release tone, readability, management language, and salient news are associated with analyst forecasts, investor attention, and market reactions (Li, 2008; Barber and Odean, 2008; Kothari et al., 2009; Loughran and McDonald, 2011; Davis et al., 2012). Related patterns appear outside finance: clinical trial reports can contain spin that affects interpretation even when claims remain formally grounded (Boutron et al., 2010), and evidence-grading frameworks such as GRADE separate evidence quality

from presentation choices because presentation can alter judgment (Guyatt et al., 2025). This literature motivates treating the source not as a neutral fact container, but as an authored representation whose salience structure may propagate downstream.

Generated-language influence and decision-facing evaluation. A parallel LLM literature shows that generated language can affect receiver judgment through framing, primacy effects, persuasion, and cognitive-bias mechanisms (Jakesch et al., 2023; Wang et al., 2023; Echtermhoff et al., 2024; Bai et al., 2025; Alessa et al., 2025). In finance, GPT-4-generated analyses can affect both amateur and professional investment decisions (Takayanagi et al., 2025), while decision-oriented text evaluation measures generated market text through the downstream decisions of human and LLM receivers (Huang et al., 2025). Multi-agent systems have also generated earnings-call reports using specialized analytical viewpoints (Goldsack et al., 2025). E2A extends this decision-facing perspective with a different estimand: it holds the underlying evidence packet fixed, treats the representation regime as the intervention, and traces its effects across Evidence Use, Rationale, Belief, Action, and Quality rather than evaluating only final decision utility or report quality.

LLM agents as controlled receivers. LLM agents are increasingly used as controlled receivers in evaluations of reasoning, decision policies, and tool-mediated workflows. They allow reproducible variation across roles, model families, prompts, and sampling seeds, but they should not be treated as direct simulations of human behavior. Prior work cautions that LLM agents do not reproduce human market trader behavior (Henning et al., 2025). We therefore use role-conditioned LLM receivers as the object of study rather than as proxies for human investors. The empirical target is the LLM-agent trace itself: what evidence the receiver cites, what rationale it states, what belief it forms, and which action it emits under a fixed evidence packet.

AI-mediated convergence and propagation evaluation. AI assistance can narrow the range of claims, opinions, perspectives, and creative artifacts produced by different users (Padmakumar and He, 2024; Doshi and Hauser, 2024; Wright et al., 2025; Srivastava, 2025; Sourati et al., 2026). Broader accounts of collective intelligence identify shared AI systems as coordination devices,

and work on algorithmic coordination shows that automated systems can converge without explicit collusion (Burton et al., 2024; Dou et al., 2025). At a broader scale, repeated reliance on LLM recommendations can translate localized model tendencies into emergent collective patterns (Cheng et al., 2025). These studies motivate measuring convergence, but they often leave open whether it arises from the model, the interface, the source, the receiver population, or receiver interaction. E2A isolates a pre-interaction path by holding the evidence packet fixed while varying its representation, separating representation effects from information and social-influence effects.

3 E2A Framework and Protocol

This section defines Evidence-to-Action Propagation (E2A), the workflow-level object measured in this controlled evaluation. E2A is not a replacement for source faithfulness. It is a complement to artifact-level evaluation: given a fixed evidence packet, it asks how different representations change what downstream LLM receivers cite, reason from, believe, and do.

3.1 Evaluation Object

Let e denote a bounded evidence packet: a timestamped source document, or set of source documents, available before the downstream decision. Let $r \in \mathcal{R}$ denote a representation regime that renders e into an agent-facing input. Primary regime contrasts hold the underlying evidence fixed; the separate E1 scope diagnostic deliberately varies evidence and context availability. Let $d \in \mathcal{D}$ denote a role-conditioned LLM receiver instance, determined by a receiver profile, model family, and sampling seed.

Each receiver produces a structured decision trace:

$$\tau(e, r, d) = (S, C, \rho, u, \hat{y}, c, a),$$

where S is the set of cited source IDs, C is the set of cited evidence categories, ρ is the stated rationale, u contains uncertainty notes, $\hat{y}$ is the expected short-window return, c is confidence, and $a \in \{\text{buy, hold, sell}\}$ is the emitted action.

E2A evaluates the distribution of traces induced by a representation regime:

$$r(e) \longrightarrow \{\tau(e, r, d) : d \in \mathcal{D}\}.$$

The central unit is therefore not the representation alone, but the event–regime cell: one fixed

Regime	Agent-facing input	Design role
T1: Raw	Original 8-K filing, earnings-release exhibit, and available XBRL fields.	Source-elicited baseline before canonical extraction or LLM synthesis.
B0: Canonical	Canonical evidence bank rendered as source-identified units in disclosure order; no narrative synthesis.	Source-faithful no-synthesis baseline; tests whether extraction preserves the raw source trace pattern.
T2: Shared summary	One LLM-generated summary from the canonical evidence bank, shared by all receivers in the event-regime cell.	Centralized synthesis; tests whether a common generated interface coordinates Evidence Use, rationale, belief, and action.
T3: Independent summaries	Independently generated LLM summaries from the same canonical evidence bank.	Decentralized synthesis; separates LLM synthesis from sharing a single representation.
T4: Structured ledger	Canonical evidence bank rendered as a structured ledger with categories, source IDs, claims, values, and source quotes.	No-synthesis format anchor; tests whether no-synthesis patterns persist under structured presentation.

Table 1: **Representation regimes.** All mediated regimes use the same canonical evidence bank.

evidence packet rendered under one regime and observed through a matched population of non-interacting LLM receivers. Regime contrasts ask whether changing the representation changes the joint distribution of Evidence Use, Rationale, Belief, Action, and Quality.

3.2 Diagnostic Layers

We decompose each LLM-agent trace into five layers, ordered from evidence routing to outcome alignment. (i) The *Evidence use layer* records which source IDs enter the trace; (ii) the *Rationale layer* records how the receiver explains its decision; (iii) the *Belief layer* records the expected return direction and dispersion; (iv) the *Action layer* records buy, hold, or sell behavior; and (v) the *Quality layer* checks alignment with the hidden short-window outcome. A regime can affect one layer without affecting another: it may diversify cited sources without changing actions, synchronize rationales without improving accuracy, or attenuate a directional skew without producing better decisions. Formal metric definitions, robustness variants, and inference details are in Appendix A.

Evidence-use metrics rely on source IDs supplied in the treatment packet. Source diversity is based on pairwise overlap of cited source sets, while per-decision uptake measures how much evidence any individual trace uses. Rationale diversity is measured with a deterministic n -gram cosine-distance proxy over stated reasons. Sentence-embedding variants for the style-added synthesis matrix are reported as a sensitivity check in Appendix C. Belief is summarized by positive-prediction rate P_+ and numeric dispersion. Action dispersion is summarized by entropy over buy, hold,

and sell decisions (Shannon, 1948), with Simpson diversity as a supporting metric (Simpson, 1949). Quality uses the hidden five-day cumulative abnormal return, $CAR_{[1,5]}$, as a diagnostic guardrail under standard event-study logic (Ball and Brown, 1968; Brown and Warner, 1985; MacKinlay, 1997).

3.3 Representation Regimes

The protocol varies the representation regime while holding fixed the source packet, receiver population, and outcome blindness. Each mediated regime is built from the same canonical evidence bank: a validated set of source-identified evidence units extracted from the raw disclosure. This design separates four factors that are often conflated in deployed evidence interfaces: direct source exposure, source-faithful extraction, LLM synthesis, and shared versus independent representation. Table 1 summarizes the five regimes and the design role of each.

The regime set is not intended to cover every possible evidence interface. It is designed to make the main contrasts interpretable. T1 measures the trace pattern elicited by the source packet before AI compression. B0 asks whether a source-faithful canonical rendering preserves that pattern. T2 and T3 add LLM synthesis on top of the same canonical substrate, either as a shared representation or as independently generated summaries. T4 removes narrative synthesis again while changing the presentation format.

The primary T2 and T3 prompts require source-grounded synthesis without assigning a neutral-analyst persona or another stylistic role. As a prompt-style ablation, T2* and T3* add only the instruction to write as a neutral analyst; the canonical

evidence bank, representation multiplicity, receiver population, and downstream prompts remain fixed.

3.4 Regime Contrasts and Controls

The primary five-regime analysis uses four contrasts specified in the analysis plan before downstream inference was run; this was an authoring-time specification, not a formal public preregistration. B0 versus T1 tests whether source-faithful extraction changes the trace pattern elicited by raw disclosure. T2/T3 versus B0 tests the effect of adding LLM narrative synthesis to the canonical evidence substrate. T3 versus T2 isolates the effect of sharing one generated representation rather than using independent summaries. T4 versus B0 tests whether no-synthesis behavior is stable under a structured format. These primary contrasts hold the underlying evidence packet fixed, so their effects concern representation rather than different information sets.

Separate diagnostic matrices probe mechanisms and scope without expanding the primary regime family. T2*/T3* add neutral analyst style; T5 applies feasible local lexical edits; T6 permutes unit order while fixing content and rendering; and E1 varies issuer, context, and evidence visibility. These controls use different baselines and are not an additive decomposition (Appendix F).

3.5 Aggregation and Inference

All metrics are computed within event–regime cells. We then aggregate to the event level before comparing regimes, so that events rather than individual traces are the unit of inference. This prevents large numbers of role-conditioned traces generated from the same source packet from being treated as independent observations.

For each regime contrast, we compute paired event-level deltas and estimate uncertainty with event-cluster bootstrap resampling using 20,000 replicates. Each bootstrap replicate samples events with replacement and recomputes the mean contrast. We report two-sided p -values and 95% confidence intervals. For families of related action-dispersion tests, we apply Benjamini–Hochberg correction; individual belief and quality contrasts are reported unadjusted. Model-family robustness is assessed by recomputing the same event-level contrasts separately for each downstream model family.

4 Experimental Design

We instantiate E2A in earnings-event workflows. The goal is not to build a stock-return predictor, nor to propose a general-purpose leaderboard task. Instead, we use earnings events as a controlled evaluation setting in which institutionally authored evidence is bounded, timestamped, source-proximate, and paired with an observable outcome guardrail. Earnings disclosures satisfy these requirements: they are produced under disclosure constraints, yet prior work shows that disclosure tone, readability, management language, and salience affect analyst forecasts and market reactions (Li, 2008; Kothari et al., 2009; Loughran and McDonald, 2011; Davis et al., 2012). Short-window market reactions provide a diagnostic quality guardrail under standard event-study logic (Ball and Brown, 1968; Brown and Warner, 1985; MacKinlay, 1997).

4.1 Event Sample

The event sample contains U.S. earnings-event Form 8-K disclosures from January to early April 2026, strictly after the knowledge cutoff of every model family used in the study. Each event includes an Item 2.02 earnings disclosure, an Exhibit 99.1 or equivalent earnings release, and available XBRL fields. The sample is balanced across 47 large-cap and 47 small/mid-cap firm-events, with at most one event per firm. A separate 12-event calibration set from November–December 2025 is used only for pipeline validation and is excluded from all primary estimates.

The sample is constructed for controlled comparison rather than market representativeness. We require extractable source packets, complete five-day return coverage, adequate trading liquidity, and one event per issuer. Events dominated by non-earnings mechanisms, such as mergers, bankruptcies, major litigation, binary biotech events, or commodity shocks, are excluded. Detailed issuer construction and filtering steps are reported in Appendix G.

4.2 Evidence Packets and Outcome Blindness

For each event, we construct a source packet from the 8-K filing, the earnings-release exhibit, and XBRL fields. From this packet we derive the canonical evidence bank used by B0, T2, T3, and T4. Each evidence unit contains a claim, source attribution, evidence category, supporting numeric value when applicable, and source ID. Source IDs provide the common citation interface across regimes,

allowing us to measure how representations route LLM receivers toward overlapping or distinct evidence subsets.

The experiment is entity-visible and outcome-blind. Company identity, ticker, filing date, fiscal period, source text, XBRL facts, pre-event price context, representation regime, receiver profile, and decision seed may be visible when relevant. Hidden from all treatment construction and downstream prompts are realized returns, $CAR_{[1,5]}$, hidden outcome valence, post-event analyst revisions, future filings or news, outcome-derived labels, and post-hoc quality metrics. The hidden outcome is joined to decision rows only after all upstream representations and downstream decision traces are frozen.

We use five-day cumulative abnormal return, $CAR_{[1,5]}$, as a quality guardrail. This outcome is not treated as investor welfare or ground truth about the intrinsic correctness of a rationale. It serves only to test whether changes in Evidence-use, Belief, or Action are accompanied by better alignment with an outcome unavailable to the agents.

4.3 Model and Receiver Pipeline

We use four model families: Claude Sonnet 4.5, GPT-5.2, Qwen3-235B-A22B, and DeepSeek-V3.1. The binding leakage constraint is the latest cutoff in the panel, August 2025. All main events occur at least four months after this cutoff, and all model calls are made with browsing and external tool use disabled. Exact model identifiers, cutoff provenance, and access conditions are reported in Appendix I.

The pipeline has two stages. In the upstream stage, models render the primary no-style T2 and T3 narrative representations from the canonical evidence bank. T2 uses one shared summary within an event-regime cell; T3 uses independently generated summaries over the same canonical bank. In the downstream stage, role-conditioned LLM receivers read one treatment packet and produce a structured decision trace containing cited evidence, key reasons, uncertainty notes, expected five-day return, confidence, action, and action strength.

Each downstream receiver is assigned one of six role-conditioned investor profiles: proprietary trader, event-driven hedge fund analyst, investment advisor, day trader, swing trader, or long-term retail investor. These profiles are controlled decision policies rather than behavioral simulations of human investors. Profile identity is applied only at the

downstream stage; no upstream summary prompt is profile-conditioned. Profile definitions and separability checks are reported in Appendix J.

The primary five-regime matrix contains 24,816 decision traces. The prompt-style, lexical, evidence-order, and entity-prior diagnostic matrices add 33,840 traces, yielding 58,656 final LLM-agent decision traces across the integrated evaluation. These counts describe corpus size rather than independent observations: many traces share an event and source packet, and all inferential contrasts are aggregated to and resampled over 94 events.

4.4 Validity Checks

Because the paper’s central claim concerns propagation rather than hallucination, we audit source packets, mediated representations, and downstream traces for grounding and leakage. Source packets are checked for numeric fidelity, claim fidelity, source alignment, and absence of post-event information. Runtime validation checks every primary decision row for schema compliance and verifies that cited source IDs exist in its treatment packet. The claim-level representation audit was conducted on the original style-added T2*/T3* matrix and is scoped accordingly below.

All 22,560 decision rows in the primary no-style T1/B0/T2/T3 bundle pass schema and source-ID validation; T4 is validated separately as a structured format anchor. The 9,024-row E1 matrix also passes source-aware validation across all four model families and four arms. A separate stratified audit of 961 rationale claims was conducted on the neutral-style T2*/T3* ablation outputs; we therefore do not transfer its claim-level rates to the primary no-style T2/T3 conditions. Malformed JSON outputs are handled by deterministic repair rules without re-querying the model. Detailed audit procedures and scope are reported in Appendix K.

The downstream actors are LLM receivers rather than human investors. We therefore interpret the collected data as a controlled test of representation effects in LLM-agent decision traces, not as a behavioral model of real investors. This distinction is important because prior work cautions that LLM agents do not reproduce human market-trader behavior (Henning et al., 2025).

Layer	Metric	T1 Raw	B0 Canonical	T2 Shared summary	T3 Independent summaries	T4 Structured ledger
EVIDENCE USE	Source diversity	0.417	0.942	0.396	0.511	0.887
	Per-decision coverage	0.159	0.059	0.164	0.164	0.072
RATIONALE	Lexical rationale distance	0.320	0.338	0.349	0.379	0.334
BELIEF	Positive-prediction rate (P_+)	0.801	0.800	0.778	0.777	0.758
	Belief MAD	0.013	0.012	0.016	0.017	0.013
ACTION	Buy share	0.680	0.662	0.598	0.598	0.564
	Action entropy	0.459	0.485	0.651	0.678	0.644
QUALITY	Directional accuracy	0.514	0.509	0.516	0.522	0.491
	Strict action accuracy	0.425	0.413	0.397*	0.400*	0.367

Table 2: **E2A results across representation regimes.** no detected T1/B0 P_+ difference ($p=0.889$); B0 paradox: highest source diversity yet lowest per-decision uptake. Primary T2/T3 use no-style synthesis; their P_+ contrasts against B0 are directionally negative but not conventionally significant ($p=.107/.130$). Bold: row maximum. * $p<.05$ vs. T1. Higher diversity/entropy = greater dispersion; P_+ is a convergence measure; quality metrics are outcome guardrails.

5 Results

Table 2 summarizes treatment means across the five E2A layers. We organize the results around the paper’s central question: when the same institutionally authored evidence packet is rendered through different representation regimes, which parts of the downstream LLM-agent trace are preserved, attenuated, synchronized, or improved?

5.1 Faithful Extraction Does Not Detectably Change the Directional Skew

Throughout the Results, P_+ denotes the event-averaged share of receiver traces predicting a positive five-day return, rather than directional accuracy. T1 yields $P_+ = 0.801$ and B0 yields $P_+ = 0.800$, with no detected difference (B0–T1: -0.18 percentage points, 95% CI $[-2.88, +2.62]$, $p = 0.889$).²

B0 replaces the raw document’s presentation with canonical, source-identified evidence units. Its failure to shift P_+ is consistent with source-framing propagation: faithful extraction need not remove the directional pattern observed under raw disclosure. This is a non-detection, not evidence of equivalence or of any single framing mechanism.

²Only 44 of 94 events have positive realized CAR, and the high P_+ persists within both realized-label groups (Appendix Figure 2). T1 and B0 also have identical source-ID sets (Jaccard = 1.00).

5.2 Synthesis Shifts Action Without a Detected Quality Improvement

The B0-relative contrasts ask what changes when narrative synthesis is added to the canonical evidence representation. Under the primary no-style prompts, T2 and T3 yield P_+ point estimates 2.14 and 2.25 percentage points below B0, respectively. Neither contrast is conventionally significant ($p = 0.107$ and $p = .130$), and T3 differs from T2 by only -0.11 points ($p = .794$). The primary matrix therefore does not establish synthesis-induced belief attenuation. The style-added results in Appendix E show that the analyst-style instruction amplifies the same directional shift.

The Action layer changes more strongly. Relative to B0, buy share falls from 0.662 to 0.598 under both T2 and T3, while hold share rises from 0.188 to 0.244 and 0.245. Action entropy increases from 0.485 under B0 to 0.651 and 0.678. No-style synthesis therefore loosens the buy-heavy action pattern even though its belief-level attenuation remains statistically unresolved.

These propagation effects are not accompanied by a detected improvement in outcome alignment. Directional accuracy remains near 0.5 across regimes (0.491–0.522), with T2 and T3 at 0.516 and 0.522; their strict action accuracies are 0.397 and 0.400. Strict action accuracy is lower than T1 by 2.79 points for T2 (95% CI $[-0.052, -0.005]$, $p = .017$) and 2.53 points for T3 (95% CI $[-0.049, -0.003]$, $p = .021$), whereas contrasts against B0 remain unresolved ($p = .270$ and $p = .395$). The

Contrast	Design question	ΔP_+ (pp)	95% CI (pp)	p	Interpretation
B0 – T1	Does source-faithful extraction change the raw-source pattern?	−0.18	[−2.88, +2.62]	0.889	No directional change is detected; this is not an equivalence result.
T2 – B0	Does shared no-style synthesis change the canonical baseline?	−2.14	[−5.09, +0.41]	0.107	The point estimate is lower, but the interval includes no change.
T3 – B0	Does independent no-style synthesis change the canonical baseline?	−2.25	[−5.53, +0.62]	0.130	The point estimate is lower, but the interval includes no change.
T3 – T2	Do shared and independent no-style summaries differ in directional skew?	−0.11	[−0.84, +0.69]	0.794	Their belief-level skew is nearly identical; the bottleneck appears in evidence paths and rationales.
T4 – B0	Does structured no-synthesis formatting change the canonical baseline?	−4.21	[−6.83, −1.73]	< 0.001	The format changes the skew, but T4 is not a clean de-framing treatment.
T6 – B0 replay	Does evidence order matter when the evidence-unit multiset is fixed?	−4.30	[−6.87, −1.91]	—	Order randomization is a candidate mitigation; its quality effect is unresolved.

Table 3: **Positive-prediction rate contrasts.** canonical extraction; directionally lower but unresolved synthesis estimates; format boundary; and separate order diagnostic. Contrasts use 20,000-replicate event-cluster bootstrap inference over 94 events. T6 is compared with its paired B0 confirmation replay; no bootstrap p -value is tabulated for that diagnostic.

realized-label balance and power analysis further characterize the unresolved quality contrasts as bounded non-detections rather than evidence of equivalence (Appendix F).

5.3 A Shared Summary Creates a Representation Bottleneck

The most robust synthesis-side result concerns representation sharing rather than synthesis itself. T2 and T3 use summaries derived from the same canonical evidence bank with the same synthesis procedure and downstream receiver population; they differ in whether receivers share one summary or receive independently generated summaries. The T3–T2 contrast therefore isolates the effect of sharing one upstream representation rather than using independent representations.

Relative to shared-summary T2, independent-summary T3 increases source diversity by +0.112 (95% CI [+0.102, +0.123]) and lexical rationale distance by +0.029 (95% CI [+0.026, +0.032]). The action-entropy difference is smaller and uncertain (+0.016, 95% CI [−0.001, +0.033], $p = .068$). The bottleneck is therefore localized to evidence paths and rationales rather than established at every downstream layer.

The style-added replication yields the same pattern: T3* exceeds T2* in source diversity by +0.103 and in rationale distance by +0.030 (Appendix E). The representation bottleneck is therefore robust to the analyst-style instruction. This coordination does not establish an improvement in

outcome alignment: T2 and T3 have similar directional accuracy (0.516 and 0.522) and strict action accuracy (0.397 and 0.400).

5.4 Format Changes Evidence Uptake More Than Evidence Availability

Evidence availability alone does not determine evidence use. B0 exposes the full canonical evidence bank without synthesis, yet receivers cite only a small subset of the available sources in each trace. Narrative regimes (T1, T2, T3) yield approximately 8.3–8.5 cited sources per decision, while structured no-synthesis regimes (B0, T4) yield approximately 2.5–3.3 cited sources per decision. The clearest case is B0: it provides the complete canonical evidence bank, while T2 compresses the bank through a summary, yet B0 receivers cite only 2.51 sources per decision compared with T2’s 8.55.

This explains the apparent paradox in Table 2. B0 has the highest source diversity (0.942) but the lowest per-decision coverage (0.059); this does not mean that each receiver uses evidence broadly. It means that different receivers anchor on different sparse subsets of the same complete source bank. Narrative formats encourage continuous reading, where several sources are connected into a single rationale thread. Structured formats encourage anchor-scanning, where receivers attach one or two source IDs to each reason. High source diversity therefore can coexist with shallow per-trace evidence uptake.

T4 supports this interpretation while requiring

caution. As a structured ledger, T4 remains close to B0 in rationale diversity (0.334 vs. 0.338), suggesting that the synthesis/no-synthesis split is not merely an artifact of plain-text canonical formatting. But T4 also changes salience cues and category grouping, so we treat it as a no-synthesis format anchor rather than a clean de-framing intervention (Appendix B). T4 therefore establishes a broad format boundary rather than identifying a single framing mechanism. This motivates the narrower T5 and T6 controls, which separately probe locally editable wording and evidence order.

5.5 Evidence Order Offers a Candidate Mitigation with an Unresolved Quality Tradeoff

T5 clarifies the interpretive boundary of T4. To isolate lexical framing, T5 changes only locally editable wording without altering the proposition or evidence structure. Human review yields 13 safe claim-level edits across 11 events. Across this narrow intervention surface, T5 yields no detected full-policy change in P_+ ($\Delta = -0.09$ pp, 95% CI $[-0.71, +0.49]$). This does not show that wording is irrelevant; it shows that proposition-preserving lexical de-framing is feasible for only a limited part of the canonical evidence bank.

T6 instead permutes the order of the same evidence-unit multiset while holding unit content and rendering fixed. Relative to its paired B0 confirmation replay, T6 reduces P_+ by 4.30 pp (95% CI $[-6.87, -1.91]$). Its strict-action accuracy estimate is -2.08 pp (95% CI $[-4.61, +0.35]$), and equivalence within ± 3 pp is not established ($p = .234$). Evidence-order randomization is therefore a candidate mitigation for directional skew, with a potential but unresolved quality cost. The full T5 feasibility audit and T6 endpoints are reported in Appendix F.

5.6 No Regime Resolves the Propagation Tradeoff

No representation regime dominates across the five E2A layers. T1 and B0 remain strongly positive and buy-heavy. T2 and T3 increase action entropy and shift action away from buy, but their belief attenuation relative to B0 is directional rather than statistically resolved, and neither yields a detected improvement in outcome alignment. Within synthesized interfaces, one shared summary creates a bottleneck in evidence paths and rationales. T4 shows that a broad format change is consequential

but confounded, while T5 shows that the surface for proposition-preserving lexical edits is narrow. T6 isolates evidence order and provides the clearest candidate mitigation, although its quality tradeoff remains unresolved.

The main result is therefore not that one interface is best. It is that source faithfulness, framing attenuation, decision diversity, and outcome alignment come apart. Artifact-level faithfulness certifies source proximity, but E2A shows that adhering to the source does not determine how that evidence is routed downstream.

6 Conclusion

This paper studied how source-faithful LLM representations route institutionally authored evidence through an agent workflow. Standard faithfulness evaluation asks whether an artifact is supported by its source. That question is necessary but incomplete because a grounded interface can still change downstream evidence use. We introduced **Evidence-to-Action Propagation (E2A)** to evaluate representation as an intervention on Evidence Use, Rationale, Belief, Action, and Quality.

In a controlled primary E2A evaluation over 94 post-knowledge-cutoff earnings events and 24,816 LLM-agent decision traces, supplemented by targeted diagnostics, raw disclosures elicited a strong positive-return skew, and canonical extraction produced no detectable change. Primary no-style synthesis shifted action away from buy but did not establish belief attenuation, and no quality improvement was detected. Its clearest effect was coordination: a shared summary made receivers cite more similar evidence and state more similar rationales.

These findings expose a *faithfulness paradox* for LLM evidence interfaces. Source fidelity can reduce hallucination risk, but the bounded T1–B0 comparison shows that it does not guarantee removal of a source-elicited directional pattern. Conversely, changing an interface can disrupt that pattern, but disruption is not evidence of debiasing unless quality is also preserved. Evaluating LLM-agent decision workflows therefore requires workflow-level propagation tests alongside artifact-level fidelity. Ultimately, **accuracy does not imply neutrality, and faithfulness does not imply safety**. Evidence-order randomization is a concrete candidate for further study, not yet a validated quality-preserving intervention.

Limitations

This study has several limitations. First, we measure source framing through downstream propagation rather than directly identifying every linguistic mechanism in the source. The absence of a detected T1–B0 difference is consistent with propagation but does not establish equivalence or prove which source cues are preserved. The T5 and T6 diagnostics narrow the mechanism space, but future work should audit tone, ordering, salience, comparison class, and risk placement jointly.

Second, the receivers in our evaluation are role-conditioned LLM agents, not human investors. This design enables controlled comparisons across regimes, profiles, model families, and seeds, but the results should be interpreted as evidence about LLM-agent decision traces rather than human decision-making. The multi-receiver setup also excludes social interaction, deliberation, imitation, and market feedback effects.

Third, earnings disclosures are a useful identification setting because they are bounded, time-stamped, institutionally authored, and paired with observable short-window outcomes. However, this also makes the study domain-specific. Other settings, such as legal briefs, clinical submissions, policy reports, or scientific articles, may involve different framing mechanisms and quality signals.

Fourth, the five-day cumulative abnormal return is only a diagnostic quality guardrail. It does not measure investor welfare, long-run firm value, or the intrinsic correctness of a rationale. The 94-event design can detect only moderate accuracy changes; non-significant quality contrasts are not evidence of equivalence. In particular, T6 may reduce directional skew at a quality cost that this study does not resolve.

Finally, post-cutoff and outcome-blind controls reduce leakage risk but cannot make the models blank slates. E1 confirms that recorded context, issuer visibility, and evidence availability affect absolute belief levels, but it does not identify every interaction between priors and representation. The regime set is also limited: T4 changes several formatting features together, T5 covers only locally editable wording, and the sample spans one early-2026 earnings window. Replication across cycles, domains, and interface designs is needed before generalizing the reported magnitudes.

Ethical Considerations

This study evaluates LLM-agent decision traces in an offline, outcome-blind research setting. It does not provide investment advice, execute trades, or recommend deployment of automated trading agents. All earnings disclosures used as source packets are public filings, and the reported quality metrics are diagnostic guardrails rather than welfare or profitability estimates. Because the results may inform the design of decision-facing evidence interfaces, we emphasize that faithful summarization alone should not be treated as sufficient for safe deployment in financial, legal, clinical, or policy settings.

Because order, grouping, and summary reuse are deployment choices, T6 motivates evaluating order randomization rather than recommending it: the unresolved quality effect requires domain-specific validation and oversight.

References

- Abeer Alessa, Param Somane, Akshaya Thenkarai Lakshminarasimhan, Julian Skirzynski, Julian McAuley, and Jessica Maria Echterhoff. 2025. [Quantifying cognitive bias induction in LLM-generated content](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2890–2910, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Hui Bai, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer. 2025. [LLM-generated messages can persuade humans on policy issues](#). *Nature Communications*, 16(1):6037.
- Ray Ball and Philip Brown. 1968. [An empirical evaluation of accounting income numbers](#). *Journal of Accounting Research*, 6(2):159–178.
- Brad M. Barber and Terrance Odean. 2008. [All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors](#). *The Review of Financial Studies*, 21(2):785–818.
- Isabelle Boutron, Susan Dutton, Philippe Ravaud, and Douglas G. Altman. 2010. [Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes](#). *JAMA*, 303(20):2058–2064.
- Stephen J. Brown and Jerold B. Warner. 1985. [Using daily stock returns: The case of event studies](#). *Journal of Financial Economics*, 14(1):3–31.

- Jason W. Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A. Bakker, Joshua A. Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, Lucie Flek, Stefan M. Herzog, Saffron Huang, Sayash Kapoor, Arvind Narayanan, Anne-Marie Nussberger, Taha Yasseri, Pietro Nickl, Abdullah Almaatouq, and 9 others. 2024. [How large language models can reshape collective intelligence](#). *Nature Human Behaviour*, 8(9):1643–1655.
- Wei-Lin Chen, Cheng-Kuang Wu, Hsin-Hsi Chen, and Chung-Chi Chen. 2023. [Fidelity-enriched contrastive search: Reconciling the faithfulness-diversity trade-off in text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 843–851, Singapore. Association for Computational Linguistics.
- Yuyang Cheng, Xingwei Qu, Tomas Goldsack, Chenghua Lin, and Chung-Chi Chen. 2025. [Observing micromotives and macrobehavior of large language models](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1258–1276, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Angela K. Davis, Jeremy M. Piger, and Lisa M. Sedor. 2012. [Beyond the numbers: Measuring the information content of earnings press release language](#). *Contemporary Accounting Research*, 29(3):845–868.
- Anil R. Doshi and Oliver P. Hauser. 2024. [Generative AI enhances individual creativity but reduces the collective diversity of novel content](#). *Science Advances*, 10(28):eadn5290.
- Winston Wei Dou, Itay Goldstein, and Yan Ji. 2025. [AI-powered trading, algorithmic collusion, and price efficiency](#). Working Paper 34054, National Bureau of Economic Research.
- Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2023. [Evaluating the tradeoff between abstractiveness and factuality in abstractive summarization](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2089–2105, Dubrovnik, Croatia. Association for Computational Linguistics.
- Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. [Cognitive bias in decision-making with LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.
- Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. [QAFactEval: Improved QA-based factual consistency evaluation for summarization](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.
- Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. [SummEval: Re-evaluating summarization evaluation](#). *Transactions of the Association for Computational Linguistics*, 9:391–409.
- Tomas Goldsack, Yang Wang, Chenghua Lin, and Chung-Chi Chen. 2025. [From facts to insights: A study on the generation and evaluation of analytical reports for deciphering earnings calls](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10576–10593, Abu Dhabi, UAE. Association for Computational Linguistics.
- Gordon Guyatt, Liang Yao, M Hassan Murad, Monica Hultcrantz, Thomas Agoritsas, Hans De Beer, Stefan Schandelmaier, Alfonso Iorio, Linan Zeng, Manya Prasad, Per Olav Vandvik, Reem A Mustafa, Arnab Agarwal, Tahira Devji, Ivan D Florez, Benjamin Djulbegovic, Derek K Chu, Bram Rochweg, Victor M Montori, and Romina Brignardello-Petersen. 2025. [Core GRADE 6: presenting the evidence in summary of findings tables](#). *BMJ*, 389:e083866.
- Thomas Henning, Siddhartha M. Ojha, Ross Spoon, Jiatong Han, and Colin F. Camerer. 2025. [LLM agents do not replicate human market traders: Evidence from experimental finance](#). *Preprint*, arXiv:2502.15800.
- Yu-Shiang Huang, Chuan-Ju Wang, and Chung-Chi Chen. 2025. Decision-oriented text evaluation. *arXiv preprint arXiv:2507.01923*.
- Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. [Co-writing with opinionated language models affects users’ views](#). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23, New York, NY, USA. Association for Computing Machinery.
- Emir Kamenica and Matthew Gentzkow. 2011. [Bayesian persuasion](#). *American Economic Review*, 101(6):2590–2615.
- S. P. Kothari, Xu Li, and James E. Short. 2009. [The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis](#). *The Accounting Review*, 84(5):1639–1670.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. [SummaC: Re-visiting NLI-based models for inconsistency detection in summarization](#). *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Feng Li. 2008. [Annual report readability, current earnings, and earnings persistence](#). *Journal of Accounting and Economics*, 45(2):221–247. Economic Consequences of Alternative Accounting Standards and Regulation.

Tim Loughran and Bill McDonald. 2011. [When is a liability not a liability? Textual analysis, dictionaries, and 10-ks.](#) *The Journal of Finance*, 66(1):35–65.

A. Craig MacKinlay. 1997. [Event studies in economics and finance.](#) *Journal of Economic Literature*, 35(1):13–39.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Vishakh Padmakumar and He He. 2024. [Does writing with language models reduce content diversity?](#) In *The Twelfth International Conference on Learning Representations*.

Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks.](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

C. E. Shannon. 1948. [A mathematical theory of communication.](#) *The Bell System Technical Journal*, 27(3):379–423.

E. H. Simpson. 1949. [Measurement of diversity.](#) *Nature*, 163(4148):688–688.

Zhivar Sourati, Alireza S. Ziabari, and Morteza Dehghani. 2026. [The homogenizing effect of large language models on human expression and thought.](#) *Trends in Cognitive Sciences*.

Shashank Srivastava. 2025. [Large language models threaten language’s epistemic and communicative foundations.](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28662–28676, Suzhou, China. Association for Computational Linguistics.

Takehiro Takayanagi, Hiroya Takamura, Kiyoshi Izumi, and Chung-Chi Chen. 2025. [Can GPT-4 sway experts’ investment decisions?](#) In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 374–383, Albuquerque, New Mexico. Association for Computational Linguistics.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. 2023. [Primacy effect of ChatGPT.](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 108–115, Singapore. Association for Computational Linguistics.

Dustin Wright, Sarah Masud, Jared Moore, Srishti Yadav, Maria Antoniak, Peter Ebert Christensen, Chan Young Park, and Isabelle Augenstein. 2025. [Epistemic diversity and knowledge collapse in large language models.](#) *Preprint*, arXiv:2510.04226.

A Formal Metric Definitions and Inference

This appendix defines the E2A metrics used in the main text. All metrics are computed within matched event–regime cells and aggregated to the event level before cross-regime comparison.

Let $b = (e, r)$ index an event–regime cell, where e is an event and r is a representation regime. Let $\mathcal{I}_b$ be the set of valid receiver traces in cell b , with $n_b = |\mathcal{I}_b|$. Let $\mathcal{P}_b = \{(i, j) : i, j \in \mathcal{I}_b, i < j\}$ denote the set of receiver pairs.

Each receiver trace is represented as

$$\tau_i = (S_i, C_i, \rho_i, u_i, \hat{y}_i, c_i, a_i), \quad (1)$$

where S_i is the cited source-ID set, C_i is the cited evidence-category set, ρ_i is the stated rationale, u_i is the uncertainty text, $\hat{y}_i$ is the expected five-day return, $c_i \in [0, 1]$ is confidence, and $a_i \in \{\text{buy, hold, sell}\}$ is the action. We code actions as $g(\text{buy}) = +1$, $g(\text{hold}) = 0$, and $g(\text{sell}) = -1$.

A.1 Evidence Use

For finite sets X and Y , Jaccard overlap is

$$J(X, Y) = \begin{cases} \frac{|X \cap Y|}{|X \cup Y|}, & |X \cup Y| > 0, \\ 1, & |X \cup Y| = 0. \end{cases} \quad (2)$$

Source diversity is one minus the mean pairwise Jaccard overlap of cited source-ID sets:

$$D_{\text{src}}(b) = 1 - \frac{1}{|\mathcal{P}_b|} \sum_{(i,j) \in \mathcal{P}_b} J(S_i, S_j). \quad (3)$$

Higher values indicate that receivers cite more distinct source subsets.

Category diversity is defined analogously over evidence-category sets:

$$D_{\text{cat}}(b) = 1 - \frac{1}{|\mathcal{P}_b|} \sum_{(i,j) \in \mathcal{P}_b} J(C_i, C_j). \quad (4)$$

Per-decision source uptake is the mean number of cited source IDs per trace:

$$U_{\text{src}}(b) = \frac{1}{n_b} \sum_{i \in \mathcal{I}_b} |S_i|. \quad (5)$$

When comparing regimes with different rendered source-bank sizes, we also report relative coverage:

$$\text{Coverage}_{\text{src}}(b) = \frac{1}{n_b} \sum_{i \in \mathcal{I}_b} \frac{|S_i|}{|S_{\text{bank}}(e, r)|}, \quad (6)$$

where $S_{\text{bank}}(e, r)$ is the set of source IDs available in the decision-facing packet for event e under regime r .

Layer	E2A question	Primary measures	Interpretation
EVIDENCE USE	Which parts of the evidence packet enter the trace?	Source diversity; per-decision source uptake	Measures whether receivers cite overlapping or distinct evidence subsets, and whether each trace uses a broad or shallow portion of the available source bank.
RATIONALE	What reasons are stated for the decision?	Lexical rationale distance; semantic robustness checks	Measures whether receivers invoke similar or varied explanations after seeing the same event under a given representation regime.
BELIEF	What directional expectation is formed?	Positive-prediction rate; belief dispersion	Measures directional skew in expected return and numeric dispersion across receivers. Positive-prediction rate is a convergence measure rather than a diversity measure.
ACTION	What action is emitted?	Action entropy; action diversity	Measures whether the regime concentrates decisions into buy, hold, or sell, or leaves action choices dispersed across the receiver population.
QUALITY	Does the trace align with the hidden outcome?	Directional accuracy; strict action accuracy	Serves as an outcome guardrail, using short-window market-reaction alignment rather than investor welfare.

Table 4: E2A diagnostic layers. Metrics are computed within matched event–regime cells and aggregated at the event level before regime contrasts. Formal metric definitions are in Appendix A.

A.2 Rationale

The primary rationale-diversity metric is a deterministic lexical-distance measure over the stated key reasons. Let $\mathbf{v}(\rho_i)$ denote a count vector constructed from word-token features and character 3-grams. Cosine distance is

$$d_{\cos}(p, q) = 1 - \cos(p, q). \quad (7)$$

Rationale diversity is

$$D_{\text{rat}}(b) = \frac{1}{|\mathcal{P}_b|} \sum_{(i,j) \in \mathcal{P}_b} d_{\cos}(\mathbf{v}(\rho_i), \mathbf{v}(\rho_j)). \quad (8)$$

This metric is deterministic and does not depend on an external embedding model.

For semantic robustness checks, we replace $\mathbf{v}(\cdot)$ with a sentence embedding $\phi(\cdot)$ from paraphrase-multilingual-mpnet-base-v2 (Reimers and Gurevych, 2019). Semantic distances are normalized to $[0, 1]$:

$$d_{\text{sem}}(p, q) = \frac{1 - \cos(p, q)}{2}. \quad (9)$$

We compute semantic reason distance over `key_reasons`, semantic uncertainty distance over `uncertainty_notes`, and a composite distance over their concatenation. These measures are used only as robustness checks.

A.3 Belief

Positive-prediction rate is the share of traces that predict a positive short-window return:

$$P_+(b) = \frac{1}{n_b} \sum_{i \in \mathcal{I}_b} \mathbf{1}\{\hat{y}_i > 0\}. \quad (10)$$

Unlike the diversity metrics, P_+ is a convergence measure: higher values indicate stronger alignment toward positive-return expectations.

Belief dispersion is reported as mean absolute deviation around the cell mean:

$$\text{MAD}_{\text{bel}}(b) = \frac{1}{n_b} \sum_{i \in \mathcal{I}_b} |\hat{y}_i - \bar{y}_b|, \quad (11)$$

where

$$\bar{y}_b = \frac{1}{n_b} \sum_{i \in \mathcal{I}_b} \hat{y}_i. \quad (12)$$

A.4 Action

Let $p_a(b)$ denote the share of traces in cell b taking action $a \in \{\text{buy, hold, sell}\}$.

Action entropy is

$$H_{\text{act}}(b) = - \sum_a p_a(b) \log_2 p_a(b), \quad (13)$$

with $0 \log 0 = 0$ by convention (Shannon, 1948). Higher entropy indicates more dispersed actions.

We also report Simpson action diversity:

$$D_{\text{act}}(b) = 1 - \sum_a p_a(b)^2. \quad (14)$$

Both action entropy and Simpson diversity are treated as action-dispersion measures.

A.5 Quality Guardrails

Quality metrics use the hidden five-day cumulative abnormal return, $\text{CAR}_{[1,5]}$, as a diagnostic guardrail. Let

$$AR_{e,d} = R_{e,d} - R_d^M \quad (15)$$

denote abnormal return for event e on day d , relative to the market return R_d^M . The hidden outcome is

$$y_e = \text{CAR}_{[1,5],e} = \sum_{d=t_0+1}^{t_0+5} AR_{e,d}. \quad (16)$$

Outcome values are joined to decision traces only after all upstream representations and downstream decisions are frozen.

Let

$$\text{sgn}_\epsilon(x) = \begin{cases} +1, & x > \epsilon, \\ 0, & -\epsilon \leq x \leq \epsilon, \\ -1, & x < -\epsilon. \end{cases} \quad (17)$$

Directional accuracy is

$$Q_{\text{dir}}(e, r, i) = \mathbf{1}\{\text{sgn}_\epsilon(\hat{y}_i) = \text{sgn}_\epsilon(y_e)\}. \quad (18)$$

Strict action accuracy is

$$Q_{\text{act}}(e, r, i) = \mathbf{1}\{g(a_i) = \text{sgn}_\epsilon(y_e)\}. \quad (19)$$

These metrics are not estimates of investor welfare. They are used to test whether convergence, framing attenuation, or increased rationale diversity is accompanied by improved alignment with a hidden short-window outcome.

A.6 Aggregation

For each metric m , we first compute an event-regime value $M_m(e, r)$. When a metric is defined at the trace level, such as directional accuracy, we average over valid receiver traces within the event-regime cell. When a metric is defined over pairs, such as source diversity or rationale distance, we compute the pairwise quantity within the event-regime cell.

Regime contrasts are paired at the event level:

$$\Delta_m(e; r_a, r_b) = M_m(e, r_a) - M_m(e, r_b). \quad (20)$$

The reported contrast is the mean over events:

$$\bar{\Delta}_m(r_a, r_b) = \frac{1}{|E|} \sum_{e \in E} \Delta_m(e; r_a, r_b). \quad (21)$$

This aggregation makes events, rather than individual traces, the unit of inference. It prevents the large number of role-conditioned traces generated from the same source packet from being treated as independent observations.

A.7 Inference

We use event-cluster bootstrap resampling with 20,000 replicates. Each replicate samples the 94 events with replacement and recomputes the mean event-level contrast. We report percentile 95% confidence intervals and two-sided bootstrap p -values.

For action-diversity families involving multiple related action-dispersion tests, we apply Benjamini-Hochberg correction and report adjusted p -values as p_{BH} . For individual contrasts in the belief and quality layers, we report unadjusted two-sided bootstrap p -values. Model-family robustness is assessed by recomputing the same event-level contrasts separately for each downstream model family.

B T4 as a No-Synthesis Format Anchor

T4 is included as a no-synthesis format anchor. It exposes the complete canonical evidence bank without LLM narrative synthesis, but renders it as a pipe-delimited structured ledger with evidence categories, source IDs, claims, values, and source quotes. This makes T4 useful for testing whether no-synthesis behavior is stable across presentation formats.

We do not treat T4 as a clean de-narrativization or de-framing treatment. The reason is that several changes occur together. First, the ledger format suppresses typographic and discourse-order cues present in the raw earnings release. Second, grouping rows by evidence category changes the local neighborhood in which evidence is encountered. Third, the structured table encourages anchor-scanning behavior: receivers often select one or two source IDs per reason rather than reading continuously across a narrative packet. Fourth, a minority of rendered rows experienced partial claim-text truncation during deterministic field rendering; source IDs and category labels remained intact, but some claim text was shortened.

For these reasons, T4 should be interpreted conservatively. It tests whether the no-synthesis cluster is robust to a structured interface, not whether source framing has been cleanly removed. In the main results, T4 is therefore used as a boundary

condition rather than as a primary causal treatment.

Empirically, T4 supports the no-synthesis interpretation at the rationale layer. Its lexical rationale diversity is close to B0 (0.334 vs. 0.338), suggesting that the main synthesis/no-synthesis split is not driven only by whether canonical evidence is rendered as plain text or as a ledger. However, T4 belief and action effects are less stable: pooled positive-prediction rate falls below B0, while action entropy is not robust across downstream model families. We therefore report T4 for completeness but avoid using it as the central evidence for source-framing attenuation.

C Semantic Rationale-Diversity Sensitivity in the Style-Added Matrix

The primary rationale-diversity metric uses deterministic lexical features so that the metric does not depend on an additional model. As a sensitivity check on the original style-added synthesis matrix, we also compute semantic distances using sentence embeddings from paraphrase-multilingual-mpnet-base-v2 (Reimers and Gurevych, 2019). We compute three semantic variants: semantic reason distance over `key_reasons`, semantic uncertainty distance over `uncertainty_notes`, and a composite distance over their concatenation.

Table 5 reports treatment means. Here T2* and T3* denote the neutral-style ablation, not the primary no-style conditions. Independent style-added summaries produce the highest semantic diversity, while B0 and T4 remain close. This result supports the same coordination interpretation as the primary lexical metric, but it is not a semantic rerun of the primary no-style matrix.

Two patterns are worth noting. First, T3* exceeds T2* on all three semantic variants, consistent with the primary lexical finding that independent summaries preserve more rationale variation than a shared summary. Second, uncertainty distance is consistently larger than reason distance within this style-added matrix.

D Quality-Threshold Sensitivity in the Style-Added Matrix

The quality metrics use short-window market-reaction alignment as a diagnostic guardrail. Because hold decisions depend on the neutral-region width, we test several thresholds ϵ in $\text{sgn}_\epsilon(\cdot)$. This sweep was run on the original style-added T2*/T3*

matrix and the evidence anchors; it is a sensitivity analysis rather than a rerun of the primary no-style conditions.

We evaluate $\epsilon \in \{0.00, 0.01, 0.02, 0.03\}$ for both directional accuracy and strict action accuracy. Across these thresholds, no condition in that matrix produces a robust accuracy gain.

The strict action metric is most sensitive to the hold threshold because it maps buy, hold, and sell directly onto positive, neutral, and negative realized outcomes. This sensitivity does not alter the qualitative result. Style-added synthesis shifts more decisions from buy to hold, but the shift does not produce a robust accuracy gain under any tested threshold.

E Prompt-Style Ablation for T2/T3 Synthesis

E.1 Motivation and Design

The primary T2/T3 synthesis prompts require source grounding, source-ID citation, no recommendation, no leakage, and a 350–500 token length bound, but do not prescribe an analyst-style register. To test whether an apparently neutral stylistic instruction changes downstream traces, we construct a style-added ablation by adding the phrase “*a neutral analyst-style summary*” to otherwise identical T2/T3 prompts. We denote the resulting conditions T2* and T3*.

This ablation isolates the contribution of the style instruction: the canonical evidence bank, upstream model, output constraints, receiver profiles, and downstream decision prompts are unchanged. It is not a no-prompt comparison; role conditioning remains a downstream receiver-profile feature in every condition.

The style-added ablation contains 18,048 decision traces (9,024 per condition) over the same 94 post-knowledge-cutoff events, four receiver model families, and six role-conditioned profiles as the core comparison. This matrix is a diagnostic ablation rather than part of the primary five-regime comparison.

E.2 Results

Belief and action. Adding the neutral analyst-style instruction lowers positive expected-return signs (P^+) by 2.6 pp for T2 and 2.2 pp for T3, and lowers buy share by 4.8 and 4.3 pp respectively, while increasing hold share by 3.3 and 2.6 pp. The

Regime	Reason distance	Uncertainty distance	Composite
T1	0.080	0.135	0.068
B0	0.079	0.144	0.066
T2*	0.092	0.147	0.075
T3*	0.105	0.162	0.086
T4	0.078	0.145	0.066
T3* – T2*	+0.013	+0.015	+0.011
Positive event share	0.714	0.670	0.722
T3* – B0	+0.026	+0.018	+0.019
Positive event share	0.777	0.689	0.755

Table 5: Semantic sensitivity checks for rationale diversity in the style-added T2*/T3* matrix. Distances are normalized to $[0, 1]$. Positive event share is the fraction of events for which the first regime exceeds the second.

style instruction is therefore associated with an additional conservative decision shift.

Attenuation decomposition. Using B0 as the canonical-evidence anchor ($P^+ = 0.800$), the primary T2 and T3 conditions reduce P^+ by 2.14 and 2.25 pp respectively (Figure 1a and its embedded table). These estimates are directionally consistent but are not conventionally significant ($p=.107$ and $p=.130$, event-cluster bootstrap, 20,000 replicates). Adding the neutral analyst-style instruction yields larger reductions of 4.75 pp for T2* and 4.43 pp for T3*. Thus, the style instruction accounts for approximately half of the larger attenuation observed in the style-added conditions. The main text therefore does not present that larger attenuation as a synthesis effect independent of prompt wording.

Action distribution. Figure 1b shows a two-step pattern. Under the primary T2/T3 conditions, buy shares are 59.8%, 6.4 pp below B0 (66.2%), while hold shares are 24.4–24.5%, above B0 (18.8%). Adding the neutral analyst-style instruction produces an additional conservative shift, raising hold to 27.1–27.6% and lowering buy to 55.0–55.6%. The style-added conditions therefore accentuate, rather than define, the observed buy–hold pattern.

Representation bottleneck. The T3-over-T2 source-diversity advantage in the primary comparison (+0.112) also appears under the style-added prompts (+0.103 for T3* over T2*), with event-cluster confidence intervals excluding zero in both cases. Rationale distance shows the same pattern (+0.029 vs. +0.030). The representation bottleneck finding is therefore not an artifact of the analyst-style instruction.

Quality and entropy. Strict action accuracy is 0.397–0.400 under primary T2/T3 synthesis, compared with 0.381–0.383 under the style-added con-

ditions and 0.425 under T1 raw disclosure. The quality-cost claim should therefore remain qualified: the style instruction is associated with a lower observed accuracy level, but the primary T2/T3 comparison does not establish a resolved quality effect relative to B0. T3–T2 action entropy (+0.016, $p=.068$) is smaller and marginal; source and rationale diversity remain the preferred bottleneck evidence.

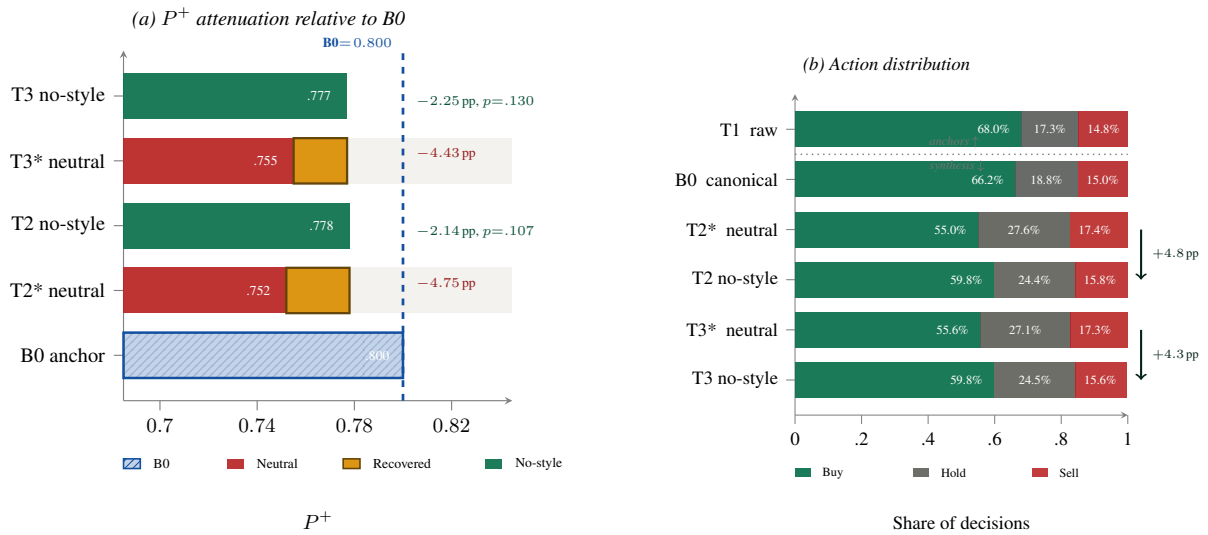
Receiver-model heterogeneity. The buy-share change associated with adding the style instruction is directionally consistent across all four downstream model families but varies in magnitude. DeepSeek-V3.1 shows the largest decrease (11.7 pp for T2 and 8.5 pp for T3); GPT-5.2 shows decreases of 5.2 and 5.7 pp; Claude Sonnet 4.5 and Qwen3-235B-A22B move in the same direction with smaller magnitudes. The direction is thus consistent across receiver families, while its magnitude is receiver-model dependent.

E.3 Claim-Level Summary

Two findings are robust to the prompt-style ablation: the representation bottleneck result (T3 yields more diverse source use and rationales than T2) and the stability of the T1/B0 anchor levels ($P^+ = 0.800$ – 0.801). The primary T2/T3 P^+ changes are directionally negative but statistically unresolved. Adding the neutral analyst-style instruction produces a larger conservative shift in both P^+ and the buy–hold distribution. The quality effect relative to B0 remains unresolved under the primary T2/T3 conditions.

F Mechanism, Scope, and Power Diagnostics

These diagnostics answer different questions and do not form an additive decomposition of the pri-



Condition	No-style primary		Neutral-style ablation	
	P^+	Residual vs B0	P^+	Attenuation vs B0
T1 raw disclosure	0.801	+0.18 pp	—	—
B0 canonical (anchor)	0.800	—	—	—
T2 (shared summary)	0.778	−2.14 pp [†]	0.752	−4.75 pp Additional: −2.62 pp (55%)
T3 (indep. summaries)	0.777	−2.25 pp [†]	0.755	−4.43 pp Additional: −2.18 pp (49%)

[†]Not conventionally significant ($p = .107$ for T2, $p = .130$ for T3; event-cluster bootstrap, 20,000 replicates, 94 events). Style ablation: 18,048 diagnostic traces. T3–T2 source diversity +0.112; rationale distance +0.029; both CIs exclude zero.

Figure 1: **Prompt-style ablation for T2/T3 synthesis.** T2*/T3* add only “a neutral analyst-style summary” to the upstream T2/T3 prompts; all other constraints are held fixed. **(a)** P^+ attenuation: the style prior explains approximately 50% of the attenuation observed in T2*/T3*; the residual attenuation in primary T2/T3 is directional but not conventionally significant. **(b)** Action distribution: T2/T3 partially recover buy share relative to T2*/T3*, but remain below the T1/B0 evidence anchors. Table: P^+ decomposition.

mary regime effects. Full-policy T5, T6, and E1 intervals use the same 94-event clustered bootstrap as the main analysis; the T5 edit-bearing interval uses the 11 events with approved edits.

T5: feasible local lexical edits. T5 scans the claim fields of the 6,026-unit canonical bank for four pre-specified forms of framing: self-evaluation, promotional superlatives, rhetorical intensifiers, and speaker stance or self-credit. The detector flags 33 candidate claim records containing 39 annotated spans. Manual review approves 13 claim-level edits covering 19 spans across 11 events and rejects 20 candidate records: 10 contain necessary comparators or legal qualifications, five leave no complete operational proposition, and five require a nonlocal rewrite. The detector records 5,993 additional non-candidate identity rows. The final four-model matrix contains 4,512 decision

T5 estimand	Events	ΔP_+	95% CI
Full policy	94	−0.09 pp	[−0.71, +0.49] pp
Edit-bearing	11	−1.89 pp	[−4.55, 0.00] pp

Table 6: T5 lexical-control results. Full-policy inference includes all 94 events; edit-bearing inference uses the 11 events containing at least one approved claim edit.

traces, split equally between a contemporaneous B0 replay and T5. The non-detection is limited to this narrow feasible-edit surface and does not characterize wording in general.

T6: fixed-content evidence order. T6 permutes canonical evidence units while preserving their content, multiset, format, and citation structure. It contributes 2,256 traces and uses the B0 confirmation replay from the T5 matrix as its paired anchor.

T6 – B0 replay endpoint	Estimate	95% CI
P_+	−4.30 pp	[−6.87, −1.91] pp
Mean expected return	−0.49 pp	[−0.71, −0.28] pp
Buy rate	−7.36 pp	[−9.84, −5.01] pp
Directional accuracy	−0.93 pp	[−3.63, +1.68] pp
Strict action accuracy	−2.08 pp	[−4.61, +0.35] pp

Table 7: T6 fixed-content evidence-order results. The strict-action accuracy TOST with a ± 3 pp margin yields $p = .234$; quality equivalence is not established.

Table 7 reports the endpoint contrasts. The intervention identifies order as one operative channel, but the unresolved quality equivalence test prevents treating it as a proven quality-preserving remedy.

E1: entity-prior hierarchy. E1 crosses 94 events, six profiles, four model families, and four arms, yielding 9,024 valid traces (2,256 per arm). The arms expose: event ID only; recorded sector, market-cap, fiscal-period, and date context with issuer masked; the same context with issuer and ticker visible; or entity-visible raw T1 evidence. Pooled P_+ rises by +46.37 pp from minimal to recorded context (95% CI [+43.57, +49.16]), by +9.00 pp when issuer identity is revealed ([+5.81, +12.15]), and by +22.43 pp when evidence is added ([+15.87, +28.68]). Model-family sensitivity is substantial: for example, the evidence contrast ranges from −0.35 to +33.81 pp in leave-one-family-out analyses. E1 characterizes absolute-level scope; it neither decomposes nor eliminates possible prior–regime interactions.

Outcome balance and detectable effects. The realized outcome contains 44 positive- and 50 negative-CAR events, so the roughly 80% predicted-positive share under T1/B0 is not a positive-outcome base-rate artifact. For directional accuracy, the 94-event design has an approximate 90% minimum detectable effect of 4.3–5.7 percentage points, depending on the paired contrast. The smaller observed primary quality differences should therefore be read as non-detections, not equivalence.

Trace ledger. Table 10 separates the primary and diagnostic matrices. T6 reuses the T5 B0 replay rather than adding another baseline arm; inference remains over 94 event clusters.

Input parity and generation budgets. T1 and B0 have identical source-ID sets for every event (Jaccard = 1.00), but differ in interface, serialization, and mean length: 25,260 versus 24,261

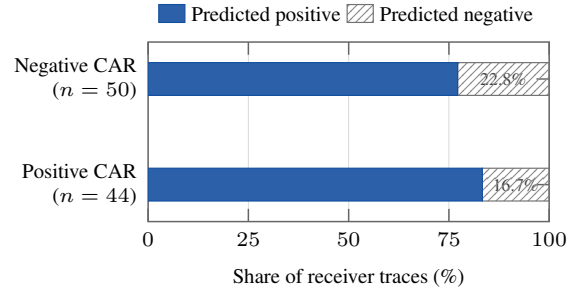


Figure 2: **Predicted-sign composition by realized outcome.** Predicted-positive traces dominate both CAR groups. Bars pool descriptive T1/B0 trace-level rates; n denotes events.

characters and 4,017 versus 3,252 words. Primary no-style T2 and T3 receive the complete canonical bank in every upstream job. Their outputs have similar mean lengths (386.42 vs. 389.68 tokens) and valid-source citation counts (21.46 vs. 21.27); the paired citation-count difference is −0.20 IDs (95% CI [−0.50, +0.06]). These are observed diagnostics, not imposed equal-length or equal-citation budgets.

G Dataset Construction

Issuer universe. We construct a stratified U.S. issuer universe of 188 NYSE/Nasdaq SEC-reporting firms with extractable earnings-related 8-K disclosures, XBRL companyfacts availability, adequate trading liquidity, and adjusted daily price data. The universe comprises 97 large-cap, S&P-500-like liquid issuers and 91 manually curated small/mid-cap issuers.

Exclusion criteria. We exclude issuer classes and event types whose price response is likely to be dominated by mechanisms outside earnings interpretation: financials, banks, insurance, REITs, utilities, OTC securities, SPACs, illiquid microcaps, binary biotech or FDA events, commodity-dominated issuers, and events dominated by merger, bankruptcy, litigation, or major regulatory shocks. We also exclude highly salient mega-cap anchors to reduce the chance that a small number of well-known issuers dominate model priors.

Event selection. We screen post-2026-01-01 8-K Item 2.02 earnings disclosures from the issuer universe, targeting filings with an EX-99.1 or equivalent earnings-release exhibit. Of 212 screened candidate filings, we assemble a 100-event preferred

E1 P_+ contrast	All	–Claude	–GPT	–Qwen	–DeepSeek
Issuer visible – issuer masked	+9.00	+12.41	+2.84	+11.70	+9.04
Issuer masked – context minimal	+46.37	+34.63	+45.39	+60.34	+45.09
Evidence replay – issuer-visible prior	+22.43	+33.81	+28.90	–0.35	+27.36

Table 8: E1 pooled and leave-one-model-family-out contrasts, in percentage points. Pooled 95% CIs are [+5.81, +12.15], [+43.57, +49.16], and [+15.87, +28.68], respectively. The arms vary information availability; the rows are diagnostic and non-additive.

Diagnostic	Events	Value
Primary directional-accuracy 90% MDE	94	4.3–5.7 pp
T5 full-policy P_+ 90% MDE	94	1.02 pp
T5 edit-bearing P_+ 90% MDE	11	4.22 pp
T6 P_+ 90% MDE	94	4.15 pp
T6 strict-action accuracy 90% MDE	94	4.07 pp

Table 9: Normal-approximation minimum detectable effects (MDEs; two-sided $\alpha = .05$, 90% power).

Matrix	Decision traces
Primary T1/B0/T2/T3/T4	24,816
Style ablation T2*/T3*	18,048
Lexical B0/T5	4,512
Order T6 (reuses lexical B0)	2,256
Entity-prior E1	9,024
Integrated corpus	58,656

Table 10: Trace ledger. Counts describe the evaluation corpus, not independent observations; all inferential comparisons use event-level aggregation.

reservoir balanced 50 large-cap / 50 small/mid-cap. We then select the final event execution set by enforcing one event per issuer and requiring complete five-day $CAR_{[1,5]}$ coverage. The final sample contains 47 large-cap and 47 small/mid-cap events across all unique issuers.

Event window. The main sample spans SEC acceptance timestamps from 2026-01-12 to 2026-04-07. A separate 12-event calibration set from November–December 2025 is used only for pipeline validation and profile separability checks; it is excluded from all primary estimates.

Source-packet construction. For each event, we construct the source packet from the 8-K filing, EX-99.1 earnings release, and XBRL company-facts data. We extract a canonical evidence bank in which each unit carries a claim, evidence category, supporting numeric data when applicable, source attribution, and source ID. T2 and T3 upstream summaries draw exclusively from this canonical evidence bank, and runtime citation validation en-

forces that all cited source IDs exist in the treatment packet.

H Highly Actionable Events

We flag 16 events containing explicit forward-looking corporate actions as highly actionable. These include announced buybacks, special dividends, spin-offs, CEO transitions, and material non-GAAP guidance. We retain these events in the main sample because they represent realistic earnings disclosures in which source presentation may directly affect downstream decision traces. This exclusion analysis was run on the original style-added T2*/T3* matrix. Removing the flagged events does not reverse its reported contrast directions, but this check is not transferred to the primary no-style T2/T3 estimates.

I Model Provenance and Leakage Controls

This appendix reports model identifiers, cutoff provenance, access conditions, and prompt visibility boundaries. All models are accessed through API calls with browsing and external tool use disabled.

Pipeline. The pipeline has two stages. In the upstream stage, models render T2/T3 narrative representations from the canonical evidence bank. T2 uses one shared summary within an event–regime cell; T3 uses independently generated summaries over the same canonical bank. In the downstream stage, role-conditioned LLM receivers read one treatment packet and emit a structured decision trace.

Stage 1 audits upstream representation validity across model families. Stage 2 crosses the shared and independent synthesis interfaces with all four downstream model families. The no-synthesis regimes T1, B0, and T4 are upstream-independent baselines. Upstream synthesis temperature is fixed at 0; downstream decision temperature is fixed at

Model family	API identifier	Cutoff	Cutoff evidence
Claude net 4.5	Son- claude-sonnet-4-5-20250929	Jul 2025	First-party / provider documentation
GPT-5.2	gpt-5.2	Aug 2025	First-party documentation
Qwen3-235B-A22B	qwen/qwen3-235b-a22b-2507	Mar 2025	Provider-reported catalog
DeepSeek-V3.1	deepseek/deepseek-chat-v3.1	Mar 2025	Provider-reported catalog

Table 11: Model provenance. The binding leakage constraint is the latest cutoff in the panel, August 2025. Main-sample events occur from 2026-01-12 to 2026-04-07.

Cluster	Profile	Style
Institutional	Proprietary trader	Aggressive, tactical
Institutional	Event-driven hedge fund analyst	Active, risk-adjusted
Institutional	Investment advisor	Conservative, suitability-aware
Retail	Day trader	Short-horizon, high-turnover
Retail	Swing trader	Directional, medium-risk
Retail	Long-term retail investor	Fundamental, hold-biased

Table 12: Receiver profiles. Profile identity is applied only at the downstream stage; no upstream summary prompt is profile-conditioned.

0.2. Downstream model family is treated as a robustness stratum.

Prompt visibility boundary. The experiment is entity-visible and outcome-blind. Treatment-generation and downstream prompts may include company identity, ticker, filing date, fiscal period, source packet text, evidence IDs, source IDs, XBRL facts, pre-event price context, representation regime, receiver profile, and decision seed. They exclude $CAR_{[1,5]}$, realized post-event returns, hidden outcome valence, post-event analyst revisions, future filings or news, outcome-derived labels, post-hoc quality metrics, and analyst recommendation labels not present in the source packet.

Leakage controls. We apply source-ID validation, canonical-bank consistency checks, outcome-blindness audits, model-family memory probes, and deterministic JSON repair rules. The realized outcome is joined to decision rows only after all upstream representations and downstream outputs are frozen.

J Profiles and Separability

We instantiate six role-conditioned receiver profiles. The profiles are controlled decision policies, not behavioral simulations of human investors.

Separability validation. We validate profile separability under T1 raw disclosure using planned action-share contrasts. Hold-heavy profiles hold

more often than buy-active profiles, and buy-active profiles buy more often than hold-heavy profiles. In the full event sample, long-term retail holds 45.7 percentage points more often than day trader, and investment advisor holds 32.4 percentage points more often than proprietary trader. Conversely, day trader buys 25.5 percentage points more often than long-term retail, and proprietary trader buys 19.9 percentage points more often than investment advisor. All four contrasts reach $p < 0.001$ under event-cluster bootstrap resampling.

K Representation and Trace Audits

We audit source packets, mediated representations, and downstream traces to support the intended propagation interpretation.

Source-packet audit. A stratified sample of canonical evidence units is checked for source alignment, numeric fidelity, claim fidelity, quote fidelity, and absence of post-event leakage. Records failing any criterion are regenerated before the main run.

Representation audit. A cross-treatment audit covers 5,181 representation records across the original style-added T2*, T3*, and T4 matrix. It checks uncited material claims, numeric fidelity, citation validity, and directional labels. These claim-level rates are not transferred to primary no-style T2/T3. Runtime citation validation separately enforces that every cited source ID exists in the corresponding treatment packet.

Downstream trace audit. A stratified audit of 961 rationale claims sampled from the style-added T2*/T3* outputs finds post-event leakage in 0.10% of claims, with 93.96% adjudicated non-hallucinated. These rates do not characterize primary no-style T2/T3. Malformed JSON outputs are handled by deterministic repair rules without

re-querying the model, preventing repair from becoming an additional sampling step.

L Toward Propagation-Aware Evaluation

Our results open a broader research agenda for evaluating language interfaces as components of decision systems. Future work should treat the source–representation–receiver pipeline as the unit of analysis. This agenda calls for controlled benchmarks that hold evidence fixed while isolating ordering, salience, grouping, comparison class, prompt style, and representation sharing; mitigation studies that test not only whether directional skew is reduced, but whether quality is preserved under non-inferiority or equivalence criteria; and replications with human and model receivers in interactive, multi-agent, and cross-domain settings. The goal is not to certify a universally neutral summary, but to make the downstream effects of evidence interfaces measurable, comparable, and governable. Under this view, source faithfulness is the starting point of evaluation, not its endpoint.