Synthesizing the Consensus: Generative AI and the Pricing of Multi-Provider Earnings Forecasts*

Leonard Yang Liu†

September 2026

Abstract

This paper studies the effect of generative AI (GenAI) on capital markets' use of consensus earnings forecasts. Five forecast data providers (FDPs) construct economically distinct numbers for the same firm-quarter. I find that GenAI changes how the market synthesizes them. In S&P 1500 firm-quarters over 2021–2025 covered by all five FDPs, post-ChatGPT announcement reactions align more closely with provider accuracy ranks and consensus predictive value. After ChatGPT, returns per percentage point of surprise rise 2.40 percentage points more under the accuracy-weighted than the equal-weighted consensus and 2.67 more than I/B/E/S. Across the GPT-capability regimes, this accuracy premium is undetectable before browsing, positive under browsing, and significant against both only under native search. Further analysis shows that the premium concentrates where provider disagreement or nonrecurring items are large and, once browsing arrives, in high-LLM-visibility firms, whose forecasts GenAI retrieves. The evidence indicates that GenAI facilitates selective cross-FDP integration of consensus forecasts.

Keywords: Generative AI, Analyst, Forecast Data Provider, Consensus Forecast, Earnings Surprise, Information Processing, Earnings Response Coefficient.

JEL classification: G11, G14, G41, M41.

*I am deeply grateful to my dissertation committee for their invaluable guidance and support: Musa Subasi (co-chair), Michael Kimbrough (co-chair), Rebecca Hann, Nick Seybert, Jingyi Qian, and Erkut Y. Ozbay. I also thank seminar participants at the University of Maryland, and Matt Ege, Antonis Kartapanis, Kimberlyn Munevar, Brady Twedt, and participants of the workshop at Texas A&M University for helpful comments. I gratefully acknowledge the Texas A&M University AI Dissertation Proposal Award in Accounting. All errors are my own.

†Department of Accounting and Information Assurance, Robert H. Smith School of Business, University of Maryland. Email: yliu337@umd.edu.

1 Introduction

Sell-side analyst forecasts are the market's primary benchmark for evaluating realized earnings (e.g., Bradshaw et al., 2017; Kothari, 2001). The earnings surprise relative to this consensus is a key determinant of announcement-window stock price reactions and post-announcement drift (e.g., Bartov et al., 2002; Bernard and Thomas, 1989; Livnat and Mendenhall, 2006), with measurable consequences for price discovery, trading volume, and firms' own disclosure choices. Yet the consensus is not one number. Multiple *forecast data providers* (FDPs) construct economically distinct consensus numbers for the same firm-quarter through proprietary aggregation rules. Larocque et al. (2026) document these differences across the five leading FDPs: I/B/E/S, FactSet, Bloomberg, S&P Capital IQ, and Zacks.[1] Although these FDPs are widely used and cited for investment decisions (Coleman et al., 2025; Xue et al., 2025), little is known about whether the market relies on FDPs selectively by their accuracy and, more importantly, how it integrates across them.

The question matters because acquiring and integrating data across multiple providers is costly.[2] Since the release of ChatGPT, generative AI (GenAI) has diffused rapidly into investor research workflows, reshaping how financial information is gathered and combined. Industry surveys document widespread adoption (CFA Institute, 2024), and recent research shows that these tools lower the costs of acquiring and integrating financial information (e.g., Bertomeu et al., 2025; Blankespoor et al., 2026; Chang et al., 2026; Cheng et al., 2025; Liu and Subasi, 2026; Xue et al., 2025). Earlier information technologies lowered particular components of these costs. EDGAR reduced the cost of acquiring filings, XBRL the cost of processing structured data, and social-media disclosure the cost of becoming aware of firm news (e.g., Drake et al., 2015; Blankespoor, 2019; Blankespoor et al., 2014). Each of these tools, however, operated within a single source of information. None of them helped an investor collect the competing consensus

[1] In a sample of 94,030 firm-quarters covered by all five FDPs over 2002–2020, Larocque et al. (2026) find that the FDPs report different earnings surprises (beyond rounding) for the same firm-quarter 58 percent of the time, and that at least one FDP signals a miss while another signals a beat in 9 percent of firm-quarters.

[2] Larocque et al. (2026) show that announcement-window returns align more with I/B/E/S and Capital IQ signals than with the other FDPs, consistent with acquisition cost limiting which FDPs the market uses.

forecasts that different providers publish for the same firm-quarter, reconcile the providers' differing methodologies, and combine the forecasts into a single expectation of earnings. This paper asks two questions in sequence. First, does the market rely on each FDP's consensus selectively, according to that FDP's accuracy? Second, because investors do not all rely on the same provider, the expectation that prices reflect is an "aggregation" of consensus forecasts across FDPs; has GenAI shifted that "aggregation" toward an accuracy-weighted multi-FDP synthesis and away from the conventional single-FDP consensus or the equal-weighted average across providers?[3] I find that both have occurred. After ChatGPT, the market's reaction to a provider's surprise depends on that provider's accuracy record, and the "aggregation" reflected in prices has shifted toward the accuracy-weighted synthesis. Before ChatGPT, announcement returns responded no more strongly to the accuracy-weighted surprise than to the I/B/E/S surprise. After ChatGPT, a one-percentage-point standardized surprise moves the announcement return 2.05 percentage points more when it is measured against the accuracy-weighted consensus than when it is measured against I/B/E/S, a post-minus-pre change of 2.67 percentage points. A quintile long–short portfolio formed on the price-scaled difference between the two consensus numbers earns 1.35 percent per announcement over the sample, so the information in that difference is economically material as well.

The potential benefits of cross-FDP synthesis lie in the substantive disagreement among providers, which differ in their screening rules, treatment of unusual items, and analyst-coverage thresholds (Bochkay et al., 2022). Although I/B/E/S is the most widely used provider and ranks highest on average, which FDP yields the most accurate consensus varies across firms and over time (Larocque et al., 2026).[4] This variation implies that synthesizing across providers can improve on any fixed choice of provider. Historically, such synthesis was challenging in practice. An effective

[3]The accuracy-weighted construction is not a claim that the market literally applies this weighting rule. It serves as a tractable proxy for the broader hypothesis that the post-ChatGPT market draws on multiple FDPs and weights them by source accuracy. If so, market reactions should align more closely with this benchmark than with the single-FDP or equal-weighted alternatives, even if the specific weights the market applies differ from mine.

[4]Accuracy and predictive value measure the same underlying quality at two different horizons, namely how well a consensus predicts the firm's street earnings. Accuracy targets the current quarter's actual, and predictive value targets the firm's year-ahead results. Section 4 tests the predictive-value dimension as well, for the I/B/E/S consensus alone, because the estimate requires up to twenty quarters of realized year-ahead outcomes and the five-provider panel is too short to supply them for every FDP.

synthesis requires investors to recognize that the FDPs disagree, to subscribe to each consensus separately, and to reconcile the providers' differing methodologies.

These frictions correspond to the awareness, acquisition, and integration costs in the disclosure-processing-cost framework of Blankespoor et al. (2020). For cross-FDP synthesis, all three costs were high. Prior single-source technologies did not reduce the cost of integrating heterogeneous consensus measures across sources. GenAI lowers all three costs, including this one, which makes accuracy-weighted aggregation across providers feasible.[5] The framework yields two empirical predictions: (i) after ChatGPT, the market should rely on each FDP's consensus more selectively according to its quality and shift its implicit consensus toward an accuracy-weighted synthesis across multiple FDPs, and (ii) this shift should concentrate where GenAI synthesis is most accessible and most valuable.

Neither prediction is obvious *ex ante*. First, the forecast-combination literature gives theoretical grounds for doubt. Combining heterogeneous, partially independent forecasts almost always dominates sub-selecting on accuracy (e.g., Ahlers and Lakonishok, 1983; Bates and Granger, 1969), and estimated combination weights rarely beat equal weights out of sample (Clemen, 1989; Timmermann, 2006). Consistent with this logic, Michaely et al. (2024) find that a consensus built from high-quality analysts within a single provider does not improve investor outcomes, and the frictions they identify apply even more strongly in the cross-FDP setting. Second, lower acquisition costs need not produce more selective integration. A richer information set yields a better consensus only if the synthesis itself is selective, and large language models (LLMs) in particular can rely on memorization instead of reasoning (e.g., Lopez-Lira et al., 2025). Whether GenAI nonetheless shifts the market toward a more selective synthesis is therefore an empirical question.

I test these predictions using a panel of S&P 1500 firm-quarters from 2021 to 2025 with consensus EPS forecasts for the same firm-quarter from all five FDPs.[6] The tests follow the order

[5] A single large language model (LLM) query of the form "What is the consensus EPS forecast for [ticker] next quarter?" typically returns a synthesized response citing at least one, and often several, free third-party distributors (e.g., Yahoo Finance, MarketBeat, TipRanks). Online Appendix B reports two worked examples (IBM and Microsoft) showing the cited distributors, their reported figures, and the disclosed upstream FDP where available.

[6] The sample begins in 2021 because it yields a balanced panel around the November 30, 2022 ChatGPT release and because Bloomberg's consensus history is accessible only for five years. Each FDP's accuracy is computed from

of the two questions. The first set takes the providers one at a time and asks whether the market's reaction to a single FDP's surprise incorporates that FDP's accuracy. After ChatGPT's release, the market reacts more strongly to forecasts from more accurate FDPs and discounts those from less accurate ones. In a within-firm-quarter test that estimates provider-specific earnings response coefficients (ERCs; e.g., Collins and Kothari, 1989; Easton and Zmijewski, 1989) by accuracy rank, the ERC increment per one-rank accuracy gain roughly doubles after ChatGPT, from 0.306 to 0.621 (post-minus-pre differential 0.316, $p < 0.01$). A firm-level design that partitions firms by whether I/B/E/S's accuracy rank rose or fell shows the same selectivity for I/B/E/S itself. Specifically, relative to the firms in which I/B/E/S held its rank, the market response to I/B/E/S-based surprises weakens significantly in the firms where I/B/E/S lost rank and does not change in the firms where it gained. This asymmetry is consistent with a pre-ChatGPT market that already overweighted I/B/E/S.[7] The same selectivity appears on a second dimension of consensus quality. For each firm, I construct a *predictive value* score that measures how closely the firm's I/B/E/S consensus has anticipated its street earnings four quarters ahead, using only comparisons that are public at the announcement date. A firm with high predictive value is one whose consensus has anticipated earnings well beyond the current quarter. The ERC gradient on this measure roughly doubles after ChatGPT.

Analysts' own revisions show the same pattern. Forecast revisions by I/B/E/S analysts provide a second measure of reliance that comes from the forecasting process itself and does not depend on market reactions. In the firms where I/B/E/S lost accuracy rank, sell-side analysts herd toward the prevailing I/B/E/S consensus significantly less after ChatGPT, a 4.8-percentage-point decline against a 32-percent base rate. Where I/B/E/S gained rank, there is no offsetting change, matching the one-sided pattern of the market's reweighting.

Given this evidence of selective use provider by provider, I next ask whether that selectivity results in a wiser "aggregation," that is, whether the market's aggregation rule itself has shifted. In

its own end-of-quarter consensus EPS and reported actual EPS.

[7]The historical prominence of I/B/E/S is documented by Larocque et al. (2026), who find that announcement-window returns align most closely with the I/B/E/S surprise, consistent with investors relying on I/B/E/S regardless of which provider is most accurate for a given firm. If the market already weighted I/B/E/S more heavily than its relative accuracy warranted, a deterioration in that accuracy calls for less weight, whereas an improvement calls for little additional weight.

practice, investors do not all rely on the same provider's consensus, so the expectation that prices reflect is an "aggregation" of consensus forecasts across FDPs, and the announcement return is the reaction to the gap between actual earnings and that "aggregation." If GenAI lowers the cost of acquiring and integrating the providers' numbers, the provider-level selectivity should carry over to the "aggregation," with heavier weights on the more accurate providers. I construct an accuracy-weighted standardized unexpected earnings (SUE) measure in which each FDP's SUE is weighted by its trailing forecast accuracy, and benchmark it against two heuristic alternatives: the single-FDP I/B/E/S consensus and the equal-weighted average across the five FDPs.[8] Before turning to pricing, I verify that the construct carries economically meaningful information. A quintile long–short portfolio on the *predicted surprise*, the price-scaled difference between the accuracy-weighted and each heuristic consensus, earns 0.63–1.35 percent per announcement (all $p < 0.01$).[9] Post-ChatGPT, ERCs on the accuracy-weighted surprise rise sharply while ERCs on the I/B/E/S and equal-weighted surprises remain flat or decline, so the resulting gap, which I term the *accuracy premium*, widens significantly against both heuristic benchmarks. The post-ChatGPT increments in the three-day announcement cumulative abnormal return (CAR) per percentage point of standardized surprise, 2.40 percentage points against the equal-weighted benchmark and 2.67 against I/B/E/S, are each about the size of the entire pre-ChatGPT ERC on the I/B/E/S consensus itself. Scaled by a typical accuracy tilt, the difference between the accuracy-weighted and the benchmark surprise, the effect is more modest. A one-standard-deviation tilt moves the three-day CAR by about 0.4 percentage points, roughly five percent of its standard deviation (Section 5.2).

[8]The composite weights each provider's surprise instead of its consensus level. The providers' consensus numbers are not comparable across providers, because each rests on a provider-specific definition of street earnings. Each provider's surprise nets its consensus against its own street actual, so provider-specific methodology differences cancel within the surprise, and a weighted average of surprises equals the difference between a composite actual and a composite consensus built on the same weights (Equation (2) in Section 3.4). The weight on FDP g is the softmax of its z-scored trailing accuracy. Figure OA1 confirms that the accuracy-weighted SUE has the smallest mean absolute forecast error of the three consensus measures, with the difference statistically significant.

[9]Following the earnings-surprise trading-strategy literature (e.g., Battalio and Mendenhall, 2005; Michaely et al., 2024), the portfolio goes long the top quintile and short the bottom quintile of the standardized predicted surprise, holding each position over the four-day $[-2,+1]$ announcement window. The per-announcement return is estimated by panel OLS with firm-clustered standard errors. The strategy earns 0.63–0.73 percent against the equal-weighted benchmark and 1.35 percent against I/B/E/S, and is robust to alternative event windows, including $[-2,+2]$ and $[-1,+1]$ (untabulated). Tradability is not a foregone conclusion, given the within-provider null of Michaely et al. (2024) noted above.

The timing and the mechanism of this shift both point to GenAI. Re-estimated within four GPT-capability regimes, the premium against I/B/E/S rises from negative in the pre-ChatGPT and text-only regimes to 1.52 percentage points under browsing and a significant 4.91 under native search.[10] The point estimates thus rise with the tools' retrieval capabilities. Integration operates only on the forecasts a model can retrieve, so acquisition is a precondition for the premium. To probe GenAI's role further, I construct a firm-level measure of *LLM visibility* from point-in-time queries to browsing-enabled models (ChatGPT with GPT-4o and Perplexity), counting how many forecast sources the model retrieves for each firm. The measure is empirically distinct from classical prominence proxies,[11] and the premium concentrates in high-visibility firms only once the model can browse. The regime pattern and the concentration of the premium in high-visibility firms together indicate that the consensus reflected in market prices has moved from heuristic alternatives toward an accuracy-weighted multi-FDP synthesis.

Two cross-sectional tests show that the accuracy premium concentrates in settings where GenAI-enabled synthesis is most valuable. Because the regime evidence localizes the premium to the browsing and native-search regimes, these tests date the shift from the onset of GPT browsing (September 27, 2023), when web retrieval first allowed a single query to collect the providers' current numbers, instead of from ChatGPT's release. First, the premium concentrates in firm-quarters with high cross-FDP disagreement, where accuracy weighting offers the largest improvement over heuristic benchmarks. Second, it concentrates in firm-quarters with large special items. Because FDPs differ most in their treatment of such nonrecurring items (Bochkay et al., 2022), this is where synthesizing across providers most effectively offsets any single provider's methodology choices.

This paper makes three contributions. First, I show that the market's reliance on FDPs

[10]The regimes are defined by the retrieval capability that ChatGPT offers the public on the announcement date, following OpenAI's release dates: pre-ChatGPT (before November 30, 2022), text-only (from ChatGPT's launch through September 26, 2023), browsing (from the permanent return of web browsing on September 27, 2023), and native search (from the launch of ChatGPT search on October 31, 2024 through the end of the sample). In the text-only regime the model can describe the providers and their street-earnings methodologies from its training data but cannot retrieve a current forecast, so the acquisition step of cross-provider synthesis remains manual. Browsing is the first regime in which the five providers' current numbers can be collected within a single query.

[11]Operationally, LLM visibility is an EPS-extraction indicator times $\log(1 + \#\text{ unique cited domains})$, averaged across ChatGPT (GPT-4o) and Perplexity responses to firm-specific EPS queries (from the DataForSEO LLM Responses API).

depends on information technology, extending recent evidence on cross-FDP differences in analyst consensus information (Larocque et al., 2026). When information-processing costs fall, the market reweights its implicit consensus toward the more accurate providers, with measurable effects on earnings-announcement returns. For researchers, this means that the FDP whose consensus best reflects market expectations depends on the information technology in use.

Second, I contribute a new mechanism to the literature on the capital-market consequences of GenAI (e.g., Bertomeu et al., 2025; Blankespoor et al., 2026; Chang et al., 2026; Cheng et al., 2025; Liu and Subasi, 2026; Lopez-Lira and Tang, 2026; Xue et al., 2025). Prior work documents that GenAI lowers investor processing costs and shifts prices around specific events such as service outages or regulatory bans. The mechanism in this paper is synthesis: GenAI reduces the market's cost of combining forecasts across multiple FDPs, and the combined consensus is now priced into earnings-announcement reactions. Establishing this channel also yields a new measurement approach. The LLM visibility measure, built from point-in-time queries to browsing-enabled models, records which firms' forecasts GenAI actually retrieves. Because the models and their retrieval behavior continue to change, the measure will need to be updated as they do, but the approach of measuring what the models actually retrieve can be used in future work on how GenAI reaches capital markets.

Third, I contribute to the disclosure-processing-cost framework of Blankespoor et al. (2020) by identifying the cost component that prior single-source technologies did not reduce and that had to fall before the market could weight consensus information by accuracy. A long line of work shows how prior single-source technologies reduce specific cost components: EDGAR for filing acquisition (Drake et al., 2015; Loughran and McDonald, 2017), XBRL for structured-data processing (Blankespoor, 2019), social media for awareness (Blankespoor et al., 2014), and robo-journalism for retail investors' processing costs (Blankespoor et al., 2018). The evidence in this paper indicates that the cost of integrating heterogeneous consensus measures across sources is that component. The earlier technologies did not reduce this cost; GenAI does, and the reduction has measurable consequences for earnings-announcement pricing.

The remainder of the paper is organized as follows. Section 2 sets out the related literature and the framework. Section 3 describes the data and constructs the accuracy-weighted surprise. Section 4 tests selective reliance on the consensus forecasts of the five FDPs one provider at a time, in market prices and in analysts' own revisions. Section 5 asks whether that provider-level selectivity results in a wiser "aggregation," measured by the accuracy premium of the accuracy-weighted synthesis over the two heuristic benchmarks, and examines its timing across GPT-capability regimes. Section 6 presents the cross-sectional tests and shows how the premium depends on LLM visibility. Section 7 concludes.

2 Related Literature and Theoretical Framework

2.1 Forecast Data Providers and the Construction of Analyst Consensus

Prior research documents that the market discriminates on forecast quality across individual analysts. Whether the market discriminates in the same way across consensus measures is not known. Sell-side analyst earnings forecasts have long served as the benchmark proxy for market expectations in research and practice (Bartov et al., 2002; Beyer et al., 2010; Bradshaw et al., 2017; Kothari, 2001), in part because analysts integrate firm-specific guidance, industry knowledge, and macroeconomic conditions in ways that simple time-series extrapolation does not, at least at short horizons (Bradshaw et al., 2012; Schipper, 1991). Individual forecast accuracy varies systematically with an analyst's experience, employer resources, and portfolio complexity (Clement, 1999; Mikhail et al., 1997), and prices react more strongly to revisions from more accurate or higher-status analysts (Clement and Tse, 2003; Park and Stice, 2000; Stickel, 1992). At the consensus level, however, a consensus built from the more accurate analysts has not been shown to dominate the pooled benchmark, even within a single provider (Michaely et al., 2024). Understanding why discrimination has not extended to the consensus level requires examining how the consensus is constructed.

The consensus number is not simply an average of the analyst forecasts it summarizes. Third-party forecast data providers (FDPs) construct it by passing each underlying analyst forecast through proprietary screening and adjustment rules.[12] Bochkay et al. (2022) document the first-order

[12]FDPs differ, among other dimensions, in (i) which broker estimates they include (in particular, when a forecast

capital-market consequences of these construction choices using the methodology change that Thomson Reuters made to I/B/E/S in 2009.[13] An FDP's treatment of unexpected charges and gains, they show, feeds directly into the time-series predictive properties of street earnings and into price discovery around announcements.[14] This evidence indicates that FDPs' construction choices themselves affect the information content of the consensus.

The same construction discretion generates pervasive disagreement across providers. Comparing the five FDPs in a large sample, Larocque et al. (2026) find that providers report different earnings surprises (beyond rounding) for the same firm-quarter 58 percent of the time, and that in 9 percent of firm-quarters at least one FDP signals a beat while at least one other signals a miss. These disagreements are associated with differences in price responsiveness, abnormal volatility, and liquidity around earnings announcements. The accuracy ranking across FDPs is itself unstable across firms and over time, so no fixed choice of provider is reliably the most accurate.

Disagreement of this magnitude implies that investors could, in principle, improve on any single provider by combining several.[15] If each FDP's accuracy could be anticipated *ex ante*, accuracy-weighted aggregation could outperform any single-provider benchmark. The historical evidence suggests that the market has not combined providers in this way. Although the most accurate provider varies by firm and period, Larocque et al. (2026) document that market reactions align most closely with the I/B/E/S surprise, consistent with investors relying on I/B/E/S regardless of which provider is most accurate for a given firm.

The closest evidence on why this opportunity went unexploited comes from Michaely et al.

---

is deemed stale or noncomparable and dropped), (ii) how they treat nonrecurring items (some report GAAP-based earnings, others a "street" figure adjusted for one-time charges), and (iii) rounding conventions.

[13] Beginning in September 2009, Thomson Reuters stopped determining the I/B/E/S street-earnings actual from analysts' *ex post* treatment of unexpected charges or gains reported in an earnings press release, adopting instead a uniform rule that immediately excludes such items from GAAP earnings. Because the change rolled out gradually, Bochkay et al. (2022) exploit the staggered variation in a difference-in-differences design.

[14] "Street" earnings are the actual EPS figures an FDP reports on the same basis as the forecasts it collects: GAAP earnings adjusted, under the provider's own rules, for items analysts exclude, such as restructuring charges, impairments, and gains on asset sales. Because both the consensus and the actual are provider-specific, each FDP's earnings surprise is internally consistent but not directly comparable to another provider's.

[15] Classical forecast-combination theory (Bates and Granger, 1969; Clemen, 1989; Timmermann, 2006) establishes that the accuracy gain from combining forecasts rises as the forecasts' errors become less correlated. Whether the gain is realizable in practice depends on the cost of locating and combining the underlying sources, a point developed below.

(2024), who examine the within-provider analog. A consensus built from high-quality analysts within I/B/E/S does not meaningfully improve investor outcomes relative to the full-analyst consensus. They attribute this result to error cancellation in the broader pool, imperfect persistence in analyst quality, and the cognitive cost of identifying high-quality forecasters.[16] The second and third of these frictions are larger across providers than within one, because accuracy rankings across providers are unstable and the providers' consensus numbers are constructed under different methodologies that must be reconciled. Whether a large enough reduction in the cost of identifying, acquiring, and integrating accurate sources leads investors to rely selectively on consensus measures is an open question.

2.2 Technological Advancements and Information Processing Costs in Capital Markets

Whether investors use available information depends on what it costs them to do so. Blankespoor et al. (2020) provide the standard economic framework, decomposing the cost of using any piece of information into three components: *awareness* of its existence, *acquisition* of the information itself, and *integration* of that information into the investor's decision. Technology affects each component, and an exogenous reduction in any one of them, holding the underlying information fixed, can change investor behavior and, in turn, asset prices and firms' reporting choices.

An extensive empirical literature documents these effects component by component, with much of the identifying variation coming from single-source technological shocks introduced by regulators or by changes in market infrastructure. For awareness, firms' dissemination of news through Twitter narrows spreads and deepens markets, especially for less-visible firms (Blankespoor et al., 2014). For acquisition, EDGAR search activity tracks firm-level information demand, and surges in retrieval precede price adjustments (Drake et al., 2015); the few filings investors actually open are associated with substantive price reactions (Loughran and McDonald, 2017). For integration, the SEC's XBRL mandate, which tagged financial-statement items in machine-readable form, lowered investors'

[16] This pattern echoes a broader theme in the analyst and disclosure literature, that how readily forecast information can be interpreted shapes how investors use it: forecast dispersion is itself a priced characteristic (Diether et al., 2002), and disclosure clarity is linked to investor behavior at the individual-disclosure level (Lawrence, 2013).

processing costs enough to shift firms' own disclosure choices (Blankespoor, 2019). On the textual side, dictionary-based sentiment measures extract return-relevant information from news columns (Tetlock, 2007) and 10-K filings (Loughran and McDonald, 2011), and robo-journalism lowers the processing cost retail investors face (Blankespoor et al., 2018).

A parallel literature approaches these costs from the demand side. Google search measures retail attention and predicts short-horizon returns (Da et al., 2011), and it rises ahead of earnings announcements and preempts part of the announcement-window price response (Drake et al., 2012). Attention is limited. Prior work documents inattention to complicated firms (Cohen and Lou, 2012) and to competing announcements (Driskill et al., 2020; Hirshleifer et al., 2009), muted responses to Friday announcements (DellaVigna and Pollet, 2009), and firms' strategic scheduling of releases (deHaan et al., 2015). This limit on attention is why the recurring cost of cross-FDP synthesis constrains investors. The awareness, acquisition, and integration steps must be repeated for every firm-quarter an investor follows (Section 2.4).

Closest to the FDP-aggregation question, broader access to firm disclosures changes price formation. Open conference calls coincide with more small-investor trading and higher return volatility during the call (Bushee et al., 2003) and with reduced post-announcement underreaction (Kimbrough, 2005), and business-press coverage is associated with narrower spreads and deeper markets around earnings releases (Bushee et al., 2010). Each of these advances lowered the cost of one or two components for a single information source. None lowered the cost of combining numbers from several sources. A technology that supports synthesis across providers must therefore lower all three cost components, and must do so for several sources at the same time. An investor who does not know that the providers disagree, cannot obtain each provider's numbers, or cannot reconcile their methodologies will not form an accuracy-weighted consensus, even if technology has lowered one of the three costs.

2.3 Generative AI in Capital Markets

GenAI changes what it costs investors to combine information across sources. Capital-markets research has already established what machine methods add to the processing of financial

information. Nonlinear, high-dimensional methods such as neural networks and tree-based models extract predictive signal that linear asset-pricing regressions miss (Goldstein et al., 2021; Gu et al., 2020). Those methods, however, operate on structured numerical data and require programming expertise to deploy, so they did not lower the cost of using information for investors at large. What distinguishes the current generation of LLMs is the conjunction of three features prior technologies lacked: they process unstructured information from heterogeneous sources, they synthesize it through natural-language queries without specialized retrieval infrastructure or programming expertise, and they operate at low marginal cost per query (e.g., Kim et al., 2023, 2024). Together, these features make cross-source synthesis, which previously required institutional teams and dedicated infrastructure, available to retail investors.

Early evidence indicates that investors' use of these tools is already price-relevant. Lopez-Lira and Tang (2026) show that ChatGPT-derived sentiment signals predict cross-sectional returns; Cheng et al. (2025) use ChatGPT service outages as exogenous shocks and find that investor responses to information degrade when the tool is unavailable; Bertomeu et al. (2025) exploit Italy's 2023 ChatGPT ban and find that bid-ask spreads widened for Italian firms during the ban, most for firms with thin institutional ownership and analyst coverage; and Chang et al. (2026) show that retail traders' alignment with AI-derived earnings-call sentiment rose sharply after ChatGPT's release, narrowing the information-processing gap between less-sophisticated retail traders and more-sophisticated short sellers.

Closer to the analyst forecasts studied here, a separate strand of work examines the production side, AI's effects on analyst research itself. Kimbrough et al. (2026) find that analysts at brokerages with greater AI integration produce more accurate earnings forecasts and that their revisions are more informative to capital markets. Bradshaw et al. (2026) document that AI-generated articles rose to 13.5 percent of Seeking Alpha's output after ChatGPT's release, and that AI-using authors cover more firms but produce less informative content per article. On the crowdsourced side, Liu and Subasi (2026) find that GenAI narrows the accuracy gap between nonprofessional and professional forecasters. These papers indicate that AI affects the production of analyst research. This paper

examines how GenAI affects the use of that research, specifically how investors aggregate the different consensus numbers that competing FDPs construct.

2.4 Theoretical Framework and Predictions

Two opposing forces shape the framework. The first comes from the forecast-combination literature. Combining heterogeneous, partially independent forecasts almost always dominates relying on any single source (Bates and Granger, 1969; Clemen, 1989; Granger and Ramanathan, 1984; Stock and Watson, 2004; Timmermann, 2006).[17] This is the wisdom-of-crowds principle popularized by Surowiecki (2004), under which averaging diversifies away idiosyncratic errors, so sub-selecting on individual accuracy can lose more from forgone diversification than it gains in precision. This force rationalizes the historical equilibrium in which the market anchors on a single pooled benchmark (typically the I/B/E/S consensus), itself an average over many analysts, and treats cross-FDP differences as noise. It is also consistent with the within-source evidence of Ahlers and Lakonishok (1983) and Michaely et al. (2024), in which sub-selecting forecasters on measured quality does not yield consistent gains. By itself, this force implies that synthesizing across providers need not yield any gain over the pooled single-provider benchmark, so the market's reliance on that benchmark should persist.

The opposing force is heterogeneity in what the providers produce: the documented cross-sectional and time-series variation in FDP accuracy (Larocque et al., 2026), together with the methodology-driven information content that Bochkay et al. (2022) identify. When providers disagree substantially, when part of that disagreement reflects differences in accuracy instead of noise, and when the more accurate provider can be identified from its observable characteristics, a sufficiently low-cost synthesis across providers can dominate the pooled benchmark. The model-based earnings-forecast literature provides a related example. Cross-sectional models

[17] The classical result also says when accuracy weighting should beat equal weighting. For two unbiased forecasts with error variances σ_1^2 and σ_2^2 and error correlation ρ, the variance-minimizing weight on the first is $(\sigma_2^2 - \rho\sigma_1\sigma_2)/(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)$ (Bates and Granger, 1969). Equal weights are thus optimal only when the two forecasts are equally accurate, and the gain from tilting toward the more accurate one rises with the accuracy gap. That gain is realized only if the gap can be estimated precisely enough to outweigh the estimation error in the weights, the "forecast combination puzzle" that makes equal weights so hard to beat in practice (Stock and Watson, 2004; Timmermann, 2006).

produce earnings forecasts with less bias and broader coverage than the analyst consensus, though not greater accuracy (Hou et al., 2012). An alternative construction of expected earnings can thus improve on the consensus along some dimensions without dominating it on all. Which force prevails depends on the cost of identifying, acquiring, and integrating accurate sources. When this cost is high, investors do not synthesize across providers regardless of the statistical gain, and the market relies on a single pooled FDP consensus. When this cost is low, accuracy-weighted synthesis becomes feasible, and the ordering reverses if the accuracy-relevant heterogeneity across providers outweighs the diversification forgone by tilting away from the pool.

Historically, all three components of the Blankespoor et al. (2020) decomposition were costly for cross-FDP synthesis, and integration was the component that no prior access technology reduced. Even when consensus data are freely redistributed through retail-facing platforms, an investor seeking the cross-FDP picture must know which platform carries which FDP's consensus (awareness), visit each individually to retrieve the number (acquisition), and reconcile numbers that differ by methodology and are presented inconsistently across platforms (integration). These steps recur for every firm-quarter an investor follows, and synthesis at scale further demands programming and data-engineering expertise. The recurring time cost and the expertise requirement limited cross-FDP synthesis to specialists, and the market relied on a single source.

GenAI lowers both barriers. A single LLM query identifies the outlets that carry a consensus for the firm, retrieves their numbers, and integrates them into one response that reports the implied consensus and the cited distributors, all without specialized infrastructure or programming skill. The effective cost of cross-FDP synthesis falls to that of issuing one query. The model does not access the FDPs' subscription products directly, but it reaches their consensus numbers through the free, retail-facing distributors that license or aggregate them.[18] Figure 1 summarizes the framework, which yields two empirical predictions:

[18]The distributors the models cite include Yahoo Finance, which redistributes the S&P Capital IQ consensus, Zacks's own free site, and independent aggregators such as MarketBeat, TipRanks, Barchart, and ChartMill. Online Appendix B reports two worked queries (IBM and Microsoft) with the actual LLM responses, each cited source, and, where disclosed, its upstream FDP.

Prediction 1. After ChatGPT's release, the consensus reflected in earnings-announcement prices shifts away from heuristic aggregation (single-FDP anchoring or equal-weighting across FDPs) toward a more sophisticated synthesis that draws on multiple FDPs and places greater weight on consensus signals of higher quality, whether quality is measured as recent forecast accuracy or as predictive value for the firm's future performance.

Prediction 2. The post-ChatGPT shift is stronger where GenAI-mediated synthesis is more accessible (LLMs can retrieve the firm's analyst forecasts) and more valuable (cross-FDP disagreement and its methodology-driven component are large).

The framework holds each provider's construction rules and the resulting street earnings fixed; only investors' awareness, acquisition, and integration costs change. This distinguishes the framework from a supply-side explanation such as the 2009 I/B/E/S methodology change studied by Bochkay et al. (2022). If the post-ChatGPT patterns instead reflected a change in how street earnings are constructed, then the predictive quality of street earnings would itself move at the ChatGPT boundary, not only the market's response to it.

3 Institutional Background, Data, Sample, and Key Measures

3.1 Institutional Background

The five providers studied here entered the consensus-forecast business from different directions, and they serve overlapping but distinct sets of clients. I/B/E/S, FactSet, Bloomberg, S&P Capital IQ, and Zacks construct and sell analyst-consensus EPS information to sell-side brokerages, buy-side investors, financial media outlets, and academic researchers. I/B/E/S is credited with creating the consensus-earnings-estimate industry in the early 1970s, and Zacks claims to have first distributed quarterly consensus forecasts in the early 1980s. Bloomberg, Capital IQ, and FactSet entered the segment from adjacent product lines (terminal data, investment-banking software, and printed "fact sets" delivered to Wall Street clients, respectively).[19]

Two sources of variation drive cross-FDP differences in consensus numbers. First, the set of contributing analysts differs. Each FDP includes a different subset of contributing sell-side analysts, applies its own recency cutoffs (e.g., Bloomberg drops estimates more than 30 days old), and runs

[19]Each FDP today operates within a broader financial-data platform. See Larocque et al. (2026) for a detailed institutional review.

proprietary statistical checks to filter outliers. Second, the methodology for constructing "street" earnings varies. Each FDP decides for itself how to treat unexpected charges, gains, and other adjustments to reported earnings (Bochkay et al., 2022). Zacks, for example, counts stock-based compensation as a recurring expense in street earnings, whereas I/B/E/S has historically excluded it. Such divergence is common. In the 2002–2020 sample of Larocque et al. (2026), 46.7 percent of firm-quarters contain a large unexpected item that FDPs treat inconsistently.

Online Appendix Table OA1 measures the methodology differences directly. For each pair of FDPs, it reports the share of firm-quarters in which the two providers publish *different* street actuals for the same reported quarter. Because both providers observe the same GAAP filing, any difference reflects only their construction rules. All five providers report the same actual in only 6.3 percent of firm-quarters, and the largest cross-provider gap reaches at least five cents per share in 77 percent (untabulated). When the actuals differ, the surprises differ even if the forecasts are identical. Provider choice therefore affects the surprise independently of any difference in analyst composition.

No single FDP, however, strictly dominates the others at the firm-quarter level. I/B/E/S ranks highest on average across composite measures of FDP quality (Larocque et al., 2026) and has long served as the de facto anchor for research and trading alike. Yet it is the most accurate provider in only 16.9 percent of firm-quarters in my S&P 1500 sample, below the 20 percent that a uniform draw across five providers would imply, and the least accurate in 12.1 percent (untabulated). Because the most accurate provider varies across firms and over time, cross-FDP synthesis is potentially valuable.

3.2 Data and Sample

Exploiting this firm-quarter variation requires parallel consensus histories from all five FDPs. I therefore collect quarterly consensus EPS forecasts and reported actuals from each provider separately. I/B/E/S and Zacks are available through WRDS (the I/B/E/S Summary History file and the Zacks Estimates database, respectively). The remaining three FDPs (FactSet, Bloomberg, and S&P Capital IQ) are accessible only through proprietary terminals or institutional portals, from which I download quarterly consensus histories firm by firm.[20]

[20] Larocque et al. (2026) follow a similar collection procedure for their 2002–2020 study, downloading or hand-collecting data from each FDP's proprietary system. Bloomberg consensus data, in particular, must be retrieved

The initial sample comprises all S&P 1500 firm-quarters with an earnings announcement over 2021–2025, 29,395 firm-quarters across 1,502 firms.[21] Announcement dates are the I/B/E/S announcement dates, on which every event window is centered.[22]

I then require complete coverage across all five FDPs, which retains the 25,027 firm-quarters that carry a consensus from every provider. This screen is applied to the announcement quarter's own consensus values. The accuracy weights of Section 3.4 additionally draw on each provider's forecast errors over the four quarters preceding the announcement, so the consensus and actual histories underlying the weights reach back before the 2021 start of the announcement window, into 2020, for the earliest sample announcements. The resulting accuracy-weighted composite is non-missing for every firm-quarter in the final sample, because each provider's rolling mean is taken over whichever of the four preceding quarters that provider reports. Dropping firm-quarters with missing market-outcome variables or controls (1,179 observations), missing constructed return or predicted-surprise variables (10), and singleton firms (5) brings the final sample to 23,833 firm-quarters across 1,334 firms. Table 1 details the full selection procedure and Table 2 reports descriptive statistics for the main variables.[23]

Firm controls and market-outcome variables are drawn from WRDS-hosted databases: accounting fundamentals from Compustat, returns and trading data from CRSP, institutional ownership from the Thomson Reuters 13F Holdings database, and analyst-coverage characteristics from I/B/E/S. The control vector comprises log market equity, book-to-market, leverage, the prior twelve-month return, log price, return volatility, analyst dispersion, announcement-day busyness, the absolute changes in income before extraordinary items and in shares outstanding, the magnitude of special items, institutional ownership, the logged number of analysts, and indicators for high stock-based compensation, unexpected items, stock splits, and fiscal fourth quarters.

ticker by ticker through Bloomberg Query Language (BQL) functions, and data access limits its consensus history to five years, which sets the sample start.

[21] Three considerations motivate the S&P 1500 universe. Provider coverage declines sharply outside the large- and mid-cap indices, the research design requires every firm-quarter to carry a consensus from all five providers, and the S&P 1500 is the universe for which ticker-by-ticker Bloomberg retrieval is feasible.

[22] The November 30, 2022 ChatGPT release falls within this window, leaving roughly two years of pre-ChatGPT and three years of post-ChatGPT announcements, a relatively balanced panel of firm-quarters.

[23] All continuous variables in the analyses that follow are winsorized at the 1st and 99th percentiles.

The definitions of all variables are provided in Appendix A.

3.3 Dependent Variables

The market's reliance on any given consensus is not directly observable, so I infer it from the price response at the earnings announcement. The surprise implied by a consensus on which the market relies more heavily should move prices more. The main outcome variables are short-window abnormal returns: a three-day cumulative abnormal return $CAR_{[-1,+1]}$ for the immediate-reaction tests and a four-day buy-and-hold abnormal return $BHAR_{[-2,+1]}$ for the trading-strategy test.[24] For each window I use two risk adjustments: a Fama-French-Carhart four-factor (FFC) abnormal return and a market-adjusted return.[25]

3.4 Constructing the Accuracy-Weighted SUE

The paper's central construct is an accuracy-weighted aggregation of the five FDP consensus signals. Let $g \in \{\text{I/B/E/S, FactSet, Bloomberg, Capital IQ, Zacks}\}$ index providers and let $SUE^g_{i,q} = (actual^g_{i,q} - consensus^g_{i,q})/p_{i,q}$ denote provider g's standardized unexpected earnings for firm i in quarter q: its own street actual less its own consensus, scaled by $p_{i,q}$, the closing share price on the last trading day before the announcement. The accuracy-weighted composite is

$$SUE^{Acc}_{i,q} = \sum_g w^g_{i,q} SUE^g_{i,q}, \tag{1}$$

with weights $w^g_{i,q}$ derived from each provider's recent forecast accuracy. The equal-weighted benchmark sets $w^g_{i,q} = 1/5$, yielding $SUE^{EW5}_{i,q}$. The single-FDP I/B/E/S benchmark, $SUE^{IBES}_{i,q}$, requires no aggregation.

Why the composite weights surprises, not consensus levels. The composite weights each provider's surprise, not its consensus level, because the five consensus numbers are not comparable across providers. Each provider builds its consensus on its own definition of street earnings, so the same firm-quarter carries five consensus levels on five bases. A weighted average of those levels

[24] Relative to the three-day CAR window, the four-day BHAR window opens one day earlier, at day -2, so that a position taken ahead of the announcement captures any pre-announcement leakage. Both windows close at day +1, which captures the next session's reaction to announcements made after the market close.

[25] FFC factor loadings are estimated on the $[-260, -11]$ pre-announcement window. The market-adjusted return subtracts the contemporaneous CRSP value-weighted index return from each daily return. All returns are expressed in percent. Inferences are unchanged under either adjustment, so tables reporting a single risk adjustment use the market-adjusted return for brevity (the FFC counterparts are available on request).

has no single actual against which it can be compared, because the providers' actuals themselves differ (Section 3.1 and Online Appendix Table OA1). Weighting surprises mitigates this concern substantially, for a simple arithmetic reason. Because $SUE^{g}_{i,q}$ nets provider g's consensus against provider g's own actual, any component of the provider's methodology that enters both its consensus and its actual, such as the treatment of a nonrecurring charge, cancels within the surprise. Each $SUE^{g}_{i,q}$ is therefore measured in the same units, percent of price, on a matched basis. The convex combination then equals

$$SUE^{Acc}_{i,q} = \frac{\sum_g w^{g}_{i,q} actual^{g}_{i,q} - \sum_g w^{g}_{i,q} consensus^{g}_{i,q}}{p_{i,q}}, \tag{2}$$

the surprise of a composite consensus relative to a composite actual built with the same weights, so consensus and actual remain on matched bases at the composite level as well. The residual concern is limited to methodology differences that enter a provider's forecasts and its actual asymmetrically, which is precisely the noise that the accuracy weights penalize. One implication for interpretation follows. Because each provider's surprise pairs its consensus with its own actual, the accuracy tilt, the difference between the accuracy-weighted surprise and a benchmark surprise, combines cross-provider differences in the consensus with cross-provider differences in the street actual. The horse race of Section 5.2 does not separate the two. A widening premium therefore shows that the market relies more on the accurate providers' matched consensus-and-actual signals, and the special-items result of Section 6.1 is consistent with reliance on both components. The predicted-surprise test of Section 5.1 isolates the consensus component and shows that it carries value on its own. The predicted-surprise trading test of Section 5.1, which compares consensus levels directly, does not enjoy this cancellation and is therefore the more conservative of the two tests.

Weight construction. The weights map each provider's recent, firm-specific track record into its share of the composite. I first compute the price-scaled absolute forecast error, $err^{g}_{i,q} = |actual^{g}_{i,q} - consensus^{g}_{i,q}|/p_{i,q}$, expressed in percent, where $consensus^{g}_{i,q}$ is provider g's end-of-quarter consensus for quarter q, observed before the announcement, and $p_{i,q}$ is the closing share price on the last trading day before the announcement. The error and the surprise $SUE^{g}_{i,q}$ are built from the same

consensus and price.[26] A backward-looking rolling mean over the four preceding quarters (or over as many of them as the provider reports) then yields a recent mean absolute forecast error, $MAFE^{g}_{i,q} = \frac{1}{4}\sum_{k=1}^{4} err^{g}_{i,q-k}$, whose sign I reverse so that larger values denote greater accuracy: $acc^{g}_{i,q} = -MAFE^{g}_{i,q}$. Because the rolling window excludes quarter q itself, the weights use only information available before the quarter-q announcement and cannot induce look-ahead bias. I standardize across the five providers within firm-quarter, $acc\_z^{g}_{i,q} = (acc^{g}_{i,q} - \overline{acc}_{i,q})/\mathrm{sd}(acc_{i,q})$, and convert the z-scores into convex weights via the softmax (multinomial-logit) transform,

$$w^{g}_{i,q} = \frac{\exp(acc\_z^{g}_{i,q})}{\sum_{h}\exp(acc\_z^{h}_{i,q})}. \tag{3}$$

The resulting weights have four properties. They are strictly positive, so no provider is dropped and the composite retains the error diversification of the full set of five. They sum to one within each firm-quarter, so the composite is a convex combination of the five providers' surprises and is measured in the same units as each of them. They are monotone in trailing accuracy, so a provider with a lower $MAFE^{g}_{i,q}$ receives a higher weight. And they depend only on the relative pattern of accuracy within the firm-quarter, not on its level, because the within-firm-quarter standardization removes any shift or rescaling of the five providers' errors that is common to all of them. When the five accuracies coincide, the within-row standard deviation is zero and the weights are set to the uniform 0.2 vector by convention. Online Appendix C works through this construction for a single hypothetical firm-quarter.

Three choices shape the weights. Standardizing within firm-quarter instead of across the pooled sample is necessary because earnings noise differs across firms and periods. What matters economically is relative accuracy for a given firm in a given quarter. The softmax transform, akin to the exponential weighting schemes of the forecast-combination literature (e.g., Bates and Granger, 1969; Timmermann, 2006), lies between equal weighting, which ignores accuracy, and winner-take-all selection, which would amplify the noise in a four-quarter MAFE estimate.[27] The

[26] Because the reported actual EPS is itself FDP-specific (Section 3.1), pairing each FDP's consensus with its own actual ensures that the forecast error reflects *ex ante* predictive accuracy instead of methodology mismatch, as in Larocque et al. (2026).

[27] Formally, the softmax is the maximum-entropy weight vector among all convex weight vectors with a given weighted-average standardized accuracy, the Gibbs distribution over providers. The weights do not, however, distinguish

four-quarter window balances recency against noise. Shorter windows are more sensitive to transient accuracy shocks, and longer windows respond more slowly to changes in provider quality.

Empirical weight distribution. The weights depart from equal weighting but remain far from winner-take-all selection. The average within-firm-quarter minimum weight is 0.028 and the average maximum is 0.414 (untabulated), against a uniform benchmark of 0.200. The typical firm-quarter thus reduces the weight on its least accurate provider to roughly 3 percent of the composite and concentrates roughly 41 percent on its most accurate. SUE^{Acc} and SUE^{EW5} are highly correlated ($\rho \approx 0.82$, untabulated) but not perfectly collinear, leaving distinct variation for the pricing tests to exploit.

4 GenAI and Selective Reliance on Consensus Forecasts Across the Five FDPs

This section asks whether the availability and capacity of GenAI are associated with selective reliance on the five FDPs' consensus forecasts, provider by provider. Under Prediction 1, once GenAI makes each provider's track record easy to look up, the market should weight a provider's surprise by how accurate that provider has recently been for the firm, and this dependence should strengthen after ChatGPT. I test it in two ways: by asking whether the ERC rises with a provider's accuracy rank within the same announcement, and by asking whether the market's reliance on one provider, I/B/E/S, moves with the change in that provider's own rank across firms. Analysts' forecast revisions provide an independent measure of reliance, so I also ask whether analysts herd toward the I/B/E/S consensus less in the firms where the market discounts it. The section closes by repeating the market test on a second measure of consensus quality, predictive value.

4.1 Selective Reliance in Announcement Returns

The first test asks whether the ERC rises with provider accuracy rank. Ranking the five FDPs within each firm-quarter by *ex ante* trailing MAFE (rank 1 = least, 5 = most accurate) and stacking the five provider observations for each firm-quarter, I estimate

$$CAR_{i,q} \;=\; \alpha + \beta\, SUE^{g}_{i,q} + \delta\, Rank^{g}_{i,q} + \gamma\left(SUE^{g}_{i,q} \times Rank^{g}_{i,q}\right) + \varepsilon_{i,q,g}, \tag{4}$$

firm-quarters in which the providers' accuracies differ trivially from those in which they differ materially. The cross-sectional tests of Section 6.1 address where the differences are large.

where γ is the ERC increment per one-rank accuracy gain.[28] Because rank is assigned within firm-quarter, identification comes from the same announcement return's differential sensitivity to the five providers' surprises. A firm- or time-level shock, including methodology drift common to all providers, shifts all five observations alike and cannot by itself generate a rank gradient. The design cannot show, however, whether the market's reliance on a given provider rises and falls with that provider's record.

The second test addresses this question directly. It fixes the provider and exploits cross-firm variation in how its accuracy rank changed around ChatGPT. I classify firms by the change in I/B/E/S's mean rank between the pre- and post-ChatGPT windows; $IBES\,Up_i$ and $IBES\,Down_i$ mark improvements and deteriorations of ≥ 2 positions. I then estimate the triple-interaction specification

$$\begin{aligned} CAR_{i,q} = \alpha + \beta_1 SUE^{IBES}_{i,q} + \beta_2 SUE^{IBES}_{i,q} \times Post_q \times IBES\,Up_i \\ + \beta_3 SUE^{IBES}_{i,q} \times Post_q \times IBES\,Down_i + \text{controls} + \text{FE} + \varepsilon_{i,q}, \end{aligned} \tag{5}$$

with the lower-order terms included but not displayed.[29] The triple interactions are the coefficients of interest. Selective reliance predicts $\beta_3 < 0$. If the pre-ChatGPT market already overweighted its anchor, a gain in I/B/E/S's rank adds little, so $\beta_2 \approx 0$.

Post-ChatGPT, the price reaction to a given surprise depends increasingly on which provider produced it. Table 3 reports both tests. In Panel A, the per-rank ERC increment γ roughly doubles, from 0.306 before to 0.621 after (both $p < 0.01$). The post-minus-pre differential is 0.316 ($p < 0.01$). The provider rows show the breadth of the pattern, with positive and significant post-ChatGPT per-rank increments for every provider except Zacks.[30] Panel B shows that, relative to the firms in which I/B/E/S held its rank, the response to the I/B/E/S surprise weakens significantly where I/B/E/S *lost* rank (the post-ChatGPT triple interaction for I/B/E/S-down firms ranges from -2.3 to -2.8 across specifications, $p < 0.05$) and does not change significantly where it gained. The baseline I/B/E/S ERC rises after ChatGPT ($SUE^{IBES} \times Post$ of about 1.0, $p < 0.01$), which

[28] The earnings response coefficient (ERC) is the slope of announcement returns on the earnings surprise. A larger ERC means the market treats the surprise as more informative.

[29] The lower-order terms are $Post_q$ and the two-way interactions among SUE^{IBES}, $Post_q$, and the two rank-change indicators. The indicators themselves are firm-level and are absorbed by the firm fixed effects.

[30] Versions of Panel A estimated cell by cell, one regression per provider–rank pair, in both absolute-value and signed form, yield the same inferences (untabulated).

is consistent with the decline in surprise magnitudes documented in Section 5.1, so the triple interaction measures a relative decline. The design has a limit. The partition is defined by the change in I/B/E/S's rank between the pre- and post-ChatGPT windows, so for I/B/E/S-down firms the I/B/E/S surprise also became a noisier measure of the news after ChatGPT, which lowers its ERC through attenuation alone. In addition, the positive pre-period coefficient on $SUE^{IBES} \times IBESDown$ shows that these firms started from above-average ERCs. Two other results weigh against attenuation as the full explanation: the within-announcement gradient in Panel A uses *ex ante* ranks and is estimated in both periods, and the analysts' revisions in Section 4.2 show the same one-sided pattern without relying on an ERC.[31] The asymmetry is consistent with a pre-ChatGPT market that already overweighted I/B/E/S.

4.2 Analyst Herding Toward the I/B/E/S Consensus

The two tests above infer the market's reliance on a consensus from the price reaction to its surprise. Analysts' forecast revisions provide a second measure of reliance. Each time a sell-side analyst revises a forecast, the analyst chooses how far to move toward the prevailing consensus (Clement and Tse, 2005), and the strength of that tendency indicates how much weight the analyst assigns to the consensus as a signal. If the post-ChatGPT information environment devalues a degraded default benchmark, analysts should herd toward the I/B/E/S consensus less in the firms where I/B/E/S lost accuracy rank in Panel B, because those are the firms in which the market discounts it. I restrict the test to the I/B/E/S forecast record.[32]

The sample comprises quarterly EPS forecast revisions for the sample firms over 2019–2025. A revision pairs consecutive forecasts by the same analyst for the same firm-quarter, issued within the 180 days preceding the announcement and no more than 120 days apart. The sample contains 170,956 revisions by 2,283 analysts covering 759 firms, of which 122,001 with the full firm and analyst control sets enter the regressions (120,868 in the analyst–firm pair specifications). The

[31] A generalized post-ChatGPT shift in risk premia or attention would move all providers' ERCs together, but neither the within-announcement rank gradient nor the cross-firm asymmetry fits that account.

[32] I/B/E/S is the only provider for which forecast-level data are available to me. Because the five providers also differ in their methodologies, a change in an analyst's revision behavior cannot be attributed to reliance on the consensus of any other provider. The test therefore asks whether analysts' reliance on the I/B/E/S consensus changes after ChatGPT, not whether analysts substitute toward the consensus forecasts of the other FDPs.

benchmark is the official prevailing I/B/E/S summary consensus that analysts actually observe, defined as the mean estimate from the most recent statistical-period summary strictly preceding the revision date (at most 45 days old, at least three estimates). *Herding* equals one when the revised forecast falls between the analyst's own prior and that consensus. The indicator is regressed on $Post_q$ and its interactions with the $IBES\,Up_i$ and $IBES\,Down_i$ indicators of Table 3 (β_1 and β_2), or on their interactions with the three post-ChatGPT GPT-capability regimes (text-only, browsing, and native search, defined in Section 5.3) ($\beta_{1,r}$ and $\beta_{2,r}$). Every column includes the firm and event controls of Equation (5), the analyst controls of Clement (1999) (*General Experience*, *Firm Experience*, *# Firms Covered*, *# Industries Covered*, *Brokerage Size*, and *Forecast Horizon*), analyst and firm fixed effects or analyst×firm pair fixed effects, revision-year fixed effects, and fixed effects for twenty ventiles of the price-scaled distance between the analyst's prior and the consensus, which absorb the mechanical tendency of a revision to agree in direction with a distant consensus. Standard errors cluster by firm.

Table 4 reports the estimates. Herding toward the consensus falls after ChatGPT in the firms where I/B/E/S lost rank. The $Post \times IBES\,Down$ coefficient β_2 is -4.8 percentage points ($t = 2.62$; -5.6, $t = 2.63$, within analyst–firm pairs; about 15 percent of the 32-percent base rate), while the $Post \times IBES\,Up$ coefficient β_1 is positive and insignificant. This is the same one-sided pattern as the market's reweighting in Table 3. The spread $\beta_1 - \beta_2$ is 6.0 percentage points ($t = 2.54$) with firm fixed effects and 7.2 ($t = 2.79$) within pairs. Split by GPT-capability regime (columns (3)–(4)), the down-firm decline is present in every post-ChatGPT regime within pairs (-5.3, -6.5, and -4.9 points; $t = 1.98$ to 2.41), with the spread largest under browsing (8.6 points, $t = 2.75$). Because the partition is defined by the rank change over the full post-period, a decline in every regime is expected and the regime split is descriptive. The pooled columns provide the formal test. The decline is already present in the text-only regime, whereas the market-level premium of Section 5.3 appears only with browsing. The difference is consistent with the analysts' position in the information environment. Sell-side analysts typically have terminal access to the other providers' numbers, so acquisition was not the binding cost for them, and a text-only model could already assist with reconciling the providers' methodologies. The text-only regime therefore serves as a placebo for

the market-level tests, not for the analyst test.

This evidence is corroborative and does not identify the mechanism on its own. It shows that the post-ChatGPT decline in reliance on a degraded historical anchor appears both in prices and in the revision behavior of the analysts whose forecasts feed the consensus.

4.3 A Second Quality Dimension: Predictive Value

Trailing accuracy is one measure of consensus quality. A second is *predictive value*, which measures how closely the consensus anticipates the firm's future street earnings. If the post-ChatGPT selectivity documented above reflects a general preference for consensus quality, and is not an artifact of the trailing-accuracy construction, the same pattern should appear on this dimension as well, as Prediction 1 anticipates.

For each firm, I score the consensus's predictive record in real time. The building block is the own-basis anticipation error, the absolute per-share distance between the quarter-q I/B/E/S consensus and the seasonally matched street actual four quarters ahead, with pairs that straddle a stock split excluded or adjusted. The error is "own basis" because the consensus is paired with the same provider's street actual. This pairing follows the evidence of Online Appendix Table OA1 that the providers' street actuals themselves disagree pervasively. I average this error over up to the twenty most recent source quarters whose year-ahead actuals are already public at the announcement (at least twelve required), scale it by the window mean $|actual|$ so it is comparable across firms, and winsorize. The measure is then $Predictive\ Value^{IBES}_{i,q} = 1 - error_{i,q}$, a continuous cross-firm variable (mean 0.56, standard deviation 0.30). It covers 22,364 of the 23,833 firm-quarters (8,730 pre-ChatGPT), of which 22,356 enter the regressions. The uncovered firm-quarters are recent listings with fewer than twelve scorable history quarters. I estimate

$$\begin{aligned} CAR_{i,q} = \alpha + \beta_1 SUE^{IBES}_{i,q} + \beta_2 SUE^{IBES}_{i,q} \times Post_q + \beta_3 SUE^{IBES}_{i,q} \times Predictive\ Value^{IBES}_{i,q} \\ + \beta_4 SUE^{IBES}_{i,q} \times Predictive\ Value^{IBES}_{i,q} \times Post_q + \text{controls} + \text{FE} + \varepsilon_{i,q}, \end{aligned} \tag{6}$$

with the lower-order terms in $Predictive\ Value^{IBES}$ and $Post_q$ included but not displayed. The coefficient β_3 is the pre-ChatGPT gradient of the ERC in predictive value, and β_4 is its post-ChatGPT increment. The test is on β_4.

The market prices the predictive-value dimension, and does so more steeply after ChatGPT. In Table 5, the ERC gradient on predictive value is already strongly positive before ChatGPT ($\beta_3 = 2.17$, $t = 4.0$). Firms whose consensus predicts well are known to have higher ERCs (Collins and Kothari, 1989), so the identified claim is the post-ChatGPT increment, $\beta_4 = 2.05$ ($t = 2.68$) market-adjusted and 2.10 ($t = 2.82$) under the FFC adjustment. The increment matches the pre-period gradient in magnitude, so the slope roughly doubles after ChatGPT.

Because the measure is close to a firm's earnings predictability, the increment is also consistent with an explanation that does not involve GenAI. The persistence–ERC relation could have steepened at the same boundary for other reasons, such as the 2022 shift in discount rates, which the specification does not rule out directly. Dispersion, volatility, and the unexpected-item flag (the characteristics most likely to proxy for predictability) are among the controls, and the $\textit{Predictive Value}^{IBES}$ level terms are indistinguishable from zero, so the effect runs through the surprise loading.

Column (3) splits the increment across the three post-ChatGPT GPT-capability regimes (text-only, browsing, and native search; Section 5). The increment is 0.92 ($t = 0.9$) under text-only, 1.44 ($t = 1.5$) under browsing, and 3.54 ($t = 3.1$) under native search. The point estimates rise with the tools' retrieval capabilities, though only the native-search step is significant at conventional levels (in column (4), the browsing increment turns marginally significant under the FFC adjustment). This pattern is consistent with a forecast track record that became retrievable only with browsing and native search. Because the native-search regime spans only the final quarters of the sample, the regime profile is descriptive. The pooled increment in columns (1)–(2) is the formal test.

Two alternative explanations for the increment receive little support. The first is announcement composition. Adding the large-unexpected-item flag of Bochkay et al. (2022) with its full set of interactions leaves the increment unchanged (2.04, $t = 2.67$), and the flag is essentially uncorrelated with predictive value (-0.02; both untabulated). The second is selection toward I/B/E/S specifically, which also does not explain the increment. In an untabulated decomposition of the surprise into the component common to all providers and the I/B/E/S-specific tilt, the common component carries

both the gradient (2.3, $t = 3.7$) and its post-ChatGPT increment (1.9, $t = 2.0$), though the increment clears the conventional threshold only narrowly.

The predictive-value evidence also indicates which quality dimension the weighting analyses should use. Applied at the current-quarter horizon, the same procedure yields a like-for-like measure of the consensus's trailing accuracy, and across firms the two are highly correlated (Pearson 0.72, Spearman 0.78; 0.82 correcting for split-half measurement error; untabulated). Predictive value and accuracy thus largely measure the same underlying quality at different horizons, and the accuracy dimension carries most of the predictive-value content.[33] The weighting analyses that follow therefore operate on accuracy, which is realized at every announcement instead of after a four-quarter delay and varies firm-quarter by firm-quarter (the weights themselves use the four-quarter $MAFE^{g}_{i,q}$ of Section 3.4).

5 GenAI and the Accuracy-Weighted Synthesis Across FDPs

Section 4 shows that, after ChatGPT, the market's response to any one provider's surprise depends on that provider's record. In practice, however, investors do not all rely on the same provider's consensus. As Section 3.1 describes, investors can obtain each provider's consensus either directly, through a subscription or a terminal, or indirectly, through the free distributors that redistribute it, and a generative AI query now returns several providers' numbers in one response. The expectation against which the market judges an announcement is therefore an "aggregation" of the consensus forecasts across FDPs, and the announcement return is, in theory, the reaction to the difference between actual earnings and that "aggregation." This section asks which "aggregation" the market uses, and in particular whether the provider-level selective reliance of Section 4 results in a wiser "aggregation." The two heuristic benchmarks anchor on the I/B/E/S consensus or weight the five providers equally. The alternative carries that selectivity into the combination itself by weighting each provider's surprise by its recent accuracy for the firm. This is the second part of Prediction 1,

[33] A provider-level version of the same test, a within-firm-quarter rank score computed for the four providers with native point-in-time history, points in the same direction but is too imprecise to support inference. Its split-half reliability is below 0.35, and its post-minus-pre shift falls to 0.099 ($t = 0.75$) when the Capital IQ observations are excluded (untabulated). The own-basis pairing also rewards internal consistency of a provider's street-earnings rules alongside genuine foresight about the firm.

that the consensus reflected in announcement prices shifts toward a synthesis that draws on multiple FDPs and weights higher-quality signals more heavily. It is also the stronger claim, because a market can favor one accurate provider without combining across providers at all. The test is whether announcement returns respond to the surprise computed from the accuracy-weighted combination of the five consensus forecasts over and above their response to the heuristic benchmarks. I use the composite SUE^{Acc} of Equation (1) in Section 3.4. Section 5.1 validates the construct, Section 5.2 runs it against the two heuristic benchmarks in a horse race, and Section 5.3 examines the timing of the resulting accuracy premium across the four GPT-capability regimes.

5.1 Validating the Accuracy-Weighted Construct

Premise and trading benefits of accuracy-weighted constructs. The design depends on one assumption, that accuracy *persists*, because weighting by *trailing* accuracy is informative only if it does. If accuracy persists, two benefits should follow. SUE^{Acc} should be more accurate than the heuristics, and it should have trading value, so that its difference from a heuristic, which is known *ex ante*, earns an *ex post* return. For the latter I form the *predicted surprise*

$$Pred.Surp._{i,q}^{X} = \left(Consensus_{i,q}^{Acc} - Consensus_{i,q}^{X}\right) / \, p_{i,q}, \qquad X \in \{EW5, IBES\}, \tag{7}$$

where $Consensus_{i,q}^{Acc} = \sum_g w_{i,q}^{g} \, consensus_{i,q}^{g}$ is the consensus EPS built with the weights of Equation (3), $Consensus_{i,q}^{X}$ is the corresponding heuristic consensus, and $p_{i,q}$ is the closing share price on the last trading day before the announcement. I min-max standardize the predicted surprise to the unit interval and test whether it forecasts the four-day announcement BHAR and whether a long–short portfolio sorted on its quintiles earns a positive return.

Two analytical properties support the construct alongside this empirical validation, and Online Appendix D develops both. First, the weights are invariant to any change in accuracy common to the five providers, so an economy-wide improvement in forecast quality leaves them untouched and only the relative accuracy pattern moves them. That relative pattern is stable across the three post-ChatGPT regimes, the period over which the premium emerges, as Table OA2 shows for the distribution of provider accuracy. Second, in a pricing model in which the market's expectation is a convex combination of the accuracy-weighted and heuristic consensus levels, the accuracy

premium equals $\theta(2\omega - 1)$, where ω is the market's relative reliance on the accuracy-weighted synthesis. Under that model, the premium measures how the market uses the providers, not how accurate they are.

The premise holds, and so do both benefits. Provider ranks are highly persistent (Spearman $\rho = 0.71$). In Table 6, Panel A, the least- and most-accurate providers retain their ranks 76 and 65 percent of the time, roughly three to four times the uniform baseline. Part of this persistence is mechanical, because adjacent four-quarter windows share three quarters, so the sharper validation is whether the trailing rank forecasts accuracy out of sample. In Figure OA1, SUE^{Acc} has the smallest mean absolute error of the three surprise measures in both the pre- and post-ChatGPT periods. In untabulated paired tests, the accuracy-weighted composite's absolute standardized surprise is smaller than either heuristic's in both periods ($t > 10$). In an untabulated stacked provider panel that regresses four-quarter-ahead earnings and cash flows on each provider's consensus, the consensus slopes are statistically indistinguishable across providers, so tilting weights toward accurate providers forgoes no detectable predictive content. The panel covers the four providers with native point-in-time history, since Bloomberg's begins only in 2020Q2, and includes provider-specific intercepts, industry effects, and firm-clustered standard errors.

Panel B confirms the trading value. The standardized predicted surprise forecasts the four-day BHAR in all four columns ($p < 0.01$). Per standard deviation of the regressor (0.116 and 0.109 in Table 2), the market-adjusted slopes imply BHAR increments of about 0.29 and 0.37 percentage points against the equal-weighted and I/B/E/S benchmarks. A long–short portfolio on the quintiles of the predicted surprise earns 0.63–0.73 percent per announcement against the equal-weighted benchmark and 1.35 percent against I/B/E/S ($p < 0.01$).

Panel (a) of Figure OA1 shows that surprise magnitudes decline over the sample for all three measures. Because ERCs are concave in surprise magnitude, part of any post-ChatGPT rise in a given loading may reflect smaller surprises, which is why the tests below rest on the contrast between composites estimated on the same announcements and not on the level of any loading.

5.2 The Accuracy Premium: Horse-Race Tests

I test the accuracy-weighted composite SUE^{Acc} against two natural benchmarks: the equal-weighted composite SUE^{EW5} and the I/B/E/S-only consensus SUE^{IBES}. The pooled specification is

$$\begin{aligned} CAR_{i,q} = \alpha + \beta_1 SUE^{Acc}_{i,q} + \beta_2 SUE^{X}_{i,q} + \beta_3 SUE^{Acc}_{i,q} \times Post_q \\ + \beta_4 SUE^{X}_{i,q} \times Post_q + \text{controls} + \text{FE} + \varepsilon_{i,q}, \end{aligned} \tag{8}$$

for $X \in \{EW5, IBES\}$. The pre-period accuracy premium is the contrast $\beta_1 - \beta_2$. Its post-minus-pre change, $\beta_3 - \beta_4$, is the central test of Prediction 1. The dependent variable is the three-day CAR under both risk adjustments, market-adjusted and FFC four-factor. The regime and cross-sectional analyses re-estimate the same contrasts within subperiods and subsamples.[34]

After ChatGPT, the market tilts toward the accuracy-weighted composite at the expense of both heuristics (Table 7). In Panel A (equal-weighted benchmark), the pre-ChatGPT accuracy premium is statistically indistinguishable from zero (-0.972 in the pooled specification, column (3)), consistent with the equal-weighted composite already capturing the pre-period information set. The premium then turns sharply positive. The accuracy-weighted slope rises from 1.376 to 3.420 while the equal-weighted slope falls from 2.324 to 1.767. The falling benchmark slope does not mean the equal-weighted signal lost information. In a horse race each slope is conditional on the other measure, so a rising accuracy-weighted loading absorbs the variation the two share. The formal test is the pooled interaction contrast, and the premium's post-minus-pre change is 2.401 ($p < 0.01$) under the market-adjusted CAR and 2.584 ($p < 0.01$) under the FFC four-factor abnormal return. Panel B repeats the test against the I/B/E/S benchmark. The pre-ChatGPT accuracy premium is negative (-0.623 in the pooled specification, column (3)) but statistically insignificant. The negative sign is consistent with the pre-ChatGPT overweighting of I/B/E/S inferred in Section 4. The post-ChatGPT premium then rises to 2.049 ($p < 0.01$), a post-minus-pre change of 2.672 ($p < 0.01$)

[34] Although SUE^{Acc} and the benchmark SUE^{X} are correlated by construction ($\rho \approx 0.82$ for $X = EW5$, a variance inflation factor near 3), the identifying quantity is the contrast $\beta_1 - \beta_2$, whose sampling variance is a property of the estimand, not of the horse-race parameterization. Any invertible linear reparameterization of SUE^{Acc} and SUE^{X} (and of their $Post_q$ interactions) fits the same model and returns the contrast with the same standard error. Because every test in this section uses the contrast, the inflated standard errors of the individual slopes do not affect the inference.

under the market-adjusted CAR and 2.688 ($p < 0.01$) under the FFC adjustment.

The magnitudes are economically large. After ChatGPT, the response to a one-percentage-point standardized surprise rises by roughly 2.4 percentage points more under the accuracy-weighted consensus than under the equal-weighted benchmark, and by 2.7 more than under I/B/E/S, increments comparable to the entire pre-ChatGPT ERC on the I/B/E/S consensus itself (2.2 in Panel B). Scaled by a typical tilt instead of a full percentage point of surprise, the magnitudes are more modest. With a post-ChatGPT premium of 1.43 against the equal-weighted benchmark and a tilt standard deviation of roughly 0.29 percent of price (implied by the Table 2 standard deviations and the 0.82 correlation), a one-standard-deviation tilt moves the three-day CAR by about 0.4 percentage points, roughly five percent of the CAR's standard deviation. Figure OA1 argues against a mechanical interpretation of the widening. A more accurate composite proxies better for the market's expectation and would earn a higher slope for that reason alone. Its accuracy advantage over each heuristic benchmark, however, is essentially unchanged from the pre- to the post-ChatGPT period (Panel (b)), so a better proxy cannot explain a premium that widens.[35] The pooled contrast, however, only locates the shift at the ChatGPT boundary. The next subsection uses the GPT-capability regimes, which differ in what the tools could retrieve, to examine the timing more finely.

5.3 The Accuracy Premium Across GPT-Capability Regimes

If the premium reflects the GenAI-mediated synthesis of Prediction 1, it should strengthen as the tools' retrieval capabilities expand. The predicted pattern is a dose–response relation across regimes, not a single jump at the ChatGPT release. I therefore re-estimate the horse race of Equation (8) within four GPT-capability regimes: pre-ChatGPT (R0), text-only (R1, from November 30, 2022), browsing (R2, from September 27, 2023), and native search (R3, from October 31, 2024 to December 31, 2025, the end of the sample).[36] The four regimes differ in

[35] Formally, the accuracy tilt $SUE^{Acc}_{i,q} - SUE^{EW5}_{i,q} = \sum_g (w^g_{i,q} - 1/5)\, SUE^g_{i,q}$ is the inner product of each provider's deviation from the uniform weight and its standardized surprise. This is the object the horse-race contrast identifies.

[36] The boundaries follow OpenAI's public release dates. Browsing became permanently available on September 27, 2023, initially for paid subscribers, after a beta that ran from May to early July 2023 had been withdrawn. Announcements are assigned to regimes by announcement date, so each regime contains the announcements made after the capability became available.

what the public tool can retrieve. In the text-only regime the model can discuss providers and their street-earnings methodologies from its training data, but it cannot fetch a current forecast, so the acquisition step of cross-provider synthesis remains manual. From September 2023, browsing allows the model to retrieve the five providers' current numbers from the web within a single query, so that every step of cross-FDP synthesis becomes inexpensive for the first time. Native search then removes the opt-in friction, building retrieval into the default product. The text-only regime also serves as a placebo period. If the premium reflected attention to AI, sentiment, or any other shock coincident with ChatGPT's release, it should already appear in the text-only months. If it reflects retrieval, it cannot. The placebo has two limits. It rules out a discrete shock at the release date, not a gradual process that happens to accelerate in late 2023, and it does not rule out shocks that coincide with the browsing boundary itself. The cross-sectional and visibility tests of Section 6 address the second limit, because a shock unrelated to retrieval would not concentrate in the firms whose forecasts the models can retrieve.

Table 8 reports the horse race regime by regime, against the equal-weighted five-FDP benchmark in Panel A and against the I/B/E/S benchmark in Panel B. In the pre-ChatGPT regime the accuracy premium is negative and statistically indistinguishable from zero on both comparisons, at -0.95 in Panel A and -0.75 in Panel B. The text-only regime looks the same, at 0.18 and -1.23, with neither estimate significant. Under browsing the premium turns positive on both comparisons, at 2.43 ($p < 0.05$) in Panel A and 1.52, which remains insignificant, in Panel B. Under native search it is significant on both, at 2.55 ($p < 0.05$) in Panel A and 4.91 ($p < 0.01$) in Panel B. The text-only regime provides the placebo result. When the chatbot cannot yet retrieve forecasts, the premium is statistically undetectable against either benchmark. The premium becomes detectable against the equal-weighted benchmark at the browsing boundary, where retrieval first becomes available, and against both benchmarks under native search. The predictive-value gradient of Section 4 rises across the same regimes (Table 5, columns (3)–(4)), with only its native-search increment significant at the five-percent level. Figure 2 plots the estimates across the four regimes.[37]

[37] Suggestive outage evidence points the same way. Using major ChatGPT outages (four incidents of at least 240 minutes through 2024) as short-duration shocks to GenAI availability, following Cheng et al. (2025) and Liu and Subasi

6 Cross-Sectional Tests and LLM Visibility

Prediction 2 holds that the premium concentrates where GenAI-mediated synthesis is more valuable and where it is feasible. This section tests both conditions. I test value using cross-sectional partitions on disagreement and special items, and I test feasibility using a direct measure of LLM retrieval.

6.1 Cross-Sectional Tests: Disagreement Across FDPs and the Magnitude of Special Items

Because the regime evidence localizes the accuracy premium to the browsing and native-search regimes, the cross-sectional tests replace the post-ChatGPT indicator $Post_q$ with the post-scraping indicator $Post_q^{\text{Scr}}$, equal to one for announcements on or after the onset of GPT browsing (September 27, 2023). I re-estimate the horse-race specification in Equation (8) within the below- and above-median halves of two partitions. Each partition tests a different channel. (i) *Concurrent cross-FDP SUE dispersion* is the interquartile range of the five FDPs' SUE for the same firm-quarter. Accuracy weighting has the greatest scope to be selective among providers when this dispersion is high. (ii) *Special-items magnitude* is the absolute value of special items scaled by the absolute value of income before extraordinary items. Because FDPs differ in their treatment of unexpected charges and gains, cross-FDP disagreement over what belongs in street earnings should be largest where such items are economically material.[38]

Table 9 reports the partition-by-partition estimates. The post-scraping change in the premium is significant only in the predicted half of each partition. In Panel A (cross-FDP dispersion), the post-minus-pre change in the accuracy premium is 2.211 ($p < 0.05$) against the equal-weighted benchmark and 2.427 ($p < 0.01$) against I/B/E/S in the high-disagreement subsample, versus insignificant estimates (2.037 and -0.691) in the low-disagreement subsample. The equal-weighted

(2026), I find that the accuracy premium against I/B/E/S falls by 3.173 ($p < 0.05$) on the 84 announcements made the day after an outage, while the contrast against the equal-weighted benchmark shows no significant change (point estimate 3.787, $t = 1.37$). This pattern is consistent with a GenAI-mediated channel. A shock that removes synthesis lowers the loadings on both composites that require it, the accuracy-weighted and the equal-weighted, so the contrast between them should be little changed, whereas the contrast against I/B/E/S, the fallback that requires no synthesis, should fall. With few treated days the evidence is suggestive, and Online Appendix A (Table OA3) reports the full test.

[38] A variant of this table that uses the post-ChatGPT boundary in place of the post-scraping indicator and adds a third partition, a median split on the LLM-visibility score of Section 6.2, yields similar inferences. The premium concentrates in the high half of each partition (untabulated).

comparison requires a caveat. The low-disagreement point estimate is close in magnitude to its high-disagreement counterpart and differs mainly in precision. The partition separates cleanly only against I/B/E/S, where the low-disagreement estimate turns negative. In that subsample SUE^{Acc} and SUE^{IBES} nearly coincide, so the two loadings are separately ill-determined and the level of the premium is imprecisely estimated. Its post-minus-pre change, -0.691, is insignificant. Panel B (special items) requires no such qualification. The premium change is 2.518 ($p < 0.05$) against the equal-weighted benchmark and 3.638 ($p < 0.01$) against I/B/E/S in the high-special-items subsample, versus point estimates of the opposite sign (-0.552 and -0.219) in the low-special-items subsample. The two partitions test different parts of the mechanism. The dispersion partition tests selectivity, which has the most scope where the five FDPs disagree most about the size of the surprise, and the special-items partition tests integration, which has the most scope where their methodological choices most affect it.

6.2 LLM Visibility and Effective Synthesis

A precondition for the premium is that GenAI actually retrieves the firm's forecasts for the investor who asks. The accessibility clause of Prediction 2 turns this precondition into a testable restriction. The premium should be stronger in firms whose forecasts the models retrieve than in firms whose forecasts they do not. Integration requires acquisition, because the model cannot synthesize forecasts it cannot retrieve. This section introduces *LLM visibility*, a direct measure of that acquisition. For each firm, I query two browsing-enabled models, ChatGPT (GPT-4o) and Perplexity, for the next-quarter consensus EPS through the DataForSEO LLM Responses API, which returns each model's answer together with the sources it cites. I record whether the response contains an EPS figure and how many unique domains it cites. The measure is the EPS-extraction indicator times $\log(1 + \#\text{ unique cited domains})$, averaged across the two models' responses.[39] It captures realized retrieval, not retrieval capability. The measure is a single firm-level classification taken at one point in time with the models then available,[40] and it is applied to every regime. It

[39]The indicator ensures that a response citing many sources but returning no EPS figure counts as no acquisition. The logarithm allows for diminishing returns to breadth, so that the difference between citing one source and two counts for more than the difference between citing nine and ten. Online Appendix B reports worked examples.

[40]The queries were issued in May 2026, after the end of the sample period.

proxies for retrievability in the browsing and native-search regimes under the assumption that the set of distributors carrying a firm's forecasts is stable over the sample. Drift in that set adds classification error, which attenuates the estimated gap and does not produce it.

Table OA4 shows how little LLM visibility shares with the classical prominence proxies. Only 47–50 percent of high-visibility firms also sit in the top half of firm size, analyst coverage, or institutional ownership, near the 50 percent benchmark implied by independent median splits,[41] whereas 80 percent of large firms also sit in the top half of analyst coverage. In Panels B and C, a pooled horse race includes every measure's high-group triple interaction in one regression, so each measure must explain where the post-scraping accuracy premium concentrates while the others are held in the specification. Against I/B/E/S, LLM visibility is the only measure significant at the five-percent level (5.84, $p < 0.01$). Against the equal-weighted benchmark, its coefficient of 3.17 is only marginally significant ($p < 0.10$), as is institutional ownership's (3.22).

A second test combines the visibility split with the regime timing in a difference-in-differences design. The first difference compares the accuracy premium across high- and low-visibility firms, and the second compares that gap across regimes, before and after the model can retrieve forecasts. If acquisition is a precondition for synthesis, the gap should be absent while retrieval is impossible and should emerge once it becomes possible. The pooled estimate of this design is the visibility triple interaction in Table OA4, Panels B and C, reported above. To examine its timing, I estimate, separately within each GPT-capability regime r,

$$\begin{aligned} CAR_{i,q} = \alpha + \beta_1 SUE^{Acc}_{i,q} + \beta_2 SUE^{X}_{i,q} + \beta_3 SUE^{Acc}_{i,q} \times High\,Visibility_i \\ + \beta_4 SUE^{X}_{i,q} \times High\,Visibility_i + \gamma High\,Visibility_i + \text{controls} + \text{FE} + \varepsilon_{i,q}, \end{aligned} \tag{9}$$

with the firm and event controls of Equation (8), firm and year-quarter fixed effects, and standard errors two-way clustered by firm and calendar year. The low-visibility premium is $\beta_1 - \beta_2$, the high-visibility premium is $(\beta_1 + \beta_3) - (\beta_2 + \beta_4)$, and the high-minus-low gap within the regime is $Premium_r^{\text{High}-\text{Low}} = \beta_{3,r} - \beta_{4,r}$. In Table 10, $Premium_r^{\text{High}-\text{Low}}$ is statistically insignificant at the five-percent level in the pre-ChatGPT and text-only regimes (R0, R1). This is the

[41] If two median splits were independent, half of the firms above the median on one would lie above the median on the other. An overlap near 50 percent therefore indicates that the two measures sort firms almost unrelatedly.

difference-in-differences analog of flat pre-trends. The one marginally significant estimate in these regimes, against I/B/E/S in the text-only regime, is negative. The framework does not predict a negative sign, and the estimate is too imprecise to interpret. A formal pre-trend test agrees. Stacking the two pre-retrieval regimes and interacting every regressor with the text-only indicator, I find the change in the gap from R0 to R1 statistically indistinguishable from zero ($p = 0.22$ against I/B/E/S, $p = 0.16$ against the equal-weighted benchmark; untabulated). Its point estimate is negative, the opposite of the opening that follows. Once the model can browse, the gap emerges. It is 2.7 to 3.0 in the browsing regime (R2, $p < 0.05$), then 3.1 to 4.2 under native search (R3), significant against I/B/E/S ($p < 0.01$) but only marginally so against the equal-weighted benchmark ($p < 0.10$). The premium is not confined to high-visibility firms once native search arrives (the low-visibility premium against I/B/E/S in R3 is 3.05, $p < 0.01$), but the gap between the two groups opens only with retrieval. Figure 3 plots the high- and low-visibility premiums regime by regime, with the gap between the two lines opening at the browsing boundary. The premium thus concentrates in firms whose forecasts are visible to the model, and it does so only once the model can retrieve them. This pattern is consistent with acquisition being a precondition for synthesis.

The regime timing also weighs against the residual concern that LLM visibility merely proxies for firm prominence. The visibility split is a single firm-level classification applied to every regime, so a prominence account resting on a static firm characteristic would predict a premium gap in the pre-ChatGPT and text-only regimes as well. The gap instead opens only when retrieval arrives. A prominence effect that itself emerged with browsing would appear in the horse race of Table OA4, where the classical proxies compete directly with LLM visibility to explain the post-scraping concentration and, against I/B/E/S, none is significant at the five-percent level.

7 Conclusion

This paper shows that after the release of ChatGPT, market reactions to earnings announcements increasingly reflect an accuracy-weighted synthesis across the five major forecast data providers instead of the heuristic benchmarks that characterized prior decades, in particular reliance on a single FDP. The post-ChatGPT market relies on the FDPs selectively, by accuracy and by the consensus's

predictive value. The within-firm-quarter per-rank ERC increment roughly doubles. The response to the I/B/E/S surprise weakens where I/B/E/S has lost relative accuracy and does not change where it has gained, and analysts' revisions herd less toward the I/B/E/S consensus in the same firms. The ERC gradient on predictive value roughly doubles. The post-ChatGPT accuracy premium relative to both heuristic benchmarks is statistically and economically significant. The premium concentrates in firm-quarters with high cross-FDP disagreement, where accuracy weighting has the most scope to be selective, and in firm-quarters with large special items, where the providers' methodologies differ most. The predicted surprise on which it rests is tradable through a quintile long–short portfolio (0.63–1.35 percent per announcement), and the premium rises across GPT-capability regimes to 4.91 against I/B/E/S under native search. It also concentrates in firms whose forecasts the model retrieves, and this visibility gap appears only in the regimes in which the model can retrieve forecasts.

These findings make three contributions. First, they show that the provider whose consensus the market prices changes with the information technology investors use. The choice of forecast data provider is therefore itself an outcome to be explained. This implication extends recent work on cross-FDP differences in consensus information (Larocque et al., 2026). For measurement practice, the benchmark against which earnings surprises should be gauged shifts with that technology. Second, they identify a specific channel through which GenAI reaches capital markets. GenAI lowers the cost of synthesizing analyst forecasts across sources, and the market now prices the synthesized forecast into earnings-announcement reactions. This channel is distinct from the broader GenAI effects documented in recent work (e.g., Bertomeu et al., 2025; Cheng et al., 2025; Liu and Subasi, 2026). Third, they contribute to the disclosure-processing-cost framework of Blankespoor et al. (2020) by identifying cross-source integration as the cost component that prior single-source technologies did not reduce and that GenAI does.

Two caveats apply. First, the accuracy-weighted construction is a tractable proxy. The evidence establishes that market reactions align more closely with an accuracy-sensitive synthesis than with the heuristic benchmarks, not that the market applies these particular weights. Second, LLM visibility is measured at a point in time, and model outputs vary across queries and evolve with the

models themselves (e.g., Chen et al., 2024; Lopez-Lira et al., 2025). To the extent that this variation is unrelated to announcement returns, it attenuates the estimated visibility gap and does not produce it.

More broadly, the market's earnings benchmark is technology-dependent. For decades, a single provider's consensus stood in for the market's expectation because integrating across providers was costly. GenAI has lowered that cost, and the benchmark has shifted. After ChatGPT, the market places measurable weight on providers' relative accuracy, weight that was undetectable before browsing became available. What investors can integrate at low cost shapes what prices reflect.

References

Ahlers, D. and J. Lakonishok. 1983. A study of economists' consensus forecasts. *Management Science* 29 (10): 1113–1125.

Bartov, E., D. Givoly, and C. Hayn. 2002. The rewards to meeting or beating earnings expectations. *Journal of Accounting and Economics* 33 (2): 173–204.

Bates, J. M. and C. W. J. Granger. 1969. The combination of forecasts. *Operational Research Quarterly* 20 (4): 451–468.

Battalio, R. H. and R. R. Mendenhall. 2005. Earnings expectations, investor trade size, and anomalous returns around earnings announcements. *Journal of Financial Economics* 77 (2): 289–319.

Bernard, V. L. and J. K. Thomas. 1989. Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research* 27 (Supplement): 1–36.

Bertomeu, J., Y. Lin, Y. Liu, and Z. Ni. 2025. The impact of generative AI on information processing: Evidence from the ban of ChatGPT in Italy. *Journal of Accounting and Economics* 80 (1): 101782.

Beyer, A., D. A. Cohen, T. Z. Lys, and B. R. Walther. 2010. The financial reporting environment: Review of the recent literature. *Journal of Accounting and Economics* 50 (2–3): 296–343.

Blankespoor, E. 2019. The impact of information processing costs on firm disclosure choice: Evidence from the XBRL mandate. *Journal of Accounting Research* 57 (4): 919–967.

Blankespoor, E., G. S. Miller, and H. D. White. 2014. The role of dissemination in market liquidity: Evidence from firms' use of Twitter. *The Accounting Review* 89 (1): 79–112.

Blankespoor, E., E. deHaan, and C. Zhu. 2018. Capital market effects of media synthesis and dissemination: Evidence from robo-journalism. *Review of Accounting Studies* 23 (1): 1–36.

Blankespoor, E., E. deHaan, and I. Marinovic. 2020. Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics* 70 (2–3): 101344.

Blankespoor, E., J. Croom, and S. M. Grant. 2026. Generative AI and investor processing of financial information. *Journal of Accounting and Economics* 82 (2): 101908.

Bochkay, K., S. Markov, M. Subasi, and E. H. Weisbrod. 2022. The roles of data providers and analysts in the production, dissemination, and pricing of street earnings. *Journal of Accounting Research* 60 (5): 1695–1740.

Bradshaw, M. T., M. S. Drake, J. N. Myers, and L. A. Myers. 2012. A re-examination of analysts' superiority over time-series forecasts of annual earnings. *Review of Accounting Studies* 17 (4): 944–968.

Bradshaw, M. T., Y. Ertimur, and P. C. O'Brien. 2017. Financial analysts and their contribution to well-functioning capital markets. *Foundations and Trends in Accounting* 11 (3): 119–191.

Bradshaw, M. T., C. Ma, B. P. Yost, and Y. Zou. 2026. Generative AI use by capital market information intermediaries: Evidence from Seeking Alpha. *Journal of Accounting Research* 64 (3): 1233–1286.

Brown, L. D., P. A. Griffin, R. L. Hagerman, and M. E. Zmijewski. 1987. An evaluation of alternative proxies for the market's assessment of unexpected earnings. *Journal of Accounting and Economics* 9 (2): 159–193.

Bushee, B. J., D. A. Matsumoto, and G. S. Miller. 2003. Open versus closed conference calls: The determinants and effects of broadening access to disclosure. *Journal of Accounting and Economics* 34 (1–3): 149–180.

Bushee, B. J., J. E. Core, W. Guay, and S. J. W. Hamm. 2010. The role of the business press as an information intermediary. *Journal of Accounting Research* 48 (1): 1–19.

CFA Institute. 2024. Employer survey results. CFA Institute survey report on AI and generative AI in the investment industry, August 2024. Available at: https://www.cfainstitute.org/sites/default/files/-/media/documents/survey/AI-disruption-employer-survey-2024.pdf.

Chang, A. Y., X. Dong, X. Martin, and C. Zhou. 2026. AI democratization and trading inequality. *Journal of Accounting Research* 64 (3): 1287–1331.

Chen, L., M. Zaharia, and J. Zou. 2024. How is ChatGPT's behavior changing over time? *Harvard Data Science Review* 6 (2).

Cheng, Q., P. Lin, and Y. Zhao. 2025. Does generative AI facilitate investor trading? Early evidence from ChatGPT outages. *Journal of Accounting and Economics* 80 (2–3): 101821.

Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* 5 (4): 559–583.

Clement, M. B. 1999. Analyst forecast accuracy: Do ability, resources, and portfolio complexity matter? *Journal of Accounting and Economics* 27 (3): 285–303.

Clement, M. B. and S. Y. Tse. 2003. Do investors respond to analysts' forecast revisions as if forecast accuracy is all that matters? *The Accounting Review* 78 (1): 227–249.

Clement, M. B. and S. Y. Tse. 2005. Financial analyst characteristics and herding behavior in forecasting. *Journal of Finance* 60 (1): 307–341.

Cohen, L. and D. Lou. 2012. Complicated firms. *Journal of Financial Economics* 104 (2): 383–400.

Coleman, B., T. Dyer, and M. Lang. 2025. The influence of financial data subscriptions on analyst research. Working paper, University of Georgia, Brigham Young University, and University of North Carolina at Chapel Hill. Available at: https://ssrn.com/abstract=5096105.

Collins, D. W. and S. P. Kothari. 1989. An analysis of intertemporal and cross-sectional determinants of earnings response coefficients. *Journal of Accounting and Economics* 11 (2–3): 143–181.

Da, Z., J. Engelberg, and P. Gao. 2011. In search of attention. *Journal of Finance* 66 (5): 1461–1499.

deHaan, E., T. Shevlin, and J. Thornock. 2015. Market (in)attention and the strategic scheduling and timing of earnings announcements. *Journal of Accounting and Economics* 60 (1): 36–55.

DellaVigna, S. and J. M. Pollet. 2009. Investor inattention and Friday earnings announcements. *Journal of Finance* 64 (2): 709–749.

Diether, K. B., C. J. Malloy, and A. Scherbina. 2002. Differences of opinion and the cross section of stock returns. *Journal of Finance* 57 (5): 2113–2141.

Drake, M. S., D. T. Roulstone, and J. R. Thornock. 2012. Investor information demand: Evidence from Google searches around earnings announcements. *Journal of Accounting Research* 50 (4): 1001–1040.

Drake, M. S., D. T. Roulstone, and J. R. Thornock. 2015. The determinants and consequences of information acquisition via EDGAR. *Contemporary Accounting Research* 32 (3): 1128–1161.

Driskill, M., M. P. Kirk, and J. W. Tucker. 2020. Concurrent earnings announcements and analysts' information production. *The Accounting Review* 95 (1): 165–189.

Easton, P. D. and M. E. Zmijewski. 1989. Cross-sectional variation in the stock market response to accounting earnings announcements. *Journal of Accounting and Economics* 11 (2–3): 117–141.

Goldstein, I., C. S. Spatt, and M. Ye. 2021. Big data in finance. *Review of Financial Studies* 34 (7): 3213–3225.

Granger, C. W. J. and R. Ramanathan. 1984. Improved methods of combining forecasts. *Journal of Forecasting* 3 (2): 197–204.

Gu, S., B. Kelly, and D. Xiu. 2020. Empirical asset pricing via machine learning. *Review of Financial Studies* 33 (5): 2223–2273.

Hirshleifer, D., S. S. Lim, and S. H. Teoh. 2009. Driven to distraction: Extraneous events and underreaction to earnings news. *Journal of Finance* 64 (5): 2289–2325.

Holthausen, R. W. and R. E. Verrecchia. 1988. The effect of sequential information releases on the variance of price changes in an intertemporal multi-asset market. *Journal of Accounting Research* 26 (1): 82–106.

Hou, K., M. A. van Dijk, and Y. Zhang. 2012. The implied cost of capital: A new approach. *Journal of Accounting and Economics* 53 (3): 504–526.

Kim, A. G., M. Muhn, and V. V. Nikolaev. 2023. Bloated disclosures: Can ChatGPT help investors process information? Becker Friedman Institute Working Paper No. 2023-59. Available at: https://arxiv.org/abs/2306.10224.

Kim, A. G., M. Muhn, and V. V. Nikolaev. 2024. Financial statement analysis with large language models. SSRN/Becker Friedman Institute. Available at: https://arxiv.org/abs/2407.17866.

Kim, O. and R. E. Verrecchia. 1991. Trading volume and price reactions to public announcements. *Journal of Accounting Research* 29 (2): 302–321.

Kimbrough, M. D. 2005. The effect of conference calls on analyst and market underreaction to earnings announcements. *The Accounting Review* 80 (1): 189–219.

Kimbrough, M. D., L. Y. Liu, and M. Subasi. 2026. AI-powered analysts. SSRN Working Paper No. 7176461. Available at: https://ssrn.com/abstract=7176461.

Kothari, S. P. 2001. Capital markets research in accounting. *Journal of Accounting and Economics* 31 (1–3): 105–231.

Larocque, S. A., J. Watkins, and E. H. Weisbrod. 2026. Consensus? An examination of differences in earnings information across forecast data providers. *Journal of Accounting Research, forthcoming* .

Lawrence, A. 2013. Individual investors and financial disclosure. *Journal of Accounting and Economics* 56 (1): 130–147.

Liu, L. Y. and M. Subasi. 2026. When does generative AI level the playing field? Evidence from crowdsourced earnings forecasts. SSRN Working Paper No. 7253378. Available at: https://ssrn.com/abstract=7253378.

Livnat, J. and R. R. Mendenhall. 2006. Comparing the post-earnings announcement drift for surprises calculated from analyst and time series forecasts. *Journal of Accounting Research* 44 (1): 177–205.

Lopez-Lira, A. and Y. Tang. 2026. Can ChatGPT forecast stock price movements? Return predictability and large language models. *Journal of Financial Economics* 184: 104335.

Lopez-Lira, A., Y. Tang, and M. Zhu. 2025. The memorization problem: Can we trust LLMs' economic forecasts? SSRN Working Paper No. 5217505. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5217505.

Loughran, T. and B. McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance* 66 (1): 35–65.

Loughran, T. and B. McDonald. 2017. The use of EDGAR filings by investors. *Journal of Behavioral Finance* 18 (2): 231–248.

Michaely, R., A. Rubin, D. Segal, and A. Vedrashko. 2024. Do differences in analyst quality matter for investors relying on consensus information? *Management Science* 70 (2): 751–772.

Mikhail, M. B., B. R. Walther, and R. H. Willis. 1997. Do security analysts improve their performance with experience? *Journal of Accounting Research* 35: 131–157. Supplement.

Park, C. W. and E. K. Stice. 2000. Analyst forecasting ability and the stock price reaction to forecast revisions. *Review of Accounting Studies* 5 (3): 259–272.

Schipper, K. 1991. Analysts' forecasts. *Accounting Horizons* 5 (4): 105–121.

Stickel, S. E. 1992. Reputation and performance among security analysts. *Journal of Finance* 47 (5): 1811–1836.

Stock, J. H. and M. W. Watson. 2004. Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting* 23 (6): 405–430.

Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York, Doubleday.

Tetlock, P. C. 2007. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance* 62 (3): 1139–1168.

Timmermann, A. 2006. Forecast combinations. In Elliott, G., C. W. J. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, chapter 4, pages 135–196. Elsevier.

Walther, B. R. 1997. Investor sophistication and market earnings expectations. *Journal of Accounting Research* 35 (2): 157–179.

Xue, J., Q. Zhang, and W. Zhu. 2025. Generative AI for analysts. SSRN Working Paper No. 5821142. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5821142.

Appendix A: Variable Definitions

Variable	Definition	Source
Panel A. Across-FDP firm-quarter outcome and surprise variables		
$CAR^{Mkt}_{i,q,[-1,+1]}$	Three-day market-adjusted cumulative abnormal return around the earnings-announcement date, $\sum_{t=-1}^{+1}(r_{i,t} - \text{vwretd}_t)$, in percent.	CRSP
$CAR^{FFC}_{i,q,[-1,+1]}$	Three-day FFC four-factor (FF3+UMD) abnormal CAR over $[-1,+1]$ with factor loadings estimated on the $[-260,-11]$ window.	CRSP
$BHAR^{Mkt}_{i,q,[-2,+1]}$	Four-day market-adjusted buy-and-hold abnormal return over $[-2,+1]$.	CRSP
$BHAR^{FFC}_{i,q,[-2,+1]}$	Four-day FFC four-factor buy-and-hold abnormal return over $[-2,+1]$.	CRSP
$SUE^{g}_{i,q}$	Provider g's standardized unexpected earnings, $g \in \{\text{FactSet, Bloomberg, Zacks, CapIQ, IBES}\}$: $(actual^{g}_{i,q} - consensus^{g}_{i,q})/\text{price}_{i,q}$, in percent of price, where $consensus^{g}_{i,q}$ is provider g's end-of-quarter consensus EPS for quarter q, $actual^{g}_{i,q}$ its own street actual EPS for the same quarter, and $\text{price}_{i,q}$ the closing share price on the last trading day before the announcement.	FactSet; Bloomberg; Zacks; S&P CapIQ; I/B/E/S; CRSP
$SUE^{Acc}_{i,q}$	Accuracy-weighted multi-FDP SUE, $\sum_g w^{g}_{i,q} SUE^{g}_{i,q}$, with the accuracy weights $w^{g}_{i,q}$ defined in Panel C.	Calculated
$SUE^{EW5}_{i,q}$	Equal-weighted multi-vendor SUE, $(1/5)\sum_g SUE^{g}_{i,q}$.	Calculated
Standardized Pred. Surp.$^{X}_{i,q}$	Min-max standardized predicted surprise $(Consensus^{Acc}_{i,q} - Consensus^{X}_{i,q})/\text{price}_{i,q}$, $X \in \{EW5, IBES\}$, where $Consensus^{Acc}_{i,q} = \sum_g w^{g}_{i,q} consensus^{g}_{i,q}$ and $Consensus^{X}_{i,q}$ is the equal-weighted five-FDP consensus (*EW5*) or I/B/E/S's own consensus (*IBES*).	Calculated
Panel B. Firm and event controls		
Log Market Equity$_{i,q-1}$	Natural log of firm i's market capitalization at the end of quarter $q-1$.	CRSP
Book-to-Market$_{i,q-1}$	Book value of equity divided by market value of equity at quarter-end $q-1$.	Compustat; CRSP
Leverage$_{i,q-1}$	Long-term debt plus debt in current liabilities, scaled by total assets, at $q-1$.	Compustat
Prior 12-Month Return$_{i,q}$	Buy-and-hold raw return over the twelve months ending one month before firm i's quarter-q announcement.	CRSP
Log Price$_{i,q-1}$	Natural log of share price at quarter-end $q-1$.	CRSP
Return Volatility$_{i,q}$	Standard deviation of daily returns over the $[-260,-11]$ window before the announcement.	CRSP
Analyst Dispersion$_{i,q}$	Standard deviation of analyst EPS forecasts divided by absolute consensus.	I/B/E/S
Busyness$_{i,q}$	Natural log of the number of S&P 1500 firms announcing earnings on firm i's announcement day.	Compustat

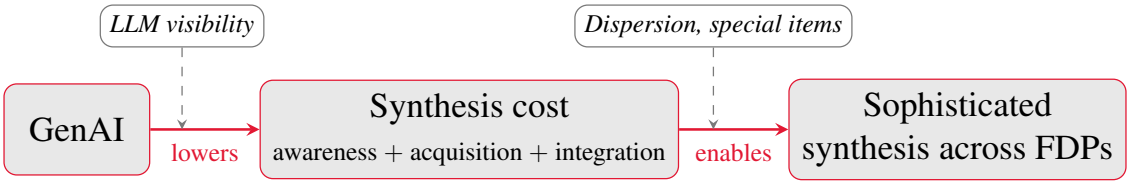
(continued on next page)

Variable	Definition	Source
$\lvert\Delta IB\rvert_{i,q}$	Absolute year-over-year change in income before extraordinary items, scaled by lagged total assets.	Compustat
$\lvert\Delta Shares\rvert_{i,q}$	Absolute year-over-year change in shares outstanding, scaled by lagged shares.	Compustat
$\lvert SpecItems\rvert_{i,q}$	Absolute value of special items, scaled by lagged total assets.	Compustat
$SBC\ High_{i,q}$	Indicator equal to one if stock-based compensation expense exceeds the sample median.	Compustat
$Unexpected\ Item_{i,q}$	Indicator equal to one if the quarter contains a nonrecurring item flagged by Compustat.	Compustat
$Stock\ Split_{i,q}$	Indicator equal to one if a stock split was effective within $[-30,+30]$ trading days of the announcement.	CRSP
$Fiscal\ Q4_{i,q}$	Indicator equal to one for fiscal fourth-quarter announcements.	Compustat
$Institutional\ Ownership_{i,q-1}$	Fraction of shares held by institutional investors at $q-1$.	Thomson Reuters 13F
$\#\ Analysts_{i,q}$	Number of unique analysts in the IBES detail file forecasting EPS for firm-quarter (i,q); enters the regressions in logs.	I/B/E/S
Panel C. Cross-sectional partitioning and identification variables		
$LLM\ Visibility_i$	EPS-extraction indicator $\times \log(1+\#$ unique cited domains$)$, averaged across ChatGPT and Perplexity for direct queries about firm i's quarterly EPS.	DataForSEO
$Cross\text{-}FDP\ SUE\ Dispersion_{i,q}$	Interquartile range of $SUE^{g}_{i,q}$ across the five FDPs in firm-quarter (i,q).	Calculated
$\lvert SpecItems\rvert_{i,q}\,(\%)$	Absolute special items as a percentage of absolute income before extraordinary items, $\lvert SpecItems\rvert_{i,q}/\lvert IB\rvert_{i,q}\times 100$.	Compustat
$IBES\,Up_i$, $IBES\,Down_i$	Indicators equal to one if I/B/E/S's rounded average per-firm-quarter accuracy rank among the five FDPs rose (fell) by at least two positions from pre- to post-ChatGPT.	Calculated
$Outage_{i,q}$	Indicator equal to one if a ChatGPT incident of at least 240 minutes occurred on the calendar day immediately preceding firm i's quarter-q announcement date.	OpenAI
Large Unexpected Item ($LUI_{i,q}$)	Indicator equal to one if firm-quarter (i,q) contains a large unexpected item: any of seven Compustat non-recurring items mapped per Bochkay et al. (2022), with restructuring, acquisition, and special items requiring top-decile magnitude. Refines *Unexpected Item* (Panel B) with a magnitude screen; $INFO_{i,q}=1-LUI_{i,q}$.	Compustat
$Post_q$, $Post^{\mathrm{Scr}}_q$, regimes R0–R3	Announcement-date indicators: $Post_q=1$ on or after 2022-11-30 (ChatGPT's public release); $Post^{\mathrm{Scr}}_q=1$ on or after 2023-09-27 (the onset of GPT browsing). The GPT-capability regimes partition the same dates: R0, pre-ChatGPT (before 2022-11-30); R1, text-only (to 2023-09-26); R2, browsing (to 2024-10-30); R3, native search (2024-10-31 to 2025-12-31, the sample end).	OpenAI

(continued on next page)

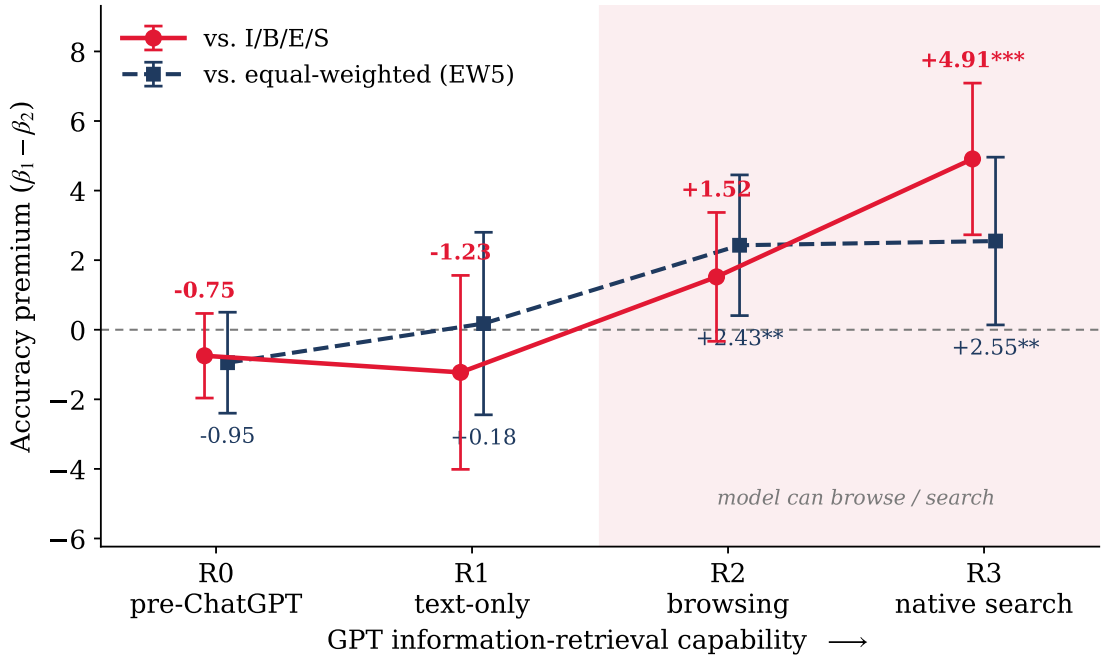
Variable	Definition	Source
$MAFE^{g}_{i,q}$	Provider g's trailing mean absolute forecast error, $(1/4)\sum_{k=1}^{4} \lvert actual^{g}_{i,q-k} - consensus^{g}_{i,q-k} \rvert / \text{price}_{i,q-k}$, over the four quarters preceding the announcement.	Calculated
$w^{g}_{i,q}$	Accuracy weight on provider g: the softmax of the within-firm-quarter z-score of $-MAFE^{g}_{i,q}$ across the five FDPs (Section 3.4); uniform when the five accuracies coincide.	Calculated
$Rank^{g}_{i,q}$	Provider g's within-firm-quarter accuracy rank by $MAFE^{g}_{i,q}$ ($1 =$ least accurate, $5 =$ most accurate).	Calculated
Predictive Value$^{IBES}_{i,q}$	One minus the I/B/E/S consensus's trailing relative anticipation error: the mean of $\lvert actual_{i,s+4} - consensus_{i,s} \rvert$ over up to the twenty most recent source quarters s whose year-ahead actuals are public at the announcement (at least twelve required; pairs straddling a stock split excluded or adjusted), scaled by the window mean $\lvert actual \rvert$ and winsorized at $(1, 99)$; defined for 22,364 firm-quarters.	I/B/E/S
High X	Median-split indicators equal to one when a measure exceeds its median: *High Visibility* (LLM Visibility, cross-firm median, ties assigned to Low), *High Cross-FDP Dispersion*, *High Special Items* ($\lvert SpecItems \rvert$ (%)), and, in Table OA4, firm size (Log Market Equity), analyst coverage (# Analysts), and Institutional Ownership.	Calculated
Panel D. Analyst revision-level variables (Section 4.2)		
Herding$_{a,i,q,t}$	Indicator equal to one when the revised forecast lands between the analyst's own prior forecast and the prevailing I/B/E/S consensus, and zero otherwise. Revisions are consecutive forecasts by the same analyst for the same firm-quarter, issued within the 180 days preceding the announcement and at most 120 days apart.	I/B/E/S
Prevailing I/B/E/S Consensus$_{i,q,t}$	Mean estimate from the latest I/B/E/S statistical-period summary strictly preceding the revision date, at most 45 days old and based on at least three estimates.	I/B/E/S
Consensus Distance$_{a,i,q,t}$	Absolute distance between the analyst's prior forecast and the prevailing consensus, scaled by the closing share price on the last trading day before the announcement, in percent; enters the regressions as twenty ventile fixed effects.	I/B/E/S; CRSP
General (Firm) Experience$_{a,t}$	Number of years since the analyst's first I/B/E/S forecast (first forecast for firm i); enters the regressions as the log of one plus the variable.	I/B/E/S
Firms (Industries) Covered$_{a,t}$	Number of distinct firms (two-digit SIC industries) for which the analyst issues forecasts in the revision year; enters the regressions as the log of one plus the variable.	I/B/E/S; CRSP
Brokerage Size$_{a,t}$	Number of analysts issuing forecasts at the analyst's brokerage in the revision year; enters the regressions as the log of one plus the variable.	I/B/E/S
Forecast Horizon$_{a,i,q,t}$	Number of days from the revision date to the earnings announcement; enters the regressions as the log of one plus the variable.	I/B/E/S

FIGURE 1. Theoretical Framework



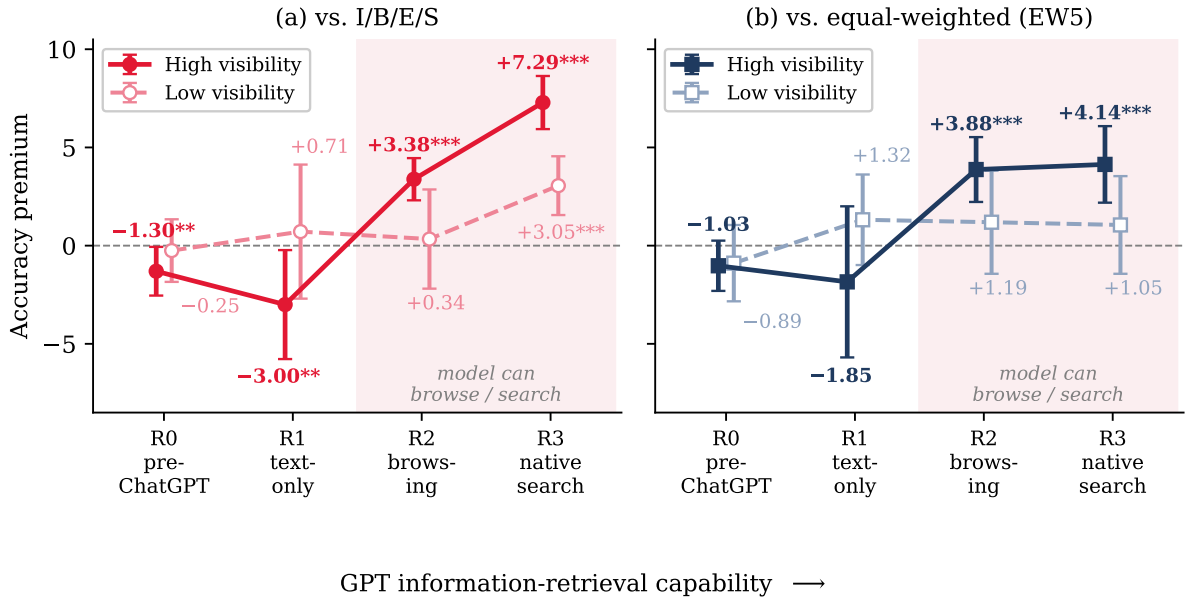
Notes. This figure summarizes the theoretical framework. Generative AI (GenAI) lowers the cost of synthesizing analyst consensus forecasts across the five forecast data providers (FDPs), the joint cost of *awareness*, *acquisition*, and *integration* in the Blankespoor et al. (2020) decomposition, which in turn enables the market to price a sophisticated, accuracy-weighted synthesis across providers in place of a single-FDP or equal-weighted heuristic. The strength of the channel is moderated by how *accessible* synthesis is for a given firm (LLM visibility, the information-acquisition channel) and by how *valuable* it is (cross-FDP disagreement and special-items magnitude). These mechanisms map to Predictions 1 and 2, developed in Section 2. Prediction 1's quality dimension covers both trailing accuracy and the consensus's predictive value for the firm's future results.

FIGURE 2. Accuracy Premium by GPT Information-Retrieval Capability



Notes. This figure plots the *accuracy premium*, the contrast $\beta_1 - \beta_2$ between the loading of the three-day market-adjusted $CAR_{[-1,+1]}$ (in percent) on the accuracy-weighted consensus surprise SUE^{Acc} and its loading on the benchmark surprise (both in percent of price), estimated separately within each GPT-capability regime. The vertical axis is therefore in percentage points of announcement return per one-percentage-point surprise. The estimates are taken directly from Table 8: the red line uses the I/B/E/S benchmark (Panel B) and the navy line the equal-weighted five-FDP benchmark (Panel A). Regimes are ordered by ChatGPT's information-retrieval capability: R0–R3, defined in Appendix A. The shaded region marks the regimes in which the model can retrieve forecasts from the web. Whiskers are 95% confidence intervals, and point labels report the premium and its significance ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$) from firm-clustered standard errors. The premium is statistically indistinguishable from zero before retrieval is available, turns positive once the model can browse, and grows monotonically thereafter, the dose-response pattern the GenAI-mediated synthesis of Prediction 1 implies.

FIGURE 3. Accuracy Premium: Firms With High vs. Low LLM Visibility



Notes. This figure plots the accuracy premium separately for high-visibility firms (solid lines, filled markers) and low-visibility firms (dashed lines, open markers) within each GPT-capability regime, from Table 10. Panel (a) uses the I/B/E/S benchmark and Panel (b) the equal-weighted five-FDP benchmark. The vertical distance between the two lines is the difference-in-differences contrast $\beta_3 - \beta_4$ reported in Table 10. Regimes are ordered by ChatGPT's information-retrieval capability (R0–R3, defined in Appendix A), and the shaded region marks the regimes in which the model can retrieve forecasts from the web. Whiskers are 95% confidence intervals, and point labels report each premium and its significance ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$) from standard errors two-way clustered by firm and calendar year. The two lines are statistically indistinguishable at the five-percent level before retrieval is available, the difference-in-differences analog of flat pre-trends, and separate once the model can browse.

TABLE 1. Sample Selection

Sample construction (firm × quarter)	**Observations**	**# Firms**
S&P 1500 firm-quarters with an earnings announcement (announcement date 2021–2025)	29,395	1,502
Retain firm-quarters with non-missing consensus from all five FDPs (FactSet, Bloomberg, Zacks, Capital IQ, and I/B/E/S)	25,027	1,353
Drop firm-quarters missing market-outcome variables or required controls	(1,179)	(13)
Drop firm-quarters missing any constructed return or predicted-surprise variable	(10)	(1)
Drop firms with only one firm-quarter (no within-firm variation for firm fixed effects)	(5)	(5)
Final Sample	**23,833**	**1,334**

Notes. This table reports the sample selection procedure for the firm-quarter sample. Counts in parentheses are the firm-quarters (firms) dropped at that step. All other rows report the number surviving. The sample is used in the main across-FDP tests and is built from CRSP, Compustat, I/B/E/S, and the five-FDP consensus panel.

TABLE 2. Descriptive Statistics

	N	Mean	SD	Min	p25	Median	p75	Max
Dependent Variables: Abnormal Returns (%)								
$CAR^{Mkt}_{[-1,+1]}$	23,833	0.204	7.739	-22.532	-3.933	0.168	4.255	23.535
$CAR^{FFC}_{[-1,+1]}$	23,833	0.196	7.635	-22.779	-3.776	0.186	4.135	23.606
$BHAR^{Mkt}_{[-2,+1]}$	23,833	0.228	7.978	-22.803	-4.165	0.177	4.418	24.992
$BHAR^{FFC}_{[-2,+1]}$	23,833	0.193	7.847	-22.935	-3.966	0.111	4.255	24.819
Per-FDP Standardized Earnings Surprise (SUE^g, %)								
$SUE^{FactSet}$	23,833	0.132	0.534	-2.190	0.000	0.074	0.238	2.484
$SUE^{Bloomberg}$	23,833	-0.222	1.659	-10.049	-0.362	-0.027	0.162	5.780
SUE^{Zacks}	23,833	0.149	0.467	-1.606	0.000	0.075	0.240	2.291
SUE^{CapIQ}	23,833	0.137	0.530	-2.185	0.001	0.076	0.243	2.477
SUE^{IBES}	23,833	0.130	0.548	-2.271	0.000	0.075	0.241	2.502
Cross-FDP Composites and Predicted Surprise								
SUE^{Acc}	23,833	0.110	0.450	-1.656	-0.015	0.059	0.212	1.972
SUE^{EW5}	23,833	0.066	0.510	-1.936	-0.064	0.041	0.201	1.942
$Standardized\ Pred.Surp.^{EW5}$	23,833	0.606	0.116	0.000	0.590	0.619	0.631	1.000
$Standardized\ Pred.Surp.^{IBES}$	23,833	0.483	0.109	0.000	0.469	0.478	0.493	1.000
Firm and Event Controls								
Log Market Equity	23,833	8.992	1.448	6.390	7.923	8.769	9.942	12.883
Book-to-Market	23,833	0.440	0.361	-0.159	0.170	0.362	0.634	1.670
Leverage	23,833	0.314	0.218	0.001	0.133	0.301	0.444	1.008
Prior 12-Month Return	23,833	0.201	0.487	-0.589	-0.101	0.112	0.379	2.400
Log Price	23,833	4.166	0.980	1.880	3.491	4.178	4.831	6.580
Return Volatility	23,833	0.021	0.009	0.009	0.015	0.019	0.025	0.054
Analyst Dispersion	23,833	0.002	0.003	0.000	0.000	0.001	0.002	0.018
Busyness	23,833	4.944	0.973	1.946	4.431	5.124	5.684	6.299
$\lvert\Delta$ *Income Before Extraordinary Items*$\rvert$	23,833	0.012	0.026	0.000	0.001	0.004	0.011	0.178
$\lvert\Delta$ *Shares Outstanding*$\rvert$	23,833	0.010	0.021	0.000	0.001	0.003	0.009	0.150
$\lvert$*Special Items*$\rvert$	23,833	0.003	0.010	0.000	0.000	0.000	0.002	0.074
High Stock-Based Compensation	23,833	0.500	0.500	0.000	0.000	0.000	1.000	1.000
Unexpected Item	23,833	0.686	0.464	0.000	0.000	1.000	1.000	1.000
Stock Split	23,833	0.003	0.051	0.000	0.000	0.000	0.000	1.000
Fiscal Q4	23,833	0.250	0.433	0.000	0.000	0.000	0.000	1.000
Institutional Ownership	23,833	0.850	0.238	0.011	0.813	0.940	1.000	1.000
# Analysts	23,833	10.914	6.536	2.000	6.000	10.000	15.000	31.000
Identification, Weighting, and Partitioning Variables								
$Predictive\ Value^{IBES}$	22,364	0.564	0.298	-0.466	0.441	0.663	0.776	0.911
Cross-FDP SUE Dispersion	23,833	0.108	0.261	0.000	0.007	0.024	0.081	1.804
Log # Analysts	23,833	2.196	0.653	0.693	1.792	2.303	2.708	3.434
Post-ChatGPT	23,833	0.604	0.489	0.000	0.000	1.000	1.000	1.000

(continued on next page)

	N	Mean	SD	Min	p25	Median	p75	Max
Post-Scraping ($Post^{Scr}$)	23,833	0.448	0.497	0.000	0.000	0.000	1.000	1.000
Regime 1 (text-only)	23,833	0.157	0.364	0.000	0.000	0.000	0.000	1.000
Regime 2 (browsing)	23,833	0.233	0.423	0.000	0.000	0.000	0.000	1.000
Regime 3 (native search)	23,833	0.215	0.411	0.000	0.000	0.000	0.000	1.000
Cross-Sectional Partitioning Indicators								
High Visibility	23,827	0.489	0.500	0.000	0.000	0.000	1.000	1.000
High Cross-FDP Dispersion	23,833	0.500	0.500	0.000	0.000	0.000	1.000	1.000
High Special Items	23,833	0.500	0.500	0.000	0.000	0.000	1.000	1.000
IBES Accuracy-Rank Up	23,833	0.099	0.299	0.000	0.000	0.000	0.000	1.000
IBES Accuracy-Rank Down	23,833	0.059	0.236	0.000	0.000	0.000	0.000	1.000
ChatGPT Outage (prior day)	23,833	0.004	0.059	0.000	0.000	0.000	0.000	1.000
Analyst Revision Panel (Section 4.2)								
Herding	122,001	0.323	0.468	0.000	0.000	0.000	1.000	1.000
Consensus Distance (% of price)	122,001	0.227	0.379	0.004	0.031	0.084	0.238	2.321
General Experience (years)	122,001	16.025	11.341	0.416	6.940	13.317	22.563	41.462
Firm Experience (years)	122,001	7.281	6.822	0.079	1.996	5.183	10.691	33.435
# Firms Covered	122,001	23.255	8.999	4.000	17.000	22.000	28.000	51.000
# Industries Covered	122,001	9.357	4.455	2.000	6.000	9.000	12.000	21.000
Brokerage Size	122,001	81.793	65.585	2.000	31.000	65.000	114.000	251.000
Forecast Horizon (days)	122,001	69.957	38.847	2.000	29.000	85.000	91.000	162.000

Notes. This table reports descriptive statistics for the across-FDP sample of S&P 1500 firm-quarters with announcement dates in 2021–2025 used in the main tests. All continuous variables are winsorized at the (1, 99) level. *High Visibility* is defined for 23,827 of the 23,833 firm-quarters, and *# Analysts* enters the regressions in logs. $Rank^g$, $MAFE^g$, and the softmax weights w^g are construction inputs to SUE^{Acc} (Section 3.4). The analyst revision panel reports the sample of Section 4.2. Experience, coverage, brokerage size, and horizon enter the regressions as logs of one plus the variable. Variable definitions are in Appendix A.

TABLE 3. Earnings Response Coefficients by Provider Accuracy Rank and by Change in I/B/E/S's Accuracy Rank

Panel A. Earnings-response coefficients by forecast-provider accuracy rank

	Pre-ChatGPT		Post-ChatGPT	
Provider	β	γ	β	γ
I/B/E/S	2.591***	0.022	2.611***	0.295**
	(6.60)	(0.17)	(5.49)	(2.09)
FactSet	2.316***	0.050	2.059***	0.380**
	(4.78)	(0.36)	(3.80)	(2.42)
Bloomberg	0.985***	-0.058	0.789***	0.162**
	(5.82)	(-0.80)	(5.01)	(2.10)
Zacks	3.603***	-0.139	4.757***	-0.020
	(8.35)	(-1.04)	(8.57)	(-0.13)
Capital IQ	1.107**	0.543***	2.483***	0.349**
	(2.40)	(3.94)	(4.94)	(2.40)
Stacked	**1.206*** **	**0.306*** **	**0.973*** **	**0.621*** **
	(9.21)	**(6.93)**	**(8.09)**	**(11.65)**
Observations		47,130		72,035
Stacked Δ (Post−Pre)			**-0.232**	**0.316*** **
			(-1.38)	**(4.91)**

Panel B. Heterogeneous post-ChatGPT effects by firm-level I/B/E/S accuracy-rank change

Dep. Var.:	Market-adjusted $CAR_{[-1,+1]}$		FFC $CAR_{[-1,+1]}$	
	(1)	(2)	(3)	(4)
SUE^{IBES}	2.859***	2.845***	2.775***	2.776***
	(13.91)	(14.03)	(13.67)	(13.84)
$SUE^{IBES} \times Post$	1.029***	0.978***	1.136***	1.065***
	(3.56)	(3.32)	(4.11)	(3.76)
$SUE^{IBES} \times IBES\,Up$	0.693	0.628	0.803	0.737
	(1.24)	(1.16)	(1.46)	(1.39)
$SUE^{IBES} \times IBES\,Down$	1.397*	1.582**	1.913***	2.080***
	(1.81)	(2.00)	(2.61)	(2.74)
$SUE^{IBES} \times Post \times IBES\,Up$	**-1.248**	**-1.135**	**-1.291**	**-1.185**
	(-1.44)	**(-1.30)**	**(-1.40)**	**(-1.26)**
$SUE^{IBES} \times Post \times IBES\,Down$	**-2.278** **	**-2.344** **	**-2.758*** **	**-2.791** **
	(-2.27)	**(-2.15)**	**(-2.70)**	**(-2.51)**
$Post \times IBES\,Up$	0.353	0.232	0.437	0.297
	(1.13)	(0.68)	(1.37)	(0.85)
$Post \times IBES\,Down$	0.544	0.713	0.678*	0.859*
	(1.34)	(1.61)	(1.68)	(1.95)
Controls	No	Yes	No	Yes
Firm FE / Year FE	Yes	Yes	Yes	Yes
Observations	23,833	23,833	23,833	23,833
Adjusted R^2	0.052	0.070	0.054	0.072

Notes. Both panels test whether the market relies more on more accurate forecasts. **Panel A** estimates Equation (4) without controls or fixed effects (FE), so that γ is the ERC increment per one-rank accuracy improvement. Each provider row is a separate regression, the *Stacked* row pools all five FDPs' surprises across firm-quarters, and the *Stacked* Δ row reports the $SUE \times Post$ and $SUE \times Rank \times Post$ coefficients from the pooled regression that interacts every term with *Post*. **Panel B** estimates Equation (5) at the firm-quarter level. The bolded triple interactions are the key coefficients, the remaining lower-order terms are included but not reported, and the specification includes the firm and event controls and firm and year fixed effects. Standard errors are clustered by firm. Continuous variables are winsorized at $(1,99)$. t-statistics in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Variable definitions are in Appendix A.

TABLE 4. Analyst Herding Toward the I/B/E/S Consensus and the Accuracy-Rank Partition

	Single post-ChatGPT indicator		GPT-capability regimes	
	(1)	(2)	(3)	(4)
Post	0.044**	0.049**		
	(2.01)	(2.23)		
Post × *IBES Up* (β_1)	0.012	0.016		
	(0.71)	(0.90)		
Post × *IBES Down* (β_2)	-0.048***	-0.056***		
	(-2.62)	(-2.63)		
R1 (text-only) × *IBES Up* ($\beta_{1,1}$)			0.002	0.006
			(0.11)	(0.30)
R1 (text-only) × *IBES Down* ($\beta_{2,1}$)			-0.049**	-0.053**
			(-2.07)	(-1.98)
R2 (browsing) × *IBES Up* ($\beta_{1,2}$)			0.016	0.021
			(0.84)	(1.09)
R2 (browsing) × *IBES Down* ($\beta_{2,2}$)			-0.059**	-0.065**
			(-2.45)	(-2.41)
R3 (native search) × *IBES Up* ($\beta_{1,3}$)			0.019	0.023
			(0.83)	(0.96)
R3 (native search) × *IBES Down* ($\beta_{2,3}$)			-0.034	-0.049**
			(-1.59)	(-2.19)
Spread ($\beta_1 - \beta_2$)	**0.060****	**0.072*****		
	(2.54)	**(2.79)**		
Spread, R1 ($\beta_{1,1} - \beta_{2,1}$)			**0.052***	**0.059***
			(1.78)	**(1.88)**
Spread, R2 ($\beta_{1,2} - \beta_{2,2}$)			**0.075****	**0.086*****
			(2.57)	**(2.75)**
Spread, R3 ($\beta_{1,3} - \beta_{2,3}$)			**0.052***	**0.072****
			(1.79)	**(2.32)**
Lower-order terms / Firm and event controls	Yes	Yes	Yes	Yes
Analyst controls	Yes	Yes	Yes	Yes
Analyst FE / Firm FE	Yes	No	Yes	No
Analyst×Firm FE	No	Yes	No	Yes
Revision-year FE / Distance-ventile FE	Yes	Yes	Yes	Yes
Observations	122,001	120,868	122,001	120,868
Adjusted R^2	0.107	0.107	0.107	0.107

Notes. This table tests whether analysts' herding toward the prevailing I/B/E/S consensus falls after ChatGPT in the firms where I/B/E/S lost accuracy rank. The sample comprises 122,001 quarterly EPS forecast revisions over 2019–2025. The dependent variable is *Herding* (base rate 0.32). $IBES\,Up_i$ and $IBES\,Down_i$ are the accuracy-rank-change indicators of Table 3, interacted with the single *Post* indicator in columns (1)–(2) and with the GPT-capability regimes in columns (3)–(4). The bolded *Spread* rows report $\beta_1 - \beta_2$ and $\beta_{1,r} - \beta_{2,r}$. All columns include the lower-order terms, the firm and event controls of Equation (5), the Clement (1999) analyst controls, and distance-ventile fixed effects (FE). Columns (1)/(3) add analyst, firm, and revision-year fixed effects, columns (2)/(4) analyst×firm pair and revision-year fixed effects. t-statistics in parentheses; standard errors cluster by firm; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Variable definitions are in Appendix A.

TABLE 5. The Predictive Value of the I/B/E/S Consensus: Measurement and Post-ChatGPT Pricing

	Single post-ChatGPT indicator		GPT-capability regimes	
Dep. Var.:	Mkt CAR	FFC CAR	Mkt CAR	FFC CAR
	(1)	(2)	(3)	(4)
SUE^{IBES} (β_1)	2.237***	2.241***	2.237***	2.241***
	(8.71)	(9.02)	(8.74)	(9.05)
$SUE^{IBES} \times Post$ (β_2)	0.534	0.565*		
	(1.54)	(1.67)		
$SUE^{IBES}\times$ Regime 1 (text-only)			0.868*	1.016**
			(1.91)	(2.29)
$SUE^{IBES}\times$ Regime 2 (browsing)			0.785	0.591
			(1.62)	(1.18)
$SUE^{IBES}\times$ Regime 3 (native search)			0.162	0.268
			(0.33)	(0.57)
$SUE^{IBES} \times Predictive\ Value^{IBES}$ (β_3)	2.172***	2.119***	2.174***	2.125***
	(4.03)	(4.11)	(4.05)	(4.13)
$SUE^{IBES} \times$ ***Predictive Value***$^{IBES} \times Post$ (β_4)	**2.049*** **	**2.098*** **		
	(2.68)	**(2.82)**		
$SUE^{IBES} \times$ ***Predictive Value***$^{IBES}\times$ **Regime 1**			**0.919**	**0.941**
			(0.93)	**(0.97)**
$SUE^{IBES} \times$ ***Predictive Value***$^{IBES}\times$ **Regime 2**			**1.435**	**1.777***
			(1.45)	**(1.80)**
$SUE^{IBES} \times$ ***Predictive Value***$^{IBES}\times$ **Regime 3**			**3.540*** **	**3.365*** **
			(3.08)	**(3.02)**
Lower-order terms / Controls	Yes	Yes	Yes	Yes
Firm FE / Year FE	Yes	Yes	Yes	Yes
Observations	22,356	22,356	22,356	22,356
Adjusted R^2	0.082	0.083	0.082	0.083

Notes. This table tests whether the post-ChatGPT strengthening of announcement-window reactions concentrates in firms whose I/B/E/S consensus has closely anticipated the firm's future street numbers. *Predictive Value*^{IBES} covers 22,364 of the 23,833 firm-quarters (8,730 pre-ChatGPT), of which 22,356 enter the regressions. The uncovered firm-quarters are recent listings with fewer than twelve scorable history quarters. Columns (1)–(2) estimate Equation (6) with the three-day market-adjusted and FFC $CAR_{[-1,+1]}$ as the dependent variable. β_3 is the pre-ChatGPT gradient of the ERC in predictive value and the bolded β_4 is its post-ChatGPT increment. Columns (3)–(4) replace the single *Post* indicator with the three post-ChatGPT GPT-capability regimes. Because the native-search regime spans only the final five sample quarters, the regime profile is descriptive and the pooled increment in columns (1)–(2) is the formal test. All columns include the lower-order terms, the firm and event controls, and firm and year fixed effects (FE). Standard errors are clustered by firm. t-statistics in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Variable definitions are in Appendix A.

TABLE 6. Accuracy Persistence and the Predicted-Surprise Trading Strategy

Panel A. Persistence of FDP accuracy ranks (trailing-MAFE rank, quarter t to $t+1$)

	$\Pr(rank_{t+1} = j \mid rank_t = i)$				
$rank_t$	$j=1$ *(least)*	$j=2$	$j=3$	$j=4$	$j=5$ *(most)*
1 *(least)*	**0.761**	0.121	0.049	0.034	0.036
2	0.121	**0.541**	0.197	0.089	0.052
3	0.047	0.195	**0.467**	0.211	0.080
4	0.034	0.091	0.208	**0.483**	0.185
5 *(most)*	0.038	0.052	0.079	0.184	**0.647**

Spearman $\rho = 0.71$; overall same-rank rate $= 0.58$; 109,915 quarter-to-quarter pairs.

Panel B. Trading strategy from predicted surprise

Dep. Var.:	Market-adj. $BHAR_{[-2,+1]}$		FFC $BHAR_{[-2,+1]}$	
Benchmark:	Equal-Weighted	IBES	Equal-Weighted	IBES
	(1)	(2)	(3)	(4)
Standardized Pred. Surp.	**2.493*****	**3.433*****	**2.440*****	**3.357*****
	(4.10)	**(5.35)**	**(4.08)**	**(5.25)**
Long−Short Return (Q5−Q1)	**0.73%*****	**1.35%*****	**0.63%*****	**1.35%*****
Controls / Firm FE / Year FE	Yes	Yes	Yes	Yes
Observations (firm-quarters)	23,833	23,833	23,833	23,833
Adjusted R^2	0.016	0.017	0.018	0.019

Notes. This table motivates the accuracy-weighted consensus. **Panel A** reports the quarter-to-quarter persistence of accuracy ranks (ranks by trailing mean absolute forecast error, MAFE): cell (i, j) is the share of provider-quarters at Rank i that move to Rank j the following quarter (rows sum to one up to rounding). **Panel B** tests whether the standardized predicted surprise against each benchmark $X \in \{EW5, IBES\}$ forecasts the four-day announcement-window abnormal return: columns (1)–(2) use the market-adjusted $BHAR_{[-2,+1]}$ and columns (3)–(4) the FFC $BHAR_{[-2,+1]}$. The Long−Short row reports the abnormal return, in percent per announcement, from going long the top quintile and short the bottom quintile of the predicted surprise over the $[-2,+1]$ window. All Panel B columns include firm and year fixed effects (FE), the firm and event controls, and firm-clustered standard errors. Continuous variables are winsorized at $(1,99)$. t-statistics in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Variable definitions are in Appendix A.

TABLE 7. The Accuracy Premium: Horse Races Between the Accuracy-Weighted and Heuristic Surprises

Panel A. Horse race: SUE^{Acc} vs. SUE^{EW5}

Dep. Var.:	Mkt CAR$_{[-1,+1]}$			FFC CAR$_{[-1,+1]}$		
	(1) Pre	(2) Post	(3) Pooled	(4) Pre	(5) Post	(6) Pooled
SUE^{Acc} (β_1)	1.376***	3.420***	1.307***	1.401***	3.519***	1.290***
	(3.45)	(9.38)	(3.60)	(3.63)	(10.01)	(3.67)
$SUE^{Acc} \times Post$ (β_3)			2.003***			2.135***
			(4.08)			(4.47)
SUE^{EW5} (β_2)	2.324***	1.767***	2.279***	2.379***	1.734***	2.284***
	(6.10)	(5.56)	(6.84)	(6.44)	(5.62)	(7.03)
SUE^{EW5} $\times Post$ (β_4)			-0.399			-0.449
			(-0.92)			(-1.06)
Premium (Pre-ChatGPT)	-0.948		-0.972	-0.978		-0.994
	(-1.28)		(-1.47)	(-1.37)		(-1.55)
Premium (Post-ChatGPT)		1.654***	1.429**		1.785***	1.590***
		(2.62)	(2.39)		(2.93)	(2.74)
Δ Premium (Post$-$Pre)			**2.401*** **			**2.584*** **
			(2.75)			**(3.03)**
Control Variables						
Log Market Equity	-5.348***	-4.964***	-2.848***	-5.585***	-5.125***	-3.238***
	(-5.35)	(-5.35)	(-6.00)	(-5.59)	(-5.74)	(-7.19)
Book-to-Market	0.413	-1.315	0.168	0.849	-1.245	0.309
	(0.36)	(-1.22)	(0.28)	(0.75)	(-1.16)	(0.53)
Leverage	-4.957**	-3.652*	-0.913	-4.561**	-4.067**	-1.312
	(-2.38)	(-1.70)	(-0.76)	(-2.17)	(-1.97)	(-1.13)
Prior 12-Month Return	-0.295	-0.868**	-0.235	-0.024	-1.038***	-0.129
	(-1.27)	(-2.52)	(-1.27)	(-0.11)	(-3.02)	(-0.71)
Log Price	-1.063	-1.047	-0.482	-0.671	-0.735	-0.029
	(-1.13)	(-1.12)	(-1.05)	(-0.71)	(-0.83)	(-0.07)
Return Volatility	-59.587***	27.445*	14.899	-53.652***	19.210	11.467
	(-3.08)	(1.66)	(1.34)	(-2.87)	(1.20)	(1.05)
Analyst Dispersion	-23.636	-40.842	-15.946	-11.993	-17.323	-2.625
	(-0.40)	(-0.71)	(-0.40)	(-0.20)	(-0.30)	(-0.07)
Busyness	-0.104	-0.245*	-0.172*	-0.132	-0.219*	-0.164*
	(-0.73)	(-1.83)	(-1.79)	(-0.93)	(-1.66)	(-1.72)
$\lvert\Delta IB\rvert$	16.729***	-4.012	2.450	17.171***	-4.609	2.027
	(2.98)	(-0.78)	(0.69)	(3.09)	(-0.91)	(0.58)
$\lvert\Delta Shares\rvert$	3.721	-1.964	1.505	3.251	-2.171	1.376
	(0.79)	(-0.51)	(0.52)	(0.72)	(-0.57)	(0.48)
$\lvert SpecItems\rvert$	-18.458	-8.991	-10.912	-21.400*	-7.112	-9.847
	(-1.50)	(-0.79)	(-1.33)	(-1.77)	(-0.64)	(-1.22)
SBC High	-0.149	0.174	0.106	-0.168	0.186	0.095
	(-0.57)	(0.77)	(0.65)	(-0.66)	(0.83)	(0.59)
Unexpected Item	-0.258	-0.036	-0.123	-0.222	-0.023	-0.103
	(-0.97)	(-0.17)	(-0.78)	(-0.89)	(-0.11)	(-0.68)
Stock Split	-2.679**	-1.110	-1.584*	-2.266*	-0.472	-0.977
	(-2.21)	(-0.92)	(-1.90)	(-1.78)	(-0.42)	(-1.18)
Fiscal Q4	0.600***	-0.045	0.276**	0.606***	0.230	0.417***
	(2.83)	(-0.24)	(2.07)	(2.89)	(1.23)	(3.19)
Institutional Ownership	3.234	-0.702**	-0.470	1.626	-0.543	-0.416
	(1.56)	(-2.02)	(-1.41)	(0.83)	(-1.57)	(-1.25)

(continued on next page)

Table 7 continued from previous page

Log # Analysts	0.508	-0.218	-0.248	0.648	-0.328	-0.371
	(0.79)	(-0.39)	(-0.65)	(1.04)	(-0.60)	(-0.99)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9,405	14,402	23,833	9,405	14,402	23,833
Adjusted R^2	0.085	0.106	0.080	0.092	0.110	0.083

Panel B. Horse race: SUE^{Acc} vs. SUE^{IBES}

Dep. Var.:	Mkt CAR$_{[-1,+1]}$			FFC CAR$_{[-1,+1]}$		
	(1) Pre	(2) Post	(3) Pooled	(4) Pre	(5) Post	(6) Pooled
SUE^{Acc} (β_1)	1.428***	3.678***	1.430***	1.527***	3.796***	1.494***
	(4.00)	(10.49)	(4.31)	(4.36)	(10.70)	(4.48)
$SUE^{Acc} \times Post$ (β_3)			2.153***			2.204***
			(4.82)			(4.98)
SUE^{IBES} (β_2)	2.176***	1.460***	2.054***	2.149***	1.407***	1.974***
	(7.08)	(4.44)	(7.31)	(7.20)	(4.20)	(7.01)
SUE^{IBES} $\times Post$ (β_4)			-0.519			-0.484
			(-1.34)			(-1.28)
Premium (Pre-ChatGPT)	-0.747		-0.623	-0.622		-0.480
	(-1.20)		(-1.09)	(-1.03)		(-0.83)
Premium (Post-ChatGPT)		2.219***	2.049***		2.389***	2.208***
		(3.56)	(3.60)		(3.75)	(3.82)
Δ Premium (Post−Pre)			**2.672*** **			**2.688*** **
			(3.44)			**(3.50)**
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9,405	14,402	23,833	9,405	14,402	23,833
Adjusted R^2	0.087	0.106	0.080	0.094	0.110	0.083

Notes. This table reports OLS estimates of the horse race of Equation (8) between the accuracy-weighted surprise SUE^{Acc} and a benchmark surprise SUE^{X}, where $Post_q$ is the post-ChatGPT indicator. **Panel A** uses the equal-weighted benchmark SUE^{EW5} and **Panel B** the I/B/E/S benchmark SUE^{IBES}. Both panels use the same firm and event controls, which Panel B suppresses for brevity. Columns (1) and (4) restrict the sample to pre-ChatGPT announcements and columns (2) and (5) to post-ChatGPT announcements, so the *Post* interactions drop out. Columns (3) and (6) pool both periods. Columns (1)–(3) use the market-adjusted $CAR_{[-1,+1]}$ and columns (4)–(6) use the Fama-French-Carhart (FFC) four-factor $CAR_{[-1,+1]}$. The *Premium* rows (Pre-ChatGPT, Post-ChatGPT) report the accuracy premium, the loading on SUE^{Acc} minus the loading on the benchmark: $\beta_1 - \beta_2$ in the single-period columns and in the pre-ChatGPT row of the pooled columns, and $(\beta_1 + \beta_3) - (\beta_2 + \beta_4)$ in the post-ChatGPT row of the pooled columns. The bolded Δ Premium row reports the post-minus-pre change, $\beta_3 - \beta_4$. All columns include firm and year fixed effects (FE) and the firm and event controls. Standard errors are clustered by firm. Continuous variables are winsorized at (1, 99). t-statistics in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Variable definitions are in Appendix A.

TABLE 8. The Accuracy Premium Across GPT-Capability Regimes

Panel A. Benchmark: equal-weighted five-FDP consensus

Regime:	(1) R0 Pre-ChatGPT	(2) R1 text-only	(3) R2 browsing	(4) R3 native search
SUE^{Acc} (β_1)	1.376***	2.597***	3.753***	3.954***
	(3.45)	(3.38)	(6.19)	(5.69)
Benchmark SUE (β_2)	2.324***	2.419***	1.325***	1.404**
	(6.10)	(3.78)	(2.63)	(2.26)
Premium ($\beta_1 - \beta_2$)	**-0.948**	**0.178**	**2.429****	**2.550****
	(-1.28)	**(0.13)**	**(2.35)**	**(2.07)**
Controls / Firm FE / Year FE	Yes	Yes	Yes	Yes
Observations	9,405	3,705	5,529	5,084
Adjusted R^2	0.085	0.158	0.151	0.140

Panel B. Benchmark: I/B/E/S consensus

Regime:	(1) R0 Pre-ChatGPT	(2) R1 text-only	(3) R2 browsing	(4) R3 native search
SUE^{Acc} (β_1)	1.428***	1.918**	3.362***	5.079***
	(4.00)	(2.35)	(6.17)	(7.87)
Benchmark SUE (β_2)	2.176***	3.144***	1.842***	0.169
	(7.08)	(4.70)	(3.74)	(0.30)
Premium ($\beta_1 - \beta_2$)	**-0.747**	**-1.226**	**1.519**	**4.909*****
	(-1.20)	**(-0.86)**	**(1.61)**	**(4.41)**
Controls / Firm FE / Year FE	Yes	Yes	Yes	Yes
Observations	9,405	3,705	5,529	5,084
Adjusted R^2	0.087	0.170	0.155	0.138

Notes. This table estimates the accuracy premium separately within each GPT-capability regime. For each regime the specification is Equation (8) without the $Post_q$ interactions, with the market-adjusted $CAR_{[-1,+1]}$ as the dependent variable, and the *Accuracy Premium* is the contrast $\beta_1 - \beta_2$. The regimes are the GPT-capability regimes R0–R3 of Appendix A. The benchmark X is the equal-weighted five-FDP consensus (Panel A) or I/B/E/S (Panel B). All columns include the firm and event controls, firm and year fixed effects (FE), and firm-clustered standard errors. Continuous variables are winsorized at $(1,99)$. t-statistics in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Variable definitions are in Appendix A.

TABLE 9. Cross-Sectional Tests of the Accuracy Premium

Panel A. Concurrent cross-FDP SUE dispersion

Benchmark:	Equal-Weighted		IBES	
	(1) Low	(2) High	(3) Low	(4) High
SUE^{Acc} (β_1)	4.138***	1.433***	1.569**	1.748***
	(6.37)	(4.23)	(2.43)	(5.57)
Benchmark SUE (β_2)	3.318***	2.024***	7.496***	1.560***
	(5.86)	(6.41)	(10.03)	(6.10)
$SUE^{Acc} \times Post^{Scr}$ (β_3)	2.748**	1.610***	1.439	1.746***
	(2.48)	(2.89)	(1.34)	(3.63)
Benchmark $SUE \times Post^{Scr}$ (β_4)	0.712	-0.602	2.130*	-0.681*
	(0.76)	(-1.20)	(1.86)	(-1.70)
Premium (Pre-Scrape)	0.821	-0.591	-5.927***	0.188
	(0.73)	(-0.96)	(-4.58)	(0.35)
Premium (Post-Scrape)	2.857*	1.620**	-6.619***	2.615***
	(1.71)	(1.98)	(-3.81)	(3.81)
Δ Premium (Post−Pre)	**2.037**	**2.211****	**-0.691**	**2.427*****
	(1.06)	**(2.21)**	**(-0.33)**	**(2.99)**
Controls / Firm FE / Year FE	Yes	Yes	Yes	Yes
Observations	11,832	11,846	11,832	11,846
Adjusted R^2	0.111	0.075	0.127	0.075

Panel B. Special-items magnitude

Benchmark:	Equal-Weighted		IBES	
	(1) Low	(2) High	(3) Low	(4) High
SUE^{Acc} (β_1)	2.345***	1.613***	1.940***	1.969***
	(4.36)	(4.49)	(4.34)	(5.48)
Benchmark SUE (β_2)	2.153***	2.078***	2.465***	1.615***
	(4.21)	(6.41)	(6.46)	(5.25)
$SUE^{Acc} \times Post^{Scr}$ (β_3)	0.247	2.035***	0.440	2.583***
	(0.31)	(3.39)	(0.58)	(4.69)
Benchmark $SUE \times Post^{Scr}$ (β_4)	0.799	-0.483	0.659	-1.055**
	(1.14)	(-0.92)	(0.98)	(-2.28)
Premium (Pre-Scrape)	0.192	-0.464	-0.525	0.354
	(0.19)	(-0.74)	(-0.68)	(0.58)
Premium (Post-Scrape)	-0.360	2.054**	-0.743	3.992***
	(-0.28)	(2.48)	(-0.65)	(5.36)
Δ Premium (Post−Pre)	**-0.552**	**2.518****	**-0.219**	**3.638*****
	(-0.39)	**(2.40)**	**(-0.16)**	**(3.92)**
Controls / Firm FE / Year FE	Yes	Yes	Yes	Yes
Observations	11,823	11,828	11,823	11,828
Adjusted R^2	0.089	0.073	0.095	0.071

Notes. This table estimates the accuracy-weighted horse race against the equal-weighted (EW5) and I/B/E/S benchmarks within two median splits: **Panel A** on *Cross-FDP SUE Dispersion* and **Panel B** on $|SpecItems|\,(\%)$ (not the asset-scaled control of the same name). Within each panel, columns (1)–(2) use SUE^{EW5} and columns (3)–(4) use SUE^{IBES}. The post-scraping indicator $Post^{Scr}$ replaces the post-ChatGPT $Post$ of the other tables. The Accuracy Premium is the loading on SUE^{Acc} minus the loading on the benchmark SUE ($\beta_1 - \beta_2$ pre-scrape; $(\beta_1 + \beta_3) - (\beta_2 + \beta_4)$ post-scrape). The bolded Δ row reports the post-minus-pre change, $\beta_3 - \beta_4$. The dependent variable is the market-adjusted $CAR_{[-1,+1]}$. All columns include the firm and event controls, firm and year fixed effects (FE), and firm-clustered standard errors. Each panel's sample comprises the firm-quarters with non-missing values for all regression variables. Continuous variables are winsorized at (1, 99). t-statistics in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Variable definitions are in Appendix A.

TABLE 10. LLM Visibility and the Accuracy Premium by GPT-Capability Regime

Panel A. Equal-weighted five-FDP benchmark

Regime:	(1) R0 Pre-ChatGPT	(2) R1 text-only	(3) R2 browsing	(4) R3 native search
SUE^{Acc} (β_1)	1.361**	3.503***	3.077***	3.055***
	(2.57)	(5.05)	(2.95)	(4.18)
SUE^{EW5} (β_2)	2.252***	2.185***	1.885***	2.002***
	(4.09)	(3.61)	(3.65)	(3.21)
$SUE^{Acc} \times High\,Visibility$ (β_3)	0.020	-2.274	1.550*	1.815*
	(0.05)	(-1.49)	(1.79)	(1.93)
$SUE^{EW5} \times High\,Visibility$ (β_4)	0.153	0.889	-1.133**	-1.269
	(0.28)	(0.79)	(-2.11)	(-1.51)
Accuracy premium = loading on SUE^{Acc} − loading on benchmark *SUE*:				
Premium (Low Vis)	-0.892	1.317	1.192	1.053
	(-0.90)	(1.12)	(0.89)	(0.83)
Premium (High Vis)	-1.025	-1.845	3.875***	4.136***
	(-1.57)	(-0.94)	(4.60)	(4.16)
Premium (High − Low)	**-0.133**	**-3.163**	**2.683****	**3.083***
	(-0.15)	**(-1.22)**	**(2.12)**	**(1.85)**
Controls / Firm FE / Year-qtr FE	Yes	Yes	Yes	Yes
Observations	9,399	3,705	5,529	5,084
Adjusted R^2	0.086	0.163	0.165	0.143

Panel B. I/B/E/S benchmark

Regime:	(1) R0 Pre-ChatGPT	(2) R1 text-only	(3) R2 browsing	(4) R3 native search
SUE^{Acc} (β_1)	1.607***	3.138***	2.737***	4.029***
	(4.07)	(3.15)	(6.05)	(8.81)
SUE^{IBES} (β_2)	1.859***	2.424***	2.403**	0.974**
	(3.60)	(2.96)	(2.28)	(2.13)
$SUE^{Acc} \times High\,Visibility$ (β_3)	-0.382	-2.353*	1.667**	2.376***
	(-1.03)	(-1.74)	(2.44)	(3.68)
$SUE^{IBES} \times High\,Visibility$ (β_4)	0.669	1.362	-1.380	-1.856***
	(0.97)	(1.45)	(-1.45)	(-3.06)
Accuracy premium = loading on SUE^{Acc} − loading on benchmark *SUE*:				
Premium (Low Vis)	-0.252	0.714	0.335	3.054***
	(-0.31)	(0.41)	(0.26)	(4.00)
Premium (High Vis)	-1.303**	-3.001**	3.382***	7.286***
	(-2.06)	(-2.12)	(6.17)	(10.56)
Premium (High − Low)	**-1.052**	**-3.715***	**3.047****	**4.232*****
	(-1.07)	**(-1.72)**	**(2.05)**	**(3.93)**
Controls / Firm FE / Year-qtr FE	Yes	Yes	Yes	Yes
Observations	9,399	3,705	5,529	5,084
Adjusted R^2	0.088	0.175	0.169	0.142

Notes. This table estimates, separately within each GPT-capability regime (R0–R3; Appendix A), the horse race of Equation (9). **Panel A** uses the equal-weighted five-FDP benchmark SUE^{EW5} and **Panel B** uses the I/B/E/S benchmark. The Accuracy Premium (the loading on SUE^{Acc} minus the loading on the benchmark) is $\beta_1 - \beta_2$ for low-visibility firms and $(\beta_1 + \beta_3) - (\beta_2 + \beta_4)$ for high-visibility firms. The bolded Premium (High − Low) row is the triple difference $\beta_3 - \beta_4$. The dependent variable is the market-adjusted $CAR_{[-1,+1]}$. All columns include the firm and event controls and firm and year-quarter fixed effects (FE), with standard errors two-way clustered by firm and calendar year. Each regime sample comprises the announcements with non-missing values for all regression variables. Continuous variables are winsorized at (1, 99). t-statistics in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Variable definitions are in Appendix A.

Online Appendix

Synthesizing the Consensus: Generative AI and the Pricing of Multi-Provider Earnings Forecasts

Leonard Yang Liu

This Online Appendix contains supplementary material for the paper:

- **Online Appendix A** reports the outage-based test of the GenAI channel (Table OA3);
- **Online Appendix B** reports two worked LLM query examples (IBM and Microsoft);
- **Online Appendix C** works through the construction of the accuracy-weighted SUE for a hypothetical firm-quarter;
- **Online Appendix D** presents a pricing model under which the accuracy premium identifies the market's reliance on accuracy-weighted synthesis, and shows that the supply-side improvement in forecast accuracy (common to the five FDPs, with the cross-provider structure stable) cannot move it (Table OA2);
- **Figure OA1** compares the accuracy of the three consensus measures before and after ChatGPT;
- **Tables OA1–OA4** report how often the five FDPs' street actuals disagree, the supply-side provider-dynamics diagnostics, the outage test, and the distinctness of LLM visibility from classical prominence proxies.

Online Appendix A: Probing the GenAI Channel with ChatGPT Outages

Sections 4 through 6 establish that the post-ChatGPT market prices an accuracy-weighted synthesis of consensus forecasts and that the resulting accuracy premium concentrates in firms whose forecasts GenAI actually retrieves, and emerges only once it can retrieve them. This section probes the GenAI channel by exploiting brief interruptions to its availability. On announcement days that follow a major ChatGPT outage, the accuracy premium declines sharply against I/B/E/S but not against the equal-weighted benchmark, the asymmetry a GenAI-mediated channel predicts. The number of treated announcements is small, so the test is suggestive instead of conclusive.

Brief interruptions to GenAI availability provide a second way to test the channel. Following Cheng et al. (2025) and Liu and Subasi (2026), I exploit major ChatGPT outages as short-duration shocks to GenAI availability. On announcement dates immediately following a major outage, investors preparing for the announcement have just lost the integration capability that GenAI provides and should revert to the one benchmark that requires no multi-FDP synthesis, the single-FDP I/B/E/S consensus. The prediction is therefore an asymmetric outage effect across the two horse-race benchmarks. Against I/B/E/S, the accuracy premium should weaken as investors substitute toward that fallback. Against the equal-weighted benchmark, it should be unchanged, because the accuracy-weighted and equal-weighted aggregates both require multi-FDP synthesis and become equally infeasible when GenAI is down.

I collect ChatGPT outage incidents from OpenAI's public status-page history and define a major outage as any incident lasting at least 240 minutes, which yields four incidents across five unique calendar days through 2024. The treatment indicator $Outage_{i,q}$ equals one if firm i's quarter-q announcement falls on the calendar day immediately following a major outage and zero otherwise. I re-estimate the horse-race specification of Equation (8) on the post-ChatGPT subsample with $Outage_{i,q}$ in place of $Post_q$, separately for the equal-weighted and I/B/E/S benchmarks, so that β_3 and β_4 become the outage-day shifts in the two slopes. Because the shock is a single-day treatment, the dependent variable is the day-0 market-adjusted abnormal return instead of the three-day CAR.

The estimates in Table OA3 match the predicted asymmetry. Against the equal-weighted

benchmark in column (1), the outage-day change in the accuracy premium is positive but statistically insignificant (3.787, $t = 1.37$), showing no sign of the decline that a reversion toward equal weighting would produce. Against the I/B/E/S benchmark in column (2), the change is -3.173 ($p < 0.05$), combining a marginally significant decline in the accuracy-weighted slope ($\beta_3 =$ -1.685, $p < 0.10$) with a rise in the I/B/E/S slope ($\beta_4 = 1.488$, $p < 0.05$). The estimates are consistent with investors reweighting toward I/B/E/S immediately after an outage, which is what a shock to multi-FDP synthesis predicts, since such a shock should leave the contrast between two synthesized aggregates untouched and move only the contrast with the single-FDP fallback.

The sample constrains what this test can establish. Five treated calendar days contribute only 84 treated announcements, and announcements in 2025 enter as untreated because the incident log ends in 2024, which can only attenuate the estimated effect. The level estimates are therefore imprecise, and identification rests on the sign pattern across the two benchmarks instead of on magnitudes. Outages are also not exogenous to everything that moves announcement returns, since they may coincide with broader disruptions to online services. I therefore read the outage evidence as suggestive corroboration of the regime and visibility results instead of as an independent test of the channel.

Online Appendix B: Worked Examples of GenAI Aggregation of Analyst Consensus

This appendix shows the prompt and ChatGPT (GPT-4o) responses for two illustrative large-cap firms, IBM and Microsoft.

Responses were obtained through the DataForSEO LLM Responses API and represent the unedited output of the model. Each firm subsection ends with a summary table mapping the cited aggregators to their upstream forecast data provider (FDP).

B.0 Prompt

The same prompt is issued for each sample firm, with [TICKER] replaced by the firm's ticker symbol.

What is the analyst consensus EPS forecast for [TICKER] for the next quarter? Please report the figures compiled by multiple third-party aggregators and cite each source.

B.1 International Business Machines Corporation (IBM)

ChatGPT (GPT-4o) response:

> Here's the latest analyst consensus EPS forecast for IBM's next quarter, as compiled from several third-party aggregators.
>
> - **MarketBeat (via Fiscal.ai)**: Q2 2026 average EPS estimate is **$2.96**, based on 2 estimates (low $2.92, high $2.99) (**marketbeat.com**).
> - **TipRanks**: next quarter's earnings estimate is **$3.02**, with a range from $2.92 to $3.09 (**tipranks.com**).
> - **Barchart**: next quarter average earnings estimate is **$2.96**, based on 6 estimates (low $2.73, high $3.50) (**barchart.com**).
> - **Yahoo Finance**: Jun 2026 average EPS estimate is **$3.04**, based on 14 analysts (low $2.80, high $3.21) (**finance.yahoo.com**).

Summary table.

Source	Forecast	Upstream FDP
MarketBeat	$2.96	Independent aggregator
TipRanks	$3.02	Independent aggregator
Barchart	$2.96	Independent aggregator
Yahoo Finance	$3.04	S&P Capital IQ

B.2 Microsoft Corporation (MSFT)

ChatGPT (GPT-4o) response:

Here are the analyst consensus EPS forecasts for Microsoft (MSFT) for the **next quarter**, as compiled by multiple third-party aggregators. Each figure is accompanied by its source.

- **MarketBeat**: for Q4 FY2026, the consensus EPS estimate is **$4.04**, based on 5 analyst forecasts (**marketbeat.com**).
- **Zacks Investment Research**: consensus estimate for the next quarter (June 2026) is **$4.13**, derived from 13 analyst estimates (**zacks.com**).
- **TipRanks**: next EPS estimate is **$4.24** (**tipranks.com**).
- **ChartMill**: consensus EPS estimate of **$4.33** for the next earnings (Q4 FY2026), based on 39 analysts (**chartmill.com**).
- **Seeking Alpha**: an earnings preview lists the consensus EPS at $4.06 for Q3 FY2026 (already reported), so not applicable for the upcoming quarter (**seekingalpha.com**).
- **Yahoo Finance**: for the period "Next qtr. (Jun 2026)", average EPS estimate is **$4.26**, based on 29 analysts (low $3.74, high $5.54) (**finance.yahoo.com**).
- **Investing.com**: I was unable to locate a clear next-quarter EPS consensus figure (**investing.com**).

Summary table.

Source	Forecast	Upstream FDP
MarketBeat	$4.04	Independent aggregator
Zacks	$4.13	Zacks
TipRanks	$4.24	Independent aggregator
Yahoo Finance	$4.26	S&P Capital IQ
ChartMill	$4.33	Independent aggregator

Online Appendix C: Numerical Illustration of the Accuracy-Weighted SUE

This appendix illustrates how the accuracy-weighted SUE is constructed for a single firm-quarter. Suppose firm i in quarter q has the following backward-looking accuracy histories ($acc^g_{i,q} = -MAFE^g_{i,q}$ over the prior four quarters, in price-scaled units) across the five FDPs (the z-scores use the sample standard deviation across the five providers), alongside their current-quarter standardized surprises $SUE^g_{i,q}$ (in percent of price).

Panel A. Weight construction from accuracy history.

Provider g	$acc^g_{i,q}$	$acc\_z^g_{i,q}$	$\exp(acc\_z^g_{i,q})$	$w^g_{i,q}$
IBES	-0.0010	1.26	3.54	0.49
FactSet	-0.0020	0.63	1.88	0.26
Capital IQ	-0.0030	0.00	1.00	0.14
Zacks	-0.0040	-0.63	0.53	0.07
Bloomberg	-0.0050	-1.26	0.28	0.04
Sum			7.24	1.00

I/B/E/S has the smallest absolute forecast error (largest acc^g) and therefore the highest within-row z-score, which the softmax transform maps to the largest weight (0.49). Bloomberg has the largest absolute forecast error and receives the smallest weight (0.04). The weights sum to one and reduce to the uniform 0.20 vector when the five accuracy scores are equal.

Panel B. Composite construction.

Provider g	$SUE^g_{i,q}$	$w^g_{i,q} \times SUE^g_{i,q}$	$0.20 \times SUE^g_{i,q}$
IBES	1.20	0.59	0.24
FactSet	0.80	0.21	0.16
Capital IQ	1.00	0.14	0.20
Zacks	1.50	0.11	0.30
Bloomberg	-0.50	-0.02	-0.10
Total		$SUE^{Acc}_{i,q} = 1.02$	$SUE^{EW5}_{i,q} = 0.80$

Entries are rounded to two decimals while the totals are computed from unrounded weights, so the displayed cells need not sum exactly to the reported totals. The accuracy-weighted composite $SUE^{Acc}_{i,q} = 1.02$ exceeds the equal-weighted composite $SUE^{EW5}_{i,q} = 0.80$ because the two most-accurate providers in this row (I/B/E/S and FactSet) reported positive surprises, while the least-accurate provider (Bloomberg) was the lone outlier on the downside. The accuracy tilt $SUE^{Acc}_{i,q} - SUE^{EW5}_{i,q} = 0.22$ captures the accuracy-driven correction. In the horse-race regression of Table 7, this tilt is the source of identification between the two composites: when the tilt is informative about announcement-window returns beyond the equal-weighted aggregate, the market is loading on provider-specific accuracy instead of treating all five FDPs symmetrically.

Online Appendix D: Supply-Side Accuracy Improvement and the Stability of Cross-Provider Dynamics

The concern. Consensus forecasts become more accurate over 2021–2025, whether from secular improvement in analyst research, from GenAI assistance reaching the analysts whose forecasts feed all five FDPs, or from shifting sample composition. If the accuracy premium merely tracked this supply-side trend, it would say nothing about how the market synthesizes providers. This appendix gives a minimal pricing model under which the premium is a demand-side object, and shows that the cross-provider structure on which the weights operate is stable throughout.

What the premium measures. Write each provider's surprise as $SUE^g_{i,q} = N_{i,q} + \eta^g_{i,q}$, where $N_{i,q}$ is the value-relevant earnings news of announcement (i,q) and the measurement components η^g may have any joint distribution, any provider-specific variances, and any drift in those variances over time. The supply side, forecast accuracy and its evolution, is captured entirely by $\{\eta^g\}$. Two properties then separate demand from supply.

First, the construction is supply-invariant. Because Equation (3) standardizes accuracy within firm-quarter before the softmax, replacing $MAFE^g_{i,q}$ by $a_{i,q}MAFE^g_{i,q} + b_{i,q}$ ($a_{i,q} > 0$) uniformly across the five providers leaves every weight $w^g_{i,q}$ unchanged. The weights respond only to the relative accuracy pattern, so any accuracy change *common* to the five FDPs, including the economy-wide improvement documented below, leaves them untouched.

Second, let the market's pre-announcement expectation of street earnings be a convex combination of the two candidate consensus benchmarks,

$$E^m_{i,q} \;=\; \omega\, c^{Acc}_{i,q} \;+\; (1-\omega)\, c^X_{i,q},$$

where $c^{Acc}_{i,q} = \sum_g w^g_{i,q}\, consensus^g_{i,q}$ is the accuracy-weighted consensus level, $c^X_{i,q}$ is the comparison benchmark's, $X \in \{EW5, IBES\}$, and ω is the market's relative reliance on the accuracy-weighted synthesis, the direct analog of the expectation mixture in Walther (1997). Each provider's consensus pairs with its own street actual (Section 3.4), so the surprise inherits the same mixture and the

return equation is

$$CAR_{i,q} \;=\; \theta\left[\omega\, SUE^{Acc}_{i,q} \;+\; (1-\omega)\, SUE^{X}_{i,q}\right] \;+\; \varepsilon_{i,q}, \tag{OA1}$$

where θ is the return response per unit of perceived surprise (Holthausen and Verrecchia, 1988; Kim and Verrecchia, 1991) and ε collects pricing components outside the span of the two benchmarks. Equation (OA1) is a pure reparameterization of the horse-race slopes applied period by period ($\theta = \beta_1 + \beta_2$ and $\omega = \beta_1/(\beta_1 + \beta_2)$ in the pre-period, with $\beta_1 + \beta_3$ and $\beta_2 + \beta_4$ taking their place in the post), so it imposes no restriction. Because the priced objects are the observable benchmarks themselves, the projection recovers $(\theta\omega,\, \theta(1-\omega))$ *exactly*, for any joint distribution of $(N, \eta^1, \ldots, \eta^5)$ and any drift in it. The accuracy premium in each period is therefore

$$\beta_{Acc} - \beta_X \;=\; \theta\,(2\omega - 1),$$

linear in ω given θ, with both parameters describing the market's pricing rule instead of the forecast supply. Neither the level of forecast accuracy nor its improvement appears in it, so its movement across regimes identifies movement in the market's information use. In these units, the pre- and post-period slopes of Table 7, Panel A, put ω at 0.37 before ChatGPT and 0.66 after.

The improvement is common, not convergent. The level improvement is real. With firm fixed effects, the log realized error is 10.1 log points lower under browsing ($t = 4.0$) and 10.6 lower under native search ($t = 3.8$), with no change in the text-only regime (untabulated). What does not move is the cross-provider structure. Table OA2 reports the share of firm-quarters in which each provider is the *ex ante* most accurate, by regime. Across the three post-ChatGPT regimes, the period over which the premium emerges, the distribution shows no significant change ($\chi^2(8) = 10.8$, $p = 0.21$). Extending the joint test to all four regimes rejects ($\chi^2(12) = 23.7$, $p = 0.022$; untabulated), driven by a one-time Bloomberg step down at the ChatGPT boundary itself ($0.172 \to 0.145$; provider-level $p = 0.004$, all other providers $p \geq 0.06$; untabulated). That step predates the premium's emergence at browsing and is precisely the kind of reordering the trailing-MAFE weights are designed to track. The weight distribution itself is nearly unchanged, its concentration $\sum_g (w^g)^2$ shifting only from 0.313 to 0.300 (untabulated). Relative differentiation if anything widens, the coefficient of variation

of $MAFE^g$ across providers rising from 0.514 to 0.612 ($t = 7.0$; untabulated). The improvement is common to the providers, not a convergence among them.

The relative structure retains its predictive validity. The weights are valid if the *ex ante* accuracy ordering continues to forecast the realized one. The within-firm-quarter rank correlation between the trailing-MAFE ordering and the realized error ordering of the announcement quarter is 0.217 in R0 and 0.215 in R3, with the R3-minus-R0 difference statistically indistinguishable from zero ($t = 0.0$; untabulated). The probability that the *ex ante* most accurate provider is also the realized most accurate is 0.30 in R0 and 0.28 in R3, against a chance rate of 0.20 (untabulated). A transient dip in the text-only regime (-0.034, $t = 2.7$; untabulated) has recovered by the regimes in which the premium appears.

The supply-side alternative, rejected on its own implications. The alternative to Equation (OA1) is that the market has always priced the latent news itself, $CAR_{i,q} = \theta N_{i,q} + \varepsilon_{i,q}$, so that the horse-race coefficients reflect nothing but proxy quality, the proxy-evaluation interpretation of unexpected-earnings horse races (Brown et al., 1987). Projection algebra then makes the premium a function of the benchmarks' measurement variances, $\beta_{Acc} - \beta_X = \theta v(\sigma_X^2 - \sigma_{Acc}^2)/D$, where $v = \mathrm{Var}(N)$ and $D > 0$ collects second moments.[42] This mechanical premium has three testable implications, and the data support none of them.

First, its sign. The composite is empirically the less noisy proxy, its median realized error in R0 being 0.126 percent of price against 0.144 for I/B/E/S and 0.146 for the equal-weighted composite (untabulated; Figure OA1 shows the composite ahead of both heuristics quarter by quarter), so the mechanical premium should have been *positive* before ChatGPT. The estimated pre-period premium is instead negative in point estimate and statistically indistinguishable from zero (-0.75 against I/B/E/S, -0.95 against EW5; Tables 7 and 8). Second, its movement. The mechanical premium can rise only if the composite's noise advantage $\sigma_X^2 - \sigma_{Acc}^2$ widens. Measured directly

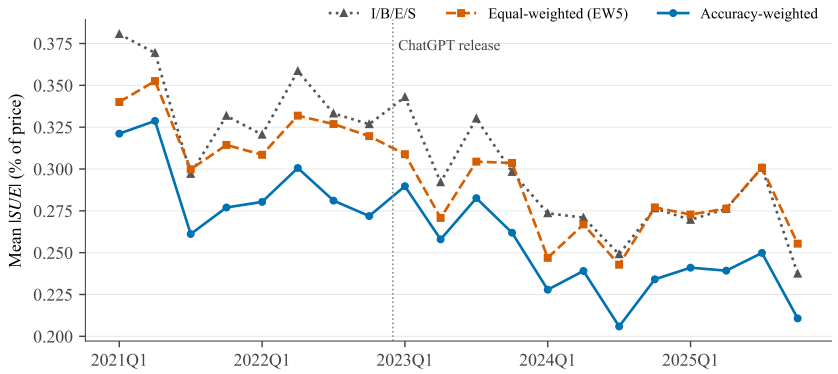
[42] With $SUE^{Acc} = N + \eta_{Acc}$ and $SUE^X = N + \eta_X$, the alternative maintains $E[\eta] = 0$, $\mathrm{Cov}(N, \eta_{Acc}) = \mathrm{Cov}(N, \eta_X) = 0$, and each benchmark's loading on N normalized to one, as both are convex combinations of the same five per-share, price-scaled provider surprises. With $CAR = \theta N + \varepsilon$, the population regression of CAR on the two benchmarks has coefficients $\theta v(\sigma_X^2 - \sigma_{AX},\ \sigma_{Acc}^2 - \sigma_{AX})/D$, where $\sigma_{AX} = \mathrm{Cov}(\eta_{Acc}, \eta_X)$ is nonzero by construction and $D = v(\sigma_{Acc}^2 + \sigma_X^2 - 2\sigma_{AX}) + \sigma_{Acc}^2\sigma_X^2 - \sigma_{AX}^2$, positive whenever the benchmarks are not collinear. The error covariance affects only D, and the premium's numerator is $\theta v(\sigma_X^2 - \sigma_{Acc}^2)$ regardless.

it does not: the median within-firm-quarter gap in realized error between the composite and each benchmark is flat across regimes, and its winsorized firm-fixed-effects shifts are weakly *positive*, at 0.011 against I/B/E/S in R3 ($t = 2.3$) and at most 0.013 against EW5 (untabulated). The advantage narrows slightly while the premium rises, the opposite of the mechanical prediction. Third, its inputs. If the overall improvement had sharpened the weights, the *ex ante* accuracy ordering would have to forecast the realized ordering better than before, and its rank validity is instead statistically indistinguishable before ChatGPT and under native search (0.217 versus 0.215). The residual route, forecasts improving and the market responding by raising ω, is itself a demand response instead of a mechanical artifact, and even it is mistimed: the accuracy improvement arrives with browsing and is flat between R2 and R3 (-10.1 versus -10.6 log points), whereas the premium's decisive step against I/B/E/S comes from R2 to R3.

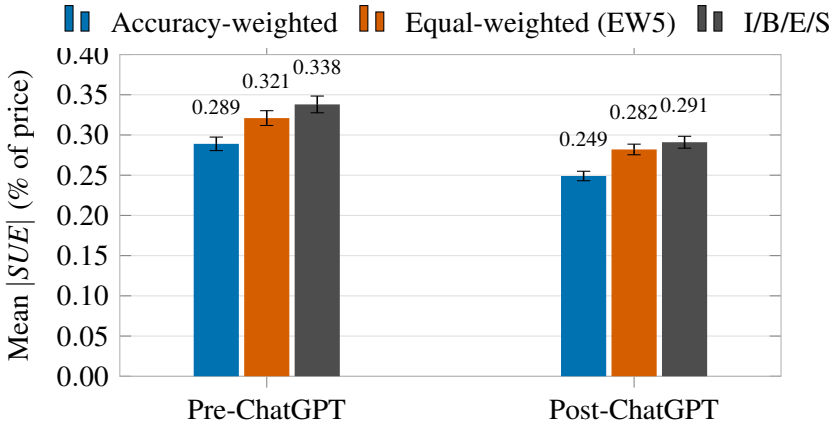
Summary. In the model of Equation (OA1) the accuracy premium equals $\theta(2\omega - 1)$, so the supply side can reach it only through ω, the market's choice. The exhibits confirm that the basis of the measure is stable, and the supply-side alternative fails each of its own predictions. What changed over 2021–2025 is ω, the market's use of the cross-provider structure, not the structure itself.

FIGURE OA1. Mean Magnitude of *SUE*: Accuracy-Weighted vs. Heuristic Benchmarks

(a) Quarterly mean *SUE* by consensus method



(b) Mean *SUE* by consensus method



Notes. This figure compares the accuracy of the accuracy-weighted multi-vendor consensus SUE^{Acc} against the equal-weighted five-FDP consensus SUE^{EW5} and the I/B/E/S consensus SUE^{IBES} across the five-FDP sample (FactSet, Bloomberg, Zacks, Capital IQ, I/B/E/S) of S&P 1500 firm-quarters with announcement dates in 2021Q1–2025Q4. $|SUE|$ is the absolute standardized unexpected earnings, scaled by the closing share price on the last trading day before the announcement and expressed in percent, winsorized at the (1, 99) level. **Panel (a)** plots the quarterly mean $|SUE|$ for each of the three methods over 20 announcement quarters. The vertical reference line marks the public release of ChatGPT (November 30, 2022). **Panel (b)** shows the pooled mean $|SUE|$ within each of the pre- and post-ChatGPT periods, with 95% confidence intervals from the cell-level standard error of the mean. The *y*-axis starts at zero so that bar height is proportional to the mean.

TABLE OA1. Methodology Differences and Disagreement in Street-Actual Earnings Across FDPs

	I/B/E/S	FactSet	Bloomberg	Zacks
FactSet	22.27%			
Bloomberg	88.50%	88.66%		
Zacks	33.63%	38.30%	90.84%	
Capital IQ	16.31%	21.06%	88.95%	32.41%

Notes. This table reports, for each pair of forecast data providers, the percentage of the 23,833 sample firm-quarters in which the two providers publish different street actuals for the same firm-quarter, where "different" means the two providers' reported actuals are not identical. All five actuals are non-missing for every sample firm-quarter. Because both providers observe the same GAAP filing, any difference in their reported actuals reflects the providers' own construction rules for street earnings, the treatment of unexpected charges and gains, stock-based compensation, and other adjustments (Bochkay et al., 2022), rather than differences in the underlying analyst pool.

TABLE OA2. Provider Accuracy Shares Across GPT-Capability Regimes

	R0 Pre-ChatGPT	R1 Text-only	R2 Browsing	R3 Native search
Share of firm-quarters in which the provider is the *ex ante* most accurate:				
I/B/E/S	0.169	0.184	0.176	0.197
FactSet	0.209	0.198	0.224	0.217
Bloomberg	0.172	0.145	0.144	0.136
Zacks	0.295	0.309	0.291	0.286
Capital IQ	0.161	0.169	0.170	0.168
Firm-quarters	4,167	3,287	5,077	4,631
Joint test, all providers: distribution equal across the post-ChatGPT regimes, $\chi^2(8) = 10.8$, $p = 0.214$.				

Notes. This table reports, by GPT-capability regime of the announcement date, the share of firm-quarters in which each provider's consensus is the *ex ante* most accurate (lowest trailing four-quarter $MAFE^g$), with a firm-quarter in which providers tie for the lowest $MAFE^g$ credited to each tied provider, so that a column can sum to slightly above one, on the firm-quarters with complete trailing-error windows for all five providers. The joint Wald test stacks the five provider-share equations, with standard errors clustered by firm.

TABLE OA3. ChatGPT Outages and the Accuracy Premium

Dep. Var.: day-0 AR_0 (market-adj., %)	Benchmark: Equal-Weighted (1)	Benchmark: IBES (2)
SUE^{Acc} (β_1)	1.631***	1.599***
	(6.93)	(7.78)
$SUE^{Acc} \times Outage$ (β_3)	1.178	-1.685*
	(0.80)	(-1.85)
Benchmark SUE (β_2)	0.626***	0.654***
	(3.13)	(4.10)
Benchmark $SUE \times Outage$ (β_4)	-2.608*	1.488**
	(-1.79)	(2.27)
Premium (non-outage)	1.005**	0.945***
	(2.48)	(2.88)
Premium (outage)	4.792*	-2.228*
	(1.77)	(-1.65)
Δ Premium (Outage−Non)	**3.787**	**-3.173****
	(1.37)	**(-2.29)**
Controls / Firm FE / Year FE	Yes	Yes
Observations	14,402	14,402
Adjusted R^2	0.051	0.052

Notes. This table tests whether the accuracy premium collapses on earnings announcements that follow a major ChatGPT outage. The sample is post-ChatGPT firm-quarters, of which 84 are outage firm-quarters ($Outage_{i,q} = 1$). The dependent variable is the day-0 market-adjusted abnormal return. Both columns include firm and year fixed effects (FE), the firm and event controls, and firm-clustered standard errors. The Accuracy Premium is the loading on SUE^{Acc} minus the loading on the benchmark SUE. The bolded row reports its outage-minus-non-outage difference. t-statistics in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Variable definitions are in Appendix A.

TABLE OA4. LLM Visibility vs. Classical Prominence Proxies

Panel A. Overlap of high-visibility classifications

	LLM Visibility	Firm Size	Analyst Coverage	Inst Ownership
LLM Visibility	1.000	0.502	0.497	0.471
Firm Size	0.510	1.000	0.796	0.407
Analyst Coverage	0.506	0.798	1.000	0.445
Inst Ownership	0.479	0.407	0.445	1.000

Panel B. Accuracy-premium concentration: equal-weighted benchmark

Visibility measure	Univariate ΔPremium		Horse-race
	High group	Low group	triple diff.
LLM Visibility	**3.263*****	**0.300**	**3.172***
	(2.59)	**(0.25)**	**(1.80)**
	[11,219]	[11,825]	
Firm Size	0.723	2.447**	0.111
	(0.55)	(2.18)	(0.05)
	[11,901]	[11,143]	
Analyst Coverage	0.109	3.076**	-3.161
	(0.08)	(2.57)	(-1.38)
	[11,861]	[11,183]	
Inst Ownership	3.224**	0.646	3.222*
	(2.51)	(0.58)	(1.86)
	[11,593]	[11,451]	
Controls / Firm FE / Year FE	Yes	Yes	Yes
Observations	23,044		23,044

Panel C. Accuracy-premium concentration: I/B/E/S benchmark

Visibility measure	Univariate ΔPremium		Horse-race
	High group	Low group	triple diff.
LLM Visibility	**5.743*****	**-0.125**	**5.841*****
	(5.67)	**(-0.12)**	**(4.03)**
	[11,219]	[11,825]	
Firm Size	2.191	2.462***	2.010
	(1.28)	(2.76)	(0.82)
	[11,901]	[11,143]	
Analyst Coverage	0.441	3.198***	-4.099*
	(0.29)	(3.59)	(-1.77)
	[11,861]	[11,183]	
Inst Ownership	2.564**	2.418**	0.625
	(2.46)	(2.08)	(0.41)
	[11,593]	[11,451]	
Controls / Firm FE / Year FE	Yes	Yes	Yes
Observations	23,044		23,044

Notes. Each measure is split at its cross-firm median: LLM visibility, firm size (log market equity), analyst

coverage (log average analyst count), and institutional ownership. The visibility score is discrete, so firms tied at the median join the Low group, which is why the split is not exactly even. **Panel A**: cell (i, j) is the proportion of firms in measure i's High group that measure j also classifies as High (diagonal $= 1$). **Panels B and C**: ΔPremium is the post-minus-pre change in the accuracy premium $\beta(SUE^{Acc}) - \beta(SUE^{X})$ against benchmark $X =$ the equal-weighted five-FDP consensus (Panel B) or I/B/E/S (Panel C), estimated within each measure's High and Low group. The horse-race triple $(SUE^{Acc} - SUE^{X}) \times Post^{\text{Scr}} \times High_m$ is from one pooled regression including the triple for every measure jointly. Throughout, $Post^{\text{Scr}}$ is the post-scraping indicator, as in Table 9. The dependent variable throughout Panels B and C is the market-adjusted $CAR_{[-1,+1]}$. All specifications include the firm and event controls, firm and year fixed effects (FE), and firm-clustered standard errors. Continuous variables are winsorized at $(1, 99)$. t-statistics in parentheses and group observation counts in brackets (each measure's High and Low groups partition the same 23,044 firm-quarters, the announcements with non-missing values for all regression variables); $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Variable definitions are in Appendix A.