When Does Generative AI Level the Playing Field? Evidence from Crowdsourced Earnings Forecasts*

Leonard Yang Liu† Musa Subasi‡

July 2026

Abstract

Whether generative AI (GenAI) levels the playing field or reinforces existing advantages depends on which constraint it relaxes. We argue that GenAI substitutes for institutionalized analytical resources while complementing human expertise. We test this distinction using crowdsourced earnings forecasts from Estimize. We find that non-professional contributors persistently underperform professionals in forecast accuracy before the release of ChatGPT; afterward, the gap is no longer present. Inferring contributors' GenAI reliance from forecasting behavior during ChatGPT service outages, we find that GPT-reliant non-professionals produce more accurate and independent earnings forecasts, while professionals are largely unaffected, consistent with GenAI relaxing institutional resource constraints. Among non-professionals, where analytical resources are relatively homogeneous, only experienced contributors improve after adopting GenAI; inexperienced adopters show no gains and may herd more. Our cross-sectional tests suggest that non-professionals benefit more when disclosures are difficult to process and, separately, when information asymmetry is low. Furthermore, stock returns around earnings announcements respond more strongly to non-professional forecast surprises in the post-GPT period. Collectively, our findings reconcile conflicting evidence on GenAI's distributional effects: the technology democratizes information production by substituting for institutionalized resources while amplifying the returns to expertise.

Keywords: Generative AI; Forecast accuracy; Information processing; Financial intermediary; Capital markets; Expertise

JEL classification: G14; G17; M41; O33; D83; J24

*We thank Michael Kimbrough, Mark Zakota, Inder Khurana, Mahfuz Chy, Hoyoun Kyung, Jeffery Piao, and workshop participants at the University of Maryland and the University of Missouri for helpful comments and suggestions. All errors are our own.

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

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

1 Introduction

Generative AI (GenAI) has spread with remarkable speed. ChatGPT reached 100 million monthly active users within two months of its late-2022 launch, making it the fastest-growing consumer application on record, and surpassed 300 million weekly active users by the end of 2024 (Bick et al., 2026). Its broad adoption reflects a key feature: unlike earlier forms of AI that required programming expertise, structured data pipelines, and proprietary computational infrastructure, large language models (LLMs) can be used through natural-language prompts. As a result, users can perform sophisticated information-processing tasks at low cost and with minimal technical training. This feature is especially important in financial settings, where decision-relevant information is often buried in dense, unstructured disclosures rather than presented in readily usable numerical form.[1] Consistent with this potential, survey evidence indicates that nearly half of AI chatbot users have relied on these tools in at least one financial decision.[2]

Whether this diffusion levels the playing field in financial information production is unsettled, and the existing evidence points in opposite directions. In general labor-market settings, GenAI compresses performance gaps, with the largest gains accruing to lower performers (Brynjolfsson et al., 2025; Noy and Zhang, 2023). In capital-market and entrepreneurial settings, by contrast, the technology appears to reward sophistication and experience, and AI assistance can even leave the weakest users worse off (Blankespoor et al., 2026; Cheng et al., 2025; Bradshaw et al., 2026; Cao et al., 2024; Otis et al., 2026). These findings are not contradictory: they reflect two distinct margins along which GenAI operates. GenAI substitutes for institutional analytical resources (e.g., proprietary data and software, research infrastructure, computational resources, and structured workflows concentrated among professionals) because it supplies routine information-processing capacity at low cost. GenAI complements expertise because extracting value from the technology requires recognizing valuable applications, steering the model, and verifying its outputs. Whether GenAI levels or widens performance gaps therefore depends on which constraint binds. We develop

[1] See, e.g., Futurism (2025), https://futurism.com/chatgpt-stocks-100-dollars; SF Standard (2025), https://sfstandard.com/2025/10/31/ai-told-stocks-buy-results-wild/.

[2] LendingTree, *AI Chatbot Users Survey* (2025), https://www.lendingtree.com/credit-cards/study/ai-chatbot-users/.

and test this distinction using crowdsourced earnings forecasts from Estimize.

Estimize provides an ideal setting for several reasons. First, the forecast accuracy of Estimize contributors is directly observable and objectively scored against realized earnings, providing a clear measure of financial information-processing quality. Second, contributors submit forecasts independently and are largely free from the investment-banking and brokerage conflicts that shape sell-side output (Jame et al., 2022). Third, the platform's blind design prevents users from observing others' forecasts at the time of submission (Da and Huang, 2020). Finally, upon registration, contributors self-identify as either financial professionals or non-professionals, providing a predetermined source of heterogeneity; Section 2.3 describes the platform in detail. This structure separates the two margins: the professional/non-professional partition varies institutional analytical resources, while accumulated forecasting experience within the non-professional pool varies expertise among contributors with similarly limited resources.

A central empirical challenge is that individual GenAI adoption by Estimize contributors is not directly observable; we address it with a ChatGPT outage-based classification that infers reliance from revealed forecasting behavior. Specifically, using 32 documented ChatGPT service outages during 2023–2024, we estimate each contributor's probability of submitting a forecast in a given time window and identify contributors whose forecasting activity drops disproportionately when ChatGPT becomes unavailable.[3] The procedure classifies 986 of the 3,876 contributors in our accuracy sample as GPT-reliant, with professionals represented at roughly twice the rate of non-professionals.[4] This approach complements recent capital-markets studies that exploit exogenous disruptions to GenAI availability but identifies contributor-level reliance within a forecasting platform (Bertomeu et al., 2025; Cheng et al., 2025).[5]

[3]A limitation of this approach is that it assumes contributors who rely on GenAI for information processing would otherwise have produced forecast submissions during the outage window. To mitigate this concern, we validate the classification by comparing the productivity trajectories of GPT-reliant and non-GPT contributors around the ChatGPT release date; see Section 5.1.2 for details.

[4]Classification requires at least one forecast in the post-GPT period (2023 or later), a condition every contributor in the accuracy sample meets by construction (Table 1, Panel B).

[5]We end the sample in 2024 to guard against the proliferation of close ChatGPT substitutes, which would blur the attribution of outage responses to ChatGPT. The falsification test in Section 5.3.3 confirms the necessity of this choice: the outage sensitivity of forecast quality vanishes once substitutes emerge in early 2025.

Our results bear out this two-margin view. Consistent with prior evidence that GenAI confers real information-processing benefits on market participants (e.g., Bertomeu et al., 2025; Cheng et al., 2025), forecast quality on Estimize deteriorates when ChatGPT is unexpectedly unavailable, confirming that GenAI is a useful input to forecast production. On the resource margin, GenAI levels the field. In the aggregate, non-professionals' long-standing accuracy disadvantage, already narrowing before ChatGPT's release, is statistically zero afterward. At the contributor level, GPT-reliant non-professionals issue more accurate forecasts and revise less frequently toward the prevailing consensus, a proxy for forecast independence, whereas professionals are largely unaffected even though they are classified as GPT-reliant at roughly twice the rate. The framework rationalizes this wedge between adoption and benefit: professionals face lower integration costs and stronger performance incentives, which raise adoption, but already command the processing capacity that GenAI supplies, which lowers its marginal value. Cross-sectional evidence supports the mechanism. The benefits of adoption concentrate among firms whose disclosures are difficult to process and, separately, among firms whose information asymmetry is low, consistent with GenAI relaxing the processing constraint that non-professionals face but not replicating the private information access that professionals retain.[6]

On the expertise margin, GenAI widens the field. Within the non-professional pool, where resources vary little, the accuracy and independence gains concentrate among experienced contributors.[7] Inexperienced adopters show no improvement and may herd more, consistent with users who lack the domain knowledge to evaluate model outputs deferring to them instead. Because resource endowments are similar within this group, the divergence is attributable to expertise: once the resource constraint ceases to bind, GenAI amplifies the returns to human capital rather than replacing it. Finally, consistent with the overall leveling, earnings announcement returns respond more strongly to earnings surprises constructed from non-professional consensus forecasts

[6]Processing difficulty is measured as the average of standardized Gunning Fog, ARI, and SMOG indices from the firm's 8-K filings; information asymmetry is the private-information component of analyst forecast dispersion from the Barron et al. (1998) decomposition, which separates differential information access from fundamental uncertainty (Section 5.2.4). The readability split is significant for revenue forecasts but not for EPS (Table 9).

[7]A contributor is experienced at the time of a forecast if at least 365 days have elapsed since the contributor's first Estimize forecast and the contributor has submitted at least 10 cumulative forecasts (Section 5.2.3).

following ChatGPT's release, suggesting that the improvement in forecast quality is impounded into prices.

Our study sits within a cluster of concurrent work on GenAI in investment research: Bradshaw et al. (2026) detect AI-generated equity research on *Seeking Alpha* and study its informativeness, Blankespoor et al. (2026) document which retail investors adopt GenAI and for which tasks, and Cao et al. (2024) benchmark AI-based analysis against human analysts. Each identifies who adopts the technology or what it produces; none separates the margins along which its use changes human output. Our design targets exactly this separation: the outage-based classification captures contributor-level variation in GPT reliance, within a platform where forecasts are objectively scored against realized earnings and whose structure supplies both the resource partition and the expertise gradient the framework requires. The comparison with Bradshaw et al. (2026) is the sharpest: both settings involve crowdsourced equity research, but text detection is suited to studying the informativeness of AI-generated content, whereas outage-based identification speaks to how GenAI use changes the accuracy and independence of human forecasts, and for whom.

Our contributions to the literature are threefold. First, we extend the literature on analyst forecasts and crowdsourced estimates. A large body of work shows that forecast quality is shaped by analyst resources, experience, cognitive constraints, and career incentives (e.g., Clement, 1999; Hong and Kubik, 2003; Cowen et al., 2006; Brown et al., 2015; Hirshleifer et al., 2019; deHaan et al., 2025). Within the crowdsourced forecasting literature, Estimize consensus forecasts contain incremental information beyond the sell-side consensus (Jame et al., 2016), and forecast independence is a central driver of the accuracy of the Estimize consensus (Da and Huang, 2020). We show how a major technological innovation reshapes both the quality and the composition of crowdsourced forecasts: GenAI brings measurable benefits to forecast quality, narrows the accuracy gap tied to differences in information-processing capacity, and is associated with greater market informativeness of non-professional contributors' forecasts.

Second, we contribute to research on how new technologies affect the processing and dissemination of financial information. Blankespoor et al. (2020) provide a comprehensive

framework for understanding disclosure processing costs and document how successive waves of technology have progressively reduced them, from EDGAR (Asthana and Balsam, 2001; Drake et al., 2015), to social media dissemination (Blankespoor et al., 2014), to XBRL tagging (Blankespoor, 2019; Bhattacharya et al., 2018). Earlier generations of AI continued this trajectory, but their heavy dependence on proprietary infrastructure limited their adoption to institutional users (Chen et al., 2022; Shanthikumar and Yoo, 2026; Kimbrough et al., 2026). The current wave of GenAI marks a departure. Recent studies show that GenAI affects investor information processing and trading behavior (Bertomeu et al., 2025; Blankespoor et al., 2026; Cheng et al., 2025; Cho et al., 2026), and that AI-generated equity research can expand coverage (Bradshaw et al., 2026). We show that GenAI yields measurable information-processing benefits in the production of crowdsourced forecasts, particularly for users without professional financial backgrounds.

Third, and most centrally, we speak to the unsettled question of whether GenAI democratizes economic activity or reinforces existing advantages, a question examined across settings as diverse as customer support and professional writing (Brynjolfsson et al., 2025; Noy and Zhang, 2023), entrepreneurship (Otis et al., 2026), knowledge work (Dell'Acqua et al., 2026), and investment research (Blankespoor et al., 2026; Cheng et al., 2025; Bradshaw et al., 2026; Cao et al., 2024; Choi and Xie, 2026). Rather than adjudicating between the two camps of evidence described above, we identify the margin along which each operates. GenAI substitutes for institutionalized resources, compressing the professional/non-professional gap; it complements expertise, widening the experienced/inexperienced gap within a resource-homogeneous pool. This boundary condition implies that the distributional consequences of GenAI turn on whether users' baseline disadvantage lies in processing capacity, which the technology supplies, or in the judgment required to deploy it, which it does not.

The remainder of the paper proceeds as follows. Section 2 reviews the related literature and develops the theoretical framework and predictions. Section 3 describes the data, sample construction, and variable measurement. Section 4 presents aggregate evidence that forecast quality depends on ChatGPT availability and that the professional/non-professional accuracy gap narrows

after ChatGPT's release. Section 5 describes the outage-based procedure for identifying GPT-reliant Estimize contributors and reports contributor-level results on forecast accuracy and herding behavior, heterogeneity by contributor experience, and cross-sectional variation across firms' information environments. Section 6 examines the capital market informativeness of crowd forecasts. Section 7 concludes.

2 Related Literature and Theoretical Framework

2.1 GenAI as an Information-Processing Technology

A large body of accounting research identifies information-processing costs as a central friction in financial markets, affecting how participants acquire, interpret, and trade on corporate disclosures (Blankespoor et al., 2020). Successive waves of technology have progressively reduced these costs (Drake et al., 2015; Blankespoor et al., 2014; Dong et al., 2016), and each wave left a distributional footprint that depended on whose constraint it relaxed: XBRL, for instance, leveled access between large and small institutions (Bhattacharya et al., 2018).

The first generation of AI in financial analysis relaxed constraints mainly for institutions: machine-learning models trained on detailed financial data could outperform professional analysts in predicting earnings changes (Chen et al., 2022), but building and maintaining such systems required proprietary data pipelines, computational capacity, and specialized expertise (Shanthikumar and Yoo, 2026; Kimbrough et al., 2026). Gated by technical barriers and the need for structured inputs, earlier waves of AI were positioned to reinforce existing institutional advantages rather than broaden access to analytical capability.

Generative AI departs from this pattern on both the capability and the accessibility margin (Choi and Xie, 2026). Large language models interpret, contextualize, and synthesize unstructured natural language without user-specified analytical rules (Eloundou et al., 2024), extending AI's reach to the corporate filings, earnings calls, and qualitative disclosures on which financial analysis depends. Just as important, LLMs are publicly available and require little technical expertise, collapsing the fixed costs of access; within a year of ChatGPT's launch, AI-generated content accounted for 13.5% of all equity research articles on Seeking Alpha (Bradshaw et al., 2026). Market participants have

come to depend on this newly available processing capacity: when ChatGPT became temporarily unavailable (banned in Italy or interrupted by service outages), information processing and informed trading measurably deteriorated (Bertomeu et al., 2025; Cheng et al., 2025). What GenAI supplies, however, is standardized processing capacity, not private information, and not the judgment required to deploy the tool well. Whether freely available processing capacity narrows or widens performance gaps therefore depends on who can convert it into performance, the question to which we turn next.

2.2 The Distributional Effects of Generative AI

New technologies usually redistribute performance rather than raise it uniformly. A long literature on automation shows that computerization substitutes for routine tasks while complementing non-routine analytical work, shifting the returns to skill across workers (Autor et al., 2003; Autor, 2015; Acemoglu and Restrepo, 2018, 2022). Recent theory sharpens this logic for artificial intelligence. Agrawal et al. (2019) model AI as a fall in the cost of prediction, shifting the locus of human value toward judgment, although they caution that not all judgment complements cheaper prediction. Autor and Thompson (2025) show that automation raises the value of remaining human expertise when it removes tasks non-experts can perform and erodes it when it reaches the expert core. Related theory shows that the same technology can either compress or widen the returns to knowledge, depending on how it slots into the organization of problem solving (Ide and Talamàs, 2025). Whether GenAI levels or widens performance gaps is therefore not a property of the technology alone, and the emerging evidence points in two directions.

In general labor-market settings, the evidence favors compression: experimental access to ChatGPT improves performance most for those who performed worst without AI assistance (Noy and Zhang, 2023), and the productivity gains from a deployed AI assistant concentrate among the least experienced customer-support agents (Brynjolfsson et al., 2025). A natural interpretation is substitution: the tool embeds the codified practices of high performers and thereby supplies capabilities that lower-performing workers lack.

Evidence from financial and entrepreneurial settings points the other way: sophistication and experience appear to govern who benefits. Experienced Seeking Alpha authors produce

higher-quality AI-generated research than opportunistic newcomers (Bradshaw et al., 2026); the sophisticated are also first through the door, adopting GenAI earlier and deploying it for more complex analytical tasks (Blankespoor et al., 2026); the trading that GenAI facilitates is informed and concentrated among transient institutional investors (Cheng et al., 2025); and the human advantage that survives AI is concentrated where institutional knowledge is essential (Cao et al., 2024). For less experienced users, the gains can even turn negative: randomized access to a GPT-4 business assistant left low-performing Kenyan entrepreneurs roughly ten percent worse off, a divergence traced to which pieces of advice users implemented rather than to the advice itself (Otis et al., 2026); GenAI interaction can inflate less sophisticated users' confidence in their own analytical ability (Croom, 2026), a concern sharpened because non-experts are more willing than experts to take algorithmic advice at face value (Logg et al., 2019). Even professionals face integration costs when adopting AI-enabled research tools (Xue et al., 2025; Fügener et al., 2026).

Existing reconciliations of these divergent findings emphasize the task. Dell'Acqua et al. (2026) characterize a "jagged technological frontier": tasks within the frontier are substantially enhanced by GenAI, while tasks beyond it may see little improvement or even degradation. In forecast production, parsing dense filings and computing comparisons lie within the frontier; evaluating management credibility, integrating proprietary intelligence, and resolving genuine uncertainty lie beyond it. A task-level frontier explains *where* GenAI creates value, not *who* captures it: it does not say why experience marks a missing capability among customer-support agents yet a prerequisite for gains among equity-research authors. Nor does existing evidence resolve the question: experiments assign GenAI use but hold the institutional environment fixed, so they cannot separate what the technology substitutes for from what it complements; in the field, where resources and expertise both vary, individual use is rarely observable, so designs recover adoption (who takes up the tool) or content (what the tool produces) rather than the effect of that use on the user's own performance. We argue that the two camps observe two distinct margins of the same technology: GenAI substitutes for the routine processing capacity that institutional resources otherwise provide, compressing gaps rooted in resource endowments; it complements expertise, widening gaps rooted in judgment;

and it leaves private-information advantages untouched. Distinguishing these margins requires a setting in which resources and expertise vary separately and performance is objectively scored; the crowdsourced forecasting setting we describe next meets both requirements.

2.3 Crowdsourced Forecasting as a Laboratory

The wisdom-of-crowds principle holds that aggregating independent judgments from diverse individuals can produce forecasts more accurate than any single expert's (Surowiecki, 2004); crowdsourced forecasting platforms operationalize this principle in financial markets with objectively scored output. Estimize is the most prominent such platform in the academic literature (see Khavis and Park (2024) for a comprehensive description of the platform). Estimize consensus forecasts contain incremental information about earnings beyond the I/B/E/S sell-side consensus, with the informational advantage increasing in the size and independence of the forecasting pool (Jame et al., 2016; Da and Huang, 2020; Brown et al., 2024). Da and Huang (2020) establish the importance of independence directly: in a field experiment on Estimize, "blind" forecasts, submitted without viewing others' estimates, produced a significantly more accurate consensus, leading the platform to permanently adopt the blind design. Estimize activity also spills over to the broader information environment, disciplining sell-side forecasts and informing investor trading (Jame et al., 2022; Schafhäutle and Veenman, 2024; Farrell et al., 2022).

Estimize contributors differ from sell-side analysts in ways central to our analysis. They submit forecasts independently rather than as part of team-based research departments, and they are free from the conflicts of interest associated with investment banking relationships or brokerage commissions (e.g., Hong and Kubik, 2003; Cowen et al., 2006; Brown et al., 2015). The platform also rewards forecast accuracy: performance dashboards build a verifiable track record of each user's historical accuracy (Figure IA.1, Panel B, in the Internet Appendix), which contributors can leverage to attract clients or enhance career prospects. These accuracy incentives arguably substitute for part of the reputational discipline present in institutional settings. At the same time, contributors vary widely in professional background.[8]

[8]Upon registration, members self-identify as either "professional" or "non-professional." Professionals further classify themselves as buy-side (e.g., hedge fund, mutual fund, pension fund), sell-side, or independent.

The professional/non-professional partition captures, above all, differences in the institutional apparatus of research, and these differences generate persistent heterogeneity along two classic dimensions emphasized in the analyst literature: information access and information processing. On the access dimension, financial professionals are more likely to command proprietary data terminals, management interactions, and dedicated research infrastructure (e.g., Clement, 1999; Brown et al., 2015; Green et al., 2014; Soltes, 2014). On the processing dimension, formal training and structured workflows enable professionals to extract signals more efficiently from complex financial disclosures (e.g., Mikhail et al., 1997; Jacob et al., 1999; Bradshaw et al., 2012), while less sophisticated participants face disproportionately higher processing costs when disclosures are complex (Blankespoor et al., 2020; Li, 2008; Miller, 2010). Consistent with these advantages, analysts with greater resources produce more accurate forecasts (e.g., Clement, 1999; Beyer et al., 2010).

Although institutional infrastructure remains largely out of reach for non-professionals as a group, differences in experience and sophistication within the pool are pronounced: it spans students and academics, employees who forecast firms in their own industries, and independent investors with years of accumulated practice. Because resource endowments are uniformly thin within the group, comparisons across non-professionals hold the resource margin largely fixed and isolate variation in expertise.[9]

On Estimize, these advantages translate into a measurable accuracy gap: non-professionals produce less accurate forecasts than professionals on average. Whether GenAI narrows or widens this gap, and through what channels, is the central question of this paper.

Three features make this setting a natural laboratory for the distributional question. First, output is objectively scored: every forecast is time-stamped and evaluated against realized earnings, so performance is measured without the rating ambiguity of content platforms. Second, as argued

Non-professionals identify the GICS sector and industry of their primary occupation, or indicate that they are students or members of academia. This self-reported classification is the basis for our professional/non-professional partition (see Figure IA.1, Panel A).

[9]Residual resource differences among non-professionals (for example, an industry employee's access to sector-specific data) cannot be ruled out. The claim is comparative: variation in institutional infrastructure is far smaller within the non-professional pool than between professionals and non-professionals.

above, the setting supplies both margins at once: the partition varies resources, while accumulated experience varies expertise within the resource-homogeneous non-professional pool. Third, the blind design mutes social influence, so changes in herding behavior can be interpreted as changes in independent information processing rather than in social dynamics (Da and Huang, 2020).[10] Section 5.1 addresses the remaining obstacle (no field setting records who uses ChatGPT) with an outage-based classification that infers reliance from forecasting behavior when ChatGPT is unexpectedly unavailable.

2.4 Theoretical Framework and Predictions

GenAI lowers information-processing costs by automating the comprehension and organization of unstructured text, allowing users to reallocate attention from routine public-information processing toward judgment-intensive tasks that remain beyond AI's current capabilities.

The distribution of these gains depends on where users' baseline disadvantages lie. Estimize contributors, including non-professionals, form a selected, performance-oriented community, so the relevant divide is one of institutionalized analytical resources rather than sophistication: professionals are more likely to possess the formal training, proprietary information access, and workflow support that non-professionals are less likely to command. The distinction matters because much of forecast production consists of routine, codifiable processing of public information, precisely the work that lies closest to GenAI's current frontier.

To organize the predicted heterogeneity, consider a stylized forecast-production technology. Contributor i's expected forecast accuracy for firm j increases in three inputs: routine processing capacity (P_i: parsing dense filings, extracting operating detail, organizing guidance, and computing comparisons); access to private information (A_i: management interaction, proprietary data, and institutional networks); and evaluative judgment (J_i: recognizing valuable applications, steering the analysis, and verifying outputs against domain knowledge).[11] The weight the forecasting problem

[10] Estimize conceals other contributors' estimates and the consensus until a contributor submits his or her own estimate for a given release, and reveals them thereafter. Herding is accordingly measured only on revisions, for which the prevailing consensus is observable (Section 5.2.2).

[11] This three-input decomposition is a useful approximation rather than an exhaustive characterization of forecast production; in practice, GenAI may also shape judgment and information acquisition indirectly. It serves to isolate the margins on which GenAI most plausibly operates.

places on processing rises with the complexity of the firm's public disclosures, and the weight on access rises with the firm's information asymmetry. GenAI supplies standardized processing capacity, denoted $\bar{P}$, at low cost, but the share of that capacity a user captures increases in judgment (Dell'Acqua et al., 2026; Otis et al., 2026). A contributor's effective processing capacity under adoption is therefore $P_i + g(J_i) \cdot \bar{P}$, with $g(\cdot)$ increasing, while A_i and J_i are unchanged by the technology. The two forces operate on different margins: across the professional/non-professional partition, differences in P_i are large and, because the marginal value of processing capacity is diminishing, dominate the comparison; within the non-professional pool, P_i varies little, so variation in the captured share $g(J_i)$ governs who gains. Figure 1, Panel A, depicts this structure.

Four implications follow (Figure 1, Panel B). First, because the marginal value of processing capacity is decreasing, adoption compresses performance gaps rooted in P: professionals, whose training and infrastructure already embed high processing capacity, gain little, while non-professionals gain substantially (substitution across resources).

Second, because captured capacity $g(J_i) \cdot \bar{P}$ rises with judgment, gains within a resource-homogeneous group concentrate among experienced users; users with little domain knowledge capture less and, if they defer to model outputs they cannot evaluate, may substitute the model's view for their own information (complementarity in expertise).

Third, gains are largest where the forecasting problem loads on processing and smallest where it loads on access, because GenAI raises no contributor's A_i. The processing load is high when public disclosures are dense and difficult to parse, as captured by readability indices from corporate filings (Bushee et al., 2018; Blankespoor et al., 2020). The access load is high when information asymmetry is high, that is, when forecast dispersion reflects differential access to private information rather than fundamental uncertainty (Diether et al., 2002; Zhang, 2006), a distinction the Barron et al. (1998) decomposition (hereafter BKLS) operationalizes.

Fourth, adoption and benefit need not coincide: adoption reflects integration costs and performance incentives, both of which favor professionals, whereas the marginal benefit reflects baseline processing capacity, which favors non-professionals. Higher adoption by professionals

alongside larger gains for non-professionals is therefore consistent with, rather than evidence against, the substitution logic.

Tasks beyond the frontier remain resistant to AI augmentation for either group. A direct corollary is that if contributors rely on GenAI for routine processing, forecast quality should deteriorate when the technology becomes unexpectedly unavailable.

This framework also clarifies why our predictions are consistent with evidence that more sophisticated users benefit more from GenAI in other settings (Blankespoor et al., 2026; Bradshaw et al., 2026; Otis et al., 2026). Those findings identify the judgment margin, $g(J)$: when the returns to GenAI depend on the judgment to deploy it well, sophistication and experience raise the value of AI. The professional/non-professional partition, by contrast, isolates the resource margin, P, on which substitution should dominate because routine public-information processing lies within GenAI's frontier.

Beyond average accuracy, an important question is whether GenAI-assisted improvements reflect more independent information processing or merely mechanical convergence toward consensus. The analyst herding literature establishes that forecasters often suppress private information and cluster around prevailing expectations because of career concerns, reputational incentives, or social influence (e.g., Scharfstein and Stein, 1990; Trueman, 1994; Welch, 2000; Clement and Tse, 2005). We expect GenAI to *reduce* herding among adopters with the judgment to evaluate its output: corporate filings report firm-specific conditions that may diverge from the consensus embedded in prior forecasts, and if GenAI helps contributors extract and process these signals, their forecasts should reflect more firm-specific information and less consensus anchoring. For contributors who lack that judgment, the same tool may operate in reverse: deference to model output anchored on publicly available expectations produces convergence toward, rather than independence from, the consensus. The distinction matters beyond the individual forecast: forecast independence is what makes crowds accurate (Da and Huang, 2020), so whether GenAI promotes independence or herding determines whether individual improvements strengthen or weaken the collective signal.

Finally, if GenAI improves the quality and independence of non-professional forecasts, the improvements should be reflected in capital market prices. Prior research shows that Estimize consensus forecasts contain incremental information for price discovery (Jame et al., 2016) and that investors use them to discount analyst biases (Schafhäutle and Veenman, 2024), so changes in crowdsourced forecast quality can affect market outcomes. As the non-professional consensus becomes more accurate, its forecast surprise becomes a more precise measure of genuinely unexpected earnings, and the market's earnings response coefficient to non-professional surprises should increase in the post-GPT period. A strengthening in this response could also reflect the crowd's improving accuracy over the sample period rather than a discrete effect of ChatGPT's release, so we interpret the evidence in Section 6 descriptively.

In summary, the framework yields four testable predictions. Prediction 1 (*adoption effects*): contributors classified as GPT-reliant should issue more accurate and more independent forecasts once ChatGPT becomes available. Prediction 2 (*resource convergence*): these gains should concentrate among non-professionals, narrowing the professional/non-professional accuracy gap, most strongly where routine public-information processing binds (complex disclosures and, separately, low information asymmetry). Prediction 3 (*expertise complementarity*): within the non-professional pool, where resources vary little, gains should concentrate among experienced contributors, while inexperienced adopters should improve little and may herd more. Prediction 4 (*market informativeness*): the market's earnings response to non-professional forecast surprises should increase in the post-GPT period. Together, these predictions trace a chain from GenAI-enabled reductions in information-processing costs, to the distribution of individual forecast gains across contributors, to aggregate market outcomes.

3 Data, Sample Selection, and Measurement

3.1 The Estimize Platform and Sample Construction

Our data come from Estimize, a crowdsourced earnings forecasting platform founded in 2011 that aggregates predictions from both financial professionals and non-professional contributors. The platform has been widely used in prior accounting and finance research (e.g., Jame et al.,

2016; Da and Huang, 2020; Jame et al., 2022; Schafhäutle and Veenman, 2024; Khavis and Park, 2024).[12] Unlike traditional analyst forecast databases such as I/B/E/S, which are restricted to sell-side analysts at brokerage firms, Estimize is open to any registered user, including buy-side analysts, independent investors, corporate finance professionals, students, and hobbyists, creating a setting where participants with varying expertise forecast the same firms.

We obtain the universe of earnings-per-share (EPS) and revenue forecasts submitted on Estimize between 2015 and 2024 and merge these data with CRSP for stock return information, Compustat for firm financial characteristics, and I/B/E/S for analyst forecast dispersion. The sample begins in 2015, coinciding with the platform's adoption of the blind forecast design (Da and Huang, 2020), which ensures that all observations reflect independent forecasting behavior. The sample ends in 2024 because our identification strategy exploits disruptions specific to OpenAI's ChatGPT service; by 2025, the rapid proliferation of competing GenAI platforms (including Google's Gemini, Anthropic's Claude, and Meta's Llama) made it increasingly difficult to attribute forecast behavior to any single tool, as users could substitute across platforms during ChatGPT outages (see Chatterji et al. (2025) for evidence of declining ChatGPT market share).[13]

We construct the full sample by applying standard filters: we exclude observations with missing values in key variables, restrict to forecasts issued within 90 days of the earnings announcement to eliminate stale predictions, require matching firm identifiers in CRSP or Compustat, require a CRSP *permno*, and require within-group variation in each firm–year-quarter cell for *PMAFE* construction.[14] These filters yield a full sample of 928,928 forecasts from 82,726 unique contributors covering 3,056 firms over 40 fiscal quarters. Table 1, Panel A summarizes this construction. Subsequent analyses require subsamples tailored to specific identification strategies.

[12] In 2021, Estimize merged with ExtractAlpha, an alternative data provider. The platform continued to operate independently and accept user submissions throughout our sample period.

[13] Industry data indicate that ChatGPT's share of AI chatbot web traffic declined from approximately 87% in January 2025 to 68% in January 2026 as competing platforms gained ground (Similarweb, *Gen AI Website Traffic Share*, January 2026); in the enterprise segment, OpenAI's share of LLM usage declined from approximately 50% in 2023 to 27% in 2025 (Menlo Ventures, *The State of Generative AI in the Enterprise*, 2025, https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/).

[14] Because most sample firms report quarterly, the 90-day window removes forecasts made before the preceding quarter's announcement, which are unlikely to reflect contributors' current information sets.

3.2 Measuring Forecast Accuracy

We measure forecast accuracy using the proportional mean absolute forecast error (*PMAFE*), following Clement (1999). For each forecast by contributor i for firm j in fiscal quarter t, we compute the absolute forecast error as $AFE_{i,j,t} = |\text{Forecast EPS} - \text{Actual EPS}|$. We then compute *PMAFE* as:

$$PMAFE_{i,j,t} = \frac{AFE_{i,j,t} - \overline{AFE}_{j,t}}{\overline{AFE}_{j,t}} \tag{1}$$

where $\overline{AFE}_{j,t}$ is the mean absolute forecast error across all contributors covering firm j in fiscal quarter t. A negative value indicates that the contributor's forecast is more accurate than the peer average for the same firm–quarter; a positive value indicates worse-than-average accuracy. *PMAFE_REV* is constructed analogously from revenue forecasts and reported revenue.

This peer-adjustment serves a central role in our research design. Because our primary question concerns how GenAI reshapes the relative performance of different types of contributors (professionals versus non-professionals, GPT-reliant versus non-GPT contributors), we require a measure that isolates contributor-level variation from the firm–quarter factors that determine forecast difficulty. The within-group demeaning in Equation (1) is equivalent to including firm $\times$ quarter fixed effects, absorbing earnings predictability, information environment, and macroeconomic conditions (e.g., Clement, 1999; Brown et al., 2015). This ensures that our estimates capture differences in contributor forecasting skill rather than differences in the firms or periods they cover. To maintain consistency with this peer-adjusted dependent variable, we min–max scale all continuous contributor-level controls (e.g., experience, effort, forecast age, firm coverage breadth) within the same firm–quarter group in specifications where *PMAFE* serves as the dependent variable.[15]

Table 2 presents descriptive statistics for each sample. We winsorize all continuous variables at the 1st and 99th percentiles. The four panels correspond to the full sample used in the aggregate tests (Panel A), the contributor–firm–quarter frequency sample (Panel B), the accuracy and herding

[15]The additional tests also employ a design that does not rely on peer adjustment: Table IA.5 in the Internet Appendix re-estimates the adoption specifications with the non-peer-adjusted absolute forecast error (*AFE*) as the dependent variable and unscaled contributor- and firm-level controls (Section 5.3.2).

sample (Panel C), and the market reaction sample (Panel D). Detailed variable definitions appear in Appendix A.

4 GenAI Availability and Forecast Quality: Aggregate Evidence

Before identifying individual GPT-reliant contributors, we document two aggregate patterns that motivate the contributor-level analyses that follow. First, GenAI has become a useful input to crowdsourced forecast production: forecast quality deteriorates when ChatGPT is unexpectedly unavailable. Second, the technology's arrival coincides with a redistribution of relative performance: the long-standing accuracy gap between professional and non-professional contributors narrows after ChatGPT's release, and non-professionals are the group most exposed to ChatGPT disruptions. The evidence draws on two sources of variation: the release of ChatGPT in late 2022 and 32 documented major ChatGPT service outages during 2023–2024.[16]

4.1 Forecast Quality During ChatGPT Outages

If GenAI enters forecast production, its unexpected unavailability should leave two footprints. Contributors who rely on the technology should issue fewer forecasts while it is down (evidence that GenAI is *used*), and the forecasts that are issued should be of lower quality (evidence that GenAI is *useful*).

Figure 2 documents the first footprint. The Estimize columns of each panel compare hourly forecast issuance during outage periods with matched benchmark periods, defined as the same time window four weeks before and after the outage (Panel A) or one day before and after (Panel B).[17] Forecasting activity on Estimize falls substantially during outages: full-sample issuance drops by 17.2% relative to the long-window benchmark and by 17.7% relative to the short-window benchmark. The rightmost column of each panel replicates the comparison for I/B/E/S analyst forecasts and

[16]The complete list of outages, including start times, end times, and durations, is reported in Internet Appendix A. Our outage data are sourced directly from OpenAI's official incident history page (status.openai.com). Cheng et al. (2025) construct a similar outage dataset but cross-reference OpenAI's records with third-party monitoring services, which may adjust reported times by several minutes. The differences between the two datasets are minor and limited to the minute level.

[17]The four-week benchmark smooths day-of-week and earnings-calendar variation around each outage; the one-day benchmark holds market conditions nearly fixed at the cost of a noisier baseline.

shows no corresponding decline.[18] The contrast suggests that Estimize forecasters incorporate ChatGPT into their forecast production, whereas sell-side analysts' workflows are largely insulated from its temporary unavailability.[19]

We then test the second footprint by estimating:

$$PMAFE_{i,j,t} = \beta_1 During\_Outage_t + \mathbf{X}'_{i,j,t}\gamma + \alpha_j + \varepsilon_{i,j,t} \tag{2}$$

where $During\_Outage_t$ equals one if the forecast was submitted during a documented ChatGPT service outage, $\mathbf{X}_{i,j,t}$ is a vector of contributor-level controls (forecast age, experience, effort, and firm coverage breadth), and α_j denotes firm fixed effects.[20] The sample is restricted to the post-GPT period (2023–2024). Saturated specifications add day-of-week $\times$ hour-of-day fixed effects to absorb temporal regularities in forecasting behavior. A positive β_1 indicates that forecast accuracy deteriorates when GPT is unavailable. Standard errors are clustered at the firm level.

Table 3, Panel A shows that forecasts submitted during outages exhibit significantly higher *PMAFE*. In the saturated specification, column (2) with controls, firm fixed effects, and day-of-week $\times$ hour-of-day fixed effects, the coefficient is 0.091 ($t = 2.27$), representing 15.8% of the *PMAFE* standard deviation in the post-GPT sample (0.091 / 0.575).[21] The effect is robust across specifications, ranging from 0.084 to 0.091 depending on the inclusion of fixed effects.[22]

[18] Sell-side analysts at brokerage firms typically operate within team-based research departments with access to proprietary analytical infrastructure, including in-house quantitative tools (e.g., Kimbrough et al., 2026; Brown et al., 2015); these institutional resources reduce dependence on any single publicly available tool.

[19] The contributor-level productivity test in Section 5.1.2 points to the same conclusion from the adoption margin: contributors we later classify as GPT-reliant sharply increase their forecasting output relative to non-GPT contributors after ChatGPT becomes available (Figure 3 and Table 5). Although we present that test as a validation of the classification procedure, it equally indicates that GenAI is being *used* in forecast production.

[20] Outage windows and Estimize submission timestamps are both recorded in Eastern Standard Time (EST), so no time-zone conversion is required. A forecast is coded as submitted during an outage if its timestamp falls within the outage's start and end times, inclusive. "Major" outages are incidents that reached full-outage status on a ChatGPT-related component of OpenAI's status page.

[21] The *PMAFE* standard deviation in the post-GPT sample (2023–2024) is 0.575 (untabulated); Table 2 reports the full-sample value of 0.613.

[22] A compositional concern is that contributors who rely on GPT may abstain during outages while others continue, so part of the quality decline could reflect selection. Controlling for contributor characteristics partially addresses this, and both channels (abstention by reliant contributors and quality deterioration among those who continue) reflect reliance on GenAI in forecast production. The contributor-level design in Section 5.1 separates the two margins by classifying reliance from the abstention margin and measuring outcomes conditional on submission.

4.2 Professional versus Non-Professional Contributors

Having established that GenAI matters for forecast quality, we next examine how relative performance between professional and non-professional contributors changes around its arrival. We estimate the following equation:

$$PMAFE_{i,j,t} = \beta_1 Non\_Professional_i + \beta_2 Post\_GPT_t + \beta_3 (Non\_Professional_i \times Post\_GPT_t) + \mathbf{X}'_{i,j,t}\gamma + \alpha_j + \varepsilon_{i,j,t} \quad (3)$$

where $Non\_Professional_i$ equals one for non-professional contributors, $Post\_GPT_t$ equals one for forecasts issued in January 2023 or later, and the controls and fixed effects are as in Equation (2).[23] Although *PMAFE* already absorbs firm–quarter variation by construction, firm fixed effects capture persistent differences in forecasting difficulty across firms. The coefficient of interest is β_3, which captures the differential change in forecast error for non-professionals relative to professionals after GPT's release. Standard errors are clustered at the firm level.

Table 4 shows that non-professionals' pre-GPT accuracy disadvantage disappears after ChatGPT's release. The coefficient on *Non_Professional* is positive and significant across all specifications ($\beta = 0.004$–0.011, $p < 0.10$), confirming that non-professionals produce less accurate forecasts than professionals in the pre-GPT period. The interaction *Non_Professional* × *Post_GPT* is negative and significant in all columns. In the most conservative specification, column (4) with controls and firm fixed effects, the interaction is –0.013 ($t = -1.96$). The pre-GPT accuracy disadvantage for non-professionals in this specification is 0.008 ($t = 3.67$); the interaction of –0.013 exceeds this gap in absolute value. The bottom row of Table 4 makes the closure explicit by testing whether the post-GPT gap, $\beta_1 + \beta_3$, equals zero. With controls, the post-GPT gap is economically small and statistically indistinguishable from zero (–0.005, $p = 0.46$, in column (4)).[24] This

[23] We begin the post-GPT period in January 2023, one month after ChatGPT's release on November 30, 2022, to allow for initial diffusion and to align the cutoff with a calendar boundary; forecasts issued in December 2022 are assigned to the pre-period.

[24] The identifying assumption of Equation (3) is that professional and non-professional accuracy would have evolved in parallel absent ChatGPT's release. In untabulated tests with year-by-year *Non_Professional* × *Year* interactions, this assumption does not hold: the accuracy gap is significantly wider than its 2022 level in every year from 2015 through 2019 (joint $p = 0.003$ for the pre-period interactions), statistically indistinguishable from that level from 2020 onward, and exhibits no incremental post-GPT change (joint $p = 0.28$ for the 2023–2024 interactions); the convergence largely predates the GPT release. We therefore interpret the aggregate convergence as a documented association rather than a

aggregate convergence provides descriptive context rather than causal evidence.

The outage evidence identifies the group most exposed to GenAI availability, non-professionals, most clearly on the production margin. The contributor-type splits in Figure 2 show that the outage-induced decline in forecast issuance is concentrated among non-professionals (–27.0% in the long window, –38.5% in the short window), while issuance by financial professionals shows no comparable decline (–5.6% in the long window and +16.5% in the short window), consistent with professionals having alternative information resources that do not depend on ChatGPT availability. On the quality margin, Panel B of Table 3 estimates the outage effect separately by contributor type: the effect is significant only for non-professionals (0.098, $t = 1.93$; professionals: 0.091, $t = 1.28$), though the point estimates are similar and not statistically distinguishable, so the differential-exposure conclusion rests primarily on the production margin.

Together, the aggregate evidence establishes the two facts that frame the rest of the paper: forecast quality depends on ChatGPT availability, and the professional/non-professional accuracy gap has narrowed in the post-GPT period, with non-professionals the group most exposed to GenAI disruptions.

5 GPT Reliance and the Effects of Adoption: Contributor-Level Evidence

This section moves from aggregate patterns to contributor-level evidence by identifying which contributors rely on GPT for forecast production. Section 5.1 develops the outage-based procedure that classifies individual contributors as GPT-reliant, validates the classification against productivity trajectories around ChatGPT's release, and constructs the samples for the contributor-level tests. Section 5.2 then estimates how GPT adoption changes forecast accuracy and herding behavior (testing Predictions 1 and 2 across contributor types and Prediction 3 across experience cohorts within the non-professional pool) and examines where the benefits are largest across firms' information environments. Section 5.3 reports robustness and falsification tests.

causal effect, and identification rests on the outage variation and on the contributor-level classification in Section 5.1, whose own parallel-trends diagnostics appear in Internet Appendix C.

5.1 Identifying GPT-Reliant Estimize Contributors

5.1.1 Classification Procedure and Distribution

We identify GPT-reliant contributors through a ChatGPT outage-based approach: contributors who depend on GenAI should exhibit disproportionate declines in forecasting activity when the service is unexpectedly unavailable. We implement this logic as a three-stage classification procedure; the full technical details, including variable and estimation definitions, are provided in Internet Appendix B; we summarize the approach here.

In the first stage, we estimate a logit model on the post-GPT sample (2023–2024), comprising approximately 1.08 billion contributor–firm–day–half-hour observations, to predict each contributor's baseline probability of submitting a forecast in a given 30-minute window, conditional on recent forecasting activity, earnings announcement proximity, stock return volatility, and a rich set of fixed effects.[25] In the second stage, we use the fitted probabilities to identify contributors who are *expected* to forecast in each 30-minute window (those in the top quartile of predicted probability within each bin). We then evaluate whether these expected forecasts materialize during documented ChatGPT outages. A contributor is recorded as having a missing forecast in an outage if the contributor is flagged as an expected forecaster during at least one window of the outage but submits no forecast throughout its entire duration. For each contributor i, we compute: $MissRatio_i$ = Number of outages with missing forecasts/Total number of outages. This ratio captures how frequently a contributor's forecasting activity disappears when ChatGPT is unavailable. In the third stage, we classify contributors in the top quartile of $MissRatio_i$, those with the most persistent outage-induced forecast absence, as GPT-reliant contributors ("GPT contributors" for short, denoted $GPT\_Contributor_i$).

[25] Specifically, we estimate:

$$\Pr(Forecast_{i,j,d,h}=1)=\Lambda\Big(\beta_1 RecentActivity_{i,d}+\beta_2 EA\_Past5_{j,d} + \beta_3 Volatility_{j,d}+\alpha_{j\times y}+\gamma_{i\times j}+\delta_w+\eta_h\Big)$$

where $\Lambda(\cdot)$ is the logistic function, $RecentActivity_{i,d}$ captures contributor i's recent forecasting behavior, $EA\_Past5_{j,d}$ indicates whether firm j had an earnings announcement within the past five days, $Volatility_{j,d}$ is the stock return volatility of firm j, and $\alpha_{j\times y}$, $\gamma_{i\times j}$, δ_w, η_h denote firm-by-year, contributor-by-firm, day-of-week, and half-hour-of-day fixed effects, respectively. Full details are in Internet Appendix B.

Table IA.1 in the Internet Appendix reports the distribution of the resulting classification. Among the 3,876 contributors in the accuracy sample, 986 (25.4%) are classified as GPT-reliant; the overall share is pinned down by the top-quartile cutoff on *MissRatio* applied in the classification, but the composition within that quartile is not.[26] Among the 986 GPT-reliant contributors, 241 are financial professionals (44.3% of the sample's professionals) and 745 are non-professionals (22.4% of non-professionals). These shares are consistent with the aggregate issuance evidence in Section 4.2: the issuance comparisons aggregate hourly counts, which the much larger non-professional pool dominates, whereas *MissRatio* is computed contributor by contributor, conditional on being an expected forecaster in the specific outage windows, so group-level issuance can remain stable even as individual expected forecasts go missing. Because the classification identifies reliance rather than dated adoption, it carries no individual adoption date; in the tests below, *Post_GPT* proxies for the availability of the technology rather than the timing of each contributor's uptake.

5.1.2 Validation

If the classification correctly identifies contributors who rely on GenAI as a forecasting tool, these contributors should exhibit a measurable increase in forecasting output after GPT becomes available. Figure 3 provides visual evidence for this prediction. Panel A plots the within-year percentile rank of annual forecast count for GPT-reliant and non-GPT contributors from 2020 to 2024. The two groups follow parallel paths in the pre-GPT period, but diverge sharply after 2022: GPT-reliant contributors increase their relative productivity while non-GPT contributors decline. Panel B reports the difference-in-differences estimate: a 9.83 percent increase in relative productivity for GPT-reliant contributors ($t = 7.48$).

Table 5 formalizes this evidence with a difference-in-differences model of forecasting frequency, estimated on the contributor–firm–quarter panel of classified contributors.[27] The

[26] GPT-reliant contributors are disproportionately active forecasters, so their share of forecasts far exceeds their share of contributors: they account for 87.2% of the frequency-sample and 91.7% of the accuracy-sample observations (Table 2, Panels B and C).

[27] The frequency measure is min–max scaled within each firm–quarter group. Contributors outside the classification of Section 5.1 do not enter this panel, so the sample mean of *GPT_Contributor* in Table 2, Panel B reflects the classified universe rather than all Estimize contributors.

coefficient on *GPT_Contributor* × *Post_GPT* is positive and highly significant ($\beta = 0.046$, $t = 7.32$ with firm fixed effects), confirming that classified GPT-reliant contributors substantially increase their forecasting output after GPT becomes available. The positive standalone coefficient on *GPT_Contributor* ($\beta = 0.074$, $t = 23.49$) indicates that these contributors were already more productive in the pre-period, a pattern consistent with the classification capturing contributors dispositionally inclined to adopt productivity-enhancing tools.

Panel B of Table 5 reveals that the productivity gains are concentrated largely among non-professionals ($\beta = 0.054$, $t = 7.95$), with no significant effect for financial professionals ($\beta = -0.006$, $t = -0.36$). This asymmetry is consistent with GenAI lowering the marginal cost of producing a forecast primarily where that cost is highest. Although the outage-based approach has limitations, which we return to in Section 7, the validation results indicate that the classification captures meaningful variation in the usage of GenAI: classified contributors sharply increase their forecasting output once ChatGPT becomes available, and the increase concentrates precisely where the technology relieves the tightest constraints.

5.1.3 Sample Construction for Contributor-Level Tests

Having established and validated the GPT-reliant contributor classification, we construct the subsamples used in the tests that follow. We restrict to forecasts issued during 2021–2024, providing a symmetric two-year window around the start of the post-GPT period (January 2023) that balances pre- and post-treatment observations; the sample retains the full-sample screens of Section 3.1, including the restriction of forecast age to $[0, 90]$ days before the earnings announcement. We end the sample in 2024 because, as discussed in Section 3.1, the post-2024 proliferation of close ChatGPT substitutes undermines the single-platform attribution on which the classification rests; Section 5.3.3 provides direct evidence. Within this window, we require contributors to have at least one forecast in the post-GPT period (2023 or later) to permit classification. The resulting accuracy sample contains 193,584 forecasts from 3,876 contributors. The herding sample further restricts to revision forecasts, excluding the first submission by a contributor within a firm–quarter, yielding 59,674 observations from 1,334 contributors. Table 1, Panel B summarizes this construction.

5.2 The Effects of GenAI Adoption

We test whether GPT adoption affects individual forecast accuracy and herding behavior, estimating difference-in-differences (DiD) models that compare outcomes for GPT-reliant contributors relative to non-GPT contributors before and after ChatGPT becomes available:

$$\begin{aligned} Y_{i,j,t} = {} & \beta_1 GPT\_Contributor_i + \beta_2 Post\_GPT_t + \beta_3 (GPT\_Contributor_i \times Post\_GPT_t) \\ & + \mathbf{X}'_{i,j,t}\gamma + \alpha_j + \varepsilon_{i,j,t} \end{aligned} \tag{4}$$

where $Y_{i,j,t}$ is alternatively *PMAFE* (for EPS and revenue forecasts) or a herding indicator (*Herding*), $GPT\_Contributor_i$ equals one for contributors classified as GPT-reliant, $Post\_GPT_t$ equals one for forecasts issued in January 2023 or later, $\mathbf{X}_{i,j,t}$ is a vector of contributor-level controls (*Effort*, *Experience*, *Forecast_Age*, and *Firm_Followed*), and α_j denotes firm fixed effects.[28] The coefficient of interest is β_3, which captures the differential change in outcomes for GPT-reliant contributors relative to non-GPT contributors after GPT becomes available. For the herding outcome, we estimate logit models and report z-statistics.[29] We further partition each sample by contributor type (financial professionals versus non-professionals) to test whether effects differ across groups.

5.2.1 GenAI Adoption and Forecast Accuracy

GenAI should improve forecast accuracy by facilitating financial information processing, with particularly large gains for non-professionals who face higher information-processing costs (Predictions 1 and 2). We test these predictions by estimating Equation (4) with *PMAFE* as the dependent variable.

Table 6 reports the results for both EPS and revenue forecast accuracy. In Panel A, the coefficient on *GPT_Contributor* × *Post_GPT* is negative and significant for both EPS ($\beta = -0.054$, $t = -2.29$) and revenue ($\beta = -0.143$, $t = -2.44$) in specifications with firm fixed effects, indicating that GPT-reliant contributors improve their relative accuracy in the post-GPT period.[30]

[28] In the herding specifications, the dependent variable is a binary indicator, so controls enter in their original units rather than min–max scaled.

[29] Standard errors in the accuracy adoption tests, the herding logits, and the aggregate tests of Section 4 are clustered by firm; the experience-cohort accuracy tests, the AFE robustness test, and the readability columns of the cross-sectional tests cluster by firm–contributor pair, and the information-asymmetry columns of the cross-sectional tests and the cohort herding logits cluster by firm, as noted in the respective tables.

[30] Clustering by firm addresses the dominant source of error correlation in forecast data, common firm-level shocks

Panel B shows that the accuracy improvement is concentrated among non-professionals. For EPS, the coefficient is –0.080 ($t = -2.25$) for non-professionals, with no significant effect for professionals ($\beta = 0.024$, $t = 0.43$). This pattern is consistent with GenAI providing information-processing capabilities that non-professionals previously lacked, narrowing the accuracy gap with professionals.

For revenue, the pattern is directionally similar but more nuanced. Non-professional GPT-reliant contributors reduce their *PMAFE* by 0.129 ($t = -1.82$). Professionals also show a marginally significant improvement ($\beta = -0.250$, $t = -1.72$), so for revenue the improvement is not confined to non-professionals.[31]

The accuracy result is not sensitive to how strictly reliance is defined. Re-estimating the EPS specifications with *GPT_Contributor* restricted to contributors who never submit a single forecast during any outage window, the sharpest revealed-reliance signal, yields a larger non-professional estimate ($\beta = -0.095$, $t = -2.64$) and leaves professionals unaffected (Table IA.3 in the Internet Appendix).

Taken together, the accuracy results support Predictions 1 and 2: GPT adoption is associated with improved forecast accuracy (Prediction 1), with the gains primarily concentrated among non-professional contributors (Prediction 2).

5.2.2 GenAI Adoption and Herding Behavior

The accuracy improvements raise a question about mechanism: do GPT-reliant contributors use the technology to extract firm-specific signals from disclosures, producing more independent forecasts, or to converge more efficiently toward existing expectations? The latter channel is grounded in classic models of social learning: when processing information is costly, rationally anchoring to the prevailing consensus can be optimal (Banerjee, 1992; Bikhchandani et al., 1992).

shared by all forecasts of the same firm. Because *GPT_Contributor* is assigned at the contributor level (Abadie et al., 2023), we also compute standard errors with two-way clustering by firm and contributor: the interactions remain significant at conventional levels (full sample: EPS $t = -1.88$, revenue $t = -2.18$; non-professionals: EPS $t = -2.03$, revenue $t = -1.65$).

[31] These estimates require that GPT-reliant and non-GPT contributors would have followed parallel accuracy paths absent GPT availability. Internet Appendix C reports event-study and Rambachan and Roth (2023) sensitivity analyses assessing this assumption.

Section 2.4 develops these two channels and their stakes for the aggregate value of the forecasting pool; here we take them to the data, using herding as a proxy for independence in forecast formation.

We distinguish these channels by estimating logit models of herding behavior using Equation (4), where the dependent variable indicates whether a forecast revision moves toward the prevailing consensus. The sample is the herding subsample of revision forecasts constructed in Section 5.1.3.[32] Table 7 reports the results, with z-statistics in parentheses.

In Panel A, the coefficient on *GPT_Contributor* × *Post_GPT* is negative and significant for both EPS ($\beta = -0.195$, $z = -2.25$) and revenue ($\beta = -0.210$, $z = -1.66$) in specifications with firm–quarter fixed effects. GPT contributors become *less* likely to herd in the post-GPT period. At the sample-mean herding rate of 71.6% for EPS, the coefficient of –0.195 implies a 17.7% ($= 1 - e^{-0.195}$) reduction in the odds of herding toward consensus. For revenue, the reduction is 18.9% ($= 1 - e^{-0.210}$).[33]

Panel B partitions by contributor type and reveals that the anti-herding effect is driven by non-professionals. For EPS revisions, non-professional GPT contributors reduce their herding by 0.239 log-odds units ($z = -2.39$), while the effect for professionals is insignificant ($\beta = -0.158$, $z = -0.78$). For revenue, the contrast is sharper: non-professionals show a reduction of 0.308 log-odds units ($z = -2.18$), while the professional coefficient is effectively zero ($\beta = -0.004$, $z = -0.02$).

Together with the accuracy results, these findings indicate that the benefits documented in Section 4 concentrate largely among non-professional contributors, who lack the institutionalized analytical resources that GenAI substitutes for: their forecasts become both more accurate and more

[32] Restricting to revisions conditions on the decision to revise, which GPT adoption may itself affect: adopters may revise less often if their initial forecasts improve, or more often if revision becomes cheaper. The estimates should therefore be read as effects on the direction of revisions, conditional on revising. A difference-in-differences on the probability that a forecast is a revision (Table IA.7 in the Internet Appendix) shows that GPT-reliant contributors revise more after ChatGPT becomes available ($\beta = 0.016$, $t = 2.95$; non-professionals 0.018, $t = 2.83$; professionals unchanged), so the herding decline occurs alongside more, not fewer, revisions.

[33] The herding results are robust to within-contributor identification: linear probability models that add contributor fixed effects to the firm–quarter fixed effects yield –0.064 ($t = -2.31$) for EPS and –0.071 ($t = -2.66$) for revenue (–0.078, $t = -2.56$, for non-professionals), so the decline in herding is identified from changes within the same contributor. Contributor-level conditional logit is computationally infeasible for high-activity contributors, which motivates the linear specification.

independent in the post-GPT period, a pattern consistent with contributors using GenAI to process firm-specific financial information rather than merely to acquire the prevailing expectations.

5.2.3 Within Non-Professional Contributors: Heterogeneity by Experience

The findings thus far indicate that GPT adoption disproportionately benefits non-professional contributors, narrowing the professional/non-professional accuracy and herding gap. This pattern supports the resource margin of the framework: GenAI levels barriers rooted in resource endowments rather than substituting for human capital itself. A sharper test comes from the expertise margin (Prediction 3): because the share of GenAI's processing capacity a user captures rises with judgment, gains within the resource-homogeneous non-professional pool should concentrate among contributors with the experience to evaluate and deploy model output, while inexperienced adopters should gain little and may herd more (e.g., Brynjolfsson et al., 2025; Choi and Xie, 2026).

To test this prediction, we partition the non-professional sample by contributor experience. We classify a contributor as *Experienced* at the time of a forecast if (i) their tenure since first forecast is at least 365 days and (ii) they have submitted at least 10 cumulative forecasts. This dual criterion captures both calendar experience (exposure to a full annual reporting cycle) and hands-on practice (accumulated forecasting iterations). Both quantities are computed from the full Estimize forecast history.[34] We then re-estimate the accuracy and herding specifications separately for inexperienced and experienced non-professional contributors.[35]

The adoption gains concentrate among experienced non-professionals, while inexperienced adopters show no improvement and may herd more (Table 8). In Panel A (accuracy), for experienced non-professionals, the *GPT_Contributor* × *Post_GPT* coefficient is significantly negative for both EPS ($\beta = -0.137$, $t = -2.42$) and revenue ($\beta = -0.219$, $t = -2.08$). In contrast, the same coefficient

[34]Because experience accrues over time, contributors can migrate from the inexperienced to the experienced group within the panel, and post-GPT observations are mechanically more likely to be experienced. The experience gradient below is therefore identified from the adoption interaction estimated within each group, not from the changing composition of the groups.

[35]Because experience transitions require the full forecast history, the cohort tests are estimated on a broader panel than Table 6: all Estimize forecasts issued during 2021–2024 with nonmissing forecasts and actuals and a CRSP *permno* (440,843 forecasts, of which 259,681 by non-professionals and 181,162 by professionals), constructed without the forecast-age and post-GPT-participation screens of the accuracy sample. Contributors outside the Section 5.1.1 classification enter the control group.

is statistically indistinguishable from zero for inexperienced contributors ($\beta = 0.060$, $t = 1.36$ for EPS; $\beta = 0.116$, $t = 0.56$ for revenue). Financial professionals show no comparable experience gradient: the adoption interaction is statistically indistinguishable from zero for both experience groups (see Table IA.2 in the Internet Appendix). Panel B of Table 8 reveals a parallel pattern for herding behavior. Experienced GPT-reliant contributors significantly reduce their herding tendency (*Herding_EPS*: $\beta = -0.211$, $z = -3.45$; *Herding_REV*: $\beta = -0.236$, $z = -3.89$), whereas inexperienced contributors show no significant change in EPS herding and *increased* herding in revenue revisions ($\beta = 0.388$, $z = 3.13$).

These results locate the boundary of the resource-margin leveling, consistent with Prediction 3. First, GenAI's accuracy and independence benefits are not uniformly distributed across non-professionals: they accrue to those who have already developed the domain familiarity needed to verify, contextualize, and selectively use GenAI-generated outputs. Second, the contrasting result for inexperienced contributors (their increased revenue herding, with no significant change in EPS herding) suggests that without sufficient domain knowledge, GenAI may amplify rather than mitigate herding tendencies, perhaps because inexperienced users defer more readily to model outputs that align with prevailing consensus. Taken together, the evidence is consistent with GenAI *complementing* rather than *replacing* human capital.[36]

5.2.4 Cross-Sectional Variation: Firm Information Environment

The results thus far establish that GPT adoption improves accuracy and reduces herding, particularly for non-professionals. We next locate *where* this non-professional benefit is largest by partitioning the sample along two dimensions of the firm's information environment, each designed to isolate a distinct source of the professional/non-professional accuracy gap identified in Section 2.3.

The first dimension is *disclosure readability*. A growing literature documents that complex, hard-to-read corporate disclosures impose disproportionate processing costs on less sophisticated

[36]Experience bundles several human-capital dimensions, including innate ability, motivation, and accumulated learning; we interpret the gradient as reflecting the judgment needed to evaluate and deploy model output, but cannot isolate these components.

market participants (e.g., Li, 2008; Miller, 2010; Loughran and McDonald, 2014; Bonsall et al., 2017; Blankespoor et al., 2020). A recent study by Kimbrough et al. (2026) finds that traditional AI tools are more effective when disclosures are already readable and well-organized. GenAI, by contrast, can process dense, unstructured text directly (Eloundou et al., 2024; de Kok, 2025). If GenAI's comparative advantage lies in comprehending complex disclosures that non-professionals struggle to parse, its accuracy benefit should be largest for firms with harder-to-read filings. We capture this dimension using the average of standardized Gunning Fog, ARI, and SMOG indices from 8-K filings (Loughran and McDonald, 2014).[37]

The second dimension is *information asymmetry*. Raw analyst forecast dispersion confounds two distinct forces: fundamental uncertainty about firm value and differential access to private information (Barron et al., 1998; Zhang, 2006). To isolate the information access channel, we employ the BKLS decomposition (Barron et al., 1998) to capture the information asymmetry from analyst forecast dispersion.[38] Because GenAI can enhance how analysts process publicly available information but cannot replicate the private information that professionals obtain through management interactions and institutional networks (e.g., Green et al., 2014; Soltes, 2014), its benefit to non-professionals should be largest in low-information-asymmetry settings, where public information contains a clear signal that processing alone can extract.

We examine these predictions by partitioning the accuracy sample at the median of each moderator: the observation-level median of *Readability* and the firm–quarter median of *Info_Asymmetry*. Table 9 reports the coefficient on *GPT_Contributor* × *Post_GPT* from separate regressions, with rows for non-professionals and financial professionals.

The results reveal heterogeneity concentrated in revenue forecasts (Panel B). Among non-professionals, the revenue accuracy improvement from GPT adoption is significant for firms

[37]Because the estimates we study are quarterly, we measure readability from 8-K filings rather than annual reports: 8-Ks arrive throughout the quarter and constitute the disclosure flow contributors process when updating quarterly estimates, whereas annual-report readability varies only once a year and cannot capture variation at the quarterly frequency of our analysis.

[38]Higher values of $1-\rho$ indicate that analysts rely more heavily on private information, reflecting environments where the professional/non-professional gap stems more from differences in information *access* than processing *ability*; the construction is reported in the note to Table 9.

with harder-to-read disclosures ($\beta = -0.093$, $t = -1.90$) but absent for firms with more readable filings ($\beta = -0.007$, $t = -0.17$); in Panel A (EPS), neither readability subsample yields a significant coefficient, so the readability evidence is confined to revenue forecasts. Where it appears, the pattern is the reverse of that documented for traditional AI tools above: GenAI's benefit is largest precisely where disclosures are most difficult for humans to process on their own.

The information asymmetry results sharpen the interpretation. GPT's accuracy benefit for non-professionals is concentrated in low-information-asymmetry environments ($\beta = -0.086$, $t = -2.66$) and statistically indistinguishable from zero in high-information-asymmetry settings ($\beta = -0.044$, $t = -1.02$). Panel A (EPS) shows a directionally consistent pattern, with the non-professional coefficient significant in the low-asymmetry subsample ($\beta = -0.047$, $t = -1.95$) and indistinguishable from zero in the high-asymmetry subsample ($\beta = 0.020$, $t = 0.59$). Because the BKLS decomposition isolates information asymmetry from total uncertainty, this pattern is consistent with the processing-versus-access mechanism: when the forecasting challenge is primarily one of processing (low asymmetry), GenAI levels the playing field across professional and non-professional contributors; when it is primarily one of access (high asymmetry), the professional advantage persists.

Among financial professionals, the information asymmetry partition yields no significant effects in either subsample, consistent with the framework's implication for professionals.

5.3 Additional Tests

This subsection reports three additional tests. The first is a placebo test that randomizes the timing of the ChatGPT outages used to classify contributors; the second replaces the peer-adjusted accuracy measure with a raw forecast-error measure; the third shows that the outage effect disappears once close substitutes to ChatGPT become available, a falsification test of the identification strategy.

5.3.1 Placebo Test: Randomized Outage Timing

The outage-based classification could, in principle, capture persistent differences in forecasting activity rather than genuine reliance on ChatGPT: contributors who forecast frequently have more opportunities to appear absent during any time window, so the classification might load on activity

rather than on outage-specific behavior. Were that the case, the accuracy effect would reflect a general post-2023 pattern among active contributors rather than GenAI use.

We address this concern with a placebo test that re-runs the entire classification on pseudo-outages. For each of the 32 documented ChatGPT outages, we construct placebo windows by shifting the outage window four weeks earlier and four weeks later, preserving the day of week and time of day but relocating it to periods when ChatGPT was operational. Applying the identical classification procedure of Internet Appendix B to these placebo windows yields a placebo classification, with which we re-estimate the accuracy regression. If the outage timing carries genuine identifying content, the placebo classification should not reproduce the effect. We focus on EPS accuracy, the outcome for which the outage timing is most informative.

Table IA.4 in the Internet Appendix reports the results, mirroring the design of Table 6: Columns 1–2 re-estimate the EPS specifications under the actual classification and Columns 3–4 under the placebo classification. The actual interaction reproduces the estimates of Table 6, negative and significant for all contributors ($\beta = -0.054$, $t = -2.29$) and for non-professionals ($\beta = -0.080$, $t = -2.25$) and null for professionals. The placebo interaction is insignificant in every column (all contributors, $\beta = -0.035$, $t = -1.38$; non-professionals, $\beta = -0.030$, $t = -1.20$; professionals, $\beta = -0.014$, $t = -0.19$). Two features support the interpretation. The placebo interaction is insignificant, so the actual effect is not a mechanical consequence of the classification selecting active contributors; and professionals show no effect under either classification, so the procedure does not manufacture effects where none exist.[39]

5.3.2 Alternative Accuracy Measure: AFE

Our main accuracy tests use *PMAFE*, a within-firm–quarter peer-adjusted measure, with min–max scaled contributor-level controls. As a robustness check, we re-estimate the adoption specifications with an accuracy measure that does not depend on peer adjustment.

Table IA.5 in the Internet Appendix replaces *PMAFE* with the absolute forecast error scaled by the absolute actual (*AFE*), defined as $|\text{Forecast} - \text{Actual}|/|\text{Actual}|$ for both EPS and revenue. This

[39]The pattern is unchanged when the placebo windows are shifted eight and twelve weeks.

measure captures forecast accuracy without peer-group benchmarking. The regressions additionally include a comprehensive set of firm-level controls.

In Panel A, the coefficient on *GPT_Contributor* × *Post_GPT* is negative and significant for both EPS ($\beta = -0.039$, $t = -2.50$) and revenue ($\beta = -0.005$, $t = -2.53$) in specifications with firm fixed effects.

Panel B partitions by contributor type and confirms the heterogeneous treatment effect. For EPS, non-professional GPT-reliant contributors reduce their *AFE* by 0.049 ($t = -2.61$), with no effect for professionals ($\beta = 0.002$, $t = 0.10$). For revenue, the pattern is consistent: $\beta = -0.006$ ($t = -2.71$) for non-professionals versus $\beta = -0.005$ ($t = -1.06$) for professionals. Consistency across two distinct accuracy measures confirms that the main findings are not artifacts of the *PMAFE* construction or omitted firm characteristics.

5.3.3 Does the GPT Outage Effect Disappear After DeepSeek R1?

Our identification strategy relies on the assumption that ChatGPT outages create binding constraints for GPT-reliant contributors. This assumption is plausible during 2023–2024, when ChatGPT was the dominant GenAI platform with limited substitutes (Chatterji et al., 2025); by 2025, close substitutes such as DeepSeek R1 (DSR1; released January 20, 2025) had achieved widespread adoption (Section 3.1).[40] If contributors substitute across platforms during ChatGPT outages, the outage-based identification should weaken.

We test this directly by comparing the *During_Outage* coefficient before and after the DeepSeek R1 release. Specifically, we extend the post-GPT sample through 2025 and partition it at January 20, 2025. The pre-DSR1 sample retains the original Table 3 construction; the post-DSR1 sample includes forecasts created on or after January 20, 2025 from the extended Estimize data, restricted to the same contributors and firms. Table IA.6 in the Internet Appendix reports the results. In the pre-DSR1 sample, the outage effect is significantly positive (Panel A: $\beta = 0.087$, $t = 2.25$), in line with the aggregate finding from Table 3. In the post-DSR1 sample, the coefficient drops to

[40] DeepSeek R1 provides a sharp, well-publicized release date for the partition, but it is not the only relevant substitute: proprietary alternatives such as Anthropic's Claude and Google's Gemini also gained substantial adoption over 2025. The post-DSR1 period therefore captures the arrival of multiple close substitutes rather than the effect of DeepSeek alone.

essentially zero ($\beta = -0.004$, $t = -0.16$). The same pattern holds for non-professionals (Panel B: pre-DSR1, $\beta = 0.091$, $t = 1.90$; post-DSR1, $\beta = -0.045$, $t = -0.69$). The two periods differ sharply in treated observations (275 outage-window forecasts pre-DSR1 versus 23 post-DSR1), so the post-DSR1 test is less powerful; the full-sample post-DSR1 coefficient is nonetheless tightly estimated near zero (s.e. $= 0.025$), and its confidence interval excludes an effect as large as the pre-DSR1 estimate.

The disappearance of the outage effect after DSR1 indicates that once comparable substitutes emerged in early 2025, contributors no longer had to wait out ChatGPT downtime, eroding the identifying variation. Other 2025 changes (fewer or milder outages, shifts in platform activity) could contribute to the attenuation, although restricting the post-DSR1 sample to the same contributors and firms holds composition fixed.[41]

6 GenAI and Market Informativeness of Crowdsourced Forecasts

The preceding analyses document that GPT-reliant contributors produce more accurate and independent forecasts (Prediction 1), that these gains concentrate among non-professionals, consistent with the narrowing of the long-standing professional/non-professional accuracy gap (Prediction 2), and that within the non-professional pool they accrue to experienced adopters (Prediction 3). The final question is whether these improvements reach price discovery. Prediction 4 posits that if the non-professional consensus becomes more accurate, its forecast surprise should become a more precise measure of genuinely unexpected earnings, and the market should respond accordingly.

We test Prediction 4 by estimating the following specification at the firm–quarter level:

$$\begin{aligned} CAR[-1,1]\%_{j,t} = {} & \beta_0 Post\_GPT_t + \beta_1 Surprise\_Prof_{j,t} + \beta_2 (Surprise\_Prof_{j,t} \times Post\_GPT_t) \\ & + \beta_3 Surprise\_NonProf_{j,t} + \beta_4 (Surprise\_NonProf_{j,t} \times Post\_GPT_t) \\ & + \mathbf{X}'_{j,t}\gamma + \alpha_j + \delta_t + \varepsilon_{j,t} \end{aligned} \tag{5}$$

where $CAR[-1,1]\%_{j,t}$ is the three-day cumulative abnormal return (in percent) centered on firm j's

[41] This pattern also supports our decision to end the main sample in 2024 and informs how researchers should think about extending outage-based designs to a multi-platform GenAI ecosystem.

earnings announcement in quarter t, computed as the firm's return minus the CRSP value-weighted market return, summed over the window $[-1,+1]$. $Surprise\_Prof_{j,t}$ and $Surprise\_NonProf_{j,t}$ are earnings surprises computed from the professional and non-professional Estimize consensus forecasts, respectively, defined as reported EPS minus the group consensus forecast, scaled by the absolute consensus forecast.[42] Controls include log market capitalization, a loss indicator, book-to-market ratio, momentum, pre-announcement return, analyst coverage (I/B/E/S), ROA, log total assets, and leverage. Saturated specifications add firm fixed effects (α_j) and year–quarter fixed effects (δ_t), which subsume $Post\_GPT_t$. The coefficient of interest is β_4, which captures whether the market responds more strongly to non-professional forecast surprises after GPT became available. The estimation sample comprises the 55,997 firm–quarter announcements during 2015–2024 with at least one professional and at least one non-professional Estimize forecast, an Estimize-reported actual, and valid CRSP returns over the announcement window.[43]

Table 10 shows that the market's response to non-professional forecast surprises strengthens after ChatGPT's release, while its response to professional surprises is unchanged; the four specifications alternate controls and fixed effects. Both professional and non-professional surprise measures are strongly associated with abnormal returns throughout, confirming that Estimize consensus forecasts contain information relevant for equity pricing (e.g., Jame et al., 2016; Schafhäutle and Veenman, 2024; Farrell et al., 2022). In the most saturated specification, column (4), the baseline earnings response coefficients are 1.075 ($t = 9.03$) for professional surprises and 1.158 ($t = 9.47$) for non-professional surprises, consistent with the incremental information content documented by Jame et al. (2016).

The key finding is the asymmetric strengthening in the post-GPT period. The interaction $Surprise\_NonProf \times Post\_GPT$ is positive and significant across all four specifications, with coefficients ranging from 0.700 to 0.799 ($p < 0.05$). In column (4), the coefficient is 0.799

[42]Reported EPS is the actual earnings figure recorded by Estimize. We use the Estimize-reported actual, rather than I/B/E/S or Compustat actuals, so that forecasts and actuals are measured on the same basis: Estimize records actuals under the same earnings definition that contributors forecast, so the surprise reflects forecast error rather than definitional differences across data sources.

[43]With firm and year–quarter fixed effects, the estimation sample is 55,896 announcements because singleton groups drop.

($t = 2.35$).[44] Relative to the pre-GPT baseline of 1.158, this is a 69% increase (0.799/1.158) in the market's sensitivity to non-professional surprises, an economically large increase. In contrast, the interaction *Surprise_Prof* × *Post_GPT* is negative and statistically insignificant across all specifications (column (4): $\beta = -0.435$, $t = -1.36$). The market does not change how it prices professional forecast surprises, consistent with the earlier finding that professional forecast quality is largely unaffected by GPT adoption. This divergent pattern is consistent with Prediction 4: as the non-professional consensus becomes more accurate and independent, the market assigns greater weight to non-professional forecast surprises.[45]

Taken together with the contributor-level results, these findings connect technology adoption, individual forecast behavior, and aggregate market outcomes: the improvements in non-professional forecast quality (Sections 4 and 5.2) appear to be impounded into equity pricing, suggesting that GenAI's leveling of the resource margin can carry consequences for capital market efficiency.

7 Conclusion

This study investigates when GenAI levels the playing field in financial information production. Using crowdsourced earnings forecasts from Estimize, we characterize ChatGPT as a technology that lowers the barriers to AI-assisted investment analysis and show that its effects run along two margins. On the resource margin, GenAI substitutes for institutionalized analytical support: exploiting documented ChatGPT service outages to identify individual GPT-reliant contributors, we find that adoption makes non-professionals' forecasts both more accurate and more independent while leaving professionals largely unaffected, consistent with GenAI expanding the information-processing capacity of contributors who lack the analytical resources and

[44] Column (4) includes two correlated size proxies, log market capitalization and log total assets; within firm fixed effects their coefficients are large and opposite in sign, a standard collinearity pattern between overlapping size measures. The interaction of interest is similar in magnitude when log total assets is dropped.

[45] This result does not require market participants to directly observe GenAI adoption; rather, it is consistent with the improved information quality of the non-professional consensus. A complementary channel operates through participation: GPT-reliant non-professionals substantially increase their forecasting output after ChatGPT's release (Table 5, Panel B), and consensus informativeness increases with the size of the forecasting pool (Jame et al., 2016). We therefore read the strengthening in the earnings response coefficient as the joint effect of better and more numerous non-professional forecasts, both consequences of GenAI lowering the cost of forecast production, rather than as accuracy gains alone. Greater investor attention to crowd forecasts in the post-GPT period could also contribute.

organizational support that professionals draw on. On the expertise margin, GenAI complements rather than replaces human capital: within the non-professional pool, where resources vary little, only experienced contributors improve, whereas inexperienced adopters show no improvement and may even herd more toward the consensus. Information access bounds the gains further: cross-sectional evidence suggests that GenAI amplifies the processing of public disclosures but does not offset the private information advantage held by professionals. Lastly, over the sample period, the market's response to non-professional forecast surprises also strengthens.

Several caveats qualify these findings. First, our outage-based classification identifies contributors whose forecasting activity co-moves with ChatGPT availability, which we interpret as GPT reliance. However, contributors who use GenAI but maintain alternative workflows during outages would not be captured, biasing the classification toward the most dependent users. Second, selection into GPT adoption is endogenous: contributors who adopt GenAI may differ in unobservable dimensions such as motivation and technological aptitude. The herding results are identified from changes within the same contributor and are therefore immune to time-invariant selection (Section 5.2.2); for the accuracy results, which are identified across contributors, the parallel pre-trends and the outage-based revealed-preference design mitigate this concern, but we cannot fully rule out that unobserved heterogeneity contributes. Finally, our sample ends in 2024 for identification reasons (Section 5.3.3), yet GenAI's "jagged technological frontier" (Dell'Acqua et al., 2026) continues to advance at an unpredictable pace. As models absorb tasks that today lie beyond the frontier and require human judgment, the expertise margin we document may itself shift; the boundary conditions we identify are therefore best read as a snapshot of a rapidly evolving technology rather than a fixed feature of it.

Our study makes three contributions. First, we extend the crowdsourced forecasting literature (e.g., Jame et al., 2016; Da and Huang, 2020; Khavis and Park, 2024) by showing how a major technological shock reshapes the formation and quality of crowd predictions: GenAI narrows the accuracy gap arising from differences in institutionalized analytical resources while amplifying the returns to forecasting experience, and improves the forecast independence of experienced

adopters, the condition that makes crowd aggregation valuable (Da and Huang, 2020). Second, we advance understanding of how new technologies affect the processing and dissemination of financial information (e.g., Blankespoor et al., 2014; Blankespoor, 2019; Blankespoor et al., 2020): unlike earlier waves of technology whose benefits concentrated within institutions (Chen et al., 2022; Shanthikumar and Yoo, 2026), GenAI delivers observable information-processing benefits directly to individual contributors, concentrated among less-resourced participants. Third, and most centrally, we speak to the debate on whether GenAI levels the playing field or reinforces existing advantages. Prior work in accounting and finance finds that sophisticated investors and experienced professionals capture larger gains (Blankespoor et al., 2026; Choi and Xie, 2026) and that experienced equity research authors produce higher-quality AI-generated content than opportunistic newcomers (Bradshaw et al., 2026). By examining a setting where contributors with varying experience and resources voluntarily adopt GenAI, we identify a boundary condition: GenAI levels the playing field when the forecasting challenge is primarily one of information processing, but not when it turns on access to private information or on the judgment required to deploy the technology well. As GenAI diffuses through financial markets, its distributional consequences will turn less on who can access the technology, which is nearly universal, than on who has the judgment to use it well.

References

Abadie, A., Athey, S., Imbens, G.W., Wooldridge, J.M., 2023. When should you adjust standard errors for clustering? Q. J. Econ. 138, 1–35. https://doi.org/10.1093/qje/qjac038.

Acemoglu, D., Restrepo, P., 2018. The race between man and machine: Implications of technology for growth, factor shares, and employment. Am. Econ. Rev. 108, 1488–1542. https://doi.org/10.1257/aer.20160696.

Acemoglu, D., Restrepo, P., 2022. Tasks, automation, and the rise in U.S. wage inequality. Econometrica 90, 1973–2016. https://doi.org/10.3982/ECTA19815.

Agrawal, A., Gans, J., Goldfarb, A., 2019. Exploring the impact of artificial intelligence: Prediction versus judgment. Inf. Econ. Policy 47, 1–6. https://doi.org/10.1016/j.infoecopol.2019.05.001.

Asthana, S., Balsam, S., 2001. The effect of EDGAR on the market reaction to 10-K filings. J. Account. Public Policy 20, 349–372. https://doi.org/10.1016/S0278-4254(01)00035-7.

Autor, D., Thompson, N., 2025. Expertise. J. Eur. Econ. Assoc. 23, 1203–1271. https://doi.org/10.1093/jeea/jvaf023.

Autor, D.H., 2015. Why are there still so many jobs? The history and future of workplace automation. J. Econ. Perspect. 29, 3–30. https://doi.org/10.1257/jep.29.3.3.

Autor, D.H., Levy, F., Murnane, R.J., 2003. The skill content of recent technological change: An empirical exploration. Q. J. Econ. 118, 1279–1333. https://doi.org/10.1162/003355303322552801.

Banerjee, A.V., 1992. A simple model of herd behavior. Q. J. Econ. 107, 797–817. https://doi.org/10.2307/2118364.

Barron, O.E., Kim, O., Lim, S.C., Stevens, D.E., 1998. Using analysts' forecasts to measure properties of analysts' information environment. Account. Rev. 73, 421–433. https://doi.org/10.2308/tar-1325287.

Bertomeu, J., Lin, Y., Liu, Y., Ni, Z., 2025. The impact of generative AI on information processing: Evidence from the ban of ChatGPT in Italy. J. Account. Econ. 80, 101782. https://doi.org/10.1016/j.jacceco.2025.101782.

Beyer, A., Cohen, D.A., Lys, T.Z., Walther, B.R., 2010. The financial reporting environment: Review of the recent literature. J. Account. Econ. 50, 296–343. https://doi.org/10.1016/j.jacceco.2010.10.003.

Bhattacharya, N., Cho, Y.J., Kim, J.B., 2018. Leveling the playing field between large and small institutions: Evidence from the SEC's XBRL mandate. Account. Rev. 93, 51–71. https://doi.org/10.2308/accr-52000.

Bick, A., Blandin, A., Deming, D.J., 2026. The rapid adoption of generative AI. Manag. Sci. In press. https://doi.org/10.1287/mnsc.2025.02523.

Bikhchandani, S., Hirshleifer, D., Welch, I., 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. J. Polit. Econ. 100, 992–1026. https://doi.org/10.1086/261849.

Blankespoor, E., 2019. The impact of information processing costs on firm disclosure choice: Evidence from the XBRL mandate. J. Account. Res. 57, 919–967.

https://doi.org/10.1111/1475-679X.12268.

Blankespoor, E., Croom, J., Grant, S.M., 2026. Generative AI and investor processing of financial information. J. Account. Econ. In press. https://doi.org/10.1016/j.jacceco.2026.101908.

Blankespoor, E., deHaan, E., Marinovic, I., 2020. Disclosure processing costs, investors' information choice, and equity market outcomes: A review. J. Account. Econ. 70, 101344. https://doi.org/10.1016/j.jacceco.2020.101344.

Blankespoor, E., Miller, G.S., White, H.D., 2014. The role of dissemination in market liquidity: Evidence from firms' use of Twitter. Account. Rev. 89, 79–112. https://doi.org/10.2308/accr-50576.

Bonsall, S.B., Leone, A.J., Miller, B.P., Rennekamp, K., 2017. A plain English measure of financial reporting readability. J. Account. Econ. 63, 329–357. https://doi.org/10.1016/j.jacceco.2017.03.002.

Bradshaw, M.T., Drake, M.S., Myers, J.N., Myers, L.A., 2012. A re-examination of analysts' superiority over time-series forecasts of annual earnings. Rev. Account. Stud. 17, 944–968. https://doi.org/10.1007/s11142-012-9185-8.

Bradshaw, M.T., Ma, C., Yost, B.P., Zou, Y., 2026. Generative AI use by capital market information intermediaries: Evidence from Seeking Alpha. J. Account. Res. 64, 1233–1286. https://doi.org/10.1111/1475-679X.70053.

Brown, L.D., Call, A.C., Clement, M.B., Sharp, N.Y., 2015. Inside the "black box" of sell-side financial analysts. J. Account. Res. 53, 1–47. https://doi.org/10.1111/1475-679X.12067.

Brown, L.D., Khavis, J., Park, H.U., 2024. Nudging towards better earnings forecasts. Working paper, available at: https://ssrn.com/abstract=4202792.

Brynjolfsson, E., Li, D., Raymond, L.R., 2025. Generative AI at work. Q. J. Econ. 140, 889–938. https://doi.org/10.1093/qje/qjae044.

Bushee, B.J., Gow, I.D., Taylor, D.J., 2018. Linguistic complexity in firm disclosures: Obfuscation or information? J. Account. Res. 56, 85–121. https://doi.org/10.1111/1475-679X.12179.

Cao, S., Jiang, W., Wang, J.L., Yang, B., 2024. From man vs. machine to man + machine: The art and AI of stock analyses. J. Financ. Econ. 160, 103910. https://doi.org/10.1016/j.jfineco.2024.103910.

Chatterji, A., Cunningham, T., Deming, D.J., Hitzig, Z., Ong, C., Shan, C.Y., Wadman, K., 2025. How people use ChatGPT. NBER Working Paper No. 34255, available at: https://www.nber.org/papers/w34255.

Chen, X., Cho, Y.J., Dou, Y., Lev, B., 2022. Predicting future earnings changes using machine learning and detailed financial data. J. Account. Res. 60, 467–515. https://doi.org/10.1111/1475-679X.12429.

Cheng, Q., Lin, P., Zhao, Y., 2025. Does generative AI facilitate investor trading? Evidence from ChatGPT outages. J. Account. Econ. 80, 101821. https://doi.org/10.1016/j.jacceco.2025.101821.

Cho, Y.H., Huang, A.H., Pacelli, J., Rennekamp, K.M., 2026. Generative AI and investor processing of financial media. Working paper (Harvard Business School Working Paper No.

26-074), available at: https://ssrn.com/abstract=6479459.

Choi, J.H., Xie, C.L., 2026. Human + AI in accounting: Early evidence from the field. J. Account. Res. 64, 1333–1373. https://doi.org/10.1111/1475-679X.70052.

Clement, M.B., 1999. Analyst forecast accuracy: Do ability, resources, and portfolio complexity matter? J. Account. Econ. 27, 285–303. https://doi.org/10.1016/S0165-4101(99)00013-0.

Clement, M.B., Tse, S.Y., 2005. Financial analyst characteristics and herding behavior in forecasting. J. Finance 60, 307–341. https://doi.org/10.1111/j.1540-6261.2005.00731.x.

Cowen, A., Groysberg, B., Healy, P., 2006. Which types of analyst firms are more optimistic? J. Account. Econ. 41, 119–146. https://doi.org/10.1016/j.jacceco.2005.09.001.

Croom, J., 2026. Interactivity and illusions of ability: How using generative AI affects investor judgments. J. Account. Res. 64, 681–719. https://doi.org/10.1111/1475-679X.70017.

Da, Z., Huang, X., 2020. Harnessing the wisdom of crowds. Manag. Sci. 66, 1847–1867. https://doi.org/10.1287/mnsc.2019.3294.

deHaan, E., Lee, C.M., Loh, W.T., 2025. Local-thinking bias. Account. Rev. 100, 87–112. https://doi.org/10.2308/TAR-2024-0453.

Dell'Acqua, F., McFowland, E., Mollick, E.R., Lifshitz, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., Lakhani, K.R., 2026. Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. Organ. Sci. 37, 403–423. https://doi.org/10.1287/orsc.2025.21838.

Diether, K.B., Malloy, C.J., Scherbina, A., 2002. Differences of opinion and the cross section of stock returns. J. Finance 57, 2113–2141. https://doi.org/10.1111/0022-1082.00490.

Dong, Y., Li, O.Z., Lin, Y., Ni, C., 2016. Does information-processing cost affect firm-specific information acquisition? Evidence from XBRL adoption. J. Financ. Quant. Anal. 51, 435–462. https://doi.org/10.1017/S0022109016000235.

Drake, M.S., Roulstone, D.T., Thornock, J.R., 2015. The determinants and consequences of information acquisition via EDGAR. Contemp. Account. Res. 32, 1128–1161. https://doi.org/10.1111/1911-3846.12119.

Eloundou, T., Manning, S., Mishkin, P., Rock, D., 2024. GPTs are GPTs: Labor market impact potential of large language models. Science 384, 1306–1308. https://doi.org/10.1126/science.adj0998.

Farrell, M., Green, T.C., Jame, R., Markov, S., 2022. The democratization of investment research and the informativeness of retail investor trading. J. Financ. Econ. 145, 616–641. https://doi.org/10.1016/j.jfineco.2021.07.018.

Fügener, A., Walzner, D.D., Gupta, A., 2026. Roles of artificial intelligence in collaboration with humans: Automation, augmentation, and the future of work. Manag. Sci. 72, 538–557. https://doi.org/10.1287/mnsc.2024.05684.

Green, T.C., Jame, R., Markov, S., Subasi, M., 2014. Access to management and the informativeness of analyst research. J. Financ. Econ. 114, 239–259. https://doi.org/10.1016/j.jfineco.2014.07.003.

Hirshleifer, D., Levi, Y., Lourie, B., Teoh, S.H., 2019. Decision fatigue and heuristic analyst

forecasts. J. Financ. Econ. 133, 83–98. https://doi.org/10.1016/j.jfineco.2019.01.005.

Hong, H., Kubik, J.D., 2003. Analyzing the analysts: Career concerns and biased earnings forecasts. J. Finance 58, 313–351. https://doi.org/10.1111/1540-6261.00526.

Ide, E., Talamàs, E., 2025. Artificial intelligence in the knowledge economy. J. Polit. Econ. 133, 3762–3800. https://doi.org/10.1086/737233.

Jacob, J., Lys, T.Z., Neale, M.A., 1999. Expertise in forecasting performance of security analysts. J. Account. Econ. 28, 51–82. https://doi.org/10.1016/s0165-4101(99)00016-6.

Jame, R., Johnston, R., Markov, S., Wolfe, M.C., 2016. The value of crowdsourced earnings forecasts. J. Account. Res. 54, 1077–1110. https://doi.org/10.1111/1475-679X.12121.

Jame, R., Markov, S., Wolfe, M.C., 2022. Can FinTech competition improve sell-side research quality? Account. Rev. 97, 287–316. https://doi.org/10.2308/TAR-2019-0266.

Khavis, J., Park, H.U., 2024. Collaborating with data aggregators and the Estimize.com setting. J. Financ. Report. 9, 71–101. https://doi.org/10.2308/JFR-2022-021.

Kimbrough, M., Liu, L.Y., Subasi, M., 2026. AI-powered analysts. Working paper, available at: https://ssrn.com/abstract=7176461.

de Kok, T., 2025. ChatGPT for textual analysis? How to use generative LLMs in accounting research. Manag. Sci. 71, 7888–7906. https://doi.org/10.1287/mnsc.2023.03253.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. J. Account. Econ. 45, 221–247. https://doi.org/10.1016/j.jacceco.2008.02.003.

Logg, J.M., Minson, J.A., Moore, D.A., 2019. Algorithm appreciation: People prefer algorithmic to human judgment. Organ. Behav. Hum. Decis. Process. 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005.

Loughran, T., McDonald, B., 2014. Measuring readability in financial disclosures. J. Finance 69, 1643–1671. https://doi.org/10.1111/jofi.12162.

Mikhail, M.B., Walther, B.R., Willis, R.H., 1997. Do security analysts improve their performance with experience? J. Account. Res. 35, 131–157. https://doi.org/10.2307/2491458.

Miller, B.P., 2010. The effects of reporting complexity on small and large investor trading. Account. Rev. 85, 2107–2143. https://doi.org/10.2308/accr.00000001.

Noy, S., Zhang, W., 2023. Experimental evidence on the productivity effects of generative artificial intelligence. Science 381, 187–192. https://doi.org/10.1126/science.adh2586.

Otis, N.G., Clarke, R.P., Delecourt, S., Holtz, D., Koning, R., 2026. The uneven impact of generative artificial intelligence on entrepreneurial performance: Evidence from a field experiment in Kenya. Manag. Sci. In press. https://doi.org/10.1287/mnsc.2024.06909.

Rambachan, A., Roth, J., 2023. A more credible approach to parallel trends. Rev. Econ. Stud. 90, 2555–2591. https://doi.org/10.1093/restud/rdad018.

Schafhäutle, S.G., Veenman, D., 2024. Crowdsourced forecasts and the market reaction to earnings announcement news. Account. Rev. 99, 421–456. https://doi.org/10.2308/tar-2021-0055.

Scharfstein, D.S., Stein, J.C., 1990. Herd behavior and investment. Am. Econ. Rev. 80, 465–479. https://www.jstor.org/stable/2006678.

Shanthikumar, D., Yoo, I.S., 2026. Beyond automation: AI and the human value of sell-side analysts. Working paper, available at: https://ssrn.com/abstract=5040339.

Soltes, E., 2014. Private interaction between firm management and sell-side analysts. J. Account. Res. 52, 245–272. https://doi.org/10.1111/1475-679X.12037.

Surowiecki, J., 2004. The Wisdom of Crowds. Doubleday, New York.

Trueman, B., 1994. Analyst forecasts and herding behavior. Rev. Financ. Stud. 7, 97–124. https://doi.org/10.1093/rfs/7.1.97.

Welch, I., 2000. Herding among security analysts. J. Financ. Econ. 58, 369–396. https://doi.org/10.1016/S0304-405X(00)00076-3.

Xue, J., Zhang, Q., Zhu, W., 2025. Generative AI for analysts. Working paper, available at: https://ssrn.com/abstract=5821142.

Zhang, X.F., 2006. Information uncertainty and analyst forecast behavior. Contemp. Account. Res. 23, 565–590. https://doi.org/10.1506/92cb-p8g9-2a31-pv0r.

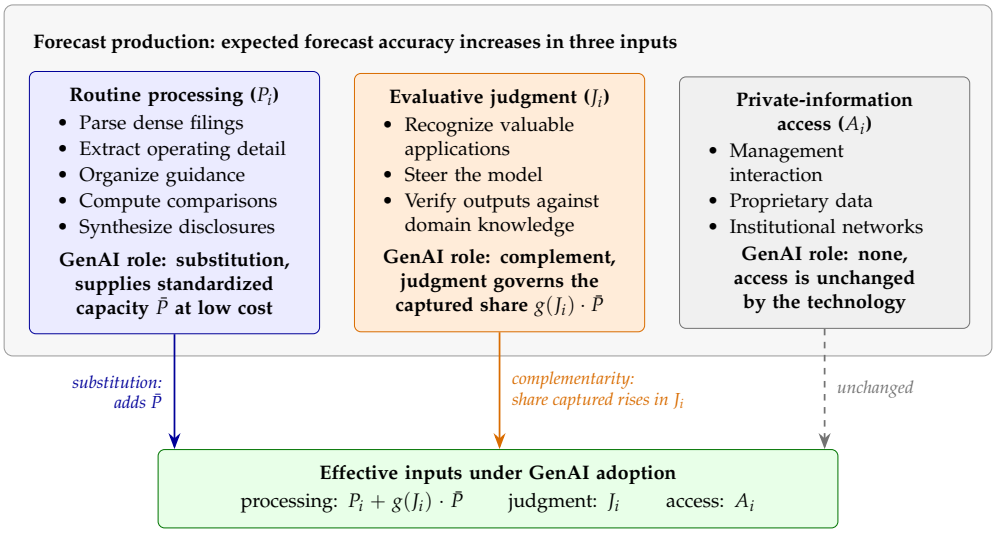
Appendix A: Variable Definitions

Variable	Definition	Source
Panel A: Dependent Variables		
$PMAFE_{i,j,t}$	Proportional mean absolute forecast error in the full-sample tests, computed from EPS forecasts as defined for *PMAFE_EPS* below; the suffix is dropped because the full-sample analyses use EPS forecasts only.	Estimize
$PMAFE\_EPS_{i,j,t}$	Proportional mean absolute forecast error for EPS, defined as the difference between contributor i's absolute forecast error (*AFE*) and the mean *AFE* for firm j in quarter t, scaled by the mean *AFE* for firm j in quarter t. *AFE* is computed as the absolute difference between the analyst's EPS forecast and actual EPS.	Estimize
$PMAFE\_REV_{i,j,t}$	Proportional mean absolute forecast error for revenue. Computed analogously to *PMAFE_EPS* using revenue forecasts.	Estimize
$AFE\_EPS_{i,j,t}$, $AFE\_REV_{i,j,t}$	Absolute forecast error scaled by the absolute actual, |Forecast − Actual|/|Actual|, for EPS and revenue, respectively.	Estimize
$Frequency_{i,j,t}$	Number of forecasts submitted by contributor i on firm j within quarter t, min–max scaled within each firm–quarter group.	Estimize
$Revision_{i,j,t}$	Indicator equal to one if contributor i's forecast for firm j in quarter t is not the contributor's first submission for that firm–quarter.	Estimize
$Herding\_EPS_{i,j,t}$	Indicator equal to one if contributor i's EPS forecast revision for firm j in quarter t moves toward the prevailing consensus, and zero otherwise. Defined only for revision forecasts (not the first submission within a firm–quarter).	Estimize
$Herding\_REV_{i,j,t}$	Indicator equal to one if contributor i's revenue forecast revision moves toward the prevailing consensus. Defined analogously to *Herding_EPS*.	Estimize
$CAR\ [-1, +1]\ (\%)_{j,t}$	Cumulative abnormal return for firm j over the 3-day window centered on the earnings announcement date of quarter t (day −1 to day +1), expressed in percent.	CRSP
Panel B: Key Independent Variables		
$GPT\_Contributor_i$	Indicator equal to one if Estimize contributor i is classified as GPT-reliant based on forecasting behavior during GPT service outages, and zero otherwise. See Internet Appendix B for identification details.	Estimize
$Post\_GPT_t$	Indicator equal to one if the forecast was created in 2023 or later (after the public release of ChatGPT in November 2022), and zero otherwise.	Estimize
$Non\_Professional_i$	Indicator equal to one if contributor i is classified as a non-professional based on Estimize user bio information, and zero otherwise (financial professional).	Estimize
$During\_Outage_t$	Indicator equal to one if the forecast was submitted during a documented GPT service outage, and zero otherwise.	Estimize
$Experienced_{i,t}$	Indicator equal to one if, at the time of the forecast, at least 365 days have elapsed since contributor i's first Estimize forecast and the contributor has submitted at least 10 cumulative forecasts; both quantities are computed from the full Estimize forecast history.	Estimize

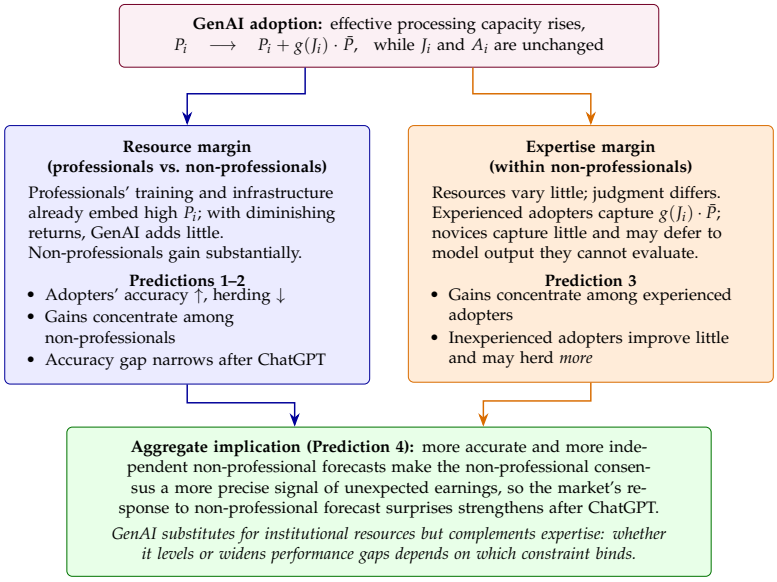
Variable	Definition	Source
$Surprise\_Prof_{j,t}$	Earnings surprise from financial professional consensus forecasts for firm j in quarter t: reported EPS minus the mean of EPS forecasts made by financial professional analysts on the firm, scaled by absolute consensus forecast.	Estimize; CRSP
$Surprise\_NonProf_{j,t}$	Earnings surprise from non-professional consensus forecasts for firm j in quarter t: reported EPS minus the mean of EPS forecasts made by non-professional analysts on the firm, scaled by absolute consensus forecast.	Estimize; CRSP
Panel C: Contributor-Level Controls		
$Experience_{i,t}$	Years since contributor i's first forecast on Estimize. Min–max scaled within firm–quarter in accuracy/herding.	Estimize
$Effort_{i,j,t}$	Number of forecasts by contributor i for firm j in quarter t. Min–max scaled or log-transformed.	Estimize
$Forecast\_Age_{i,j,t}$	Days between contributor i's forecast and firm j's earnings announcement in quarter t. Min–max scaled or raw.	Estimize
$Mean\_Forecast\_Age_{i,j,t}$	Average of *Forecast_Age* across contributor i's forecasts for firm j in quarter t; used in the contributor–firm–quarter frequency tests.	Estimize
$Firm\_Followed_{i,t}$	Log of total unique firms contributor i has covered in quarter t.	Estimize
$Estimize\_Coverage_{j,t}$	Log of unique Estimize contributors covering firm j in fiscal quarter t; used in the contributor-level tests.	Estimize
$Analyst\_Coverage_{j,t}$	Log of the number of I/B/E/S analysts covering firm j in quarter t; used in the market reaction tests.	I/B/E/S
Panel D: Firm-Level Controls		
$Log\_Mktcap_{j,t}$	Log of market capitalization of firm j (stock price $\times$ shares outstanding).	Compustat
$Loss_{j,t}$	Indicator equal to one if firm j reports negative net income in quarter t, and zero otherwise.	Compustat
$Book\_to\_Market_{j,t}$	Book value of equity divided by market capitalization for firm j in quarter t.	Compustat
$Momentum_{j,t}$	Cumulative stock return of firm j over the prior 12 months.	CRSP
$Pre\_Return_{j,t}$	Buy-and-hold return of firm j over the four trading days ending the day before the earnings announcement (days –4 to –1).	CRSP
$Size_{j,t}$	Log of total assets of firm j in quarter t.	Compustat
$Leverage_{j,t}$	Total liabilities divided by total assets for firm j in quarter t.	Compustat
$ROA_{j,t}$	Net income divided by total assets for firm j in quarter t (return on assets).	Compustat
Panel E: Cross-Sectional Moderators		
$Readability_{j,t}$	Average of standardized Gunning Fog, ARI, and SMOG indices from 8-K filings of firm j prior to the earnings announcement. Higher = harder to read.	SEC Analytics Suite
$Info\_Asymmetry_{j,t}$	Information asymmetry component from the Barron et al. (1998) decomposition of I/B/E/S analyst forecast properties: $1-\rho = D/(D+N\times SE)$, where D is forecast variance, SE is squared consensus forecast error, and N is analyst coverage. Higher values indicate greater reliance on private information.	I/B/E/S

FIGURE 1. Theoretical Framework: GenAI in Forecast Production

Panel A: How GenAI Enters Forecast Production

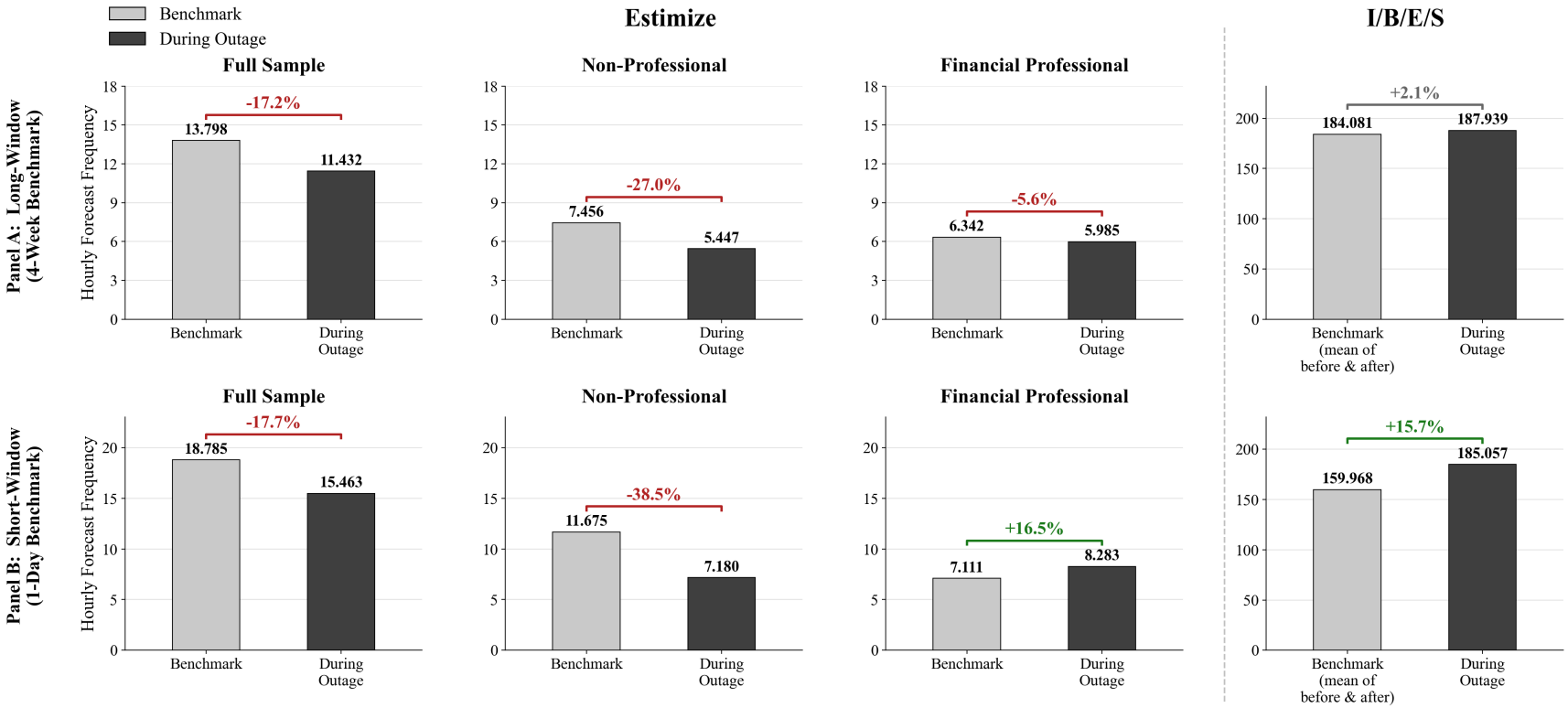


Panel B: The Two Margins and Testable Predictions



This figure summarizes the framework developed in Section 2.4. **Panel A** depicts forecast production as combining three inputs: routine processing capacity (P_i), evaluative judgment (J_i), and private-information access (A_i). GenAI supplies standardized processing capacity $\bar{P}$ at low cost (substitution), the share of that capacity a user captures, $g(J_i) \cdot \bar{P}$, rises with judgment (complementarity), and access is unchanged by the technology. **Panel B** traces the predictions that follow along the two margins of heterogeneity: across resource endowments (professionals versus non-professionals), adoption compresses performance gaps; within the resource-homogeneous non-professional pool, gains concentrate among experienced contributors. Predictions 1–4 correspond to the summary at the end of Section 2.4.

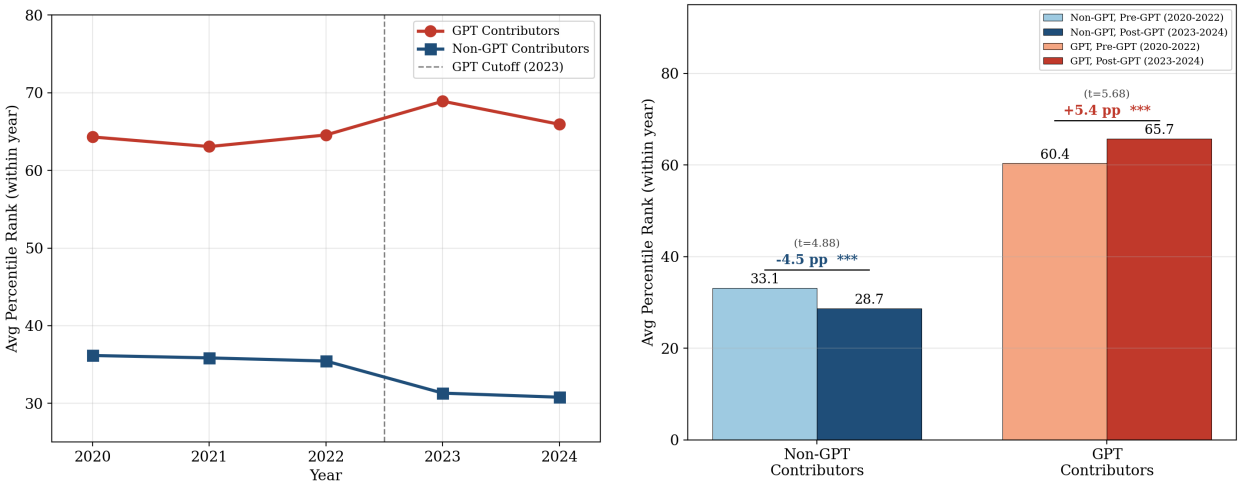
FIGURE 2. GPT Outage and Forecast Issuance Frequency



This figure compares the hourly forecast issuance frequency during GPT outage hours with that during benchmark hours. The left three columns present results for the Estimize sample: the full sample, non-professional contributors, and financial professional contributors. The right column presents results for the I/B/E/S sample. Panel A uses a long-window benchmark, defined as the same time window four weeks before and after the outage; the outage window includes the hour after resolution, and rates are pooled across outages. Panel B uses a short-window benchmark, defined as the same time window one day before and after the outage; rates are computed for each outage and averaged across events. All values are standardized to an hourly frequency. Percentage changes between outage and benchmark periods are shown above each pair of bars.

FIGURE 3. GPT Contributors and Relative Forecasting Productivity

Panel A: Relative Productivity Over Time **Panel B: Pre- vs. Post-GPT**



This figure illustrates the relative forecasting productivity of GPT contributors compared to non-GPT contributors around ChatGPT's release (November 2022); the post-GPT period begins in January 2023. **Panel A** plots the annual time series for each group from 2020 to 2024, with the dashed vertical line marking the start of the post-GPT period (2023). GPT contributors consistently outperform non-GPT contributors across the sample period, and the gap widens noticeably after GPT's release. **Panel B** presents a pre- (2020–2022) versus post-GPT (2023–2024) comparison for each group. Among non-GPT contributors, average productivity declines from 33.1 to 28.7 percent after GPT's release ($t = 4.88$, $p < 0.01$). In contrast, GPT contributors experience a significant increase from 60.4 to 65.7 percent ($t = 5.68$, $p < 0.01$). The difference-in-differences estimate of +9.83 percent ($t = 7.48$, $p < 0.001$) indicates that GPT contributors improve their relative productivity significantly more than non-GPT contributors following ChatGPT's release, consistent with GPT contributors leveraging generative AI tools to increase their forecasting output.

TABLE 1. Sample Selection

Panel A: Full Sample

Procedure	Forecasts	Contributors
Full Estimize sample (2015–2024)	1,872,609	113,097
Less: observations with missing values	(479)	(15)
Subtotal	1,872,130	113,082
Less: forecast age outside [0, 90] days before earnings announcement	(514,004)	(2,341)
Subtotal	1,358,126	110,741
Less: missing firm identifiers in CRSP or Compustat	(427,928)	(27,992)
Subtotal	930,198	82,749
Less: no within-group variation in firm × year-quarter	(1,270)	(23)
Full sample	**928,928**	**82,726**

Panel B: Contributor-Level Test Samples

Procedure	Forecasts	Contributors
Full sample (from Panel A)	928,928	82,726
Less: forecasts outside 2021–2024	(705,872)	(70,516)
Subtotal	223,056	12,210
Less: contributors without forecasts in the post-GPT period (2023+)	(29,472)	(8,334)
Accuracy sample	**193,584**	**3,876**
Less: non-revision forecasts	(133,910)	(2,542)
Herding sample	**59,674**	**1,334**

This table presents the sample selection process. **Panel A** presents the construction of the full sample from the universe of Estimize EPS forecasts: observations with missing values are excluded, forecasts must be issued within 90 days of the earnings announcement, firms must match to CRSP or Compustat identifiers, and each firm–year-quarter cell must exhibit within-group variation to construct the proportional mean absolute forecast error (*PMAFE*). **Panel B** presents the construction of the subsamples used in the contributor-level tests: forecasts issued during 2021–2024 by contributors with at least one forecast in the post-GPT period (2023 or later), which permits the outage-based classification of Internet Appendix B; the herding sample further restricts to revision forecasts, excluding the first submission by a contributor within a firm–quarter. Parenthetical values indicate observations removed at each step.

TABLE 2. Descriptive Statistics

Panel A: Full Sample (2015–2024, Forecast-Level, N=928,928)

Variable	N	Mean	Std	Min	P25	Median	P75	Max
PMAFE	928,928	−0.013	0.613	−1.000	−0.365	−0.038	0.204	2.547
Non_Professional	928,928	0.645	0.479	0.000	0.000	1.000	1.000	1.000
Post_GPT	928,928	0.104	0.305	0.000	0.000	0.000	0.000	1.000
During_Outage‡	96,238	0.003	0.053	0.000	0.000	0.000	0.000	1.000
Forecast_Age	928,928	0.239	0.337	0.000	0.000	0.059	0.328	1.000
Experience	928,928	0.395	0.356	0.000	0.032	0.323	0.699	1.000
Effort	928,928	0.263	0.392	0.000	0.000	0.000	0.500	1.000
Firm_Followed	928,928	0.222	0.348	0.000	0.003	0.039	0.210	1.000

Panel B: Frequency Sample (2015–2024, Contributor–Firm–Quarter Level, N=289,383)

Variable	N	Mean	Std	Min	P25	Median	P75	Max
Frequency	289,383	0.221	0.371	0.000	0.000	0.000	0.431	1.000
GPT_Contributor	289,383	0.872	0.334	0.000	1.000	1.000	1.000	1.000
Post_GPT	289,383	0.223	0.416	0.000	0.000	0.000	0.000	1.000
Non_Professional	289,383	0.666	0.472	0.000	0.000	1.000	1.000	1.000
Experience	289,383	0.484	0.337	0.000	0.169	0.490	0.773	1.000
Mean_Forecast_Age	289,383	0.221	0.313	0.000	0.011	0.061	0.311	1.000
Firm_Followed	289,383	0.486	0.315	0.000	0.237	0.488	0.715	1.000
Estimize_Coverage	289,383	3.426	1.187	1.099	2.565	3.332	4.190	6.232

Panel C: Accuracy and Herding Sample (2021–2024, Forecast-Level, N=193,584)

Variable	N	Mean	Std	Min	P25	Median	P75	Max
PMAFE_EPS	193,558	−0.026	0.610	−1.000	−0.400	−0.053	0.206	2.547
PMAFE_REV	193,571	−0.070	0.701	−0.991	−0.537	−0.129	0.171	3.334
Herding_EPS†	59,674	0.716	0.451	0.000	0.000	1.000	1.000	1.000
Herding_REV†	59,674	0.699	0.459	0.000	0.000	1.000	1.000	1.000
GPT_Contributor	193,584	0.917	0.276	0.000	1.000	1.000	1.000	1.000
Post_GPT	193,584	0.469	0.499	0.000	0.000	0.000	1.000	1.000
Non_Professional	193,584	0.671	0.470	0.000	0.000	1.000	1.000	1.000
Effort	193,584	0.417	0.432	0.000	0.000	0.356	1.000	1.000
Experience	193,584	0.536	0.345	0.000	0.217	0.603	0.836	1.000
Forecast_Age	193,584	0.243	0.341	0.000	0.011	0.066	0.318	1.000
Firm_Followed	193,584	0.454	0.311	0.000	0.209	0.446	0.675	1.000
Estimize_Coverage	193,584	3.016	1.133	1.099	2.197	2.833	3.714	5.557

Panel D: Market Reaction Sample (2015–2024, Firm–Quarter Level, N=55,997)

Variable	N	Mean	Std	Min	P25	Median	P75	Max
CAR [-1, +1] (%)	55,997	0.167	8.774	−26.386	−4.427	0.143	4.798	26.096
Surprise_Prof	55,997	0.056	0.802	−3.917	−0.060	0.029	0.164	4.000
Surprise_NonProf	55,997	0.039	0.802	−4.123	−0.064	0.027	0.157	3.812
Post_GPT	55,997	0.145	0.352	0.000	0.000	0.000	0.000	1.000
Log_Mktcap	55,997	8.314	1.690	4.582	7.136	8.207	9.440	12.483
Loss	55,997	0.252	0.434	0.000	0.000	0.000	1.000	1.000

(continued on next page)

Table 2 (continued from the previous page)

Panel D: Market Reaction Sample (continued)

Variable	N	Mean	Std	Min	P25	Median	P75	Max
Book_to_Market	55,997	0.468	0.529	–0.472	0.165	0.341	0.608	3.339
Momentum	55,997	0.124	0.448	–0.722	–0.144	0.072	0.306	2.010
Pre_Return	55,997	0.003	0.047	–0.135	–0.021	0.003	0.026	0.152
Analyst_Coverage	55,997	2.141	0.767	0.000	1.609	2.197	2.708	3.497
ROA	55,997	0.006	0.036	–0.167	–0.000	0.010	0.022	0.089
Size	55,997	8.240	1.729	4.685	7.011	8.089	9.334	12.812
Leverage	55,997	0.598	0.250	0.098	0.433	0.592	0.746	1.438

This table presents summary statistics for the four estimation samples. **Panel A** reports the forecast-level full sample used in the difference-in-differences and outage analyses (2015–2024); ‡*During_Outage* is defined only for forecasts in the post-GPT period (2023–2024), hence the smaller count. **Panel B** reports the contributor–firm–quarter frequency sample (2015–2024). **Panel C** reports the forecast-level accuracy and herding sample (2021–2024); †the herding variables are defined only for revision forecasts. The sample means of *GPT_Contributor* in Panels B and C (0.872 and 0.917) far exceed the contributor-level share of GPT-reliant contributors (986 of 3,876, or 25.4%; Table IA.1) because GPT-reliant contributors forecast far more actively than non-GPT contributors. **Panel D** reports the firm–quarter market reaction sample. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A.

TABLE 3. GPT Outage and Forecast Accuracy (Post-GPT)

Panel A: All Contributors

Dependent Var:	(1) *PMAFE*	(2) *PMAFE*
During_Outage	0.084**	0.091**
	(2.117)	(2.269)
Forecast_Age	0.244***	0.238***
	(21.244)	(19.507)
Experience	−0.043***	−0.042***
	(−5.282)	(−4.408)
Effort	−0.042***	−0.043***
	(−4.589)	(−4.494)
Firm_Followed	−0.086***	−0.080***
	(−12.770)	(−9.954)
Adj. R^2	0.030	0.025
Observations	96,238	96,217
Firm FE	No	Yes
DOW × Hour FE	No	Yes

Panel B: Financial Professionals vs. Non-Professionals

Dependent Var:	(1) *PMAFE*	(2) *PMAFE*
Sample split:	Fin. Prof.	Non-Prof.
During_Outage	0.091	0.098*
	(1.277)	(1.925)
Adj. R^2	0.027	0.021
Observations	26,082	69,913
Controls	Yes	Yes
Firm FE	Yes	Yes
DOW × Hour FE	Yes	Yes

This table presents OLS estimates of Equation (2). The sample is restricted to the post-GPT period (2023–2024), and the dependent variable is *PMAFE*. *During_Outage* equals one if the forecast is submitted during a documented major ChatGPT service outage (Internet Appendix A). Column (2) and the Panel B subsamples drop singleton fixed-effects groups, so their counts fall below the 96,238 forecasts in Column (1). **Panel A** reports full-sample estimates. **Panel B** partitions the sample by contributor type. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. *t*-statistics based on standard errors clustered by firm are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE 4. Forecast Accuracy Dynamics in Estimize: Full Sample Exploration

Dependent Var:	(1)	(2)	(3)	(4)
		PMAFE		
Non_Professional (β_1)	0.008***	0.004*	0.011***	0.008***
	(4.157)	(1.954)	(5.137)	(3.672)
Post_GPT	0.041***	0.009*	0.043***	0.010*
	(8.190)	(1.759)	(8.490)	(1.935)
Non_Professional × *Post_GPT* (β_3)	−0.056***	−0.012*	−0.060***	−0.013*
	(−8.549)	(−1.950)	(−8.950)	(−1.956)
Forecast_Age		0.265***		0.266***
		(32.818)		(32.042)
Experience		−0.074***		−0.080***
		(−26.699)		(−28.526)
Effort		−0.009***		−0.012***
		(−2.638)		(−3.811)
Firm_Followed		−0.056***		−0.067***
		(−21.255)		(−24.694)
Constant	−0.019***	−0.035***	−0.020***	−0.033***
	(−12.063)	(−8.154)	(−14.540)	(−9.931)
Test: $\beta_1 + \beta_3 = 0$	−0.048***	−0.008	−0.049***	−0.005
Adj. R^2	0.000	0.027	0.000	0.028
Observations	928,928	928,928	928,887	928,887
Controls	No	Yes	No	Yes
Firm FE	No	No	Yes	Yes

This table presents OLS estimates of Equation (3). The sample includes Estimize EPS forecasts over 2015–2024, and the dependent variable is *PMAFE*. In the fixed-effects specifications, a small number of singleton observations are dropped. *Non_Professional* equals one for self-identified non-professional contributors, and *Post_GPT* equals one for forecasts created in 2023 or later. The row Test: $\beta_1 + \beta_3 = 0$ reports the linear combination of the *Non_Professional* and interaction coefficients (the post-GPT accuracy gap between non-professional and professional contributors), with significance based on a Wald test. The control variables are those in Table 3. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. t-statistics based on standard errors clustered by firm are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE 5. Validation: GPT Reliance Measure and Post-GPT Productivity

Panel A: All Contributors

Dependent Var:	(1) *Frequency*	(2) *Frequency*
GPT_Contributor	0.087***	0.074***
	(28.704)	(23.492)
Post_GPT	−0.061***	−0.071***
	(−9.143)	(−11.900)
GPT_Contributor × *Post_GPT*	0.032***	0.046***
	(4.754)	(7.320)
Experience	0.151***	0.152***
	(44.006)	(40.835)
Mean_Forecast_Age	0.063***	0.030***
	(12.872)	(5.778)
Firm_Followed	−0.167***	−0.184***
	(−27.717)	(−32.983)
Constant	0.147***	0.172***
	(30.565)	(54.208)
Adj. R^2	0.037	0.054
Observations	289,383	289,383
Controls	Yes	Yes
Firm FE	No	Yes

Panel B: Financial Professionals vs. Non-Professionals

Dependent Var:	(1) *Frequency*	(2) *Frequency*
Sample split:	Fin. Prof.	Non-Prof.
GPT_Contributor × *Post_GPT*	−0.006	0.054***
	(−0.364)	(7.945)
Adj. R^2	0.087	0.071
Observations	96,373	192,751
Controls	Yes	Yes
Firm FE	Yes	Yes

This table reports OLS difference-in-differences estimates of forecasting frequency, interacting *GPT_Contributor* with *Post_GPT*. The sample spans 2015–2024 at the contributor–firm–quarter level and is restricted to contributors classified by the outage-based procedure of Section 5.1; the dependent variable is *Frequency*, min–max scaled within firm–quarter. The control variables are *Experience*, *Mean_Forecast_Age*, and *Firm_Followed*. **Panel A** reports full-sample estimates. **Panel B** partitions the sample by contributor type. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. *t*-statistics based on standard errors clustered by firm are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE 6. GPT Adoption and Forecast Accuracy

Panel A: All Contributors

Dependent Var:	(1) *PMAFE_EPS*	(2) *PMAFE_EPS*	(3) *PMAFE_REV*	(4) *PMAFE_REV*
GPT_Contributor	0.010	−0.017	−0.045	−0.089***
	(0.589)	(−0.862)	(−1.432)	(−2.672)
Post_GPT	0.052**	0.050**	0.155***	0.149**
	(2.439)	(2.144)	(2.640)	(2.535)
GPT_Contributor × *Post_GPT*	−0.051**	−0.054**	−0.143**	−0.143**
	(−2.385)	(−2.291)	(−2.448)	(−2.444)
Effort	0.015*	0.003	0.010	−0.011
	(1.797)	(0.413)	(0.996)	(−1.062)
Experience	−0.032***	−0.035***	−0.028*	−0.030*
	(−3.411)	(−3.961)	(−1.694)	(−1.749)
Forecast_Age	0.188***	0.171***	0.335***	0.313***
	(18.509)	(16.959)	(15.655)	(14.763)
Firm_Followed	−0.089***	−0.104***	−0.112***	−0.135***
	(−8.828)	(−11.070)	(−7.050)	(−9.430)
Constant	−0.245***	−0.202***	−0.325***	−0.256***
	(−11.546)	(−10.573)	(−9.844)	(−8.809)
Adj. R^2	0.003	0.006	0.009	0.017
Observations	193,558	193,556	193,571	193,569
Firm FE	No	Yes	No	Yes

Panel B: Financial Professionals vs. Non-Professionals

Dependent Var:	(1) *PMAFE_EPS*	(2) *PMAFE_EPS*	(3) *PMAFE_REV*	(4) *PMAFE_REV*
Sample split:	Fin. Prof.	Non-Prof.	Fin. Prof.	Non-Prof.
GPT_Contributor	−0.018	−0.021	−0.104	−0.084***
	(−0.485)	(−0.904)	(−1.103)	(−2.875)
Post_GPT	−0.035	0.081**	0.267*	0.141*
	(−0.676)	(2.513)	(1.925)	(1.908)
GPT_Contributor × *Post_GPT*	0.024	−0.080**	−0.250*	−0.129*
	(0.433)	(−2.253)	(−1.720)	(−1.815)
Adj. R^2	−0.009	0.006	0.025	0.016
Observations	63,492	129,857	63,497	129,865
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

This table presents OLS estimates of Equation (4) with forecast accuracy as the outcome. Of the 193,584 forecasts in the accuracy sample, *PMAFE_EPS* is undefined for 26 and *PMAFE_REV* for 13 that lack a within-firm–quarter peer, leaving 193,558 and 193,571 observations; the fixed-effects and split-sample columns drop additional singleton groups and need not sum to the pooled counts. The sample spans 2021–2024 at the forecast level; the dependent variable is *PMAFE_EPS* in Columns 1–2 and *PMAFE_REV* in Columns 3–4. *GPT_Contributor* equals one for contributors classified as GPT-reliant via the outage-based procedure of Internet Appendix B, and *Post_GPT* equals one for forecasts created in 2023 or later. **Panel A** reports full-sample estimates. **Panel B** partitions the sample by contributor type, with controls included. The control variables are those in Table 3. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. *t*-statistics based on standard errors clustered by firm are in parentheses; the full-sample and non-professional interactions remain significant at conventional levels under two-way clustering by firm and contributor. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE 7. GPT Adoption and Contributors' Herding Behavior

Panel A: All Contributors

Dependent Var:	(1) *Herding_EPS*	(2) *Herding_EPS*	(3) *Herding_REV*	(4) *Herding_REV*
GPT_Contributor	−0.442***	−0.424***	−0.183***	−0.138**
	(−7.489)	(−6.140)	(−3.729)	(−2.524)
Post_GPT	0.319***	−0.112	0.328***	0.257
	(4.346)	(−0.543)	(3.254)	(1.312)
GPT_Contributor × *Post_GPT*	−0.176**	−0.195**	−0.220**	−0.210*
	(−2.355)	(−2.246)	(−2.223)	(−1.664)
Pseudo R^2	0.032	0.023	0.011	0.008
Observations	59,674	43,858	59,674	44,357
Controls	Yes	Yes	Yes	Yes
Firm-Qtr FE	No	Yes	No	Yes

Panel B: Financial Professionals vs. Non-Professionals

Dependent Var:	(1) *Herding_EPS*	(2) *Herding_EPS*	(3) *Herding_REV*	(4) *Herding_REV*
Sample split:	Fin. Prof.	Non-Prof.	Fin. Prof.	Non-Prof.
GPT_Contributor × *Post_GPT*	−0.158	−0.239**	−0.004	−0.308**
	(−0.781)	(−2.387)	(−0.019)	(−2.181)
Pseudo R^2	0.017	0.021	0.002	0.009
Observations	11,108	27,350	11,224	27,721
Controls	Yes	Yes	Yes	Yes
Firm-Qtr FE	Yes	Yes	Yes	Yes

This table presents logit estimates of Equation (4) with herding as the outcome. The sample includes revision forecasts over 2021–2024, excluding the first submission by a contributor within a firm–quarter; the dependent variable is *Herding_EPS* in Columns 1–2 and *Herding_REV* in Columns 3–4. Columns 2 and 4 use conditional (fixed-effects) logit with firm–quarter groups. The control variables are those in Table 3, entered unscaled given the binary dependent variable. **Panel A** reports full-sample estimates. **Panel B** partitions the sample by contributor type. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. z-statistics based on standard errors clustered by firm are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE 8. GPT Adoption Effect by Experience Within Non-Professional Contributors

Panel A: Forecast Accuracy (PMAFE)

Dependent Var:	(1) *PMAFE_EPS*	(2)	(3) *PMAFE_REV*	(4)
Sample split:	Non-Exp.	Exp.	Non-Exp.	Exp.
GPT_Contributor	−0.038	0.080***	0.003	−0.008
	(−1.054)	(3.073)	(0.022)	(−0.131)
Post_GPT	−0.065***	0.058	0.345***	0.294***
	(−2.585)	(1.090)	(4.215)	(3.207)
GPT_Contributor × *Post_GPT*	0.060	−0.137**	0.116	−0.219**
	(1.357)	(−2.416)	(0.560)	(−2.079)
Adj. R^2	0.047	0.020	0.195	0.220
Observations	50,906	208,540	50,867	208,030
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

Panel B: Herding (Logit)

Dependent Var:	(1) *Herding_EPS*	(2)	(3) *Herding_REV*	(4)
Sample split:	Non-Exp.	Exp.	Non-Exp.	Exp.
GPT_Contributor	0.025	−0.200***	−0.110	−0.189***
	(0.353)	(−4.748)	(−1.539)	(−4.339)
Post_GPT	−0.015	−0.139***	−0.119	0.011
	(−0.193)	(−2.583)	(−1.372)	(0.186)
GPT_Contributor × *Post_GPT*	0.125	−0.211***	0.388***	−0.236***
	(0.916)	(−3.446)	(3.131)	(−3.890)
Pseudo R^2	0.023	0.047	0.040	0.071
Observations	12,278	84,029	12,278	84,029
Controls	Yes	Yes	Yes	Yes
Firm-Qtr FE	No	No	No	No

This table presents the results from estimating Equation (4) separately for inexperienced (*Non-Exp.*) and experienced (*Exp.*) non-professional contributors. The sample comprises all Estimize forecasts issued during 2021–2024 with nonmissing forecasts and actuals and a CRSP *permno*, a broader panel than Table 6 (see Section 5.2.3). The difference between the 259,681 non-professional forecasts in this panel and the Panel A observation counts reflects singletons dropped by the fixed-effects estimator and, in the revenue columns, forecasts with missing *PMAFE_REV*. *Experienced* equals one if, at the time of the forecast, the contributor's tenure since the first forecast is at least 365 days and cumulative forecasts number at least 10, both computed from the full Estimize history. **Panel A** reports OLS estimates of forecast accuracy with firm fixed effects; the dependent variable is *PMAFE_EPS* in Columns 1–2 and *PMAFE_REV* in Columns 3–4. **Panel B** reports logit estimates of herding on the subsample of revision forecasts; the dependent variable is *Herding_EPS* in Columns 1–2 and *Herding_REV* in Columns 3–4, equal to one if the revision moves the forecast strictly closer to the running Estimize consensus. The control variables are those in Table 3. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. t-statistics (Panel A) and z-statistics (Panel B) are in parentheses, based on standard errors clustered by firm–contributor pair in Panel A and by firm in Panel B. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE 9. Cross-Sectional Tests: GPT Adoption, Firm Information Environment, and Forecast Accuracy

Panel A: PMAFE_EPS

Dependent Var:	*PMAFE_EPS*			
Moderator:	Readability		Info. Asymmetry	
Sample split →	(1)	(2)	(3)	(4)
Contributor group ↓	Low	High	Low	High
Non-Professionals:				
GPT_Contributor × *Post_GPT*	0.041	–0.033	–0.047*	0.020
	(1.247)	(–0.776)	(–1.950)	(0.590)
Financial Professionals:				
GPT_Contributor × *Post_GPT*	0.055	0.092	–0.015	0.008
	(1.013)	(1.395)	(–0.400)	(0.140)
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

Panel B: PMAFE_REV

Dependent Var:	*PMAFE_REV*			
Moderator:	Readability		Info. Asymmetry	
Sample split →	(1)	(2)	(3)	(4)
Contributor group ↓	Low	High	Low	High
Non-Professionals:				
GPT_Contributor × *Post_GPT*	–0.007	–0.093*	–0.086***	–0.044
	(–0.170)	(–1.897)	(–2.660)	(–1.020)
Financial Professionals:				
GPT_Contributor × *Post_GPT*	0.015	–0.155	–0.079	–0.092
	(0.190)	(–1.578)	(–1.360)	(–1.580)
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

This table reports the coefficient on *GPT_Contributor* × *Post_GPT* from separate estimations of Equation (4) on subsamples defined by contributor type and the firm's information environment; the readability subsamples split at the observation-level median of *Readability*, and the information-asymmetry subsamples split at the firm–quarter median of *Info_Asymmetry*. The sample spans 2021–2024 at the forecast level. *Readability* is the average of the standardized Gunning Fog, ARI, and SMOG indices from 8-K filings (higher values denote harder-to-read disclosures); *Info_Asymmetry* is the information-asymmetry component of the Barron et al. (1998) decomposition of I/B/E/S analyst forecast properties, $1-\rho = D/(D+N\times SE)$, where D is forecast variance, SE is squared consensus forecast error, and N is analyst coverage (higher values denote greater reliance on private information). **Panel A** reports results for *PMAFE_EPS*; **Panel B** for *PMAFE_REV*. All specifications include firm fixed effects. The control variables are those in Table 3. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. *t*-statistics based on standard errors clustered by firm–contributor pair in the readability columns and by firm in the information-asymmetry columns are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE 10. Market Reaction to Surprises from Consensus Forecasts by Professional vs. Non-Professional Contributors on Estimize

Dependent Var:	(1)	(2)	(3)	(4)
	CAR [-1, +1] (%)			
Surprise_Prof	1.041***	1.050***	1.139***	1.075***
	(8.693)	(8.894)	(9.265)	(9.026)
Surprise_Prof × *Post_GPT*	−0.354	−0.432	−0.393	−0.435
	(−1.136)	(−1.393)	(−1.210)	(−1.362)
Surprise_NonProf	1.282***	1.146***	1.268***	1.158***
	(10.534)	(9.627)	(10.026)	(9.468)
Surprise_NonProf × *Post_GPT*	0.709**	0.795**	0.700**	0.799**
	(2.144)	(2.398)	(2.018)	(2.350)
Log_Mktcap		0.661***		2.907***
		(7.963)		(15.685)
Loss		−0.390***		−0.049
		(−2.806)		(−0.298)
Book_to_Market		−0.157		−0.947***
		(−0.997)		(−3.438)
Momentum		−1.057***		−2.859***
		(−10.438)		(−21.510)
Pre_Return		19.995***		18.541***
		(20.465)		(18.622)
Analyst_Coverage		−0.555***		−1.663***
		(−7.483)		(−9.454)
ROA		7.167***		12.865***
		(4.260)		(5.590)
Size		−0.523***		−3.245***
		(−6.943)		(−13.708)
Leverage		0.711***		3.399***
		(3.112)		(6.147)
Constant	0.117***	−0.088	0.052***	4.843***
	(2.856)	(−0.310)	(12.681)	(3.317)
Adj. R^2	0.044	0.063	0.056	0.092
Observations	55,997	55,997	55,896	55,896
Firm FE	No	No	Yes	Yes
Year-Quarter FE	No	No	Yes	Yes

This table presents OLS estimates of Equation (5). The sample comprises firm–quarter earnings announcements over 2015–2024, and the dependent variable is the three-day cumulative abnormal return around the announcement, *CAR [-1, +1] (%)*. *Surprise_Prof* and *Surprise_NonProf* are reported EPS minus the professional and non-professional Estimize consensus forecasts, scaled by the absolute consensus forecast. Columns (1) and (3) include no controls; Columns (2) and (4) add the firm-level controls shown. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. *t*-statistics based on standard errors clustered by firm are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Internet Appendix

Internet Appendix A: GPT Major Outage History (2023–2025)

No.	Start Date	Start Time	End Date	End Time	Duration (min)
Panel A: 2023–2024 (32 events, used in the classification and the main outage tests)					
1	Feb 14, 2023	14:30	Feb 14, 2023	14:37	7
2	Feb 21, 2023	14:40	Feb 21, 2023	19:03	263
3	Feb 23, 2023	15:56	Feb 23, 2023	18:15	139
4	Feb 27, 2023	10:00	Feb 27, 2023	13:21	201
5	Mar 15, 2023	13:55	Mar 15, 2023	14:20	25
6	Mar 20, 2023	05:18	Mar 20, 2023	05:42	24
7	Apr 23, 2023	23:23	Apr 23, 2023	23:37	14
8	May 18, 2023	19:50	May 18, 2023	20:20	30
9	May 25, 2023	07:23	May 25, 2023	08:27	64
10	Jun 30, 2023	14:41	Jun 30, 2023	15:20	39
11	Aug 3, 2023	19:36	Aug 3, 2023	19:50	14
12	Aug 31, 2023	09:52	Aug 31, 2023	12:29	157
13	Sep 1, 2023	00:36	Sep 1, 2023	01:11	35
14	Nov 8, 2023	08:54	Nov 8, 2023	10:46	112
15	Nov 9, 2023	15:03	Nov 9, 2023	15:56	53
16	Nov 12, 2023	00:57	Nov 12, 2023	01:44	47
17	Nov 23, 2023	09:31	Nov 23, 2023	09:37	6
18	Dec 13, 2023	20:52	Dec 13, 2023	21:24	32
19	Jan 29, 2024	19:30	Jan 29, 2024	19:49	19
20	Apr 1, 2024	20:28	Apr 1, 2024	20:45	17
21	May 3, 2024	02:00	May 3, 2024	02:28	28
22	May 3, 2024	03:18	May 3, 2024	03:39	21
23	May 21, 2024	14:40	May 21, 2024	15:48	68
24	Jun 4, 2024	03:21	Jun 4, 2024	07:45	264
25	Jun 4, 2024	10:33	Jun 4, 2024	13:17	164
26	Jul 5, 2024	18:23	Jul 5, 2024	18:33	10
27	Jul 12, 2024	06:03	Jul 12, 2024	06:52	49
28	Aug 21, 2024	12:27	Aug 21, 2024	13:02	35
29	Oct 17, 2024	13:31	Oct 17, 2024	13:43	12
30	Nov 8, 2024	19:13	Nov 8, 2024	20:47	94
31	Dec 11, 2024	18:42	Dec 11, 2024	22:53	251
32	Dec 26, 2024	14:00	Dec 26, 2024	23:38	578
Panel B: 2025 (5 events, used only in the DeepSeek R1 falsification test)					
1	Apr 24, 2025	14:17	Apr 24, 2025	14:53	36
2	Jun 23, 2025	14:08	Jun 23, 2025	14:59	51
3	Jul 16, 2025	16:07	Jul 16, 2025	16:11	4
4	Jul 18, 2025	18:03	Jul 18, 2025	18:36	33
5	Nov 18, 2025	07:12	Nov 18, 2025	10:18	186

This table documents the ChatGPT service outages used in the paper. **Panel A** lists the 32 major outages during 2023–2024 that underlie the GPT-contributor classification and the main outage tests; **Panel B** lists the five major outages during 2025, used only in the DeepSeek R1 falsification test (Section 5.3.3). All start and end times are in Eastern Standard Time (EST), matching the time basis of Estimize submission timestamps. Outage events are sourced from OpenAI's status-page incident history; the 2025 events are recovered from archived snapshots of the status page.

Internet Appendix B: Procedure to Identify GPT-Reliant Contributors

This appendix provides the technical details underlying the classification procedure summarized in Section 5.1.1. We identify contributors who rely on generative AI tools ("GPT contributors") by exploiting exogenous variation in the availability of ChatGPT. The identification rests on a revealed-preference argument: contributors who depend on GPT for forecast generation should exhibit a disproportionate decline in forecasting activity during periods when the service is unexpectedly unavailable. We operationalize this intuition through a three-stage procedure: Stage 1 estimates each contributor's baseline propensity to forecast in a given time window; Stage 2 measures how often a contributor's expected forecasts fail to materialize during documented ChatGPT outages; and Stage 3 converts this measure into a discrete classification. The outage windows are the 32 major ChatGPT service disruptions during 2023–2024 listed in Internet Appendix A, Panel A.

Stage 1: Estimating Baseline Forecasting Propensity

We first estimate each contributor's probability of submitting a forecast within a given 30-minute window, conditional on observable determinants of forecasting activity. The estimation sample is the universe of contributor–firm–day–half-hour observations over the post-GPT period (2023–2024), comprising approximately 1.08 billion cells. On this sample we estimate the following logit model:

$$\Pr(Forecast_{i,j,d,h}=1)=\Lambda\Big(\beta_1 RecentActivity_{i,d}+\beta_2 EA\_Past5_{j,d} + \beta_3 Volatility_{j,d}+\alpha_{j\times y}+\gamma_{i\times j}+\delta_w+\eta_h\Big) \tag{IA.1}$$

where $\Lambda(\cdot)$ denotes the logistic function, i indexes contributors, j indexes firms, d indexes calendar days, and h indexes 30-minute intervals. $Forecast_{i,j,d,h}$ equals one if contributor i submits a forecast for firm j during half-hour window h of day d. The covariates capture the three first-order determinants of forecast timing: $RecentActivity_{i,d}$ measures the contributor's forecasting intensity in the preceding week, $EA\_Past5_{j,d}$ indicates whether firm j had an earnings announcement in the prior five days, and $Volatility_{j,d}$ is the 30-day return volatility of firm j's stock. The model absorbs firm-by-year fixed effects ($\alpha_{j\times y}$) to control for time-varying firm-level demand for coverage, contributor-by-firm fixed effects ($\gamma_{i\times j}$) to capture persistent contributor–firm match quality, and day-of-week (δ_w) and half-hour-of-day (η_h) fixed effects to account for temporal regularities in forecasting behavior.

The scale of the estimation problem precludes standard in-memory computation, and exact maximum-likelihood estimation of Equation (IA.1) with fixed effects of this dimension is computationally infeasible. We implement an out-of-core approximation: the covariates are residualized with respect to the fixed effects by iterative within-group demeaning (alternating

projections), applied across partitioned data files until convergence (residual group means $< 10^{-6}$), and a logit of the binary outcome on the residualized covariates is then estimated via Newton–Raphson optimization, with the gradient and Hessian accumulated in streaming fashion across data partitions. Linear residualization absorbs fixed effects exactly in a linear model but not in a logit, so the procedure is a computational approximation to Equation (IA.1) rather than its maximum-likelihood estimator.[46] Fitted probabilities are generated by applying the logistic transformation to the estimated linear index.

Stage 2: Constructing the Missing Forecast Ratio

Using the fitted probabilities from Stage 1, we identify contributor–window observations in which a contributor is *expected* to issue a forecast: those whose fitted probability exceeds the sample mean of the fitted probabilities by more than one standard deviation. The threshold is global, applied uniformly across contributors and windows, so expected-forecaster status reflects a high predicted propensity rather than a relative ranking within a particular window. If an expected forecaster does not submit a forecast in that window, the observation is recorded as a *miss*.

We then evaluate contributor behavior during the documented ChatGPT outages. A contributor is recorded as having a *missing forecast in an outage* if (i) the contributor is flagged as an expected forecaster in at least one 30-minute bin during the outage, and (ii) the contributor does not submit any forecast throughout the entire duration of the outage. For each contributor i, we compute a continuous GPT-reliance measure:

$$MissRatio_i = \frac{\text{Number of outages with missing forecasts}}{\text{Total number of outages}} \tag{IA.2}$$

This ratio captures the frequency with which a contributor's forecasting activity disappears precisely when ChatGPT is unavailable, relative to the number of opportunities to observe such behavior. Higher values indicate a stronger revealed dependence on GPT.

Stage 3: Classification

For use in regression analyses requiring a discrete treatment indicator, we classify contributors whose $MissRatio_i$ falls in the top quartile of the distribution as GPT contributors ($GPT\_Contributor_i = 1$), with all remaining contributors assigned to the control group ($GPT\_Contributor_i = 0$). The top-quartile threshold balances two considerations: it is stringent enough to isolate contributors with a pronounced and persistent pattern of outage-induced forecast absence, while retaining sufficient statistical power for cross-sectional tests. By construction, the quartile rule fixes the overall share of classified contributors at roughly one quarter; the composition of the classified group across contributor types, which underlies the tests in Section 5, is not

[46]The approximation is innocuous for our purpose. No Stage 1 coefficient enters any subsequent inference: the stage serves only to produce a propensity index for the classification in Stage 2, which requires a monotone ranking of contributor–window observations rather than consistent structural coefficients.

mechanical.

Discussion

Several features of this identification strategy merit discussion. First, ChatGPT service outages are plausibly exogenous to individual forecasting decisions: they arise from infrastructure failures unrelated to capital markets, and their timing is unpredictable to users. Second, by conditioning on the rich set of fixed effects and time-varying controls in Stage 1, the procedure nets out predictable variation in forecasting behavior, so the classification loads on the incremental absence of forecasts when ChatGPT is down rather than on general differences in activity levels. Third, the contributor-by-firm fixed effects ensure that expected-forecaster status is assessed within each contributor–firm relationship, guarding against confounds arising from systematic differences in coverage patterns across contributor types.

A potential concern is that the top-quartile cutoff introduces classification error at the margin: contributors near the threshold may be misclassified in either direction. Such measurement error attenuates the difference-in-differences estimates toward zero, biasing against finding significant adoption effects. The validity of the classification is corroborated by three independent pieces of evidence. First, classified GPT contributors sharply increase their forecasting output relative to non-GPT contributors after ChatGPT's release, following parallel paths beforehand (Figure 3 and Table 5; Section 5.1.2). Second, forecast quality deteriorates during outages in the aggregate (Section 4.1), consistent with the reliance the classification is designed to detect. Third, the outage sensitivity of forecast quality vanishes after close ChatGPT substitutes become available in early 2025 (Section 5.3.3), the pattern the revealed-preference logic implies if ChatGPT's availability is the binding constraint.

Internet Appendix C: Parallel Trends Test for the DiD Analysis

The identifying assumption for Equation (4) is that GPT-reliant and non-GPT contributors would have followed parallel outcome paths absent GPT availability. We assess this assumption by estimating an event-study specification that replaces $GPT\_Contributor_i \times Post\_GPT_t$ with year-specific interactions, using 2022 as the reference year:

$$\begin{aligned} Y_{i,j,t} = & \sum_{s \neq 2022} \delta_s (GPT\_Contributor_i \times \mathbf{1}[t = s]) + \beta_1 GPT\_Contributor_i \\ & + \mu_t + \mathbf{X}'_{i,j,t}\gamma + \alpha_j + \varepsilon_{i,j,t} \end{aligned} \tag{IA.3}$$

where μ_t denotes year fixed effects. The pre-treatment coefficient δ_{2021} tests whether GPT-reliant and non-GPT contributors were on differential accuracy paths before ChatGPT became available; the post-treatment coefficients δ_{2023} and δ_{2024} capture the dynamic treatment effect.

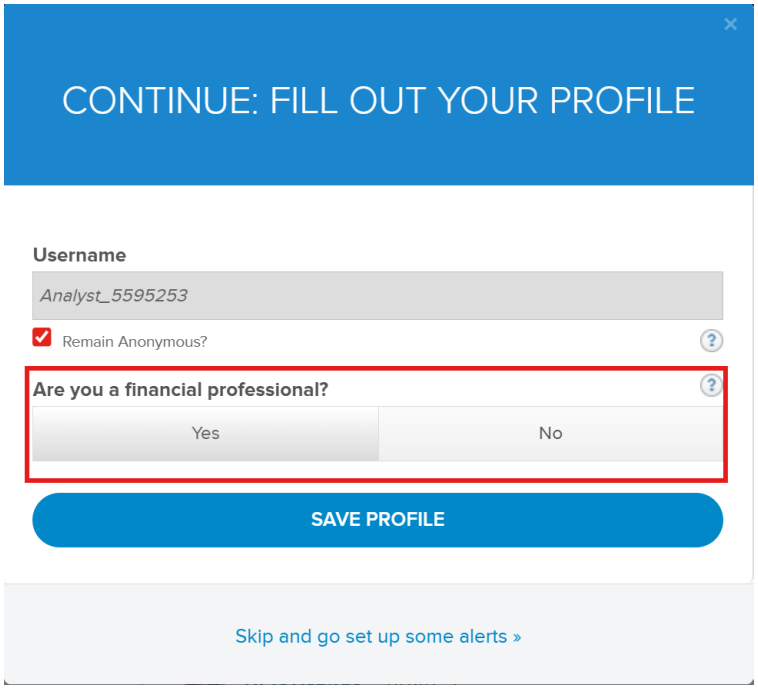
Figure IA.2 plots the estimated δ_s coefficients with 95% confidence intervals for both EPS and revenue accuracy, estimated on the accuracy sample and specification of Table 6. For revenue accuracy, the 2021 pre-trend coefficient is small and statistically indistinguishable from zero ($\hat{\delta}_{2021} = 0.027$, $t = 0.47$), supporting parallel pre-treatment paths. For EPS accuracy, the pre-trend coefficient is negative and marginally significant ($\hat{\delta}_{2021} = -0.051$, $t = -1.83$), so a portion of the EPS gap may predate ChatGPT's availability and the EPS estimates should be read with this caveat. The sample provides two pre-treatment years (2021 and 2022), of which only one yields an estimable pre-trend coefficient because 2022 serves as the reference year, which limits the power of the test. The restriction arises from our symmetric two-year window around the start of the post-GPT period (2021–2024); extending the pre-period further back would introduce additional confounds from platform evolution and user-base composition changes.

The post-treatment coefficients trace the dynamics of the adoption effect. For EPS accuracy, GPT-reliant contributors diverge from non-GPT contributors in both post-treatment years ($\hat{\delta}_{2023} = -0.068$, $t = -3.05$; $\hat{\delta}_{2024} = -0.104$, $t = -2.20$; joint $p = 0.002$), with the effect growing over time. For revenue accuracy, the year-specific coefficients are larger in magnitude ($\hat{\delta}_{2023} = -0.135$; $\hat{\delta}_{2024} = -0.115$) but individually and jointly insignificant (joint $p = 0.20$), as splitting the pooled effect across years reduces power; the pooled interaction in Table 6 remains negative and significant.

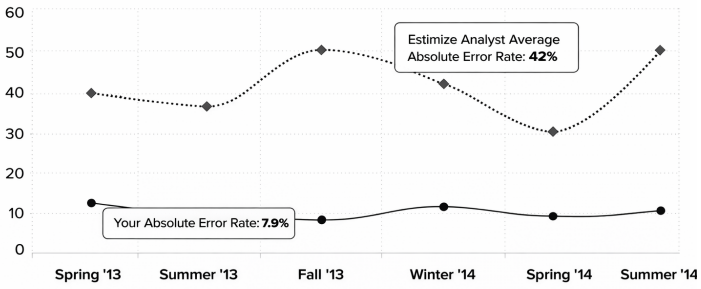
To assess sensitivity, we implement the relative-magnitudes analysis of Rambachan and Roth (2023), which widens the confidence interval for the post-treatment effect to allow violations of parallel trends up to $\bar{M}$ times the observed pre-treatment deviation. For EPS, the untabulated confidence interval for $\hat{\delta}_{2023}$ continues to exclude zero for violations up to half the pre-treatment deviation ($\bar{M} = 0.5$) but not beyond; for revenue, the year-specific intervals include zero throughout, consistent with the limited power of the year-by-year splits.

FIGURE IA.1. Estimize Platform Features

Panel A: Professional Status Self-Identification at Registration



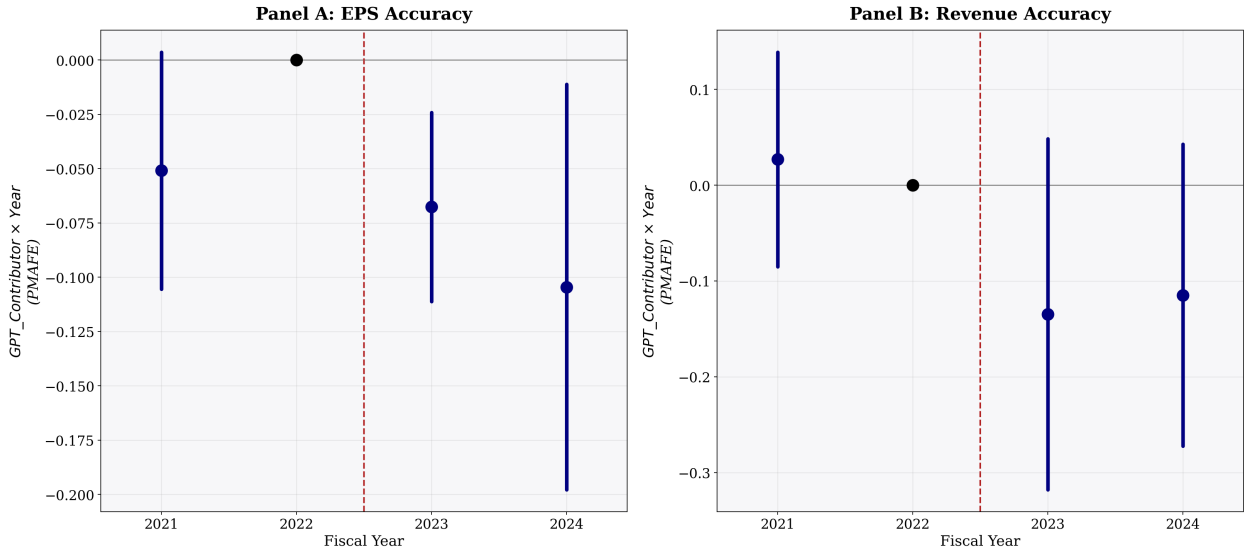
Panel B: Individual Performance Tracker



Abs Error Rate	12.6%	9.6%	7.9%	11.2%	8.9%	9.0%
Accuracy Pctile	36%	48%	50%	39%	54%	36%
Coverage	63	413	389	224	213	109

Panel A displays the Estimize registration page, where users self-report whether they are financial professionals when signing up for the Estimize platform. **Panel B** shows the individual performance dashboard, which tracks a user's absolute error rate over time and benchmarks it against the Estimize contributor population average, providing incentives to improve forecast accuracy.

FIGURE IA.2. Parallel Trends Test: Event-Study Estimates of Forecast Accuracy



This figure plots the estimated coefficients $\hat{\delta}_s$ from the event-study specification in Equation (IA.3), with 2022 as the reference year and 95% confidence intervals. **Panel A** reports results for EPS forecast accuracy (*PMAFE_EPS*) and **Panel B** for revenue forecast accuracy (*PMAFE_REV*). The dashed vertical line marks the beginning of the post-GPT period (January 2023). For revenue accuracy, the 2021 pre-treatment coefficient is small and statistically indistinguishable from zero ($\hat{\delta}_{2021} = 0.027$, $t = 0.47$); for EPS accuracy, it is negative and marginally significant ($\hat{\delta}_{2021} = -0.051$, $t = -1.83$). The post-treatment coefficients are negative for both outcomes: individually and jointly significant for EPS ($\hat{\delta}_{2023} = -0.068$; $\hat{\delta}_{2024} = -0.104$; joint $p = 0.002$), and larger in magnitude but insignificant year by year for revenue ($\hat{\delta}_{2023} = -0.135$; $\hat{\delta}_{2024} = -0.115$; joint $p = 0.20$). The specification is that of Table 6: controls include *Effort*, *Experience*, *Forecast_Age*, and *Firm_Followed*, with firm and year fixed effects and standard errors clustered by firm.

TABLE IA.1. Distribution of GPT Contributors Across Financial Professionals and Non-Professionals

	GPT Contributors	Non-GPT Contributors	Total
Financial Professional	241 (44.3%)	303 (55.7%)	544
Non-Professional	745 (22.4%)	2,587 (77.6%)	3,332
Total	986	2,890	3,876

This table reports the distribution of the outage-based *GPT_Contributor* classification (Internet Appendix B) across contributor types. The sample comprises the 3,876 unique contributors in the accuracy sample (Table 1, Panel B). Percentages in parentheses are within-contributor-type (row) shares; the two columns in each row sum to 100%. All variables are defined in Appendix A.

TABLE IA.2. GPT Adoption Effect by Experience Within Financial Professionals

Panel A: Forecast Accuracy (PMAFE)

Dependent Var:	(1) *PMAFE_EPS*	(2)	(3) *PMAFE_REV*	(4)
Sample split:	Non-Exp.	Exp.	Non-Exp.	Exp.
GPT_Contributor	0.068	0.036	0.011	−0.883***
	(1.303)	(1.226)	(0.052)	(−4.254)
Post_GPT	0.062	−0.066*	0.175	−0.368
	(0.634)	(−1.761)	(0.914)	(−1.091)
GPT_Contributor × *Post_GPT*	−0.118	−0.008	−0.060	0.384
	(−1.001)	(−0.188)	(−0.190)	(1.101)
Adj. R^2	0.140	0.022	0.212	0.240
Observations	32,528	148,375	32,395	147,918
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

Panel B: Herding (Logit)

Dependent Var:	(1) *Herding_EPS*	(2)	(3) *Herding_REV*	(4)
Sample split:	Non-Exp.	Exp.	Non-Exp.	Exp.
GPT_Contributor	0.091	0.076*	−0.019	−0.043
	(1.156)	(1.805)	(−0.232)	(−0.906)
Post_GPT	0.175	0.073	0.134	0.098
	(1.597)	(1.334)	(1.189)	(1.572)
GPT_Contributor × *Post_GPT*	0.010	−0.005	0.192	0.074
	(0.079)	(−0.087)	(1.365)	(1.101)
Pseudo R^2	0.011	0.007	0.016	0.012
Observations	15,915	84,543	15,915	84,543
Controls	Yes	Yes	Yes	Yes
Firm-Qtr FE	No	No	No	No

This table presents the results from estimating Equation (4) separately for inexperienced (*Non-Exp.*) and experienced (*Exp.*) financial professional contributors. The sample comprises all Estimize forecasts issued during 2021–2024 with nonmissing forecasts and actuals, a broader panel than Table 6 (see Section 5.2.3). *Experienced* is defined as in Table 8. **Panel A** reports OLS estimates of forecast accuracy with firm fixed effects; the dependent variable is *PMAFE_EPS* in Columns 1–2 and *PMAFE_REV* in Columns 3–4. **Panel B** reports logit estimates of herding on the subsample of revision forecasts; the dependent variable is *Herding_EPS* in Columns 1–2 and *Herding_REV* in Columns 3–4, equal to one if the revision moves the forecast strictly closer to the running Estimize consensus. The control variables are those in Table 3. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. t-statistics (Panel A) and z-statistics (Panel B) are in parentheses, based on standard errors clustered by firm–contributor pair in Panel A and by firm in Panel B. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE IA.3. Alternative Classification: Strict GPT Reliance and EPS Forecast Accuracy

Dependent Var:	(1)	(2) *PMAFE_EPS*	(3)
Sample split:	All Contributors	Non-Professionals	Fin. Professionals
GPT_Contributor	–0.009	–0.007	–0.021
	(–0.466)	(–0.319)	(–0.513)
Post_GPT	0.047**	0.082**	–0.042
	(2.047)	(2.520)	(–0.769)
GPT_Contributor × *Post_GPT*	–0.056**	–0.095***	0.059
	(–2.412)	(–2.644)	(0.997)
Adj. R^2	0.007	0.000	0.016
Observations	124,114	91,902	31,885
Controls	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes

This table re-estimates the EPS accuracy specifications of Table 6 (Equation (4)) under a stricter reliance classification, in which GPT-reliant contributors fall in the top quartile of the outage miss ratio (Internet Appendix B). *GPT_Contributor* is redefined to equal one only for contributors who are in the top quartile of the outage miss ratio and never submit a forecast during any outage window (891 of the 986 GPT-reliant contributors in the sample); the 95 who submit during at least one outage are ambiguous under the stricter criterion and are excluded from the estimation sample, which shrinks accordingly because these contributors are disproportionately active. The dependent variable is *PMAFE_EPS*; Column (1) reports full-sample estimates, and Columns (2)–(3) partition the sample by contributor type. The sample spans 2021–2024 at the forecast level, and all specifications include the controls of Table 3 and firm fixed effects. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. t-statistics based on standard errors clustered by firm are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE IA.4. Placebo Test: Randomized Outage Timing and EPS Forecast Accuracy

Panel A: All Contributors

Dependent Var:	(1)	(2)	(3)	(4)
	PMAFE_EPS			
Classification:	Actual outages		Placebo outages	
GPT_Contributor	0.010	−0.017	0.010	−0.018
	(0.589)	(−0.862)	(0.561)	(−1.009)
Post_GPT	0.052**	0.050**	0.037	0.032
	(2.439)	(2.144)	(1.245)	(1.153)
GPT_Contributor × *Post_GPT*	−0.051**	−0.054**	−0.034	−0.035
	(−2.385)	(−2.291)	(−1.284)	(−1.378)
Adj. R^2	0.003	0.006	0.003	0.006
Observations	193,558	193,556	193,558	193,556
Controls	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes

Panel B: Financial Professionals vs. Non-Professionals

Dependent Var:	(1)	(2)	(3)	(4)
	PMAFE_EPS			
Classification:	Actual outages		Placebo outages	
Sample split:	Fin. Prof.	Non-Prof.	Fin. Prof.	Non-Prof.
GPT_Contributor	−0.018	−0.021	0.012	−0.038*
	(−0.485)	(−0.904)	(0.500)	(−1.662)
Post_GPT	−0.035	0.081**	−0.000	0.036
	(−0.676)	(2.513)	(−0.001)	(1.271)
GPT_Contributor × *Post_GPT*	0.024	−0.080**	−0.014	−0.030
	(0.433)	(−2.253)	(−0.187)	(−1.195)
Adj. R^2	−0.009	0.006	−0.009	0.006
Observations	63,492	129,857	63,492	129,857
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

This table reports a placebo test of the outage-based classification, re-estimating the EPS specifications of Table 6 (Equation (4)). In Columns 1–2, *GPT_Contributor* is the actual classification, assigned from forecasting behavior during the 32 documented ChatGPT outages. In Columns 3–4, it is a placebo classification obtained by re-running the procedure of Internet Appendix B on pseudo-outage windows, each true outage window shifted four weeks earlier and four weeks later, preserving day of week, time of day, and duration while relocating it to periods when ChatGPT was operational; results are unchanged with eight- and twelve-week shifts. Of the 986 GPT-reliant contributors, 676 (68.6%) are also flagged under the placebo windows (1,059 in total), reflecting the activity screen common to both procedures. A stacked test of the difference between the placebo and actual interactions yields 0.019 ($t = 0.66$) for all contributors and 0.049 ($t = 1.44$) for non-professionals; because nearly two-thirds of the placebo group is genuinely GPT-reliant, the placebo coefficient mechanically inherits a comparable share of the actual effect, making the contrast conservative. The dependent variable is *PMAFE_EPS*; the sample spans 2021–2024 at the forecast level. **Panel A** reports full-sample estimates. **Panel B** partitions the sample by contributor type. Controls are included in all specifications. The control variables are those in Table 3. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. t-statistics based on standard errors clustered by firm are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE IA.5. Alternative Accuracy Measure: AFE and GPT Adoption

Panel A: All Contributors

Dependent Var:	(1) *AFE_EPS*	(2) *AFE_EPS*	(3) *AFE_REV*	(4) *AFE_REV*
GPT_Contributor	−0.037***	−0.014	−0.007***	−0.002
	(−2.993)	(−1.442)	(−3.153)	(−1.309)
Post_GPT	0.023	0.005	−0.007***	−0.008***
	(1.321)	(0.355)	(−3.130)	(−3.802)
GPT_Contributor × Post_GPT	−0.040**	−0.039**	−0.007***	−0.005**
	(−2.282)	(−2.495)	(−3.139)	(−2.528)
Effort	0.017***	0.015***	0.006***	0.003***
	(3.498)	(4.099)	(7.311)	(5.403)
Experience	−0.003***	−0.002***	−0.000	−0.000
	(−2.875)	(−3.779)	(−1.588)	(−1.174)
Forecast_Age	0.001***	0.001***	0.000***	0.000***
	(8.176)	(10.778)	(15.317)	(16.919)
Firm_Followed	−0.002	−0.001	−0.002***	−0.002***
	(−1.185)	(−1.165)	(−5.734)	(−8.088)
Estimize_Coverage	0.086***	0.032***	−0.004***	0.001
	(20.426)	(3.097)	(−5.523)	(0.776)
ROA	−0.847***	−1.152***	0.142***	0.055***
	(−6.689)	(−6.823)	(5.907)	(2.741)
Size	−0.000	0.005	0.003***	−0.012***
	(−0.053)	(0.360)	(3.572)	(−5.807)
Leverage	0.037***	0.120**	−0.007***	−0.012
	(2.802)	(2.407)	(−3.066)	(−1.554)
Log_Mktcap	−0.066***	−0.071***	−0.005***	0.006***
	(−13.441)	(−6.405)	(−6.149)	(4.419)
Loss	0.268***	0.114***	0.019***	0.006***
	(25.239)	(9.923)	(14.216)	(6.105)
Book_to_Market	0.105***	0.136***	0.004**	0.010***
	(6.582)	(4.213)	(2.155)	(2.929)
Constant	0.671***	0.792***	0.096***	0.107***
	(24.723)	(4.930)	(23.302)	(5.129)
Adj. R^2	0.090	0.275	0.057	0.311
Observations	173,991	173,988	174,132	174,128
Firm FE	No	Yes	No	Yes

(continued on next page)

Table IA.5 (continued from the previous page)

Panel B: Financial Professionals vs. Non-Professionals

Dependent Var:	(1)	(2)	(3)	(4)
	AFE_EPS		*AFE_REV*	
Sample split:	Fin. Prof.	Non-Prof.	Fin. Prof.	Non-Prof.
GPT_Contributor	–0.010	–0.017	–0.000	–0.002*
	(–0.565)	(–1.479)	(–0.058)	(–1.679)
Post_GPT	–0.042*	0.019	–0.009**	–0.005**
	(–1.680)	(0.989)	(–1.994)	(–2.427)
GPT_Contributor × *Post_GPT*	0.002	–0.049***	–0.005	–0.006***
	(0.098)	(–2.605)	(–1.056)	(–2.710)
Adj. R^2	0.283	0.273	0.315	0.360
Observations	57,377	116,418	57,431	116,510
Controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

This table re-estimates Equation (4) (the specification of Table 6) using the absolute forecast error as the dependent variable, with unscaled contributor- and firm-level controls. *AFE* is |forecast – actual|/|actual|, for EPS in Columns 1–2 and revenue in Columns 3–4. The sample spans 2021–2024 and excludes observations with missing controls. **Panel A** reports full-sample estimates. **Panel B** partitions the sample by contributor type. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. *t*-statistics based on standard errors clustered by firm–contributor pair are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE IA.6. GPT Outage and Forecast Accuracy: Pre vs. Post DeepSeek R1

Panel A: All Contributors

Dependent Var:	(1)	(2)
	PMAFE	
Sample split:	Pre-DSR1	Post-DSR1
During_Outage	0.087**	−0.004
	(2.254)	(−0.161)
Adj. R^2	0.015	0.016
Observations	96,217	49,041
Controls	Yes	Yes
Firm FE	Yes	Yes
DOW × Hour FE	Yes	Yes

Panel B: Non-Professionals

Dependent Var:	(1)	(2)
	PMAFE	
Sample split:	Pre-DSR1	Post-DSR1
During_Outage	0.091*	−0.045
	(1.897)	(−0.687)
Adj. R^2	0.017	0.067
Observations	69,913	29,652
Controls	Yes	Yes
Firm FE	Yes	Yes
DOW × Hour FE	Yes	Yes

This table re-estimates Equation (2) (the outage specification of Table 3), partitioning the sample by the public release of DeepSeek R1 (January 20, 2025). The pre-DSR1 sample is the 2023–2024 post-GPT sample of Table 3; the post-DSR1 sample comprises forecasts created on or after January 20, 2025 in the extended Estimize data, restricted to the same contributors and firms. Forecasts submitted during outage windows number 275 in the pre-DSR1 period (167 by non-professionals) and 23 in the post-DSR1 period (7 by non-professionals). The dependent variable is *PMAFE*, and *During_Outage* equals one if the forecast is submitted during a documented major ChatGPT service outage, defined as an incident that reached full-outage status on a ChatGPT-related component of OpenAI's status page (32 events in 2023–2024 and 5 events in 2025, all listed in Internet Appendix A). **Panel A** reports full-sample estimates. **Panel B** restricts the sample to non-professional contributors. The control variables are those in Table 3, min–max scaled within firm–fiscal quarter; the slight differences from the Table 3 estimates reflect this scaling. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. t-statistics based on standard errors clustered by firm are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE IA.7. Revision Propensity and GPT Adoption

Dependent Var:	(1)	(2)	(3)
		1[forecast is a revision]	
Sample split:	All Contributors	Non-Professionals	Fin. Professionals
GPT_Contributor	0.011***	0.011***	0.006
	(3.323)	(3.446)	(0.781)
Post_GPT	−0.022***	−0.020***	−0.009
	(−3.928)	(−2.982)	(−0.972)
GPT_Contributor × *Post_GPT*	0.016***	0.018***	−0.002
	(2.946)	(2.832)	(−0.230)
Adj. R^2	0.406	0.400	0.407
Observations	193,582	129,876	63,499
Controls	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes

This table reports OLS linear probability estimates of Equation (4) with an indicator for whether a forecast is a revision as the outcome. The dependent variable equals one if the forecast is not the contributor's first submission for the firm–quarter (sample mean 0.338), constructed by ordering each contributor's submissions within a release by timestamp. Column (1) reports full-sample estimates, and Columns (2)–(3) partition the sample by contributor type. The sample is the accuracy sample of Table 6, spanning 2021–2024 at the forecast level. All specifications include the controls of Table 3 and firm fixed effects. All continuous variables are winsorized at the 1st and 99th percentiles. All variables are defined in Appendix A. *t*-statistics based on standard errors clustered by firm are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.