

When Summaries Distort Decisions: Information Fidelity in LLM-Compressed Financial Analysis

Hoyoung Lee^{1,2} Suhwan Park¹ Seunghan Lee³ Jun Seo³ Jaehoon Lee³
Sungdong Yoo³ Minjae Kim³ CheolWon Na⁴ Zhangyang Wang⁵ V.Zach Golkhou⁶
Minkyu Kim⁷ Sotirios Sabanis^{8,9,10} Alejandro Lopez-Lira¹¹ Dhagash Mehta¹²
Soonyoung Lee³ Chanyeol Choi² Wonbin Ahn^{3,†} Yongjae Lee^{1,2,†}

¹UNIST ²LinqAlpha ³LG AI Research ⁴Sungkyunkwan University

⁵University of Texas at Austin ⁶J.P. Morgan Chase ⁷State Street Investment Management

⁸University of Edinburgh ⁹National Technical University of Athens

¹⁰Archimedes/Athena Research Centre ¹¹University of Florida ¹²BlackRock

[†]Correspondence: wonbin.ahn@lgresearch.ai, yongjaelee@unist.ac.kr

Abstract

Financial decision-makers face more information than they can directly inspect, making context compression necessary. Yet when large language models (LLMs) compress financial source material, they can alter the investment judgment supported by the original source. We frame this problem as information fidelity: compression loses fidelity when it changes the decision induced by the source. In agentic systems, such losses may recur across intermediate steps and amplify throughout the decision process. Across financial filings and earnings-call transcripts, we find that LLM-based compression can produce fluent and factually plausible compressed contexts that nevertheless alter downstream decisions. We analyze two diagnostic patterns associated with fidelity loss: decontextualization, where salient evidence is retained but separated from the caveats and contextual qualifiers needed for correct interpretation, and model dependency, where different compressors expose different views of the same source. We then propose Agentic Context Compression, which generates multiple candidate compressions and audits their disagreements against the original source. Our results suggest that financial compression should be evaluated not only by efficiency or factuality, but also by its ability to preserve decision-relevant context.

1 Introduction

Large language models (LLMs) are increasingly used to compress long documents into shorter contexts before those contexts are passed to another model or a human decision-maker. In high-stakes domains such as finance, the resulting short context is useful only if it preserves the source evidence needed to support the decision.

A preprint of this work is available at <https://arxiv.org/abs/2606.29251>.

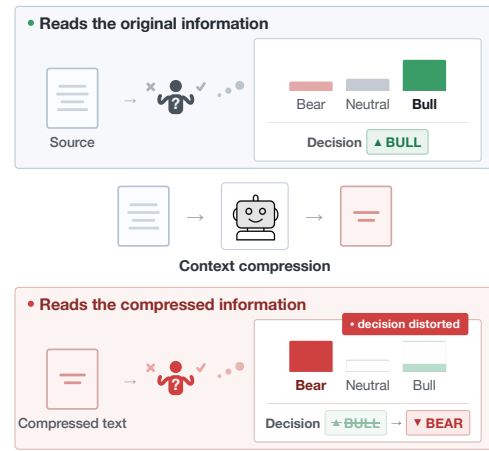


Figure 1: **Compression-induced decision flip.** Context compression can cause a decision-maker to reach a different decision from the one supported by the original source. For example, information that supports a *bullish* decision in the original source may lead to a *bearish* decision after compression.

Long-context compression is also inherently open-ended. A financial source text has no single correct summary because different summaries can select different evidence while remaining fluent and plausible. This paper studies a simple but consequential failure mode: an LLM-compressed representation can lead a decision-maker to make a financial judgment that differs from the judgment supported by the full source text, as shown in Figure 1. The issue is not limited to hallucination or factual inconsistency. It also depends on which evidence remains visible, how it is contextualized, and how competing signals are balanced.

Existing compression evaluations often measure reconstruction, factual consistency, or surface-level information preservation (Łajewska et al., 2025), but they do not directly test whether compression

preserves the decision supported by the original source. We frame this problem as information fidelity, where a compressed representation has high information fidelity if it preserves the decision implied by the original source. A summary may retain headline facts but omit the caveats or offsets needed to interpret them, remaining factually correct while inducing a different decision.

We examine two structured patterns that help characterize fidelity loss. Decontextualization builds on work showing that isolated claims need context to remain meaningful (Choi et al., 2021; Gunjal and Durrett, 2024). Compression may preserve headline claims while stripping away decision-relevant caveats, comparisons, or expectation framing. When such context is removed, a factually correct summary can still carry a different decision implication.

Model dependency is motivated by evidence that investment analyses vary across LLMs (Lee et al., 2025b). In our setting, the relevant variation is not only how models analyze the same evidence, but also which evidence each compressor preserves from the same source. These patterns suggest that decision change is not merely random degradation. Instead, it can reflect systematic differences in the evidence made visible after compression.

These observations motivate a source-grounded compression approach that preserves contextual qualifiers and compares multiple compressed views of the same source. We instantiate this approach as agentic context compression, in which an agent generates competing compressed contexts and audits their disagreements against the original source.

We evaluate information fidelity and compare several compression methods on quarterly filings and earnings-call transcripts. Our main contributions are as follows:

- We introduce *information fidelity*, a decision-centered criterion for evaluating LLM context compression against the original source.
- Using real-world quarterly filings and earnings-call transcripts, we show that LLM-based compression can alter decisions, with decontextualization and model dependency emerging as recurring diagnostic patterns.
- We propose agentic context compression, a source-grounded procedure that audits multiple candidate compressed contexts, and show that it reduces compression-induced decision flips.

2 Related Work

2.1 Context Compression

Context compression has become a central problem in LLM-based systems. Prior work shows that LLM-mediated representations can degrade source information during agentic context updating (Zhang et al., 2026) and delegated document workflows (Laban et al., 2026). Compressed contexts may preserve the general topic while dropping essential details, motivating grounding and information-preservation metrics beyond efficiency (Łajewska et al., 2025). A related line asks what compression should retain: LLM salience aligns only weakly with human judgment (Trienès et al., 2025), and conformal summarization formalizes coverage of critical content in high-stakes settings (Kuwahara et al., 2025). Even recent generative prompt compression that targets coherence and key information (Zhang et al., 2025) is evaluated mainly by task quality or summary metrics, not by preservation of the source-implied decision.

2.2 Summarization Distortion

Repeated generation through LLMs can cumulatively distort source information (Mohamed et al., 2025). Summarization can likewise distort a source by selectively reweighting content and failing to preserve perspectives or polarity balance. FairSumm measures gaps between source and summary perspectives (Zhang et al., 2024); opinion summarization can amplify majority polarity and weaken minority polarity (Lei et al., 2024); and LLMs may implicitly resolve disagreement during generation, suppressing minority viewpoints (Aghaebie et al., 2026). LLM-generated news summaries also introduce framing through selective emphasis and omission (Pastorino and Moosavi, 2026).

LLM-generated summaries can also change readers’ decisions while remaining fluent and plausible. Alessa et al. (2025) find that LLM-generated content can induce framing and primacy biases and shift human purchase decisions, while Peters and Chin-Yee (2025) find that LLM summaries of scientific studies often omit scope-limiting details and overgeneralize conclusions.

These studies show that summarization is not neutral compression but a process of selecting and aggregating perspectives (Mayilvaghanan et al., 2025), where selective omission is itself a form of bias. We extend this insight to financial texts by asking which decision-relevant evidence survives.

2.3 LLMs in Financial Decision-Making

LLMs are increasingly embedded in financial workflows, from semantic stock representations for thematic investing (Lee et al., 2025a) and text-paired time-series construction (Lee et al., 2026) to screening of trading signals (Kim et al., 2026) and alpha extraction from corporate disclosures (Choi et al., 2025). At the same time, research on financial LLMs shows that the investment judgments they produce can be systematically biased by company names, asset classes, prior knowledge, or evaluation design. Glasserman and Lin (2023) show that company names and general knowledge can interfere with financial sentiment prediction. Nakagawa et al. (2024) quantify company-specific bias in financial sentiment analysis. Lee et al. (2025b) analyze model-specific investment preferences and confirmation bias in LLM investment analysis. Kong et al. (2026) place these issues in a broader evaluation framework and argue that financial LLM evaluation must account for bias.

3 Information Fidelity

3.1 Problem Formulation

Task. We define information fidelity as source-relative decision preservation: a compression has high fidelity when the decision-model belief induced by the compressed text remains close to the belief induced by the original source. Given a source text s , a compression model $\mathcal{M}$ produces a compressed text c whose length $|c|$, measured in bullets, must respect a fixed budget B :

$$c = \mathcal{M}(s) \quad \text{subject to} \quad |c| \leq B. \quad (1)$$

We hold B fixed across methods so that fidelity differences reflect what each method retains rather than how much it writes. The realized token ratio $\tau = \text{tok}(c)/\text{tok}(s)$ is reported descriptively.

A decision model $\mathcal{E}$ maps any text to a belief distribution over the label set $V = \{\text{bear}, \text{neutral}, \text{bull}\}$, that is, $\mathcal{E}(\cdot) \in \Delta(V)$, the probability simplex over V . For each fixed text, we average $R = 3$ independent decision runs, yielding a source belief p_s and a compressed belief p_c ,

$$p_s = \frac{1}{R} \sum_{r=1}^R \mathcal{E}^{(r)}(s), \quad p_c = \frac{1}{R} \sum_{r=1}^R \mathcal{E}^{(r)}(c), \quad (2)$$

where $c = \mathcal{M}(s)$ is fixed and $\mathcal{E}^{(r)}$ is the r -th decision-model run. Their top decisions are the most likely labels

$$\hat{v}_x = \arg \max_{v \in V} p_x(v), \quad x \in \{s, c\}. \quad (3)$$

We quantify its loss with two metrics. The Decision Flip rate is the fraction of source documents whose top decision changes after compression,

$$\text{Flip} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\hat{v}_c^{(i)} \neq \hat{v}_s^{(i)}], \quad (4)$$

where i indexes the N source documents, while Total Variation Distance (TVD), $d_{\text{TV}}(p_c, p_s)$, measures how far the belief distribution moves even when the top decision is unchanged,

$$d_{\text{TV}}(p_c, p_s) = \frac{1}{2} \sum_{v \in V} |p_c(v) - p_s(v)|. \quad (5)$$

The flip rate captures outright decision changes, while TVD captures smaller belief shifts that may not cross the decision boundary. We report TVD as its mean over source texts for each method.

3.2 LLM Context Compression

We realize the compression operator $\mathcal{M}$ with calls to a language model. Following Kang et al. (2026), we write $\text{LLM}(x; \theta, \mathcal{P})$ for a model with pretrained weights θ that maps an input x to an output under a natural-language prompt $\mathcal{P}$.

Naive prompt. The naive baseline is a single pass $c = \text{LLM}(s; \theta, \mathcal{P}_{\text{comp}})$: the model receives the full source and a fixed bullet budget, then selects the facts it judges most material. The full prompt is shown in Appendix H.

Contextualization. Contextualization uses the same fixed budget but changes what counts as worth keeping. Instead of choosing facts that look important on their own, it asks the compressor to preserve the details that make each material point interpretable: the relevant comparison, caveat, offsetting signal, or qualifier. Under the same model, source, and budget, it uses $\mathcal{P}_{\text{ctx}}$ to generate candidate contexts $\tilde{c} = \text{LLM}(s; \theta, \mathcal{P}_{\text{ctx}})$. The mechanism is unchanged; only the selection criterion shifts from isolated salience to interpretable evidence.

3.3 Agentic Context Compression

We further consider Agentic Context Compression (ACC), which realizes the same operator $\mathcal{M}$ as a two-stage agentic compression process.

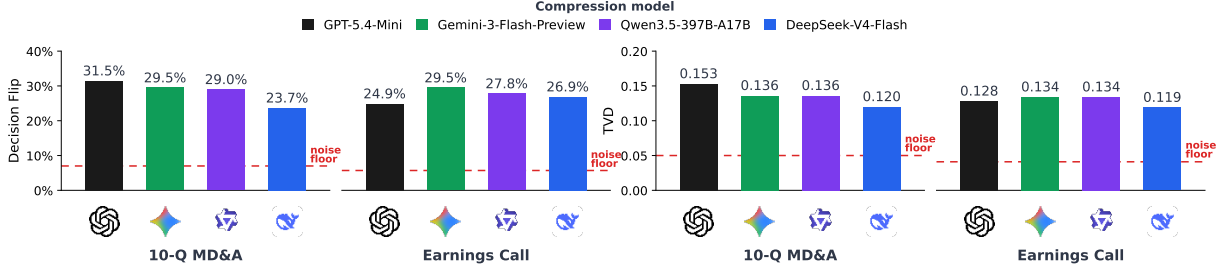


Figure 2: Decision change under one-shot compression. Bars show Decision Flip and source-relative TVD for four compressors across MD&A and earnings calls, averaged across the two decision models. Dashed lines show the decision-model randomness floor: the same decision model reads the original source again, with no compression.

First, the agent calls LLM candidate-generation tools to produce contextualized candidate contexts $c_k = \text{LLM}(s; \theta_k, \mathcal{P}_{\text{ctx}})$ from the same source s , where k indexes a distinct compression model. Second, the agent audits disagreements among the set $\{c_k\}$ against the source. At step t , it emits an action $a_t = \text{LLM}(o_t, \mathbf{h}_{t-1}; \theta, \mathcal{P}_{\text{agent}})$ from the latest tool observation o_t and history $\mathbf{h}_{t-1} = (o_1, a_1, \dots, o_{t-1}, a_{t-1})$. For a claim or omission that differs across candidate contexts c_k , the action can grep the source for the relevant terms and then read only a short source snippet around the matching span, following the direct-corpus-interaction view of agentic search (Li et al., 2026). The agent uses these snippets to test whether a candidate context c_k overclaims the source, for example by making the evidence more bullish, bearish, or certain than the source supports, then scores each c_k by overclaim risk and source-locus coverage, and commits to the intact compressed context that the source best supports (Appendix H.4). In preliminary experiments, allowing the agent to edit candidate text tended to break the preserved context and degrade decision fidelity, so we restrict the agent to selecting among intact candidates rather than revising them. This connects ACC to multi-agent factuality verification (Ning et al., 2025) and financial atomic-claim grounding (Guo et al., 2026), but targets decision-changing omissions and overclaims rather than factuality alone.

4 Experiments

4.1 Experimental Setup

Datasets. Our main panel covers two source types from S&P 100 firms over fiscal 2025 Q1–Q3: 10-Q MD&A sections ($N = 300$), collected from structured SEC filings following EDGAR-Crawler (Loukas et al., 2025), and quarterly earnings call transcripts ($N = 297$).

	10-Q MD&A	Earnings Call
Self-agreement	$90.2 \pm 9.3\%$	$90.5 \pm 7.8\%$
Fleiss’ κ	0.839 ± 0.135	0.819 ± 0.140

Table 1: Reliability of the decision models. We report pairwise top-decision agreement and Fleiss’ κ across three independent reads of the same raw source, as the mean $\pm$ s.d. across the two decision models.

Decision models. We use two decision models, GEMINI-3.1-FLASH-LITE and CLAUDE-HAIKU-4.5, under the same decision prompt: given a source s or compressed text c , each produces the belief distribution used to compute p_s or p_c .

Baselines. We compare against compression baselines under the same budget. **Chunking** partitions s into fixed-size chunks, applies the same compression to each chunk, and concatenates the outputs. For **Token Pruning**, we use LLMingua (Jiang et al., 2023), which drops low-information tokens, and LongLLMingua (Jiang et al., 2024), its long-context extension; both use FINANCE-LLAMA3-8B as the token-scoring model. **Integrator** obtains compressed texts from multiple models and combines them into one final context.

4.2 Fidelity Loss Under Compression

Compression changes the decision model’s output beyond rereading noise. In Figure 2, the dashed line is the decision-model randomness floor: the same model rereads the original source, with no compression. Averaged across the two decision models, rereading still flips 7.0% of MD&A documents and 5.7% of earnings calls, so this floor captures rereading randomness alone.

All four one-shot compressors exceed this floor by more than a factor of two on both source types. The excess is the compression-induced component: summaries remove or reweight decision-relevant

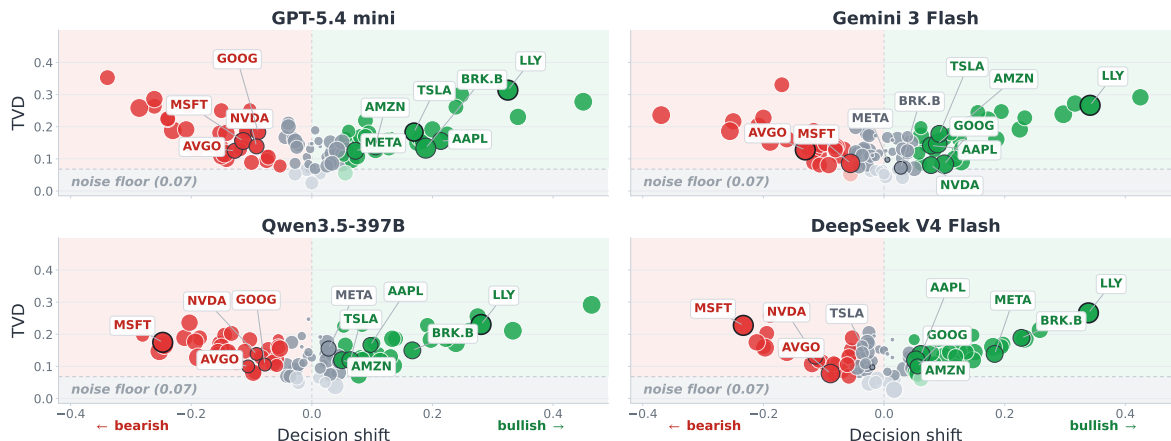


Figure 3: Model-specific decision movement under one-shot compression. Each panel shows one compressor; points are source-relative movements for individual disclosures. The horizontal axis shows signed decision shift from bearish to bullish, the vertical axis shows TVD magnitude, and the dashed line marks the noise floor.

evidence. The parallel rise in TVD shows that compression also shifts the decision model’s belief distribution even without a top-label flip.

Table 1 addresses the possibility that these flips come from unstable decision models. Across three independent reads of the same raw source, both models reproduce their top decisions at high rates, with Fleiss’ κ in the substantial-agreement range or above on both disclosure types. This makes it unlikely that the gap in Figure 2 is only a judge artifact, and supports interpreting the no-compression line as a rereading-noise floor. Per-model results are reported in Appendix D, and the main conclusions persist when the label space is refined to five return bands (Appendix F).

4.3 Model Dependency

All compressors lose fidelity in aggregate, but flip rate reduces movement to a binary outcome (Figure 2). We call this model dependency when the size or direction of belief shifts varies with the compressor. To examine this pattern, we plot source-relative movement by ticker for each compressor.

Figure 3 separates these shifts by compressor. The horizontal axis shows direction, from bearish to bullish, and the vertical axis shows magnitude as source-relative TVD. If compression were model-invariant, the panels would look similar. Instead, compressors occupy different regions: some move tickers toward bearish decisions, others toward bullish decisions, and high-TVD points differ across panels. This pattern suggests that compression-induced decision change is not

merely information loss; single-model summaries can carry a model-specific tilt.

If part of the tilt is compressor-specific, aggregating models may reduce it. The Integrator provides a simple check: it differs from single-model compression only by merging a second model’s summary, and Table 2 shows this alone lowers the MD&A flip rate from 31.5% to 25.6%. ACC reduces it further by drawing on both models and checking their disagreements against the source.

4.4 Decontextualization

We diagnose decontextualization with an offline fact inventory following FActScore and SAFE (Min et al., 2023; Wei et al., 2024); details of the decomposition and role-labeling prompts are in Appendix H.6. We split each source disclosure into atomic facts, link each fact to its source sentence, and tag it as headline, context, or boilerplate. Headline facts carry the main investment signal; context facts calibrate that signal through caveats, offsets, comparisons, or expectation framing; boilerplate facts are background or procedural text.

The second diagnostic pattern is decontextualization: under a fixed budget, compression often keeps headline results while dropping the caveats, offsets, comparisons, and forward-looking qualifiers needed to interpret them. Figure 4A shows that this stripping is selective: under MD&A compression, the context share falls from 25% in the source to 9% in the one-shot compression. This disproportionate loss can make a factually accurate summary read less calibrated than the source

Method	10-Q MD&A				Earnings Call			
	Flip ↓	TVD ↓	Cost ↓	τ	Flip ↓	TVD ↓	Cost ↓	τ
<i>No compression</i>	7.0±5.6%	0.050	-	100%	5.7±3.3%	0.041	-	100%
LLM (GPT-5.4-MINI)								
Naive Prompt	31.5±2.2%	0.153	\$6.69	4.1%	24.9±1.4%	0.128	\$3.83	4.6%
Chunking	30.1±9.8%	0.133	\$6.03	4.0%	29.8±2.6%	0.138	\$4.26	4.0%
Contextualization	21.2±0.2%	0.096	\$8.06	4.2%	22.1±0.2%	0.111	\$3.82	4.2%
Token Pruning (FINANCE-LLAMA3-8B)								
LLMLingua	49.8±4.5%	0.218	-	4.0%	59.7±12.6%	0.276	-	4.0%
LongLLMLingua	43.0±9.9%	0.200	-	4.0%	40.8±2.4%	0.191	-	4.0%
Multi-LLM (GPT-5.4-MINI + DEEPSEEK-V4-FLASH)								
Integrator	25.6±1.3%	0.121	\$9.76	4.3%	21.7±1.7%	0.114	\$6.52	4.7%
ACC	19.5±1.1%	0.096	\$9.07	4.3%	19.9±1.9%	0.106	\$7.52	4.7%

Table 2: Main results for information fidelity under compression. Flip and TVD are averaged across the two decision models; the $\pm$ on Flip is the s.d. across the two models. *No compression* is the identity floor, where the decision model rereads the source. Flip denotes the rate at which compression changes the source-based decision. TVD denotes source-relative probability drift. Cost denotes total inference cost.

it came from. The pattern is milder for shorter earnings transcripts, where more facts fit under the same budget and context is largely preserved.

We then run an add-back diagnostic on flip cases (Figure 4B). Starting from one-shot summaries that flipped the decision model’s output, we append omitted facts and have the model reread the summary. Restoring dropped context facts recovers the source decision in 37% of MD&A flips, compared with 16–19% for boilerplate or random facts from the same filing. Both placebo conditions append the same number of facts per document as the context condition, so the recovery gap reflects which facts are restored rather than how much text is added. Earnings calls show the same ordering but a narrower gap: 33% for context versus 19–27% for the placebos. This pattern is consistent with missing context being an important contributor, though the add-back exceeds the budget and remains diagnostic rather than deployable.

The gain is largest where the decontextualization diagnostic is strongest: for 10-Q MD&A, where dropped context produces the clearest add-back recovery, contextualization lowers the flip rate from 31.5% to 21.2%. For earnings calls, where context is already better preserved under the budget, the gain is smaller, from 24.9% to 22.1%.

4.5 Source-Grounded Selection and Cost

Table 3 holds ACC’s two candidate compressions fixed and varies only the selection rule. Audited

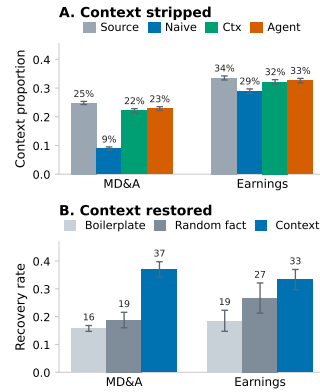


Figure 4: (A) Context facts are dropped much more under MD&A than earnings-call. (B) Re-adding dropped context recovers more flipped decisions than boilerplate or random facts.

selection lowers the flip rate over a seeded random choice by 2.0pp on MD&A and 3.7pp on earnings calls, and on MD&A it matches the best fixed candidate. ACC recovers the best candidate without knowing it in advance (worked example in Appendix G). Paired comparisons (Appendix E) show that the ACC–Naive reduction is significant on both source types and the audited-versus-random gain on earnings calls, while the ACC–Contextualization difference is not.

ACC carries cost overhead relative to a single compression pass. Against the budget-matched Integrator that consumes the same two compressors, it achieves lower flip rates at comparable aggregate cost. Because compressed representations are

Selection variant	10-Q MD&A	Earnings Call
Contextualization A (GPT)	21.3 [16.7, 26.0]	21.5 [17.2, 26.3]
Contextualization B (DEEPSEEK)	20.3 [15.7, 25.0]	22.6 [17.8, 27.3]
Random candidate	22.3 [17.7, 27.3]	22.2 [17.8, 26.9]
ACC (audited selection)	20.3 [15.7, 25.0]	18.5 [14.1, 22.9]

Table 3: Ablation of the selection rule. The two candidate compressions are held fixed, and each row changes only how the final compressed context is chosen. Cells report Flip (%) with 95% bootstrap confidence intervals over documents.

commonly reused downstream, fidelity errors can propagate beyond the initial call, so the additional audit cost is small relative to the cost of decisions made on distorted context.

5 Industry Case Study

To test whether these fidelity effects translate to production forecasting systems, we apply ACC in a globally deployed commercial equity-forecasting product used by institutional investors. The study is conducted on S&P 100 disclosures from fiscal 2025 Q1–Q3. Forecasting quality is measured by the information coefficient (IC), the cross-sectional Spearman correlation between predicted and realized next-quarter returns, holding the forecasting model, stock universe, and evaluation period fixed and varying only the disclosure representation.

Relative to the original disclosures, both naive compression and ACC reduced token consumption by more than 90%. However, naive compression degraded forecasting IC, suggesting that compression-induced decision changes remove forecast-relevant signals. In contrast, ACC improved forecasting IC by 8.3% over the original-source baseline and by 23.8% relative to naive compression, while reducing source-relative decision flips by 59.5% (a relative reduction versus naive compression, not the rate of distorted outputs in production). In production, outputs that change the source-based decision are routed through ACC, and the original disclosure is retained when the discrepancy remains unresolved.

Preserving decision-relevant context thus matters not only for compression quality, but also for retaining downstream forecasting signal.

6 Conclusion

This paper examines fidelity loss in LLM context compression for financial decision-making: compressed contexts can remain fluent and factually

plausible while still changing downstream investment judgments. Our results point to two recurring sources of such loss. One is decontextualization: when key caveats, offsets, or comparative framing are omitted, preserving isolated facts is insufficient. The other is model dependency: compressors shift decisions on the same source in different directions and magnitudes, so a single-model summary can carry a model-specific tilt. To address both, we propose ACC, which audits contextualized candidates from different models against the source and adopts the best-supported one. Under the same budget, ACC better preserves source-relative decisions than one-shot baselines. Overall, compression should be evaluated not only by efficiency or factuality, but also by its ability to preserve decision-relevant context.

7 Limitations

Our experiments focus on financial disclosures from S&P 100 firms, so the findings may not directly generalize to other domains, firm types, or market conditions. In addition, fidelity is measured through decision models rather than human analysts or realized market outcomes, although the main conclusions are consistent across two independent decision models and a refined label space. Future work should validate these findings on broader domains and against human judgments.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-24803208) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2020-II201336, Artificial Intelligence Graduate School Program (UNIST)).

References

- Favour Yahdii Aghaebe, Tanefa Apekey, Elizabeth Williams, and Nafise Sadat Moosavi. 2026. [Faithful Summarisation under Disagreement via Belief-Level Aggregation](#). *Preprint*, arXiv:2601.04889.
- 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.
- Chanyeol Choi, Yoon Kim, Yu Yu, Young Cha, V. Zach Golkhou, Igor Halperin, Georgios Papaioannou, Minkyu Kim, Zhangyang Wang, Jihoon Kwon, Minjae Kim, Alejandro Lopez-Lira, and Yongjae Lee. 2025. [From Text to Alpha: Can LLMs Track Evolving Signals in Corporate Disclosures?](#) *Preprint*, arXiv:2510.03195.
- Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. [Decontextualization: Making Sentences Stand-Alone](#). *Transactions of the Association for Computational Linguistics*, 9:447–461.
- Paul Glasserman and Caden Lin. 2023. [Assessing Look-Ahead Bias in Stock Return Predictions Generated By GPT Sentiment Analysis](#). *Preprint*, arXiv:2309.17322.
- Anisha Gunjal and Greg Durrett. 2024. [Molecular Facts: Desiderata for Decontextualization in LLM Fact Verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768, Miami, Florida, USA. Association for Computational Linguistics.
- Dongxin Guo, Jikun Wu, and Siu Ming Yiu. 2026. [FinGround: Detecting and Grounding Financial Hallucinations via Atomic Claim Verification](#). *Preprint*, arXiv:2604.23588.
- Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. [LLMLingua: Compressing prompts for accelerated inference of large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore. Association for Computational Linguistics.
- Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. [LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand. Association for Computational Linguistics.
- Minki Kang, Wei-Ning Chen, Dongge Han, Huseyin A. Inan, Lukas Wutschitz, Yanzhi Chen, Robert Sim, and Saravan Rajmohan. 2026. [ACON: Optimizing Context Compression for Long-horizon LLM Agents](#). In *Forty-third International Conference on Machine Learning*.
- Sumin Kim, Minjae Kim, Jihoon Kwon, Yoon Kim, Nicole Kagan, Joo Won Lee, Oscar Levy, Alejandro Lopez-Lira, Yongjae Lee, and Chanyeol Choi. 2026. [LLM as a Risk Manager: LLM Semantic Filtering for Lead-Lag Trading in Prediction Markets](#). *Preprint*, arXiv:2602.07048.
- Yaxuan Kong, Hoyoung Lee, Yoontae Hwang, Alejandro Lopez-Lira, Bradford Levy, Dhagash Mehta, Qingsong Wen, Chanyeol Choi, Yongjae Lee, and Stefan Zohren. 2026. [Evaluating LLMs in Finance Requires Explicit Bias Consideration](#). In *Forty-third International Conference on Machine Learning*.
- Bruce Kuwahara, Chen-Yuan Lin, Xiao Shi Huang, Kin Kwan Leung, Jullian Arta Yapeter, Ilya Stanevich, Felipe Pérez, and Jesse C. Cresswell. 2025. [Document Summarization with Conformal Importance Guarantees](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 67107–67152. Curran Associates, Inc.
- Philippe Laban, Tobias Schnabel, and Jennifer Neville. 2026. [LLMs Corrupt Your Documents When You Delegate](#). *Preprint*, arXiv:2604.15597.
- Weronika Łajewska, Momchil Hardalov, Laura Aina, Neha Anna John, Hang Su, and Lluís Marquez. 2025. [Understanding and Improving Information Preservation in Prompt Compression for LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17520–17541, Suzhou, China. Association for Computational Linguistics.
- Hoyoung Lee, Wonbin Ahn, Suhwan Park, Jaehoon Lee, Minjae Kim, Sungdong Yoo, Taeyoon Lim, Woohyung Lim, and Yongjae Lee. 2025a. [THEME: Enhancing Thematic Investing with Semantic Stock Representations and Temporal Dynamics](#). In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25*, pages 5797–5804, New York, NY, USA. Association for Computing Machinery.
- Hoyoung Lee, Junhyuk Seo, Suhwan Park, Junhyeong Lee, Wonbin Ahn, Chanyeol Choi, Alejandro Lopez-Lira, and Yongjae Lee. 2025b. [Your AI, Not Your View: The Bias of LLMs in Investment Analysis](#). In *Proceedings of the 6th ACM International Conference on AI in Finance, ICAIF '25*, pages 150–158, New York, NY, USA. Association for Computing Machinery.
- Jaehoon Lee, Suhwan Park, Taeyoon Lim, Seunghan Lee, Jun Seo, Dongwan Kang, Hwanil Choi, Minjae Kim, Sungdong Yoo, Soonyoung Lee, Yongjae Lee, and Wonbin Ahn. 2026. [FinTextTS: Financial Text-Paired Time-Series Dataset via Semantic-Based and Multi-Level Pairing](#). In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD '26*, pages 9266–9277, New York, NY, USA. Association for Computing Machinery.
- Yuanyuan Lei, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Ruihong Huang, and Dong Yu. 2024. [Polarity Calibration for Opinion Summarization](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5211–5224, Mexico City, Mexico. Association for Computational Linguistics.

- Zhuofeng Li, Haoxiang Zhang, Cong Wei, Pan Lu, Ping Nie, Yi Lu, Yuyang Bai, Shangbin Feng, Hangxiao Zhu, Ming Zhong, Yuyu Zhang, Jianwen Xie, Yejin Choi, James Zou, Jiawei Han, Wenhui Chen, Jimmy Lin, Dongfu Jiang, and Yu Zhang. 2026. [Beyond semantic similarity: Rethinking retrieval for agentic search via direct corpus interaction](#). *Preprint*, arXiv:2605.05242.
- Lefteris Loukas, Fabian Billert, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2025. [EDGAR-CRAWLER: From Raw Web Documents to Structured Financial NLP Datasets](#). In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, pages 761–764, New York, NY, USA. Association for Computing Machinery.
- Kawin Mayilvahanan, Siddhant Gupta, and Ayush Kumar. 2025. [Spot the BlindSpots: Systematic Identification and Quantification of Fine-Grained LLM Biases in Contact Center Call Summarization](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1299–1340, Suzhou (China). Association for Computational Linguistics.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Amr Mohamed, Mingmeng Geng, Michalis Vazirgianis, and Guokan Shang. 2025. [LLM as a broken telephone: Iterative generation distorts information](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7493–7509, Vienna, Austria. Association for Computational Linguistics.
- Kei Nakagawa, Masanori Hirano, and Yugo Fujimoto. 2024. [Evaluating Company-specific Biases in Financial Sentiment Analysis using Large Language Models](#). In *2024 IEEE International Conference on Big Data (BigData)*, pages 6614–6623. IEEE.
- Yucheng Ning, Xixun Lin, Fang Fang, and Yanan Cao. 2025. [MAD-Fact: A Multi-Agent Debate Framework for Long-Form Factuality Evaluation in LLMs](#). *Preprint*, arXiv:2510.22967.
- Valeria Pastorino and Nafise Sadat Moosavi. 2026. [Frame In, Frame Out: Measuring Framing Bias in LLM-Generated News Summaries](#). In *Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 378–384, San Diego, California, United States. Association for Computational Linguistics.
- Uwe Peters and Benjamin Chin-Yee. 2025. [Generalization Bias in Large Language Model Summarization of Scientific Research](#). *Royal Society Open Science*, 12(4):241776.
- Jan Trienes, Jörg Schlötterer, Junyi Jessy Li, and Christin Seifert. 2025. [Behavioral Analysis of Information Saliency in Large Language Models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23428–23454, Vienna, Austria. Association for Computational Linguistics.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. [Long-form Factuality in Large Language Models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 80756–80827. Curran Associates, Inc.
- Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. 2026. [Agentic context engineering: Evolving contexts for self-improving language models](#). In *The Fourteenth International Conference on Learning Representations*.
- Tinghui Zhang, Yifan Wang, and Daisy Zhe Wang. 2025. [SCOPE: A Generative Approach for LLM Prompt Compression](#). *Preprint*, arXiv:2508.15813.
- Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, Kathleen McKeown, and Rui Zhang. 2024. [Fair Abstractive Summarization of Diverse Perspectives](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3404–3426, Mexico City, Mexico. Association for Computational Linguistics.

A Budget Sensitivity

As a brief robustness check, Figure 5 repeats one-shot compression with budgets from 5 to 40 bullets on a 150-document subset, evaluated under GEMINI-3.1-FLASH-LITE only. Larger budgets reduce source-relative TVD, especially from 20 to 40 bullets, but decision flips remain well above that model’s no-compression floor. Extra space therefore smooths belief drift more reliably than it eliminates top-label changes.

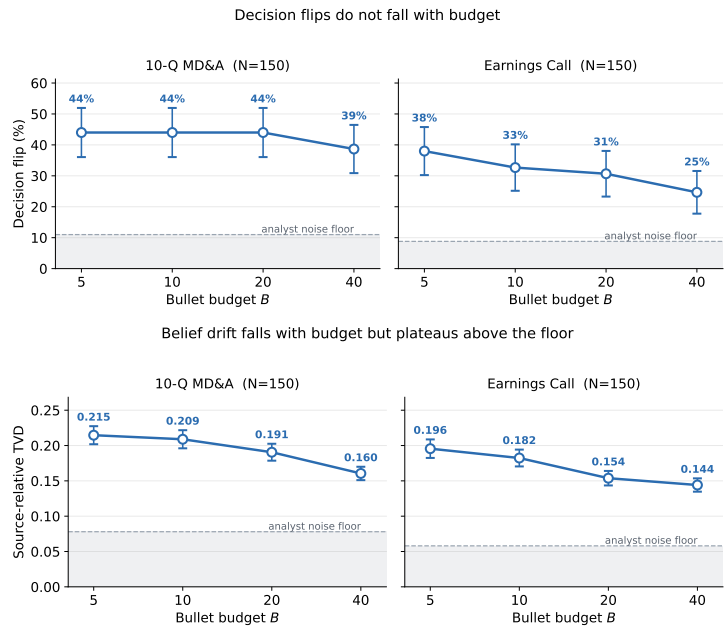


Figure 5: Budget sensitivity under GPT-5.4-MINI one-shot compression. Top: decision flips remain above the decision-model noise floor even at larger budgets. Bottom: source-relative TVD declines with budget but plateaus above the floor.

B Inter-Model Agreement

Aggregate flip rates can hide whether compressors fail on the same filings. Figure 6 reports pairwise top-decision agreement across four one-shot compressors on the 597 source documents pooled across MD&A and earnings-call Q&A. The mean off-diagonal agreement is 0.75, so compressors disagree on roughly one filing in four even when their marginal flip rates are similar. This supports treating compression fidelity as model-dependent rather than as a property of the source alone.

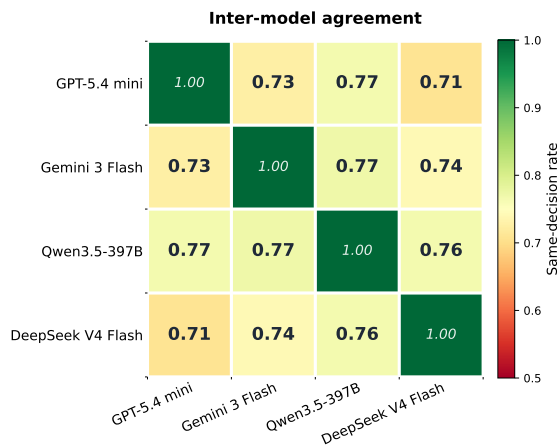


Figure 6: Pairwise top-decision agreement among one-shot compressors on the same source filings. Cells show the share of filings for which two compressors induce the same downstream bear/neutral/bull decision; off-diagonal values near 0.75 indicate that compressor choice changes the induced decision on about one quarter of filings.

C Decision Movement

Figures 7 and 8 show the same movement diagnostic separately for MD&A and earnings-call Q&A. The black vertical marker is the source-implied stance for each ticker, and each colored arrow shows where a one-shot compressor moves the downstream decision. Wider horizontal spread within a ticker indicates stronger model dependence.

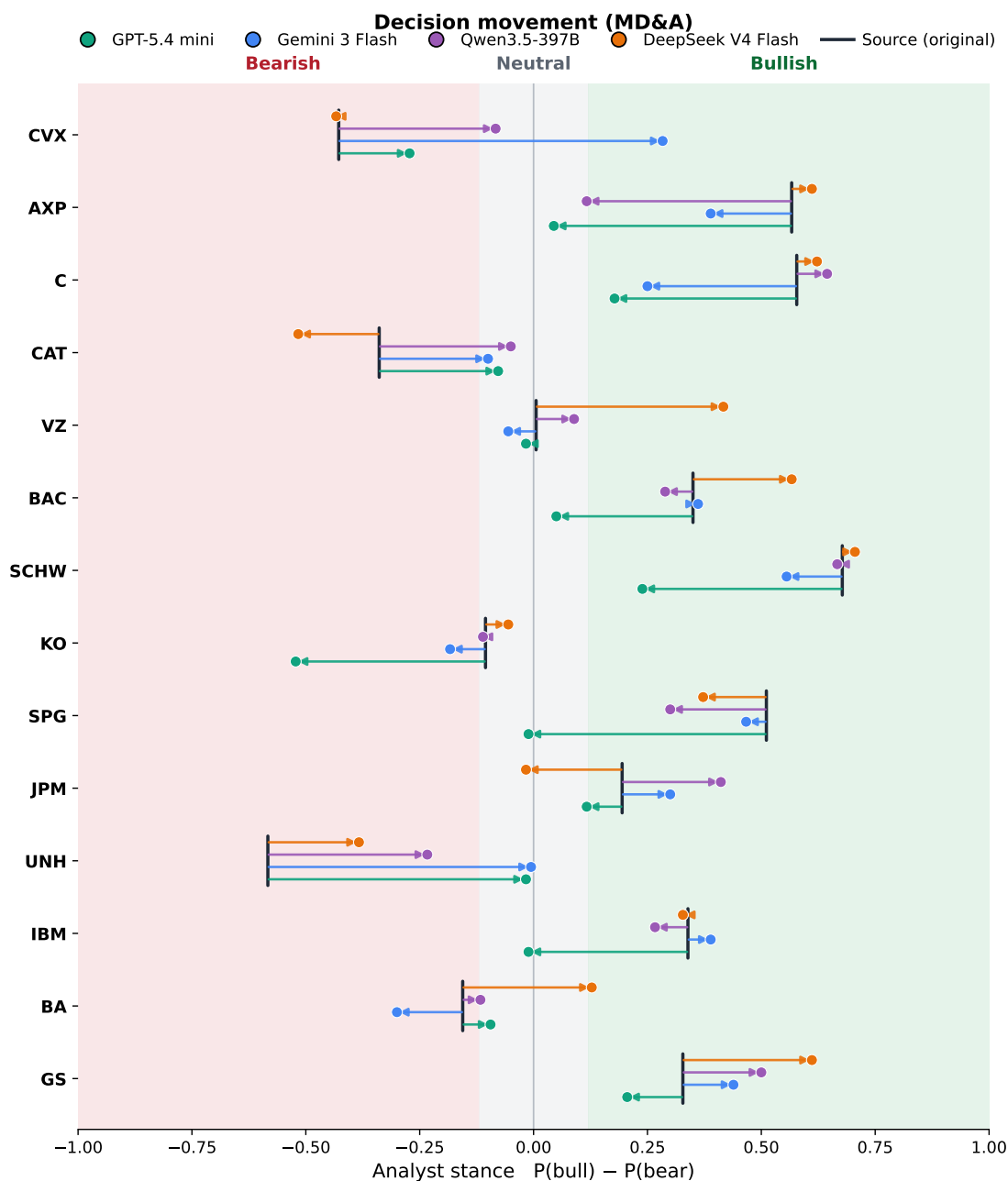


Figure 7: Decision movement for MD&A disclosures.

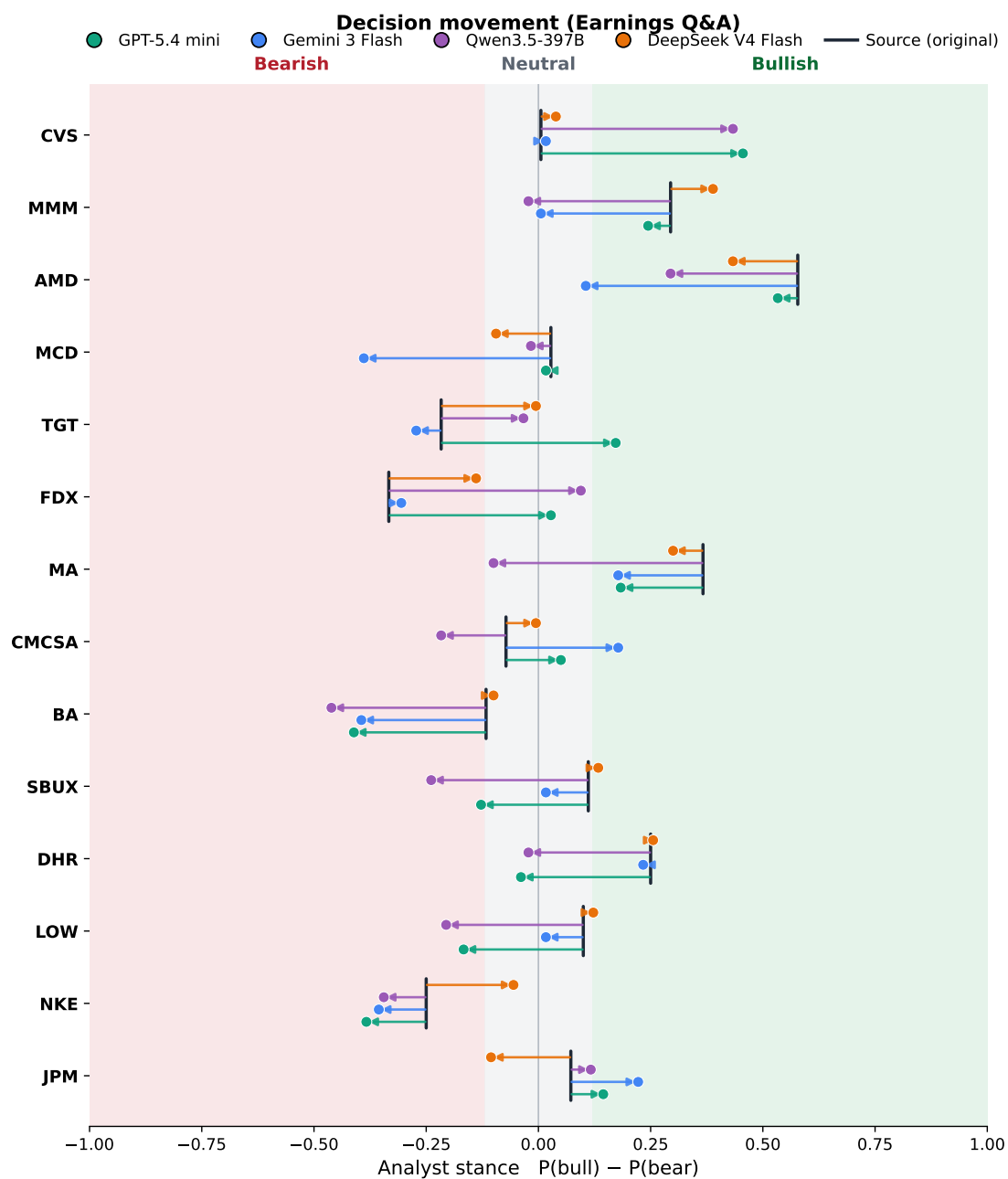


Figure 8: Decision movement for earnings-call Q&A disclosures.

D Per-Evaluator Results

Table 4 reports the components of Table 2 separately for the two decision models; Flip is shown with the s.d. across each model’s three decision runs. The main contrasts hold under both evaluators: Contextualization and ACC lower the flip rate relative to the Naive prompt on both source types, and the token-pruning baselines remain weakest. The full ordering is not identical, however: on MD&A, Chunking is the weakest LLM method under GEMINI-3.1-FLASH-LITE but outperforms the Naive prompt under CLAUDE-HAIKU-4.5, and on earnings calls the Integrator attains a lower flip rate than ACC under CLAUDE-HAIKU-4.5 (20.5% vs. 21.2%), consistent with framing ACC as source-grounded selection. The candidate-A row of Table 3 re-evaluates the same artifacts under fresh decision reads; its gap from the 22.2% entry here is decision-run resampling noise, well inside the bootstrap CI.

Method	10-Q MD&A		Earnings Call	
	Flip ↓	TVD ↓	Flip ↓	TVD ↓
GEMINI-3.1-FLASH-LITE				
Naive Prompt	33.0 ± 1.6	0.172	23.9 ± 1.4	0.137
Chunking	37.0 ± 1.3	0.155	31.6 ± 0.5	0.147
Contextualization	21.3 ± 1.8	0.099	22.2 ± 3.7	0.116
LLMLingua	53.0 ± 0.7	0.231	50.8 ± 1.8	0.247
LongLLMLingua	50.0 ± 1.0	0.226	39.1 ± 1.9	0.196
Integrator	24.7 ± 1.0	0.128	22.9 ± 1.9	0.128
ACC	20.3 ± 1.7	0.103	18.5 ± 0.4	0.111
CLAUDE-HAIKU-4.5				
Naive Prompt	29.9 ± 0.9	0.135	25.9 ± 0.9	0.118
Chunking	23.1 ± 1.2	0.111	27.9 ± 1.9	0.129
Contextualization	21.1 ± 0.5	0.094	21.9 ± 1.2	0.106
LLMLingua	46.6 ± 1.7	0.205	68.7 ± 3.7	0.306
LongLLMLingua	36.1 ± 1.1	0.174	42.4 ± 1.2	0.186
Integrator	26.5 ± 0.7	0.113	20.5 ± 1.6	0.100
ACC	18.7 ± 2.1	0.089	21.2 ± 0.5	0.101

Table 4: Per-evaluator information fidelity results. Flip (%) is reported ± the s.d. across the three decision runs of each decision model; TVD is the mean source-relative total variation distance.

E Bootstrap Confidence Intervals and Paired Comparisons

Table 5 reports 95% bootstrap confidence intervals (5,000 document resamples) for the main methods under each decision model, and Table 6 reports paired flip deltas over the same documents. Under GEMINI-3.1-FLASH-LITE, the ACC–Naive reduction is significant on both source types and the audited-versus-random gain on earnings calls, while the ACC–Contextualization difference is not; CLAUDE-HAIKU-4.5 shows the same qualitative pattern.

Method	10-Q MD&A		Earnings Call	
	Flip [95% CI]	TVD [95% CI]	Flip [95% CI]	TVD [95% CI]
GEMINI-3.1-FLASH-LITE				
Naive Prompt	33.0 [27.7, 38.3]	0.172 [0.157, 0.186]	23.9 [19.2, 29.0]	0.137 [0.124, 0.150]
Contextualization	21.3 [16.7, 26.0]	0.098 [0.089, 0.108]	22.2 [17.5, 27.3]	0.116 [0.105, 0.127]
Integrator	24.7 [20.0, 29.7]	0.128 [0.117, 0.140]	22.9 [18.2, 27.6]	0.128 [0.116, 0.140]
ACC	20.3 [15.7, 25.0]	0.103 [0.093, 0.112]	18.5 [14.1, 22.9]	0.111 [0.101, 0.122]
CLAUDE-HAIKU-4.5				
Naive Prompt	29.9 [24.8, 35.0]	0.135 [0.121, 0.148]	25.9 [20.9, 31.0]	0.118 [0.106, 0.131]
Contextualization	21.1 [16.7, 25.9]	0.094 [0.082, 0.106]	21.9 [17.2, 26.6]	0.106 [0.095, 0.118]
Integrator	26.5 [21.8, 31.6]	0.113 [0.100, 0.127]	20.5 [15.8, 25.3]	0.100 [0.089, 0.112]
ACC	18.7 [14.3, 23.1]	0.089 [0.078, 0.100]	21.2 [16.5, 25.9]	0.101 [0.091, 0.113]

Table 5: 95% bootstrap confidence intervals (5,000 document resamples) for Flip (%) and TVD under each decision model.

Contrast	10-Q MD&A	Earnings Call
GEMINI-3.1-FLASH-LITE		
Naive → Contextualization	−11.7 [−17.7, −5.3]	−1.7 [−7.4, +4.0]
Contextualization A → random candidate	+1.0 [−2.7, +4.7]	+0.7 [−3.4, +4.7]
Random candidate → ACC	−2.0 [−5.3, +1.3]	−3.7 [−7.4, −0.3]
Naive → ACC	−12.7 [−19.0, −6.3]	−5.4 [−10.4, −0.3]
Contextualization → ACC	−1.0 [−5.7, +3.7]	−3.7 [−8.8, +1.0]
Integrator → ACC	−4.3 [−10.0, +1.3]	−4.4 [−9.8, +0.7]
CLAUDE-HAIKU-4.5		
Naive → Contextualization	−8.8 [−14.6, −3.1]	−4.0 [−9.8, +1.7]
Naive → ACC	−11.2 [−17.0, −5.1]	−4.7 [−10.4, +1.0]
Contextualization → ACC	−2.4 [−7.1, +2.4]	−0.7 [−4.4, +3.0]
Integrator → ACC	−7.8 [−13.3, −2.4]	+0.7 [−4.7, +6.4]

Table 6: Paired flip deltas in percentage points with 95% bootstrap CIs over documents. The candidate-routing contrasts (rows two and three) are defined by the fixed-candidate ablation and reported under GEMINI-3.1-FLASH-LITE only.

F Five-Class Label Space

Keeping the compressed artifacts fixed, we re-run GEMINI-3.1-FLASH-LITE with five next-quarter return bands (strong bear, bear, neutral, bull, strong bull), averaging two independent decision runs. The main conclusion persists (Table 7); on MD&A, contextualization attains the lowest five-class flip rate (18.7% vs. 21.3% for ACC), consistent with the source-grounded-selection framing.

Method	10-Q MD&A		Earnings Call	
	Flip ↓	TVD ↓	Flip ↓	TVD ↓
Naive Prompt	36.0	0.185	30.7	0.165
Contextualization	18.7	0.125	30.7	0.151
Integrator	26.7	0.151	27.3	0.161
ACC	21.3	0.125	23.3	0.149

Table 7: Information fidelity under a five-class next-quarter return label space, with the compressed artifacts held fixed and GEMINI-3.1-FLASH-LITE averaged over two independent decision runs.

G Worked Example of the ACC Audit

In this logged case from a Lockheed Martin earnings call, management describes the Golden Dome opportunity as fluid and not yet quantifiable and warns that charges on legacy programs cannot guarantee zero future issues; the two candidate compressions treat this evidence differently:

Worked audit example (Lockheed Martin, earnings call)

Candidate A (GPT-5.4-MINI) writes:

- Munitions and Golden Dome are the main new upside.

Candidate B (DEEPSEEK-V4-FLASH) writes:

- Golden Dome unquantifiable until architecture and budget set; Lockheed well-positioned.

- Management took charges on legacy helicopter and missile programs to reduce future risk, but cannot guarantee zero issues.

Audit. Candidate A elevates the fluid, not-yet-quantifiable opportunity into the main new upside and omits the legacy-program warning; the audit rates A as higher overclaim risk against the retrieved source spans. Candidate B retains both qualifications and is selected verbatim.

Outcome. Without observing any evaluator outputs, ACC selects candidate B. Under GEMINI-3.1-FLASH-LITE, the source and candidate B induce neutral decisions, whereas candidate A induces a bull decision.

H Prompt Templates

The boxed templates below reproduce the prompts used in our experiments. Template variables are instantiated with the values used for the main results: the bullet budget is $B=20$ and word caps are stated per box. Compression prompts are sent as system messages; the user message supplies the source document and, for the integrator, the model views.

H.1 One-Shot Compression Prompt

One-shot compression

System.

Compress the document below into exactly 20 bullets. This count is STRICT; do not exceed it. Each bullet must be on its own line starting with '- '. Do not write headers or blank lines. Each bullet captures exactly ONE fact or figure. Do not combine multiple facts into a single bullet. If the document contains more facts than 20, select only the 20 most material facts and omit the rest. Output only the bullets.

User.

```
[[{ticker}} {company} {document_label}  
<document>  
{source}  
</document>
```

H.2 Contextualization Compression Prompt

Contextualization compression

Compress the document below into exactly 20 bullets. This count is STRICT; do not exceed it. Each bullet must be on its own line starting with '- '. Do not write headers or blank lines. Each bullet captures exactly ONE decision-relevant point. Do not combine unrelated points into one bullet. The 20 bullets together should be about 240-250 words total. Do not exceed 250 words, but do not make the summary much shorter unless the source lacks material decision-relevant information.

A downstream decision model will make a bear / neutral / bull investment judgment from your bullets alone. Your goal is to preserve the source's decision direction and confidence as closely as possible. The compressed bullets should induce the same overall investment interpretation as the full source.

Selection rules:

- Preserve the source's dominant signal when the source is directionally positive, negative, or neutral.
- Preserve facts that explain why that signal is strong, weak, improving, deteriorating, uncertain, or mixed.
- When including a number, keep the comparison or frame that gives it meaning: prior year, prior quarter, guidance, consensus, management expectation, segment trend, margin impact, liquidity impact, or risk exposure.
- Do not write standalone figures without explaining what they imply.
- Preserve management's framing when it changes interpretation.
- Preserve forward-looking guidance, demand signals, margin or cost pressure, capital allocation, liquidity, and material risks when they affect the investment judgment.
- Include caveats or offsetting evidence only when they materially change the interpretation or confidence.
- Do not mechanically include both positive and negative facts.
- Do not force neutrality if the source has a clear directional signal.
- Avoid boilerplate, immaterial details, and facts that are vivid but unlikely to affect the investment judgment.

Use only what the source actually says; do not add outside information.

Output only the bullets.

H.3 Multi-LLM Integrator Prompt

Multi-LLM integrator

You are a view-only financial-disclosure compression integrator.

You will receive one-shot compressed summaries from different LLMs. Treat the summaries as compressed model views, not as truth. Integrate them into one 20-bullet summary that keeps the decision-relevant information visible across the views.

Rules:

- Use only the provided compressed model views and task metadata.
- Do not use source decision labels, decision-model probabilities, realized outcomes, or external knowledge.
- Do not take a majority vote across views.
- Do not invent facts to fill gaps that no view mentions.
- If views differ, use cautious wording or keep both material perspectives.
- Output exactly 20 bullets.
- Each bullet must start with ‘-’ and capture one view-supported fact or qualifier.
- Output only the bullets, with no heading or commentary.
- The entire 20-bullet summary must total at most 330 words. Keep every bullet concise; drop lower-value detail rather than exceed the word budget.

H.4 Agentic Context Compression

Agent runtime contract

Candidates. The agent adjudicates two one-shot Contextualization summaries of the same disclosure. Candidate A and candidate B are produced independently by two compression models.

Source access. The raw `source.txt` is not placed in the prompt, and the agent does not read the disclosure end-to-end. It uses targeted tools: `get_metadata` identifies the document and source length, `grep_source` searches for terms or claims where candidates disagree, and `read_source_span` opens short local passages around retrieved evidence. These inspections are span-limited; the agent must perform at least four targeted source checks before deciding.

Audit criteria. The audit focuses on disagreements between candidates and checks them against retrieved source spans for *overclaim risk* and *source-locus coverage*. Each checked claim is labeled *supported*, *unclear*, or *overclaiming* and each material source locus *covered*, *partial*, or *missed*; these labels are aggregated into categorical overclaim-risk and coverage scores per candidate.

Selection rule. The agent must select exactly one candidate, A or B. It compares the two candidates using retrieved source spans, chooses the candidate with lower overclaim risk, and uses broader source-locus coverage as the deciding criterion when overclaim risk is comparable.

H.5 Decision Model Prompt

Decision model

System.

You are a financial decision model. You are given a corporate disclosure (or a compressed summary of one). Do not use prior knowledge of the company or any external context.

Estimate the disclosure-implied update, not a full investment recommendation. Use ONLY the provided disclosure evidence. Do not use valuation, stock price, prior returns, external company knowledge, or market context.

Based on the disclosure evidence alone, estimate how the disclosure would update an investor's belief about the stock's NEXT-QUARTER total-return regime:

bear = disclosure evidence points to downside risk ($< -5\%$ if realized)

neutral = evidence is mixed, immaterial, or balanced (-5% to $+5\%$ if realized)

bull = disclosure evidence points to upside potential ($> +5\%$ if realized)

Output ONLY a JSON object with three probabilities that sum to 1.0, no other text:

`{"bear": 0.xx, "neutral": 0.xx, "bull": 0.xx}`

User.

Ticker: {ticker} ({company}, {sector})

Period of report: {period} ({q_label} {year})

Disclosure type: {level_label}

Disclosure:

{material}

Disclosure-implied probability distribution over next-quarter regime (JSON only):

H.6 Fact Decomposition and Role Labeling

Fact decomposition

You are a financial-document fact extractor. Decompose the 10-Q MD&A below into atomic facts. For each fact, return both the atomic claim and the source sentence it was extracted from.

Rules:

- Each fact is ONE atomic claim or figure. Do not combine multiple claims.
- Preserve quantitative detail verbatim (numbers, percentages, dates, entities).
- "source" is the full sentence (or short adjacent span) from the MD&A that contains the fact. Copy it VERBATIM from the document – do not paraphrase or summarize.
- Do NOT assign sentiment, polarity, or positivity/negativity. Sentiment is labeled in a separate step.
- Skip boilerplate (forward-looking disclaimers, table-of-contents text, GAAP definitions) and strictly neutral structural sentences.
- Return ONLY a JSON object: `{"facts": [{"content": str, "source": str}]}`. No prose, no markdown fences.

Role labeling

You are a financial fact-role labeling assistant.

You are labeling atomic facts from a financial disclosure for an information-fidelity study.

The study asks whether compression preserves the context needed to interpret headline investment signals.

You will receive a numbered list of atomic facts from one disclosure. For each atomic fact, assign exactly one evidence-role label:

Use exactly one label:

- "headline": a primary investment signal that could directly move the decision model's assessment. This includes standalone facts about revenue, sales, demand, margins, profitability, costs, cash flow, guidance, growth, decline, market share, customer traction, operational performance, capital allocation, liquidity, or material risk when the fact itself is the main directional signal.
- "context": a fact that calibrates how a headline should be read. Label a fact as context only if it changes the interpretation of a headline signal by providing a caveat or limitation, offsetting evidence, year-over-year/sequential/segment/geography/customer/product comparison, magnitude/baseline/expectation-relative framing, causal driver, timing/durability/one-time nature, forward-looking qualifier, or risk condition/dependency.
- "boilerplate": generic background, corporate description, accounting/legal boilerplate, procedural text, vague management language, or scene-setting that does not materially affect how a headline investment signal should be interpreted.

Important: Do not label a fact as context merely because it is detailed, neutral, hedged, explanatory, legal/regulatory, or generally relevant to the business. A context fact must help interpret the direction, strength, scope, durability, cause, or expectation-relative meaning of a headline.

Decision rule: Ask: "If this fact were removed, would the decision model interpret the relevant headline signal differently?" If yes, and the fact calibrates another signal, label it "context". If the fact itself is the main directional signal, label it "headline". If it does not affect interpretation of a decision-relevant signal, label it "boilerplate".

Tie-break rules:

1. If a fact contains both a headline result and a qualifier that changes how the result should be read, choose "context".
2. If it is a standalone KPI/result/metric/guidance/capital-allocation fact, choose "headline".
3. If it is generic, procedural, definitional, or only broadly relevant, choose "boilerplate".

Output exactly one JSON object and no prose:

```
{ "labels": [{ "fact_id": 0, "role": "<headline|context|boilerplate>", "context_function": "<none|caveat|offset|comparison|expectation_framing|causal_driver|timing_durability|scope_magnitude|risk_condition>", "target_headline": "<short description or null>", "rationale": "<one short sentence>" } ] }
```