Limited Attention and Generative AI in Financial Reporting: Theory and Experimental Evidence

Guangzhe Wang*

February 27, 2026

Abstract

This paper studies the trade-off between the efficiency gains from AI-assisted disclosure processing and the cognitive costs of verifying when investors use generative AI tools in financial reporting. The model predicts that a capacity threshold governs diagnostic activation, that higher hallucination rates shift attention from parsing to verification, and that the traditional summary regime dominates when AI information loss is small. I test these predictions with an incentivized experiment in which 353 Prolific participants analyze an AI-generated summary under randomly assigned hallucination levels. Consistent with the model, higher-capacity investors are more likely to verify AI output, the high-hallucination group devotes more time to diagnostics and less to parsing, and the hallucination–capacity interaction amplifies diagnostic behavior. Diagnostic engagement substantially improves task accuracy, and investors dynamically increase verification after discovering hallucinations. These findings offer the first theoretical and experimental account of the cognitive costs of AI-assisted financial analysis.

Keywords: Rational Inattention, Generative AI, AI Hallucination, Information Disclosure

JEL Classification: D83, G11, G14, G41, M41

*School of Accounting and Finance, The Hong Kong Polytechnic University

I thank Vivian Fang, Jinzhi Lu for their constructive comments and suggestions. I gratefully acknowledge financial support from The Hong Kong Polytechnic University. All errors are my own.

1 Introduction

The rapid integration of generative artificial intelligence (AI) into financial markets represents a paradigm shift in how investors access and process corporate disclosures. Large language models (LLMs) now routinely summarize earnings reports, analyze financial statements, and generate investment recommendations, dramatically reducing the time required to process lengthy regulatory filings (Cao et al., 2023; Kim et al., 2024; Korinek, 2023). Yet this technological convenience comes with a fundamental cost: AI systems inevitably produce hallucinations, that is, plausible-sounding but factually incorrect outputs that can mislead investors who rely on them (Ji et al., 2023; Kalai and Vempala, 2024; Huang et al., 2025). This paper investigates a central question at the intersection of information economics and artificial intelligence: How do investors allocate their limited attention when processing AI-generated financial information that may contain hallucinations?

The answer to this question has first-order implications for financial markets. If investors blindly trust AI summaries without verification, hallucinated information could distort asset prices, amplify mispricing, and undermine the information efficiency of capital markets. Conversely, if investors expend substantial cognitive resources verifying AI outputs, the purported efficiency gains from AI adoption may be significantly diminished or even reversed. Understanding the equilibrium between these forces, specifically, the convenience of AI-assisted information processing and the cognitive cost of verifying its output, is essential for regulators, disclosure designers, and market participants.

I address this question through both theory and experiment. On the theoretical side, I develop a rational inattention model that extends the framework of Lu (2022) to incorporate generative AI. The model features a representative investor who processes a firm's financial disclosure under two possible information regimes. In Regime S (Summary), the investor reads a traditional financial summary without AI assistance. In Regime D+AI (Details disclosure plus AI processing), the investor uses generative AI to process detailed disclosures but faces the risk of AI hallucination. The key innovation is that in Regime D+AI, the

investor allocates limited attention across two channels: a parsing channel that extracts informational content from the AI-generated summary, and a diagnostic channel that verifies the reliability of the AI output against original disclosures.

The model builds on the canonical rational inattention framework of Sims (2003), Maćkowiak and Wiederholt (2009), and Veldkamp (2011), in which investors face a capacity constraint on the total amount of information they can process. The investor optimally trades off between these two activities to minimize the uncertainty about the firm's fundamentals.

The theoretical analysis yields several results. Proposition 1 establishes the existence of a capacity threshold that determines whether an investor activates the diagnostic channel. When the investor's information processing capacity falls below this threshold, she devotes all capacity to the parsing channel, effectively trusting the AI output without verification. When capacity exceeds the threshold, the investor allocates attention to both channels, simultaneously extracting content from the AI summary and verifying its reliability. Intuitively, verification is a "luxury good" in cognitive terms: only investors with sufficient processing resources find it worthwhile to spend some of those resources checking AI reliability, because the opportunity cost of diverting attention from content extraction is too high for capacity-constrained investors.

Corollary 1 characterizes the comparative statics with respect to the hallucination rate. When the hallucination rate increases, the capacity threshold decreases, making it more likely that an investor activates the diagnostic channel. Moreover, for investors who have already activated diagnostic attention, a higher hallucination rate leads to a strict increase in diagnostic investment and a strict decrease in parsing investment. The intuition is that as AI becomes less reliable, the expected value of verification rises because the investor is more likely to encounter misleading content. This increased need for fact-checking crowds out attention that would otherwise be devoted to extracting useful information from the AI summary, creating a "double cost" of hallucination, that is, the direct cost of inaccurate content plus the indirect cost of diverted cognitive resources.

Corollary 2 reveals a more subtle finding regarding the effect of capacity on parsing. While diagnostic investment always increases monotonically with capacity, higher capacity does not necessarily translate into more parsing when the hallucination rate is not sufficiently high. This occurs because additional capacity may be first absorbed by the diagnostic channel before any "trickles down" to parsing. This result has important implications for understanding heterogeneity in AI adoption: professional investors may not only verify more but may also parse more effectively, while retail investors may be trapped in a regime of blind trust.

Proposition 2 addresses the regime choice between the summary and AI-assisted regimes. The traditional summary regime uniformly dominates the AI-assisted regime if and only if the information loss in the summary is sufficiently small. Intuitively, when a summary already captures most of the relevant information, the marginal benefit of processing more detailed disclosures through AI is small, while the cognitive cost of hallucination verification remains substantial. As hallucination rates rise, the region where traditional summaries are preferred expands, because the verification burden further erodes the informational advantage of AI-processed details.

On the empirical side, I design and conduct an incentivized online experiment to test the model's predictions. Using the Prolific platform, I recruit 353 participants who analyze an AI-generated summary of PepsiCo's FY2022 annual report (10-K filing). Participants are randomly assigned to one of three hallucination levels with low, medium, or high that vary the number of paragraphs containing AI-introduced inaccuracies. The experimental interface records granular behavioral data: the time participants spend reading the AI summary, the time they spend viewing the original financial documents, the number of clicks on links to original text, and their accuracy on five comprehension questions. Before the main task, participants complete a seven-item Cognitive Reflection Test (CRT), which serves as the empirical proxy for the model's capacity parameter.

The experimental results strongly support the model's predictions. First, consistent with

Proposition 1, higher-capacity investors are significantly more likely to activate the diagnostic channel. A one-standard-deviation increase in CRT-based capacity raises the probability of diagnostic activation by 6.3 percentage points and increases total verification clicks by 0.675. Second, consistent with Corollary 1, participants in the high-hallucination group spend significantly more time on diagnostic activities and significantly less time parsing the AI summary, with the diagnostic ratio rising by 5.9 percentage points and the parsing ratio falling by 5.7 percentage points relative to the low-hallucination group. Third, consistent with the interaction between hallucination rate and capacity, the amplifying effect of capacity on diagnostic behavior is concentrated in the high-hallucination group, where a one-standard-deviation increase in capacity nearly doubles the treatment effect on diagnostic time.

Beyond validating the model, I document two additional empirical findings that deepen our understanding of AI-assisted financial analysis. First, diagnostic channel engagement substantially improves investment performance. Activating the diagnostic channel raises the number of correct answers by approximately 0.4, a 30% improvement relative to the sample mean. This performance effect operates within individuals: at the question level with individual fixed effects, additional diagnostic time on a specific paragraph significantly increases the probability of answering the corresponding question correctly. Second, I find evidence of real-time Bayesian belief updating. Using a within-individual panel that tracks verification behavior before and after participants first discover a hallucination, I show that post-discovery verification increases dramatically on subsequent paragraphs. This suggests that investors dynamically revise their assessment of AI reliability as they accumulate experience.

This paper makes several contributions to the literature. First, this is the first paper to develop a formal theoretical framework for investor limited attention in the age of generative AI. By extending the rational inattention framework of Sims (2003) and Lu (2022) to incorporate AI hallucination as a source of information unreliability, I provide a tractable model that generates testable predictions about how the hallucination rate, investor capacity, and

information structure jointly determine attention allocation. The model suggests that AI assistance introduces a new dimension of attention allocation that competes with traditional information extraction, which has broad implications for understanding how AI reshapes information processing in financial markets.

Second, this paper provides the first micro-level empirical evidence on the cognitive costs that investors incur when using generative AI for financial analysis. While prior work has documented AI's potential to improve productivity (Noy and Zhang, 2023; Dell'Acqua et al., 2023), reduce information processing costs (Blankespoor et al., 2020), and alter corporate disclosure strategies (Cao et al., 2023), the question of how investors actually allocate their attention when AI outputs may be unreliable has remained unaddressed. My experiment fills this gap by directly measuring the time and effort investors devote to parsing AI content versus diagnosing its reliability.

Third, this paper contributes to the growing literature on AI hallucination by providing experimental evidence on how hallucination affects investor behavior in a controlled setting. While prior work has characterized the causes and prevalence of AI hallucination (Ji et al., 2023; Kalai and Vempala, 2024), and has begun to study AI's effects on financial decisions (Croom, 2025; Kim et al., 2024), the behavioral consequences of hallucination for investor attention allocation have not been explored. My finding that investors dynamically reallocate attention in response to hallucination discovery provides novel evidence on the real-time cognitive costs of AI unreliability.

Fourth, this paper contributes to the experimental finance literature by introducing a novel process-tracing methodology for studying AI-assisted financial decision-making. Drawing on the tradition of information acquisition measurement in behavioral economics (Payne et al., 1993; Gabaix et al., 2006; Willemsen and Johnson, 2011), I design an experimental interface that continuously records participants' attention allocation across information sources. This approach enables researchers to observe not just the outcomes of AI-assisted decisions but the cognitive process through which investors navigate the tension between AI

convenience and verification.

Fifth, the findings have practical implications for the design of AI tools in finance and for regulatory policy. The existence of a capacity threshold below which investors do not verify AI output suggests that retail investors may be particularly vulnerable to AI hallucination. The crowding-out effect, whereby verification demands reduce the cognitive resources available for actual information processing, implies that the net value of AI assistance may be lower than commonly assumed, especially when hallucination rates are high.

The remainder of this paper proceeds as follows. Section 2 reviews the related literature. Section 3 develops the theoretical model. Section 4 describes the experimental design and data. Section 5 presents the empirical results. Section 6 concludes.

2 Literature Review

This paper connects to several strands of the literature in finance, accounting, and economics. I discuss each in turn, emphasizing how my work extends and integrates these research areas.

2.1 Disclosure Processing Costs and Limited Investor Attention

A large body of work in accounting and finance establishes that investors have limited capacity to process information and that this limitation affects asset prices and market outcomes. Hirshleifer and Teoh (2003) provide an early theoretical treatment, arguing that limited attention affects how financial disclosures are incorporated into stock prices, with information that is "recognized" in financial statements receiving more attention than information that is merely "disclosed" in footnotes. Empirically, DellaVigna and Pollet (2009) show that earnings announcements on Fridays, the time investor attention is lower, produce smaller immediate price reactions and larger post-earnings drift, while Hirshleifer et al. (2009) demonstrate that simultaneous earnings announcements by multiple firms dilute the attention available for each, leading to underreaction. Da et al. (2011) introduce Google

Search Volume as a direct measure of retail investor attention and link it to IPO pricing and abnormal trading volume. Barber and Odean (2008) document that attention-grabbing events drive individual investor purchasing behavior, and Cohen and Frazzini (2008) show that limited attention to economic links between firms generates return predictability along supply chains.

Blankespoor et al. (2020) provide a comprehensive review that organizes the literature on disclosure processing costs into three categories: awareness costs, acquisition costs, and integration costs. Their framework is particularly relevant to my setting, in which generative AI dramatically reduces awareness and acquisition costs by summarizing lengthy disclosures, but simultaneously introduces a new form of integration cost, that is, the need to verify AI output for hallucinations. My theoretical model formalizes this tradeoff by explicitly modeling the parsing and diagnostic channels through which investors process AI-generated information.

More recently, Gao and Huang (2020) study how modern information technologies affect information production in financial markets, finding that technological improvements can both facilitate and distort information flows. Ben-Rephael et al. (2017) use EDGAR search traffic to measure institutional investor attention and document systematic underreaction when attention is low. This study extends this literature by examining how AI, the latest and most transformative information technology, fundamentally alters the structure of investor attention allocation.

2.2 Rational Inattention

Sims (2003) models economic agents as having finite information processing capacity measured by shannon mutual information and shows that this capacity constraint generates sluggish responses to economic shocks. Maćkowiak and Wiederholt (2009) extend this framework to pricing decisions, showing that firms optimally allocate limited attention across multiple signals. Veldkamp (2011) provides a comprehensive treatment of information choice in

macroeconomics and finance, establishing the mathematical toolkit I employ.

In financial economics, the rational inattention framework has been applied to portfolio choice (Van Nieuwerburgh and Veldkamp, 2010; Huang and Liu, 2007; Mondria, 2010), mutual fund management (Kacperczyk et al., 2016), and home bias (Van Nieuwerburgh and Veldkamp, 2009). Peng and Xiong (2006) model investors who allocate limited attention across asset-specific and category-level information, generating excess comovement in returns. Kacperczyk et al. (2016) show that fund managers rationally shift attention between stock picking and market timing over the business cycle, providing evidence that professional investors indeed optimize their attention allocation as the rational inattention framework predicts.

Most directly, Lu (2022) applies the rational inattention framework to financial reporting, modeling how investors allocate attention across multiple components of periodic disclosures. I extend Lu's framework in a fundamental way by introducing the D+AI regime, in which generative AI processes detailed disclosures but introduces hallucination risk. This extension creates the novel two-channel structure including parsing and diagnostic channel that is the theoretical core of my paper. While Lu (2022) suggests the investor allocates attention across disclosure components, in this study, the investor faces the additional challenge of allocating attention between extracting and verifying AI-processed information, which is a distinction that becomes economically relevant only in the age of generative AI.

2.3 AI Hallucination and AI in Finance

The phenomenon of AI hallucination defined as the generation of plausible but factually incorrect outputs by large language models has received growing attention in computer science and, more recently, in economics and finance. Ji et al. (2023) provide a comprehensive survey categorizing hallucination types in natural language generation systems, while Huang et al. (2025) survey the principles, taxonomy, and challenges of hallucination in large language models. Most strikingly, Kalai and Vempala (2024) prove theoretically that calibrated

language models must hallucinate on certain queries, establishing that hallucination is an inherent feature of generative AI rather than a fixable engineering problem. This theoretical inevitability directly motivates my modeling choice of a strictly positive hallucination parameter.

In my model, AI hallucination is characterized following the framework of Bertomeu et al. (2024), who model AI-generated misinformation as a random "hallucination" unrelated to the firm's fundamentals. Specifically, the AI output is a mixture: it is an informative signal about the firm's fundamentals in the non-hallucination state, but an independent noise signal in the hallucination state. This formulation captures the key economic feature of hallucination, that is, hallucinated output is statistically indistinguishable from truthful output *ex ante* but contains no useful information about the underlying state.

The broader literature on AI in finance provides context for this paper. Cao et al. (2023) study how firms strategically alter their disclosures when they anticipate that AI tools will process them, while Kim et al. (2024) evaluate GPT-4's ability to analyze financial statements, documenting both impressive performance and systematic error patterns. Lopez-Lira and Tang (2023) test whether ChatGPT can predict stock returns from news headlines, highlighting the risks of relying on LLM outputs for financial decisions. Korinek (2023) surveys the use cases and implications of generative AI for economic research, discussing hallucination as a key limitation. Grennan and Michaely (2020) study how AI tools affect financial analysts' work quality and efficiency, providing evidence that AI complements rather than substitutes for human expertise. The emerging picture from this literature is that AI is a powerful but imperfect tool whose value depends critically on how users interact with it, which it is also a conclusion that my theoretical model formalizes and my experiment validates.

2.4 Cognitive Reflection and Investor Sophistication

The Cognitive Reflection Test (CRT), introduced by Frederick (2005), serves as the empirical proxy for the investor's capacity. The CRT measures the tendency to override an intuitive but incorrect response with a reflective and correct one, capturing what Toplak et al. (2014) term "miserly information processing." I employ a seven-item CRT battery drawn from three sources: the original three items from Frederick (2005), additional items from Toplak et al. (2014), and alternate-form items from Thomson and Oppenheimer (2016). Using multiple items mitigates the ceiling effects and familiarity concerns documented in the CRT literature (Toplak et al., 2011).

Several features make the CRT an appropriate proxy for the information processing capacity in my model. First, Toplak et al. (2011) demonstrate that CRT performance strongly predicts susceptibility to cognitive biases across a wide range of heuristics-and-biases tasks, suggesting it captures a general capacity for deliberative reasoning. Second, Oechssler et al. (2009) and Hoppe and Kusterer (2011) show that higher CRT scores are associated with fewer behavioral biases in economic decision-making, including reduced susceptibility to anchoring, framing, and the conjunction fallacy. Third, Pennycook and Rand (2019) find that CRT scores predict the ability to distinguish true from false information online, a finding directly relevant to my setting where investors must distinguish accurate AI output from hallucinations. Fourth, Stagnaro et al. (2018) demonstrate that CRT performance is stable over time, supporting its use as a between-subjects measure of a stable individual trait.

2.5 Process Tracing and Information Acquisition

My experimental methodology builds on the process-tracing tradition in behavioral economics and decision science. Payne et al. (1993) establish the foundational framework for studying adaptive decision strategies through information acquisition patterns, arguing that observing *how* people search for information reveals their underlying cognitive strategies. The Mouselab paradigm (Willemsen and Johnson, 2011) and its online successor MouselabWEB

have been widely adopted in economics for recording which information cells participants open, in what order, and for how long.

In a seminal contribution, Gabaix et al. (2006) use Mouselab-style information acquisition data to test a boundedly rational model of directed cognition, showing that experimental subjects acquire information selectively in a manner consistent with the marginal value of information. Costa-Gomes et al. (2001) use information lookup patterns to identify strategic reasoning levels in normal-form games. Arieli et al. (2011) track decision-makers under uncertainty, providing evidence that information search patterns reveal underlying preferences and beliefs.

My experiment extends this tradition to the study of AI-assisted financial analysis. Rather than using a traditional Mouselab information board, I design a custom interface that tracks attention allocation across three information sources: the AI-generated summary which indicates the parsing channel, the original financial documents which indicates the diagnostic channel, and the comprehension questions. The key innovation is that the "information acquisition" in my setting is not just about gathering facts but about *verifying* the reliability of an AI intermediary, which is a distinction that maps directly onto the two-channel structure of my theoretical model. The continuous recording of time-on-task, click events, and panel focus enables within-individual analysis that goes beyond aggregate treatment effects.

2.6 Experimental Finance and Online Platforms

Finally, this paper contributes to the growing body of experimental work in accounting and finance that uses online platforms. Libby et al. (2002) provide foundational guidance for experimental research design in financial accounting, establishing standards for internal validity, treatment design, and measurement. Farrell et al. (2017) validate the use of online labor platforms for accounting experiments, demonstrating that online subjects produce data of comparable quality to traditional laboratory participants. Palan and Schitter (2018)

specifically evaluate the Prolific platform, which I use, documenting its advantages in data quality, participant attentiveness, and demographic diversity relative to alternatives such as Amazon Mechanical Turk. Peer et al. (2017) provide additional evidence that Prolific produces more attentive and reliable participants.

In the specific context of AI and investor behavior, Croom (2025) studies how the interactivity inherent in generative AI affects investor judgments, finding that AI creates "illusions of ability" whereby investors misattribute AI capabilities to themselves and become overconfident. My paper complements his work by examining how AI hallucination risk restructures the allocation of investor attention between content extraction and verification. The combination of my theoretical model with granular behavioral data from a carefully designed experiment enables a more structural analysis of the cognitive costs of AI-assisted financial decision-making.

3 Theoretical Model

This section develops a rational inattention model of investor information processing in the presence of generative AI. I consider a representative investor who must form beliefs about a firm's fundamental value using either a traditional summary or an AI-assisted processing regime. The model extends the framework of Lu (2022) by introducing AI hallucination and a dual-channel attention allocation structure.

3.1 Model Setup

There is a firm with fundamental state $\theta \sim \mathcal{N}(0, \sigma_\theta^2) \equiv \mathcal{N}(0, A)$. In the summary regime (Regime S), the report discloses a noisy summary of the firm's fundamentals: $r_S = \theta + \gamma$, where $\gamma \sim \mathcal{N}(0, \sigma_\gamma^2)$ is an independent noise term representing information loss in the summary. A representative investor processes the summary with noise: $z = r_S + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is also an independent noise term. Under Lu (2022) rational inattention (RI)

constraints, the expected posterior variance (ex ante Bayes risk) for Regime S is:

$$V_S(C) = \frac{A(C^2\sigma_\gamma^2 + A)}{C^2(\sigma_\gamma^2 + A)}, \quad C \geq 1. \tag{3.1}$$

This is a direct rewrite of equation (11) in Lu (2022).

In the detailed disclosure plus AI processing regime (D+AI regime), a representative investor uses generative AI to handle several detailed disclosures and asks AI to output one signal, named s_{AI}, reflecting the firm's fundamental state. However, the AI output, s_{AI}, is mixed. Specifically, with probability $(1 - \lambda)$, the AI output is no-hallucination (NH) state: $s_{NH} = \theta + \varepsilon_{AI}$, where $\varepsilon_{AI} \sim \mathcal{N}(0, \sigma_{AI}^2) \equiv \mathcal{N}(0, B)$. With probability λ, the AI output is hallucination (H) state: $s_H \perp \theta$ and $s_H \sim \mathcal{N}(0, \sigma_\theta^2 + \sigma_{AI}^2) \equiv \mathcal{N}(0, A + B)$, which means s_H follows the same distribution as s_{NH} and is independent of fundamental, following Bertomeu et al. (2024). The investor then allocates her attention into two channels:

In parsing channel, the investor observes the AI output with additional noise,

$$\tilde{s} = s_{AI} + \nu, \quad \nu \sim \mathcal{N}(0, \sigma_p^2). \tag{3.2}$$

In diagnostic channel, the consistency, or reliability, of the AI summary compared with the original detailed reporting is a continuous latent variable:

$$c \sim \mathcal{N}(\mu_c, \sigma_c^2). \tag{3.3}$$

The "no-hallucination" (NH) state is a hard threshold, and $\bar{c}$ is a fixed threshold:

$$Z = \mathbf{1}\{c \geq \bar{c}\}. \tag{3.4}$$

Therefore, the prior hallucination probability, which is also denoted as λ, is:

$$1 - \lambda \equiv \Pr(Z = 1) = \Pr(c \geq \bar{c}) = \Phi\left(\frac{\mu_c - \bar{c}}{\sigma_c}\right). \tag{3.5}$$

Then we can have:

$$d \equiv \frac{\mu_c - \bar{c}}{\sigma_c} = \Phi^{-1}(1 - \lambda). \tag{3.6}$$

The investor uses attention to verify consistency c, obtaining a noise-diagnostic signal t:

$$t = c + \eta, \quad \eta \sim \mathcal{N}(0, \sigma_d^2). \tag{3.7}$$

3.2 Mutual Information and Capacity Constraints

The investor allocates attention to process two independent sources of uncertainty: the content of the AI signal (s_{AI}) and the consistency of the AI signal (c). The observed signals are $\tilde{s}$ and t, respectively. Using the chain rule for mutual information, the total information processed is:

$$I\Big((s_{AI}, c); (\tilde{s}, t)\Big) = I\Big((s_{AI}, c); \tilde{s}\Big) + I\Big((s_{AI}, c); t \mid \tilde{s}\Big). \tag{3.8}$$

Consider the first term. Since $\tilde{s} = s_{AI} + \nu$ depends only on s_{AI} and independent noise ν, and because the marginal distribution of s_{AI} is independent of the consistency state c (as s_H and s_{NH} follow the same distribution), $\tilde{s}$ provides no information about c. Thus:

$$I\Big((s_{AI}, c); \tilde{s}\Big) = I(s_{AI}; \tilde{s}) + \underbrace{I(c; \tilde{s} \mid s_{AI})}_{0} = I(s_{AI}; \tilde{s}). \tag{3.9}$$

Now consider the second term. The diagnostic signal $t = c + \eta$ depends only on c and independent noise η. Given that $\tilde{s}$ and s_{AI} are independent of c and η, they provide no

information about t. Thus:

$$I\Big((s_{AI}, c); t \mid \tilde{s}\Big) = I(c; t \mid \tilde{s}) + \underbrace{I(s_{AI}; t \mid \tilde{s}, c)}_{0} = I(c; t \mid \tilde{s}). \tag{3.10}$$

Combining these, the total mutual information is the sum of the information gained from each channel:

$$I\Big((s_{AI}, c); (\tilde{s}, t)\Big) = I(s_{AI}; \tilde{s}) + I(c; t \mid \tilde{s}). \tag{3.11}$$

The distribution of s_{AI} in the hallucination (H) state is exactly the same as the distribution in the non-hallucination (NH) state, which means $\tilde{s}$ and s_{AI} are statistically independent with c and t. In other words, $s_{AI} \perp c$ and $\tilde{s} \perp c$ make $H(c \mid \tilde{s}) = H(c)$ and $H(c \mid t, \tilde{s}) = H(c \mid t)$. Therefore:

$$I(c; t \mid \tilde{s}) = H(c \mid \tilde{s}) - H(c \mid t, \tilde{s}) = H(c) - H(c \mid t) = I(c; t). \tag{3.12}$$

To capture limited attention, studies in the rational inattention literature (e.g., Sims, 2003; Maćkowiak and Wiederholt, 2009) assume that the reduction in uncertainty (or equivalently, mutual entropy) is bounded by some capacity term. That is,

$$I\Big((s_{AI}, c); (\tilde{s}, t)\Big) = I(s_{AI}; \tilde{s}) + I(c; t) \leq \log_2 C, \tag{3.13}$$

where $\log_2 C$ is the investor's information processing capacity. Then this constraint simplifies into the following inequality following Veldkamp (2011) and Maćkowiak and Wiederholt (2009):

$$\underbrace{\left(1 + \frac{A+B}{\sigma_p^2}\right)}_{a_p} \cdot \underbrace{\left(1 + \frac{\sigma_c^2}{\sigma_d^2}\right)}_{a_d} = C^2, \quad 1 \leq a_p, a_d \leq C^2. \tag{3.14}$$

The terms a_p and a_d represent the information gains from the parsing and diagnostic channels, respectively. The product structure of the capacity constraint reflects the additive separability of mutual information across independent channels: the investor's total capacity C^2 must be distributed across the two channels, but the constraint operates multiplicatively

in terms of signal-to-noise ratios.

3.3 Objective Function and Optimization

There is a firm with fundamentals θ, which is normally distributed with mean 0 and variance σ_θ^2. A representative investor is choosing investment k to match the fundamentals. The investor's utility is given by

$$u = \theta k - \frac{1}{2}k^2. \tag{3.15}$$

Using Ω to denote the information set of the investor, the first-order condition implies that the optimal investment is equal to the conditional expectation of the fundamentals: $k^* = E[\theta|\Omega]$. Plugging k^* into the utility function, it can be easily shown that the investor's ex ante expected utility can be written as

$$\mathbb{E}[u] = \frac{1}{2}(\text{Var}(\theta) - \text{Var}(\theta|\Omega)). \tag{3.16}$$

Therefore, the investor's expected utility is proportional to the difference between the prior variance and the posterior variance of the fundamentals. It becomes clear that the investor's preferred regime is the one that allows her to minimize the posterior variance of θ.

The posterior variance of θ in S, $V_S(C)$, has been shown in equation (S). The expected posterior variance of θ in D+AI, V_{D+AI}, can be written as follows:

$$V_{D+AI}(C; a_d) = A - \frac{A^2(C^2 - a_d)}{C^2(A+B)} f(a_d), \quad C > 1, \quad 1 \leq a_d \leq C^2, \tag{3.17}$$

where

$$f(a_d) = \Phi_2(d, d; \rho(a_d)), \qquad \rho(a_d) = 1 - \frac{1}{a_d}, \qquad d = \frac{\mu_c - \bar{c}}{\sigma_c} = \Phi^{-1}(1-\lambda), \tag{3.18}$$

with Φ_2 denoting the bivariate normal CDF with correlation ρ.

The function $f(a_d)$ captures the probability that the investor correctly identifies the AI output as non-hallucinated, accounting for the precision of the diagnostic signal. As a_d increases (more attention to the diagnostic channel), the correlation ρ approaches 1 and $f(a_d)$ approaches $\Phi(d)^2$, reflecting near-perfect diagnostic ability. When $a_d = 1$ (no diagnostic attention), $f(1) = \Phi(d)^2$, and the investor cannot update her beliefs about AI reliability beyond the prior.

3.4 Optimal Attention Allocation

The investor's problem is to choose $a_d \in [1, C^2]$ to minimize $V_{D+AI}(C; a_d)$, or equivalently, to maximize $(C^2 - a_d)f(a_d)$. The following lemma establishes the concavity of the objective:

Lemma 1. *If the prior hallucination rate λ is not too extreme, specifically lying between 0.07865 and 0.92135, equivalently $|d| \leq \sqrt{2}$ where $d = \Phi^{-1}(1-\lambda)$, then $f(a_d)$ is concave for all $a_d \in [1, C^2]$.*

This condition is empirically well-satisfied. According to the Vectara Hallucination Leaderboard, which tracks hallucination rates across major large language models, mainstream models exhibit hallucination rates well within this range. For instance, Google Gemini 2.5 Flash at approximately 7.8%, OpenAI GPT-5.2 Low around 8.4%, Anthropic Claude Sonnet 4 at approximately 10.3%, and OpenAI GPT-5.1 High at about 12.1%.

The main result on attention allocation is:

Proposition 1. *Given a moderate hallucination rate as shown in Lemma 1, in Regime D+AI, there exists a capacity threshold*

$$C^* = \sqrt{1 + \left(\frac{\Phi(d)}{\phi(d)}\right)^2}, \qquad d = \Phi^{-1}(1-\lambda). \tag{3.19}$$

When capacity is small enough so that $C < C^*$, the investor devotes all his capacity to the parsing channel and ignores the diagnostic channel. When capacity is large so that $C > C^*$,

the investor allocates attention to both channels and chooses a unique pair of $a_d^*, a_p^* \in (1, C^2)$ *satisfying:*

$$f(a_d^*) = (C^2 - a_d^*)f'(a_d^*), \qquad a_p^* = \frac{C^2}{a_d^*}. \tag{3.20}$$

Proposition 1 characterizes the investor's attention strategy when using generative AI to process detailed information disclosure in financial reporting. This result mirrors and extends the findings in Lu (2022) regarding attention allocation between components.

In the D+AI regime, the tradeoff is structural rather than additive. The investor chooses between parsing or extracting the signal content $\tilde{s}$ and diagnosing or verifying the signal's reliability t. In the low attention regime that the capacity is lower than the capacity threshold, investors are essentially blindly trusting or blindly listening to the AI's output. They lack the surplus cognitive resources to adjust weights based on AI reliability through the received diagnostic signal. This behavior resembles what Lu (2022) describes, that is, when information is overly complex like too many details and attention is limited, investors may be forced to ignore certain dimensions. Here, what is ignored is the verification of information quality.

When the capacity exceeds the diagnostic threshold, the marginal benefit of further reducing parsing noise begins to diminish, while the value of identifying hallucinations through the diagnostic signal becomes prominent. High-capability investors not only hear more clearly, but are also better at distinguishing truth from falsehood. This dual enhancement explains why professional investors can better leverage generative AI, while retail investors may be more easily trapped by AI hallucinations—because the latter may not allocate any resources to activate the diagnostic channel at all.

3.5 Comparative Statics: Hallucination Rate and Capacity

Corollary 1. *The investor is more likely to start the diagnostic channel when the AI hallucination rate is higher. For an investor who has already activated diagnostic attention, an*

increase in hallucination rate leads to a strict increase in diagnostic investment and a strict decrease in parsing investment.

The parameter λ represents the probability that AI produces hallucinations. When AI hallucination rate increases, the risk of blind trusting the AI output increases significantly. To mitigate this heightened risk, investors feel a more urgent need to verify the authenticity of the information. Consequently, even investors with lower processing capacity are compelled to activate the diagnostic channel earlier to filter out hallucinations. In addition, a higher hallucination rate will lead investors to pay more attention to fact-checking, thereby crowding out the attention paid to interpreting the AI's output. This crowding-out effect implies that as AI hallucination rates increase, the net informational value of AI assistance is doubly eroded, because the hallucinated content not only reduces the quality of the AI signal but also creates a need for verification that diverts cognitive resources away from extracting its informational content.

Corollary 2. *Given the hallucination rate, for investors who have already activated diagnostic attention, a higher capacity does not necessarily imply more parsing. When hallucination rate is sufficiently low, additional capacity is first reallocated toward verification, so parsing can be temporarily crowded out. When hallucination rate is not that low, parsing investment increases monotonically with capacity. The diagnostic investment always increases monotonically with capacity.*

Corollary 2 reveals a perhaps surprising result: giving an investor more cognitive capacity does not always result in better parsing of AI output. The intuition is as follows. When the hallucination rate is sufficiently high, both the diagnostic and parsing channels benefit from additional capacity, and the returns to parsing are high enough that parsing increases monotonically with capacity. However, when the hallucination rate is relatively low, the investor initially underinvests in the diagnostic channel, since the perceived risk is moderate, and additional capacity is first channeled toward verification to reach an optimal diagnostic

level. Only after the diagnostic channel reaches a sufficient precision does surplus capacity flow into parsing. This result has important implications for the heterogeneous effects of AI adoption across investor types. For professional investors like institutional investors and analysts, the model predicts that both channels are well-funded. They extract substantial information from AI summaries while simultaneously maintaining effective fact-checking. For retail investors, the model predicts two possible outcomes depending on where they fall relative to diagnostic threshold. Either they are below the threshold and blindly trust AI, or they are just above the threshold and devote most of their incremental capacity to diagnostics, leaving parsing relatively impoverished.

3.6 Regime Choice

Proposition 2. *For all finite capacities $C > 1$, the regime S uniformly strictly dominates the regime D+AI if and only if the information loss in the summary are sufficiently small. Formally, there exists threshold*

$$T = \frac{A+B}{\Phi(d)} = \frac{A+B}{1-\lambda} - A \tag{3.21}$$

such that regime S uniformly strictly dominates the regime D+AI if and only if $\sigma_\gamma^2 \leq T$. Besides, regime S is more likely to dominate regime D+AI when the AI hallucination rate is higher.

Proposition 2 addresses a fundamental question: when should an investor prefer the traditional summary regime over the AI-assisted regime? The answer depends on the relative magnitudes of two forms of information loss. The summary regime loses information through compression , while the D+AI regime loses information through hallucination and the associated need to divert attention to verification.

When the summary is sufficiently informative, the information loss from compression is small, and the investor is better off reading the summary directly rather than using AI and

bearing the cognitive cost of hallucination verification. This means that as AI hallucination rates rise, the region where the traditional summary dominates expands. This result provides a theoretical foundation for understanding when AI adoption is value-destroying for investors. Even though AI can process more detailed information than a human-readable summary, the cognitive cost of verifying AI output may outweigh the informational gains. This is particularly relevant as firms produce increasingly detailed disclosures that investors struggle to process (Blankespoor et al., 2020): generative AI offers to bridge this gap, but Proposition 2 shows that the bridge may collapse if the hallucination rate is too high.

The theoretical model generates the following testable predictions, which I take to the experimental data:

Prediction 1 (Capacity Threshold). *Investors with higher information processing capacity are more likely to activate the diagnostic channel.* This follows directly from Proposition 1, that is, the probability of diagnostic activation is $\Pr(C > C^*)$, which is increasing in C for any given C^*.

Prediction 2 (Hallucination and Attention Allocation). *A higher AI hallucination rate increases diagnostic investment and decreases parsing investment.* This follows from Corollary 1.

Prediction 3 (Interaction Effect). *The positive effect of hallucination rate on diagnostic investment is amplified by investor capacity.* This follows from the combination of Proposition 1 and Corollary 1, that is, higher hallucination rates lower the activation threshold, bringing more investors into the diagnostic regime, and conditional on activation, both hallucination rate and capacity increase diagnostic investment.

Prediction 4 (Performance). *Diagnostic channel activation improves investment performance.* This follows from the model's implication that investors who diagnose AI reliability can condition their inference on the hallucination state, reducing posterior variance.

Prediction 5 (Belief Updating). *Investors who discover hallucinations increase diagnostic*

investment on subsequent information. This follows from Bayesian updating: discovering a hallucination increases the posterior hallucination probability, which by Corollary 1 increases diagnostic investment.

4 Experimental Design and Data

This section describes the experimental design used to test the model's predictions. I conduct an incentivized online experiment in which participants analyze an AI-generated summary of a publicly listed company's annual report under randomly assigned hallucination levels.

The experiment is conducted on the Prolific platform, which has been validated for behavioral research in accounting and finance (Palan and Schitter, 2018; Peer et al., 2017; Farrell et al., 2017). Participants are presented with an AI-generated summary of PepsiCo's FY2022 annual report (10-K filing) and are asked to answer five multiple-choice comprehension questions about the company's financial performance, operations, and risk factors. The AI summary is divided into five paragraphs, each corresponding to a different topic area (business overview, financial performance, divisional results, management discussion of impairments, and risk factors). Crucially, the AI summary may contain hallucinations that will lead to incorrect answers of the questions and participants can verify the AI's claims by clicking "Link to Original Text" buttons that open scanned excerpts from the actual annual report.

The experiment features a between-subjects design with three treatment conditions that vary the hallucination rate: Low, Medium, and High. Treatment assignment uses balanced block randomization, where every block of three consecutive participants receives exactly one of each level in a randomly shuffled order, ensuring near-equal allocation across conditions. Participants are explicitly informed of their assigned hallucination level, which is displayed as a colored badge (green for Low, yellow for Medium, red for High) in the test interface header. This design choice ensures that the treatment variable corresponds to the model's

parameter λ, which represents both the actual and perceived hallucination rate.

4.1 Treatment: Hallucination Levels

The three hallucination levels vary the number of paragraphs in the AI summary that contain inaccuracies. The hallucinations are implemented as subtle omissions or distortions rather than fabricated content, mimicking realistic AI hallucination patterns:

- **Low:** Two paragraphs contain hallucinations. The second paragraph, EBITDA paragraph, uses an incorrect operating profit figure ($10,080M instead of the reported $11,512M), yielding a wrong margin calculation. The third paragraph, divisional performance paragraph, incorrectly attributes QFNA revenue growth partly to "organic volume growth," which is not supported by the original report.
- **Medium:** Three paragraphs contain hallucinations. In addition to the Low-level errors, the fourth paragraph, impairment paragraph, omits the critical qualifier "in Russia" when describing the discontinuation of juice and dairy brands, making it harder to identify that both management decisions and the Russia-Ukraine conflict contributed to European segment impairments.
- **High:** Four paragraphs contain hallucinations. In addition to the Medium-level errors, the first paragraph, business overview paragraph, omits "Europe" from the list of PepsiCo's seven reportable segments, listing only six segments while still stating the total is seven.

Importantly, the fifth paragraph (risk factors and legal proceedings) is accurate in all three conditions, providing a clean baseline for within-individual analysis. This nested design ensures that the treatment effect operates through the extensive margin of hallucination exposure, consistent with the model's parameter λ representing the prior probability of encountering a hallucination in any given paragraph.

4.2 Experiment Procedure

The experiment proceeds through four stages: In the first stage, participants receive detailed instructions explaining that they will read an AI-generated summary of a U.S. publicly listed company's annual report, that the summary may contain inaccuracies due to AI hallucination, and that they can verify the AI's claims by clicking links to view original report excerpts. They are informed of their randomly assigned hallucination level and provide informed consent before proceeding. Next, participants need to provide demographic information including gender, age, education, finance background, industry experience, and investment experience and then complete a seven-item Cognitive Reflection Test (CRT). The CRT items are drawn from Frederick (2005), Toplak et al. (2014), and Thomson and Oppenheimer (2016), and measure the participant's tendency to override intuitive but incorrect responses with reflective and correct ones. The CRT score serves as the empirical proxy for the model's capacity parameter C. Upon starting the test, the browser enters fullscreen mode and a 10-minute countdown timer begins. The test interface features a split-panel layout: the left panel (55% of screen width) displays the AI-generated summary with five paragraphs and "Link to Original Text" buttons, while the right panel (45%) displays five collapsible multiple-choice questions in an accordion format. Clicking a link opens a modal overlay displaying scanned images from the actual PepsiCo 10-K filing. The experiment continuously records: (i) the time the participant's cursor spends on the AI summary panel (parsing time), (ii) the time spent viewing original text modals (diagnostic time), (iii) the time spent on the question panel, (iv) the number and timing of link clicks, and (v) mouse position and scroll events sampled at regular intervals. Last, the test ends when the participant submits all answers or the 10-minute timer expires. Participants who exit fullscreen mode during the test are immediately disqualified to ensure data integrity.

Several anti-cheating measures are implemented: text selection is disabled, right-click and keyboard shortcuts (Ctrl+C, Ctrl+A, F12) are blocked, and the use of external AI tools is prohibited. Each participant receives a base payment of $2.50 for approximately 15 minutes

of participation (equivalent to $10 per hour), plus a bonus of $0.10 for each correct answer, providing financial incentives for accurate responses. The experiment is hosted online and can be accessed at https://singular-truffle-59ec1c.netlify.app.

4.3 Measurement

The experimental design yields several key variables that map directly onto the theoretical model. Capacity is measured by the CRT score (number of correct answers out of seven), standardized to create the variable *Capacity*. The CRT captures the investor's ability to engage in deliberative, reflective thinking, which corresponds to the information processing capacity C in the model. *Diagnostic Duration* records the total seconds spent viewing original text modals. *Diagnostic Ratio* equals Diagnostic Duration divided by *Test Duration*, capturing the share of total time devoted to verification. *Diagnostic Activation* is an indicator variable equal to one if the participant clicked at least one "Link to Original Text" button, corresponding to whether $C > C^*$ in the model. *Total Clicks* counts the number of times links to original text were opened. *Parsing Duration* records the net time spent on the AI summary panel (total AI panel time minus time when original text modals were open, to avoid double-counting). *Parsing Ratio* equals *Parsing Duration* divided by *Test Duration*. *Total Correct Answer* counts the number of correctly answered questions (out of five). *Correct Dummy* indicates whether the participant answered at least one question correctly. *High* and *Medium* are indicator variables for the high and medium hallucination treatment groups, with Low as the omitted baseline. The definitions of other variables are shown in the APPENDIX A.

4.4 Sample Construction

I aimed to recruit 300 participants through the Prolific platform. Ultimately, 461 individuals initiated the experiment. Of these, the Prolific platform automatically returned 139 responses (although some participant records were saved and remained valid), and 5 participants timed out. I rejected 26 responses with total test times below 5 minutes or above 30

minutes. After further cleaning, I excluded participants who exited fullscreen mode during the test (resulting in automatic disqualification) and those with total test times below 2 minutes. The final sample consists of 353 valid responses: 120 in the Low group, 119 in the Medium group, and 114 in the High group, closely reflecting the balanced block randomization design.

4.5 Summary Statistics and Balance Checks

Table 1 presents the balance check across the three treatment groups. The randomization is well-balanced, in which none of the six individual characteristics including CRT Score, Finance Background, Industry Background, Male, Education, and Investment Experience differs significantly across groups at the 10% level, and Age is not significant at 5% level. The robust F-statistics from joint tests of the Medium and High indicators are uniformly small, with p-values ranging from 0.070 (Age) to 0.975 (CRT Score). The marginal significance of Age ($F = 2.68$, $p = 0.070$) warrants the inclusion of individual-level controls in the regression analysis, though the economic magnitude of the age difference (approximately 3 years between the Low and High groups) is modest.

Table 2 Panel A reports the full-sample summary statistics. The average CRT score is 3.93 out of 7 (s.d. = 2.03), indicating substantial heterogeneity in cognitive reflection across participants. The average test duration is approximately 433 seconds (7.2 minutes) out of the 10-minute maximum. Participants spend an average of 86.6 seconds viewing original text, 167.9 seconds on the AI summary, and 176.9 seconds on questions. The mean diagnostic ratio is 17.1%, while the mean parsing ratio is 38.5%, indicating that participants devote roughly twice as much time to parsing as to diagnostics. The mean number of link clicks is 3.01, and 52.7% of participants activate the diagnostic channel at least once. The average number of correct answers is 1.35 out of 5.

Table 2 Panel B presents t-tests comparing the Low and High hallucination groups. Consistent with the model's predictions, the High group spends significantly more time on

diagnostic activities: Diagnostic Duration is 33.2 seconds higher ($t = -2.08$, $p < 0.05$) and the Diagnostic Ratio is 6.2 percentage points higher ($t = -2.06$, $p < 0.05$) in the High group relative to the Low group. The Parsing Ratio is 5.2 percentage points lower ($t = 1.87$, $p < 0.10$) in the High group, suggesting attention reallocation from parsing to diagnostics. Total Clicks and Total Correct Answers do not differ significantly between groups, consistent with the model's prediction that hallucination rate shifts the composition rather than the total level of attention.

5 Empirical Results

This section presents the main empirical findings. I organize the results around the five predictions derived from the theoretical model, proceeding from diagnostic channel activation (Prediction 1) through attention allocation (Predictions 2–3), performance effects (Prediction 4), and belief updating (Prediction 5).

5.1 Capacity and Diagnostic Channel Activation

Prediction 1 states that investors with higher information processing capacity are more likely to activate the diagnostic channel. Table 3 tests this prediction using three specifications. Column (1) reports average marginal effects from a logit regression of the Diagnostic Activation indicator on Capacity, the Medium and High treatment indicators, and individual-level controls including age, finance background, industry background, gender, education, and investment experience. Column (2) reports the corresponding OLS regression. Column (3) uses Total Clicks as the dependent variable.

The results strongly support Prediction 1. In Column (1), a one-standard-deviation increase in Capacity raises the probability of diagnostic activation by 6.3 percentage points ($t = 2.40$, $p < 0.05$). This effect is economically meaningful. It corresponds to a 12% increase relative to the baseline activation rate of 52.7%. The OLS model in Column (2) yields a

virtually identical estimate of 6.3 percentage points ($t = 2.33$, $p < 0.05$). In Column (3), Capacity has a highly significant positive effect on Total Clicks: a one-standard-deviation increase in capacity raises total verification clicks by 0.675 ($t = 3.30$, $p < 0.01$).

Notably, the treatment indicators (Medium and High) are not individually significant in these regressions. This is consistent with the model's prediction that the hallucination rate affects the *threshold* at which diagnostic activation occurs rather than having a direct effect conditional on Capacity. In other words, the hallucination rate interacts with capacity to determine activation, a prediction I test directly in Section 5.3.

These findings are consistent with Proposition 1: investors with higher cognitive capacity are better able to surpass the threshold C^* and engage in verification of AI output. Low-capacity investors, by contrast, tend to accept the AI summary at face value, which is a pattern that the model interprets as rational given their binding capacity constraints, but one that leaves them vulnerable to hallucinated information.

5.2 Hallucination Rate and Attention Allocation

Prediction 2 states that a higher AI hallucination rate increases diagnostic investment and decreases parsing investment. Table 4 tests this prediction using OLS regressions of the four main attention allocation variables on the treatment indicators, Capacity, and individual-level controls. Columns (1) and (2) use the time-based measures including Diagnostic Duration and Parsing Duration, winsorized at the 1st and 99th percentiles, while Columns (3) and (4) use the ratio-based measures including Diagnostic Ratio and Parsing Ratio.

The results provide strong support for Prediction 2 across all specifications. In Column (1), participants in the High hallucination group spend 28.96 more seconds on diagnostic activities relative to the Low group ($t = 1.86$, $p < 0.10$), while in Column (2), they spend 24.22 fewer seconds parsing the AI summary ($t = -1.79$, $p < 0.10$). The ratio-based specifications in Columns (3) and (4) yield sharper results: the High group has a Diagnostic Ratio that is 5.9 percentage points higher ($t = 1.98$, $p < 0.05$) and a Parsing Ratio that is

5.7 percentage points lower ($t = -2.07$, $p < 0.05$). The Medium treatment coefficients are positive for diagnostic and negative for parsing but not statistically significant, consistent with the medium hallucination level representing an intermediate treatment intensity.

Capacity is highly significant across all specifications. A one-standard-deviation increase in Capacity raises Diagnostic Duration by 24.09 seconds ($t = 3.62$, $p < 0.01$), reduces Parsing Duration by 13.27 seconds ($t = -2.17$, $p < 0.05$), increases the Diagnostic Ratio by 4.3 percentage points ($t = 3.57$, $p < 0.01$), and reduces the Parsing Ratio by 3.0 percentage points ($t = -2.49$, $p < 0.05$). These findings confirm that capacity is a fundamental determinant of attention allocation, consistent with the model's prediction that higher-C investors allocate more attention to both channels but especially to the diagnostic channel.

Taken together, the results in Table 4 directly validate Corollary 1: higher hallucination rates crowd out parsing in favor of diagnostics. The economic magnitudes are substantial: the 5.9 percentage point shift in the Diagnostic Ratio from Low to High represents a 34.5% increase relative to the Low group mean (0.143), while the 5.7 percentage point decrease in the Parsing Ratio represents a 14.5% decrease relative to the Low group mean (0.394). These findings indicate that AI hallucination imposes a significant cognitive cost on investors by forcing them to reallocate attention from information extraction to information verification.

5.3 Interaction Effect Between Hallucination Rate and Capacity

Prediction 3 states that the positive effect of hallucination rate on diagnostic investment is amplified by investor capacity. Table 5 tests this prediction by augmenting the Table 4 specifications with interaction terms between the treatment indicators and Capacity.

The interaction effects are significant and economically meaningful for the high-hallucination group. In Column (1), the High $\times$ Capacity interaction coefficient for Diagnostic Duration is 32.11 ($t = 2.16$, $p < 0.05$), indicating that the effect of the high-hallucination treatment on diagnostic time is substantially larger for high-capacity investors. In Column (3), the corresponding interaction for the Diagnostic Ratio is 0.060 ($t = 2.17$, $p < 0.05$). These in-

teraction effects are quantitatively large: the direct effect of High on Diagnostic Duration is 29.09 ($t = 1.88$, $p < 0.10$), so the interaction coefficient implies that a one-standard-deviation increase in capacity nearly doubles the high-hallucination treatment effect on diagnostic time.

By contrast, the Medium × Capacity interactions are not statistically significant in any specification (Diagnostic Duration: $\beta = 15.58$, $t = 1.05$; Diagnostic Ratio: $\beta = 0.019$, $t = 0.70$). This pattern is consistent with the model's prediction that the interaction between hallucination rate and capacity is most pronounced at higher hallucination levels, where the diagnostic channel becomes more valuable and the capacity constraint more binding.

These findings validate Prediction 3 and provide micro-level evidence for the mechanism underlying Proposition 1 and Corollary 1. High-capacity investors respond more strongly to higher hallucination rates because they have the cognitive resources to intensify their diagnostic efforts, while low-capacity investors are already constrained and cannot meaningfully increase verification even when the need is greater.

5.4 Diagnostic Channel and Task Performance

Prediction 4 states that diagnostic channel activation improves investment performance. I test this prediction at both the individual level shown in Table 6 Panel A and the question level shown in Table 6 Panels B and C, using increasingly demanding identification strategies.

Table 6 Panel A reports individual-level regressions of performance on diagnostic channel measures. In Column (1), the coefficient on Diagnostic Activation for Total Correct Answer is 0.402 ($t = 8.43$, $p < 0.01$), indicating that participants who activated the diagnostic channel answered approximately 0.4 more questions correctly—a 29.8% improvement relative to the sample mean of 1.35. In Column (2), each additional click on a link to original text improves performance by 0.068 correct answers ($t = 12.06$, $p < 0.01$). Columns (3) and (4) report average marginal effects from logit regressions using the Correct Dummy (at least one correct answer) as the dependent variable, and the results are qualitatively similar: Diagnostic Activation raises the probability of answering at least one question correctly

by 11.0 percentage points ($t = 5.48$, $p < 0.01$), and each additional click raises it by 1.6 percentage points ($t = 5.23$, $p < 0.01$). Interestingly, after controlling for diagnostic behavior, the Medium treatment indicator becomes significantly negative ($\beta = -0.248$, $t = -4.52$, $p < 0.01$ in Column 1; $\beta = -0.255$, $t = -4.75$, $p < 0.01$ in Column 2), while the High indicator is not significant. This pattern suggests that the additional hallucination in the Medium group (relative to Low) reduces performance for participants who do not adequately verify, while the even higher hallucination in the High group triggers sufficient verification behavior to offset the direct negative effect on accuracy. Capacity is not significant after controlling for diagnostic behavior, suggesting that the positive effect of capacity on performance operates primarily through the diagnostic channel—consistent with the model's mechanism.

Table 6 Panel B exploits within-individual variation across five questions using conditional logit models with individual fixed effects. This specification eliminates all individual-level heterogeneity including capacity, demographics, and unobserved traits and identifies the effect of paragraph-specific diagnostic behavior on question-specific accuracy. In Column (1), the coefficient on Diagnostic Time (seconds spent viewing the original text for the specific paragraph corresponding to each question) is 0.002 ($t = 3.21$, $p < 0.01$). This result demonstrates that even within the same individual, spending more time verifying the original source material for a specific paragraph significantly increases the probability of answering the corresponding question correctly. Columns (2) and (3) use alternative diagnostic measures (an indicator for any clicks on that paragraph, and the number of clicks), which are positive but not statistically significant at conventional levels. The Hallucinated indicator, which equals one if the paragraph contains a hallucination for the participant's treatment group, is not significant in any specification. This null finding is important: it indicates that, conditional on diagnostic effort, whether a paragraph is hallucinated does not independently affect accuracy. This is consistent with the model's prediction that the cost of hallucination is not direct reflecting participants are not uniformly misled but indirect indicating it requires attention reallocation to diagnostics.

Table 6 Panel C provides the most demanding test by decomposing diagnostic behavior into between-individual and within-individual components. The key variables are Mean Clicks measured by the participant's average clicks across all five paragraphs, Excess Mean Clicks measured by clicks on a specific paragraph minus the participant's average, Diagnostic Time, and Excess Diagnostic Time measured by diagnostic time on a specific paragraph minus the participant's average. The results reveal that both between- and within-individual variation in diagnostic behavior predict performance. Mean Clicks is highly significant across specifications ($\beta = 0.063$, $t = 5.71$, $p < 0.01$ in Column 4), confirming that participants who click more overall perform better. Excess Diagnostic Time is significant when included ($\beta = 0.002$, $t = 3.04$, $p < 0.01$ in Column 3), indicating that within-individual deviations, spending relatively more time verifying a specific paragraph compared to one's own average could independently predict accuracy on that question. These decomposition results provide compelling evidence that the performance effect of diagnostic behavior is not driven solely by individual-level confounders (e.g., motivation or ability) but reflects the mechanism that verifying specific paragraphs against original sources enables more accurate comprehension, consistent with the model's prediction that the diagnostic channel reduces the posterior variance of the firm's fundamentals.

5.5 Belief Updating After Hallucination Discovery

Prediction 5 states that investors who discover hallucinations increase diagnostic investment on subsequent information. Table 7 tests this prediction using a within-individual panel that tracks verification behavior before and after participants first successfully discover a hallucination. The identification relies on comparing the same participant's behavior on the same paragraph across two time phases. The pre-discovery phase refers to the period before a participant first correctly identifies a hallucinated paragraph, while the post-discovery phase refers to the period after this first correct identification.

The key variables are: *Post*, an indicator equal to one for events occurring after the

discovery index; *Later Paragraph*, an indicator equal to one for paragraphs sequenced after the discovered hallucination paragraph; and the interaction *Post* × *Later Paragraph*, which captures the difference-in-differences effect of discovery on subsequent verification behavior. The discovery events shown in the first paragraph and last hallucination paragraph are excluded to ensure that there exist events both before and after.

Columns (1)–(4) report results with individual-paragraph fixed effects. The Post × Later Paragraph interaction is highly significant across all outcome variables. In Column (1), the interaction coefficient for Clicks is 1.240 ($t = 4.36$, $p < 0.01$), indicating that after discovering a hallucination, participants make 1.24 additional clicks on subsequent paragraphs relative to their pre-discovery behavior. In Column (2), the interaction for Duration is 26.43 seconds ($t = 3.37$, $p < 0.01$), representing a substantial increase in verification time. The rate-based measures in Columns (3) and (4) tell a similar story: the Click Rate increases by 0.389 ($t = 5.28$, $p < 0.01$) and the Duration Rate by 9.60 ($t = 3.08$, $p < 0.01$). Columns (5)–(6) extend the analysis by interacting Post × Later Paragraph with the High treatment indicator to test whether the belief-updating effect differs across hallucination levels. The triple interaction (Post × Later × High) is not statistically significant for either Clicks ($\beta = -0.035$, $t = -0.04$) or Duration ($\beta = 2.31$, $t = 0.08$), suggesting that the belief-updating mechanism operates similarly across treatment groups. This finding is consistent with the model: the hallucination rate λ is a prior that affects the initial allocation, but once a hallucination is discovered, the Bayesian update toward higher diagnostic investment is driven by the discovery event itself rather than the prior level.

These results provide strong evidence that investors engage in real-time updating of their beliefs about AI reliability. The discovery of a hallucination serves as a "wake-up call" that triggers a substantial and persistent increase in diagnostic behavior on subsequent information.

6 Conclusion

This paper develops a theoretical and empirical framework for understanding how investors allocate limited attention when processing AI-generated financial information. The central trade-off I formalize is between the efficiency gains from AI-assisted disclosure processing and the cognitive costs of verifying AI output for hallucinations. Through a rational inattention model and an experiment, I document several findings that advance our understanding of micro investor behavior in the age of generative AI.

This paper makes several contributions to the literature. First, to my knowledge, this is the first theoretical model of investor limited attention in the age of generative AI. By formalizing the dual-channel attention allocation between parsing information and diagnostics information, I provide a tractable framework for understanding the cognitive costs investors pay while enjoying the convenience of artificial intelligence.

Second, this paper provides the first micro-level empirical evidence from a controlled experiment on the potential risks, hallucinations, that investors face when using AI tools. The granular behavioral data on time allocation, click patterns, and accuracy outcomes offers a detailed picture of how investors navigate the tension between AI convenience and reliability. The finding that over 47% of participants never activated the diagnostic channel (i.e., never clicked to view the original text) suggests that a substantial fraction of investors may be vulnerable to AI hallucination in naturalistic settings.

Third, this paper demonstrates the real-time dynamics of investor attention allocation. The belief-updating results show that investors' diagnostic behavior is not fixed but adapts in response to experience, in other words, discovering a hallucination triggers a lasting increase in verification behavior. This finding bridges the rational inattention literature, which typically treats capacity constraints as static, with the literature on learning and belief updating.

Fourth, the findings have practical implications for the design of AI tools in finance and for regulatory policy. The existence of a capacity threshold below which investors do

not verify AI output suggests that most retail investors may be particularly exposed to AI-generated misinformation. Regulators may need to consider mandatory disclosure of AI hallucination rates, standardized verification interfaces, or investor education programs that lower the cognitive cost of AI verification. For AI developers, the crowding-out effect implies that reducing hallucination rates has a double benefit: it not only improves the direct quality of AI output but also frees up investor attention for more productive information extraction.

Table 1: Balance Check Across Hallucination Rate Levels

	Total	Low			Medium			High			Robust	
	Obs	Obs	Mean	s.d.	Obs	Mean	s.d.	Obs	Mean	s.d.	F	p-value
CRT Score	353	120	3.958	2.159	119	3.899	1.955	114	3.930	1.994	0.025	0.975
Age	353	120	41.058	11.543	119	44.151	13.024	114	44.386	14.478	2.677	0.070
Finance BG	353	120	0.483	0.502	119	0.429	0.497	114	0.439	0.498	0.405	0.667
Industry BG	353	120	0.408	0.494	119	0.336	0.474	114	0.325	0.470	1.033	0.357
Male	353	120	0.658	0.476	119	0.605	0.491	114	0.623	0.487	0.378	0.685
Education	353	120	2.225	0.750	119	2.193	0.762	114	2.263	0.705	0.266	0.766
Investment	353	120	3.308	0.807	119	3.168	0.986	114	3.254	0.881	0.724	0.486

Notes: This table reports means, standard deviations, and sample sizes for baseline characteristics across the three treatment groups (Low, Medium, High hallucination rate). Robust F-statistics and p-values are obtained by regressing each baseline characteristic on treatment indicators with robust standard errors and testing the joint significance of the Medium and High coefficients. Variable definitions are provided in Appendix A.

Table 2: Descriptive Statistics and Univariate Tests

Panel A: Full Sample Summary Statistics

	Obs	Mean	s.d.	Min	Max
Individual Characteristics					
CRT Score	353	3.929	2.033	0.000	7.000
Age	353	43.176	13.097	19.000	84.000
Male	353	0.629	0.484	0.000	1.000
Finance Background	353	0.450	0.498	0.000	1.000
Industry Background	353	0.357	0.480	0.000	1.000
Education	353	2.227	0.738	1.000	4.000
Investment	353	3.244	0.894	1.000	4.000
Experimental Variables					
Test Duration	353	433.146	142.753	123.032	632.432
Diagnostic Duration	353	86.629	122.397	0.000	549.490
Parsing Duration	353	167.934	108.054	0.000	598.930
Question Duration	353	176.866	113.178	0.980	577.230
Diagnostic Ratio	353	0.171	0.229	0.000	0.916
Parsing Ratio	353	0.385	0.210	0.000	0.998
Question Ratio	353	0.440	0.258	0.002	1.000
Total Clicks	353	3.011	4.354	0.000	24.000
Diagnostic Activation	353	0.527	0.500	0.000	1.000
Total Correct Answer	353	1.348	1.039	0.000	4.000

Panel B: t-Test Between Low vs. High Hallucination Rate

	Low			High				
	Obs	Mean	s.d.	Obs	Mean	s.d.	Diff	t
Test Duration	120	416.278	140.055	114	440.228	142.484	−23.950	−1.297
Diagnostic Duration	120	70.973	103.143	114	104.200	139.853	−33.227**	−2.075
Parsing Duration	120	167.373	109.931	114	149.762	100.508	17.612	1.277
Question Duration	120	174.542	105.099	114	185.245	125.195	−10.703	−0.710
Diagnostic Ratio	120	0.143	0.199	114	0.204	0.257	−0.062**	−2.055
Parsing Ratio	120	0.394	0.222	114	0.343	0.201	0.052*	1.870
Question Ratio	120	0.455	0.259	114	0.450	0.275	0.005	0.132
Total Clicks	120	2.817	4.067	114	3.184	4.326	−0.368	−0.670
Diag Activation	120	0.500	0.502	114	0.553	0.499	−0.053	−0.804
Total Correct	120	1.450	1.099	114	1.439	1.113	0.011	0.079

Notes: Panel A reports summary statistics for the full sample of 353 participants. Panel B reports means, standard deviations, and two-sample t-tests comparing the Low ($N = 120$) and High ($N = 114$) hallucination groups. Diff is calculated as Mean(Low) − Mean(High). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Variable definitions are provided in Appendix A.

Table 3: Investor Capacity and Diagnostic Channel Activation

	(1) Diagnostic Activate (Logit AME)	(2) Diagnostic Activate (OLS)	(3) Total Clicks (OLS)
Capacity	0.063**	0.063**	0.675***
	(2.396)	(2.334)	(3.297)
Medium	0.016	0.017	0.202
	(0.258)	(0.263)	(0.343)
High	0.040	0.041	0.293
	(0.629)	(0.620)	(0.527)
Constant		0.562***	3.531***
		(3.601)	(2.775)
Controls	YES	YES	YES
N	353	353	353

Notes: This table reports the relationship between investor capacity and diagnostic channel activation. Column (1) reports average marginal effects from a logit regression. Columns (2)–(3) report OLS estimates. The dependent variables are Diagnostic Activate (indicator for any link clicks) in Columns (1)–(2) and Total Clicks in Column (3). Capacity is the standardized CRT score. Medium and High are indicators for the medium and high hallucination treatment groups (Low is the omitted baseline). Individual-level controls include age, finance background, industry background, gender, education, and investment experience. Robust t-statistics in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: AI Hallucination Rate, Investor Capacity, and Attention Allocation

	(1) Diagnostic Duration	(2) Parsing Duration	(3) Diagnostic Ratio	(4) Parsing Ratio
Capacity	24.089***	−13.269**	0.043***	−0.030**
	(3.617)	(−2.174)	(3.567)	(−2.492)
Medium	12.195	13.094	0.025	0.015
	(0.826)	(0.924)	(0.882)	(0.561)
High	28.960*	−24.220*	0.059**	−0.057**
	(1.861)	(−1.788)	(1.979)	(−2.068)
Constant	75.818*	105.428***	0.176**	0.340***
	(1.942)	(3.258)	(2.531)	(5.489)
Controls	YES	YES	YES	YES
N	353	353	353	353

Notes: This table reports OLS regressions of attention allocation measures on treatment indicators and capacity. Diagnostic Duration and Parsing Duration (Columns 1–2) are winsorized at the 1st and 99th percentiles. Diagnostic Ratio and Parsing Ratio (Columns 3–4) are the respective durations divided by Test Duration. Capacity is the standardized CRT score. Medium and High are treatment indicators (Low is the omitted baseline). Individual-level controls include age, finance background, industry background, gender, education, and investment experience. Robust t-statistics in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Interaction Between Hallucination Rate and Investor Capacity

	(1) Diagnostic Duration	(2) Parsing Duration	(3) Diagnostic Ratio	(4) Parsing Ratio
High × Capacity	32.111**	−18.542	0.060**	−0.034
	(2.164)	(−1.310)	(2.168)	(−1.238)
Medium × Capacity	15.582	−16.223	0.019	−0.016
	(1.045)	(−1.161)	(0.696)	(−0.581)
Medium	12.233	13.141	0.025	0.015
	(0.832)	(0.928)	(0.891)	(0.559)
High	29.088*	−24.258*	0.059**	−0.057**
	(1.881)	(−1.792)	(1.999)	(−2.074)
Capacity	9.496	−2.658	0.019	−0.014
	(0.933)	(−0.278)	(1.070)	(−0.761)
Constant	80.878**	100.361***	0.182***	0.334***
	(2.068)	(3.076)	(2.617)	(5.362)
Controls	YES	YES	YES	YES
N	353	353	353	353

Notes: This table augments the Table 4 specifications with interactions between treatment indicators and Capacity. All variables are defined as in Table 4. Robust t-statistics in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Diagnostic Channel and Task Performance

Panel A: Individual-Level Evidence

	(1) Total Correct (OLS)	(2) Total Correct (OLS)	(3) Correct Dummy (Logit AME)	(4) Correct Dummy (Logit AME)
Diagnostic Activation	0.402***		0.110***	
	(8.429)		(5.478)	
Total Clicks		0.068***		0.016***
		(12.055)		(5.229)
Medium	−0.248***	−0.255***	−0.027	−0.023
	(−4.520)	(−4.749)	(−1.133)	(−0.960)
High	0.010	0.007	−0.008	−0.006
	(0.168)	(0.112)	(−0.316)	(−0.263)
Capacity	0.039	0.018	0.002	−0.001
	(1.527)	(0.738)	(0.229)	(−0.074)
Constant	1.589***	1.575***		
	(11.896)	(11.815)		
Individual FE	NO	NO	NO	NO
Question FE	NO	NO	NO	NO
Controls	YES	YES	YES	YES
N	1765	1765	1765	1765

Panel B: Question-Level Evidence with Individual Fixed Effects

	(1) Question Correct	(2) Question Correct	(3) Question Correct
Diagnostic Time	0.002***		
	(3.212)		
Diagnostic Clicks		0.074	
		(1.591)	
Clicks Numbers			0.023
			(1.142)
Hallucinated	0.005	0.006	0.008
	(0.091)	(0.127)	(0.150)
Individual FE	YES	YES	YES
Question FE	YES	YES	YES
Controls.	NO	NO	NO
N	1365	1365	1365

Panel C: Within-Individual Decomposition

	(1)	(2)	(3)	(4)	(5)	(6)
			Question Correct			
Mean Clicks				0.063***	0.051***	0.079***
				(5.712)	(3.821)	(3.625)
Excess Mean Clicks	0.023	−0.016	−0.016	0.012	−0.001	−0.011
	(1.145)	(−0.695)	(−0.694)	(0.896)	(−0.052)	(−0.694)
Diagnostic Time		0.002***			0.001	
		(3.066)			(1.433)	
Excess Diag Time			0.002***			0.001**
			(3.044)			(2.473)
Mean Diag Time						−0.001
						(−0.842)
Hallucinated	0.007	0.004	0.004	0.005	0.002	0.005
	(0.150)	(0.083)	(0.083)	(0.153)	(0.058)	(0.172)
Individual FE	YES	YES	YES	NO	NO	NO
Question FE	YES	YES	YES	YES	YES	YES
Controls	NO	NO	NO	YES	YES	YES
N	1365	1365	1365	1365	1365	1365

Notes: This table examines the relationship between diagnostic channel engagement and task performance. Panel A reports individual-level regressions ($N = 353 \times 5 = 1765$ question-level observations) of performance measures on diagnostic channel variables, treatment indicators, capacity, and controls. Panel B reports conditional logit models with individual fixed effects, using paragraph-specific diagnostic measures. Panel C decomposes diagnostic behavior into between-individual (Mean) and within-individual (Excess) components. Columns (1)–(3) include individual fixed effects; Columns (4)–(6) replace individual fixed effects with individual-level controls. Robust t-statistics (clustered at the individual level in Panels B–C) in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Investor Belief Updating After Hallucination Discovery

	(1) Clicks	(2) Duration	(3) Click Rate	(4) Duration Rate	(5) Clicks	(6) Duration
Post	0.133	−2.215	−0.162***	−4.534*	0.216	−1.492
	(0.613)	(−0.371)	(−4.712)	(−1.904)	(0.854)	(−0.235)
Post × Later Para	1.240***	26.429***	0.389***	9.599***	1.252***	26.080***
	(4.359)	(3.371)	(5.282)	(3.081)	(4.053)	(3.310)
Post × High					−0.510	−4.467
					(−1.505)	(−0.243)
Post × Later × High					−0.035	2.308
					(−0.043)	(0.084)
Constant	0.483***	14.224***	0.200***	5.902***	0.483***	14.224***
	(4.478)	(7.159)	(39.232)	(8.526)	(4.564)	(7.131)
Ind-Para FE	YES	YES	YES	YES	YES	YES
N	360	360	360	360	360	360

Notes: This table examines whether investors update their verification behavior after discovering a hallucination. The sample includes participants who successfully discovered at least one hallucination during the experiment. Post is an indicator for events after the participant's first hallucination discovery. Later Paragraph indicates paragraphs sequenced after the discovered hallucination paragraph. Clicks and Duration are the total clicks and diagnostic time on a specific paragraph in a given phase. Click Rate and Duration Rate are normalized by total events. Columns (5)–(6) include triple interactions with the High treatment indicator. Individual-paragraph fixed effects absorb time-invariant heterogeneity. t-statistics in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

References

Amos Arieli, Yaniv Ben-Ami, and Ariel Rubinstein. Tracking decision makers under uncertainty. *American Economic Journal: Microeconomics*, 3(4):68–76, 2011.

Brad M. Barber and Terrance Odean. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies*, 21(2):785–818, 2008.

Azi Ben-Rephael, Zhi Da, and Ryan D. Israelsen. It depends on where you search: Institutional investor attention and underreaction to news. *Review of Financial Studies*, 30(9): 3009–3047, 2017.

Jeremy Bertomeu, Yupeng Lin, Yibin Liu, and Zhenghui Ni. Voluntary disclosure, misinformation, and AI information processing: Theory and evidence. SSRN Working Paper No. 5026360, 2024.

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

Sean Cao, Wei Jiang, Baozhong Yang, and Alan L. Zhang. How to talk when a machine is listening: Corporate disclosure in the age of AI. *Review of Financial Studies*, 36(9): 3603–3642, 2023.

Lauren Cohen and Andrea Frazzini. Economic links and predictable returns. *Journal of Finance*, 63(4):1977–2011, 2008.

Miguel Costa-Gomes, Vincent P. Crawford, and Bruno Broseta. Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69(5):1193–1235, 2001.

Joe Croom. Interactivity and illusions of ability: How using generative AI affects investor

judgments. *Journal of Accounting Research*, 2025. Forthcoming, DOI: 10.1111/1475-679X.70017.

Zhi Da, Joseph Engelberg, and Pengjie Gao. In search of attention. *Journal of Finance*, 66 (5):1461–1499, 2011.

Fabrizio Dell'Acqua, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Working Paper No. 24-013, 2023.

Stefano DellaVigna and Joshua M. Pollet. Investor inattention and friday earnings announcements. *Journal of Finance*, 64(2):709–749, 2009.

Anne M. Farrell, Jonathan H. Grenier, and Justin Leiby. Scoundrels or stars? theory and evidence on the quality of workers in online labor markets. *The Accounting Review*, 92(1): 93–114, 2017.

Shane Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42, 2005.

Xavier Gabaix, David Laibson, Guillermo Moloche, and Stephen Weinberg. Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review*, 96(4):1043–1068, 2006.

Meng Gao and Jiekun Huang. Informing the market: The effect of modern information technologies on information production. *Review of Financial Studies*, 33(4):1367–1411, 2020.

Jillian Grennan and Roni Michaely. Artificial intelligence and high-skilled work: Evidence from analysts. Swiss Finance Institute Research Paper No. 20-84, 2020.

David Hirshleifer and Siew Hong Teoh. Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, 36(1–3):337–386, 2003.

David Hirshleifer, Sonya Seongyeon Lim, and Siew Hong Teoh. Driven to distraction: Extraneous events and underreaction to earnings news. *Journal of Finance*, 64(5):2289–2325, 2009.

Eva I. Hoppe and David J. Kusterer. Behavioral biases and cognitive reflection. *Economics Letters*, 110(2):97–100, 2011.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Lixin Huang and Hong Liu. Rational inattention and portfolio selection. *Journal of Finance*, 62(4):1999–2040, 2007.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Eshaan Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Marcin Kacperczyk, Stijn Van Nieuwerburgh, and Laura Veldkamp. A rational theory of mutual funds' attention allocation. *Econometrica*, 84(2):571–626, 2016.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, pages 160–171. ACM, 2024. arXiv:2311.14648.

Alex G. Kim, Maximilian Muhn, and Valeri V. Nikolaev. Financial statement analysis with large language models. *Working Paper, University of Chicago Booth*, 2024.

Anton Korinek. Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317, 2023.

Robert Libby, Robert Bloomfield, and Mark W. Nelson. Experimental research in financial accounting. *Accounting, Organizations and Society*, 27(8):775–810, 2002.

Alejandro Lopez-Lira and Yuehua Tang. Can ChatGPT forecast stock price movements? return predictability and large language models. *Working Paper*, 2023.

Jinzhi Lu. Limited attention: Implications for financial reporting. *Journal of Accounting Research*, 60(5):1991–2027, 2022.

Bartosz Maćkowiak and Mirko Wiederholt. Optimal sticky prices under rational inattention. *American Economic Review*, 99(3):769–803, 2009.

Jordi Mondria. Portfolio choice, attention allocation, and price comovement. *Journal of Economic Theory*, 145(5):1837–1864, 2010.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.

Jörg Oechssler, Andreas Roider, and Patrick W. Schmitz. Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization*, 72(1):147–152, 2009.

Stefan Palan and Christian Schitter. Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

John W. Payne, James R. Bettman, and Eric J. Johnson. *The Adaptive Decision Maker*. Cambridge University Press, 1993.

Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

Lin Peng and Wei Xiong. Investor attention, overconfidence and category learning. *Journal of Financial Economics*, 80(3):563–602, 2006.

Gordon Pennycook and David G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188: 39–50, 2019.

Christopher A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690, 2003.

Michael N. Stagnaro, Gordon Pennycook, and David G. Rand. Performance on the cognitive reflection test is stable across time. *Judgment and Decision Making*, 13(3):260–267, 2018.

Keela S. Thomson and Daniel M. Oppenheimer. Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, 11(1):99–113, 2016.

Maggie E. Toplak, Richard F. West, and Keith E. Stanovich. The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39(7): 1275–1289, 2011.

Maggie E. Toplak, Richard F. West, and Keith E. Stanovich. Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, 20(2): 147–168, 2014.

Stijn Van Nieuwerburgh and Laura Veldkamp. Information immobility and the home bias puzzle. *Journal of Finance*, 64(3):1187–1215, 2009.

Stijn Van Nieuwerburgh and Laura Veldkamp. Information acquisition and under-diversification. *Review of Economic Studies*, 77(2):779–805, 2010.

Laura L. Veldkamp. *Information Choice in Macroeconomics and Finance*. Princeton University Press, 2011.

Martijn C. Willemsen and Eric J. Johnson. Visiting the decision factory: Observing cognition with MouselabWEB and other information acquisition methods. In Michael Schulte-Mecklenbeck, Anton Kühberger, and Rob Ranyard, editors, *A Handbook of Process Tracing Methods for Decision Research*, pages 21–42. Psychology Press, 2011.

A Variable Definitions

Appendix A. Variable Definitions

Variable	Definition
Individual Characteristics	
CRT Score	Number of correct answers of participants in the CRT test (0–7).
Capacity	Standardized CRT Score (mean zero, unit variance).
Age	The age of participants.
Male	Indicator equals 1 if participants are male, 0 otherwise.
Finance Background	Indicator equals 1 if participants have finance, accounting, economics, or statistics background, 0 otherwise.
Industry Background	Indicator equals 1 if participants have work experience in finance, accounting, economics, or statistics related industry, 0 otherwise.
Education	Highest education of participants (1 = high school or below, 2 = bachelor's degree, 3 = master's degree, 4 = doctoral degree).
Investment	Investment experience of participants (1 = no experience, 2 = less than 1 year, 3 = 1 to 5 years, 4 = more than 5 years).
Treatment Variables	
High / Medium / Low	The AI summary hallucination level that is randomly assigned to participants.
Experimental Variables	
Test Duration	Total test time of participants (seconds).
Diagnostic Duration	Total time spent by participants viewing the original text (seconds).
Parsing Duration	Total time spent by participants on the AI summary (seconds).
Question Duration	Total time spent by participants on the questions (seconds).
Diagnostic Ratio	Diagnostic Duration divided by Test Duration.
Parsing Ratio	Parsing Duration divided by Test Duration.
Question Ratio	Question Duration divided by Test Duration.
Total Clicks	Number of times the participants clicked to view the original content.
Diagnostic Activation	Indicator equals 1 if participants have at least 1 click to view the original content, 0 otherwise.
Total Correct Answer	Number of correct answers of participants in this test (0–5).
Correct Dummy	Indicator equals 1 if participants have at least 1 correct answer, 0 otherwise.

Appendix A. Variable Definitions (Continued)

Variable	Definition
Question-Level Variables (Panel Data)	
Diagnostic Time$_i$	Time spent by the participant viewing the original text corresponding to paragraph i, $i = 1, \ldots, 5$.
Diagnostic Clicks$_i$	Indicator equals 1 if participants view at least 1 original text about paragraph i, 0 otherwise.
Clicks Numbers$_i$	Number of times the participants clicked to view the original content about paragraph i.
Question Correct$_i$	Indicator equals 1 if participants have the correct answer to question i, 0 otherwise.
Hallucinated	Indicator equals 1 if the paragraph corresponding to question i contains a hallucination for the participant's assigned treatment level, 0 otherwise.
Mean Clicks	Average number of clicks across all five paragraphs.
Excess Mean Clicks$_i$	Number of clicks on paragraph i minus the average number of clicks across all five paragraphs.
Mean Diagnostic Time	Average time spent viewing the original text across all five paragraphs.
Excess Diagnostic Time$_i$	The time spent viewing the original text for paragraph i minus the average time across all five paragraphs.
Belief Updating Variables	
Post	Indicator equals 1 if the click event index is larger than the discovery index. The discovery index is the sequence number of the event in which the participant first successfully discovered the hallucination.
Later Paragraph	Indicator equals 1 if the sequence of this paragraph is after the paragraph which is discovered as hallucination paragraph, 0 otherwise.
Clicks	The total number of events a participant performed on a specific paragraph during a given time phase (pre- or post-first discovery).
Duration	The total diagnostic time (seconds) that a participant spent interacting with a specific paragraph during a given time phase.
Click Rate	The total number of clicks on a specific paragraph divided by the total number of events across all paragraphs during that same time phase.
Duration Rate	The total time spent on a specific paragraph divided by the total number of events across all paragraphs during that same time phase.

B Mathematical Proofs

Throughout the appendix we adopt the following notation. Let $A = \sigma_\theta^2$, $B = \sigma_{AI}^2$, and $d = \Phi^{-1}(1-\lambda)$, where Φ is the standard normal CDF and ϕ is its density. Define

$$\rho(a_d) \equiv 1 - \frac{1}{a_d}, \qquad f(a_d) \equiv \Phi_2\Big(d, d; \rho(a_d)\Big), \tag{A.1}$$

where $\Phi_2(\cdot,\cdot;\rho)$ denotes the standard bivariate normal CDF with correlation ρ, and $\phi_2(\cdot,\cdot;\rho)$ denotes the corresponding density. We write $\phi_2(a_d) \equiv \phi_2\Big(d, d; \rho(a_d)\Big)$ for brevity. The capacity constraint is $a_p \cdot a_d = C^2$ with $1 \le a_d \le C^2$.

Proof of Equation (D+AI). The investor observes two signals: a parsed AI signal $\tilde{s} = s_{AI} + \nu$ and a diagnostic signal $t = c + \eta$. We derive the expected posterior variance $V_{D+AI}(C; a_d)$.

By normal conjugate updating, $c \mid t \sim \mathcal{N}(m(t), v)$ where

$$m(t) = \mu_c + \frac{\sigma_c^2}{\sigma_c^2 + \sigma_d^2}(t - \mu_c), \qquad v = \frac{\sigma_c^2 \sigma_d^2}{\sigma_c^2 + \sigma_d^2} = \frac{\sigma_c^2}{a_d}. \tag{A.2}$$

The posterior non-hallucination probability is

$$r(t) \equiv \Pr(c \ge \bar{c} \mid t) = \Phi\left(\frac{m(t) - \bar{c}}{\sqrt{v}}\right). \tag{A.3}$$

Since $m(t) = \mu_c + \sigma_c\sqrt{1 - 1/a_d} \cdot X$ with $X \sim \mathcal{N}(0,1)$, substitution into (A.3) yields

$$r(t) = \Phi\Big(\sqrt{a_d}\, d + \sqrt{a_d - 1}\, X\Big). \tag{A.4}$$

Let $\alpha = \sqrt{a_d}\, d$ and $\beta = \sqrt{a_d - 1}$. Introducing independent standard normals $\varepsilon_1, \varepsilon_2$ and using $\Phi(y) = \Pr(\varepsilon \le y)$,

$$f(a_d) = \mathbb{E}\Big[\Phi(\alpha + \beta X)^2\Big] = \Pr(\varepsilon_1 - \beta X \le \alpha,\ \varepsilon_2 - \beta X \le \alpha).$$

Setting $Y_i = \varepsilon_i - \beta X$, we have $\mathrm{Var}(Y_i) = 1 + \beta^2 = a_d$ and $\mathrm{Cov}(Y_1, Y_2) = \beta^2 = a_d - 1$. Hence the correlation is $\rho = (a_d - 1)/a_d = 1 - 1/a_d$ and the standardized threshold is $\alpha/\sqrt{a_d} = d$, so

$$f(a_d) = \Phi_2\Big(d, d; \rho(a_d)\Big). \tag{A.5}$$

In the non-hallucination state, $\tilde{s} = \theta + \varepsilon_{AI} + \nu$ gives $\mathbb{E}[\theta \mid \tilde{s}, Z = \mathrm{NH}] = \beta_0 \tilde{s}$ with $\beta_0 = A/(A+\sigma_T^2)$, where $\sigma_T^2 = B + \sigma_p^2$. In the hallucination state, $\tilde{s} \perp \theta$ so $\mathbb{E}[\theta \mid \tilde{s}, Z = \mathrm{H}] = 0$. By iterated expectations, $\mathbb{E}[\theta \mid \tilde{s}, t] = \beta_0 \, \tilde{s} \, r(t)$. Since $\tilde{s} \perp t$,

$$\mathrm{Var}\Big(\mathbb{E}[\theta \mid \tilde{s}, t]\Big) = \beta_0^2 \, \mathbb{E}[\tilde{s}^2] \, \mathbb{E}[r(t)^2] = \frac{A^2}{A + \sigma_T^2} \, f(a_d). \tag{A.6}$$

From the capacity constraint, $a_p = 1 + (A + B)/\sigma_p^2$ and $a_p = C^2/a_d$, one obtains

$$\frac{1}{A + \sigma_T^2} = \frac{1}{A + B} \cdot \frac{C^2 - a_d}{C^2}. \tag{A.7}$$

Substituting into $V_{D+AI} = A - \mathrm{Var}\Big(\mathbb{E}[\theta \mid \tilde{s}, t]\Big)$ yields

$$V_{D+AI}(C; a_d) = A - \frac{A^2(C^2 - a_d)}{C^2(A + B)} \, f(a_d), \qquad C > 1, \quad 1 \le a_d \le C^2. \quad \blacksquare \tag{A.8}$$

Proof of Lemma 1. By Plackett's identity $\partial_\rho \Phi_2 = \phi_2$ and the chain rule $\rho'(a_d) = a_d^{-2}$,

$$f'(a_d) = \frac{\phi_2(a_d)}{a_d^2} > 0. \tag{A.9}$$

For the second derivative, define

$$M(a_d) \equiv \frac{\partial \log \phi_2(d, d; \rho)}{\partial \rho} = \frac{\rho}{1 - \rho^2} + \frac{d^2}{(1 + \rho)^2}. \tag{A.10}$$

Using $1-\rho^2 = (2a_d-1)/a_d^2$ and $1+\rho = (2a_d-1)/a_d$, this simplifies to

$$M(a_d) = \frac{a_d(a_d-1)}{2a_d-1} + \frac{d^2\, a_d^2}{(2a_d-1)^2}. \tag{A.11}$$

Differentiating (A.9) via the product rule,

$$f''(a_d) = \frac{\phi_2(a_d)}{a_d^4}\Big[M(a_d) - 2a_d\Big]. \tag{A.12}$$

A direct computation shows

$$M(a_d) - 2a_d = -\frac{a_d\, Q(a_d)}{(2a_d-1)^2}, \qquad Q(a_d) \equiv 6a_d^2 - (5+d^2)a_d + 1, \tag{A.13}$$

so that

$$f''(a_d) = -\frac{\phi_2(a_d)\, Q(a_d)}{a_d^3(2a_d-1)^2}. \tag{A.14}$$

Hence $f''(a_d) \le 0$ if and only if $Q(a_d) \ge 0$ for all $a_d \ge 1$. The quadratic Q is convex with minimum at $a_d^{\min} = (5+d^2)/12$. Under $|d| \le \sqrt{2}$, we have $a_d^{\min} \le 7/12 < 1$, so Q is increasing on $[1,\infty)$ with $Q(1) = 2 - d^2 \ge 0$. Therefore $Q(a_d) \ge 0$ for all $a_d \ge 1$.

To verify that $|d| \le \sqrt{2}$ is also necessary, note that $Q(a_d) \ge 0$ for all $a_d \ge 1$ requires the larger root $a_+ = \Big[(5+d^2) + \sqrt{(5+d^2)^2 - 24}\Big]/12$ to satisfy $a_+ \le 1$. Squaring the resulting inequality yields $d^2 \le 2$. Since $d = \Phi^{-1}(1-\lambda)$, the condition $|d| \le \sqrt{2}$ is equivalent to $\lambda \in [1-\Phi(\sqrt{2}),\, \Phi(\sqrt{2})] \approx [0.0787, 0.9213]$. ■

Proof of Proposition 1. Minimizing V_{D+AI} in (A.8) is equivalent to maximizing

$$\Psi(a_d) \equiv (C^2 - a_d)\, f(a_d), \qquad a_d \in [1, C^2]. \tag{A.15}$$

The first two derivatives are

$$\Psi'(a_d) = -f(a_d) + (C^2 - a_d)\, f'(a_d), \tag{A.16}$$

$$\Psi''(a_d) = (C^2 - a_d)\, f''(a_d) - 2f'(a_d). \tag{A.17}$$

Under $|d| \leq \sqrt{2}$, Lemma 1 gives $f''(a_d) \leq 0$ and $f'(a_d) > 0$. Since $C^2 - a_d \geq 0$, equation (A.17) yields $\Psi''(a_d) < 0$. Hence Ψ is strictly concave on $[1, C^2]$.

Evaluating the derivative at the left boundary,

$$\Psi'(1) = -f(1) + (C^2 - 1)\, f'(1). \tag{A.18}$$

Using $\rho(1) = 0$ gives $f(1) = \Phi(d)^2$ and $f'(1) = \phi_2(d, d; 0) = \phi(d)^2$. The condition $\Psi'(1) \leq 0$ yields $C^2 \leq 1 + \Phi(d)^2/\phi(d)^2$, defining the activation threshold

$$C^* = \sqrt{1 + \left(\frac{\Phi(d)}{\phi(d)}\right)^2}. \tag{A.19}$$

If $C \leq C^*$, strict concavity implies the maximum is at $a_d^* = 1$ (corner solution: all capacity to parsing). If $C > C^*$, then $\Psi'(1) > 0$ and the unique interior maximizer $a_d^* \in (1, C^2)$ satisfies

$$f(a_d^*) = (C^2 - a_d^*)\, f'(a_d^*), \qquad a_p^* = C^2/a_d^*. \quad \blacksquare \tag{A.20}$$

Proof of Corollary 1.

Part (i): $\partial C^*/\partial \lambda < 0$.

Define $R(d) \equiv \Phi(d)/\phi(d) > 0$, so that $C^*(d) = \sqrt{1 + R(d)^2}$ and

$$\frac{dC^*}{dd} = \frac{R(d)\, R'(d)}{\sqrt{1 + R(d)^2}}. \tag{A.21}$$

Using $\phi'(d) = -d\,\phi(d)$,

$$R'(d) = 1 + d\,R(d). \tag{A.22}$$

For $d \geq 0$, $R'(d) > 0$ trivially. For $d < 0$, the Mills ratio inequality $\Phi(d) < \phi(d)/(-d)$ gives $d\,R(d) > -1$, so $R'(d) > 0$. Hence $dC^*/dd > 0$.

Since $\Phi(d) = 1 - \lambda$, differentiating gives $dd/d\lambda = -1/\phi(d) < 0$. By the chain rule,

$$\frac{dC^*}{d\lambda} = \underbrace{\frac{dC^*}{dd}}_{>0} \cdot \underbrace{\frac{dd}{d\lambda}}_{<0} < 0. \tag{A.23}$$

Part (ii): $\partial a_d^*/\partial\lambda > 0$ and $\partial a_p^*/\partial\lambda < 0$.

For $C > C^*$, the FOC (A.20) can be written as $F(a_d, d) \equiv -f(a_d, d) + (C^2 - a_d)\,f_a(a_d, d) = 0$, where subscripts denote partial derivatives. By the implicit function theorem,

$$\frac{da_d^*}{d\lambda} = \frac{F_d}{F_a\,\phi(d)}. \tag{A.24}$$

Sign of F_a. Differentiating F with respect to a_d,

$$F_a = (C^2 - a_d)\,f_{aa} - 2f_a. \tag{A.25}$$

Since $f_a > 0$, $f_{aa} \leq 0$ (Lemma 1), and $C^2 - a_d^* > 0$, we have $F_a < 0$.

Sign of F_d. Standard bivariate normal identities give

$$f_d(a_d, d) = 2\phi(d)\,\Phi\left(d\sqrt{\tfrac{1-\rho}{1+\rho}}\right) > 0, \tag{A.26}$$

and, since ρ is independent of d,

$$f_{ad}(a_d, d) = f_a(a_d, d) \cdot \Big(- \frac{2d}{1+\rho}\Big). \tag{A.27}$$

Using the FOC to substitute $C^2 - a_d = f/f_a$,

$$F_d = -f_d - \frac{2d}{1+\rho}\, f. \tag{A.28}$$

For $d \geq 0$, both terms are non-positive with at least one strictly negative, so $F_d < 0$. For $d < 0$, it suffices to show $f_d > -\frac{2d}{1+\rho} f$. The bivariate density satisfies $(x+y)\,\phi_2(x,y;\rho) = -(1+\rho)\Big(\frac{\partial \phi_2}{\partial x} + \frac{\partial \phi_2}{\partial y}\Big)$. Integrating over $(-\infty, d]^2$ yields

$$\iint_{(-\infty,d]^2} (x+y)\,\phi_2(x,y;\rho)\,dx\,dy = -(1+\rho)\, f_d(d). \tag{A.29}$$

On $(-\infty, d]^2$, $x + y < 2d$ almost everywhere, so $-(1+\rho)\, f_d(d) < 2d\, f(d)$. Dividing by $-(1+\rho) > 0$ reverses the inequality, giving $f_d > -\frac{2d}{1+\rho} f$ and hence $F_d < 0$.

Combining: $F_a < 0$, $F_d < 0$, $\phi(d) > 0$, so $da_d^*/d\lambda > 0$ by (A.24). Since $a_p^* = C^2/a_d^*$,

$$\frac{da_p^*}{d\lambda} = -\frac{C^2}{(a_d^*)^2}\,\frac{da_d^*}{d\lambda} < 0. \quad \blacksquare \tag{A.30}$$

Proof of Corollary 2. Assume $|d| \leq \sqrt{2}$ and $C > C^*$ so that $a_d^* \in (1, C^2)$ is interior. The FOC (A.20) is equivalent to

$$C^2 = h(a_d^*), \qquad h(a_d) \equiv a_d + \frac{f(a_d)}{f'(a_d)}. \tag{A.31}$$

Part (a): $da_d^*/dC > 0$. Differentiating (A.31) gives

$$\frac{da_d^*}{dC} = \frac{2C}{h'(a_d^*)}. \tag{A.32}$$

Since $h'(a_d) = 2 - f\, f''/(f')^2$ and $f'' \leq 0$ (Lemma 1), we have $h'(a_d) \geq 2 > 0$. Hence $da_d^*/dC > 0$.

Part (b): Sign of da_p^*/dC. Since $a_p^* = C^2/a_d^*$,

$$\frac{da_p^*}{dC} = \frac{2C}{(a_d^*)^2\, h'(a_d^*)}\Big[a_d^*\, h'(a_d^*) - h(a_d^*)\Big].$$

The prefactor is strictly positive, so

$$\operatorname{sign}\left(\frac{da_p^*}{dC}\right) = \operatorname{sign}\Big(G(a_d^*)\Big), \qquad G(a_d) \equiv a_d\, h'(a_d) - h(a_d). \tag{A.33}$$

Substituting the expressions for h and h', and using (A.9)–(A.14), we obtain after algebraic simplification

$$G(a_d) = a_d + \frac{f(a_d)\, a_d^3\, (2a_d - 1 - d^2)}{\phi_2(a_d)\, (2a_d - 1)^2}. \tag{A.34}$$

Define

$$H(a_d) \equiv \frac{a_d^2(2a_d - 1 - d^2)}{(2a_d - 1)^2}, \qquad g(a_d; d) \equiv \phi_2(a_d) + f(a_d)\, H(a_d). \tag{A.35}$$

At $a_d = 1$: $\rho(1) = 0$, $f(1) = \Phi(d)^2$, $\phi_2(1) = \phi(d)^2$, $H(1) = 1 - d^2$, so

$$\psi(d) \equiv g(1; d) = \phi(d)^2 + \Phi(d)^2(1 - d^2). \tag{A.38}$$

For $d \le 1$, $1 - d^2 \ge 0$ implies $\psi(d) \ge \phi(d)^2 > 0$. For $d \ge 1$, one verifies $\psi'(d) < 0$, $\psi(1) = \phi(1)^2 > 0$, and $\psi(\sqrt{2}) = \phi(\sqrt{2})^2 - \Phi(\sqrt{2})^2 < 0$. Hence there exists a unique $d_0 \in (1, \sqrt{2})$ such that $\psi(d_0) = 0$, and we set $\lambda_0 \equiv \Phi(-d_0) \approx 0.15$.

Case $-\sqrt{2} \le d \le d_0$ (equivalently $\lambda_0 \le \lambda \le \Phi(\sqrt{2})$): Then $\psi(d) \ge 0$. Since $g(\cdot; d)$ is strictly increasing and $a_d^* > 1$, $g(a_d^*; d) > g(1; d) \ge 0$. By (A.36), $da_p^*/dC > 0$ for all $C > C^*$.

Case $d_0 < d \le \sqrt{2}$ (equivalently $\Phi(-\sqrt{2}) \le \lambda < \lambda_0$): Then $\psi(d) < 0$, i.e. $g(1; d) < 0$. Since $H(\bar{a}) = 0$ at $\bar{a} = (1 + d^2)/2$ gives $g(\bar{a}; d) = \phi_2(\bar{a}) > 0$, the strict monotonicity of g

implies a unique zero $a_0 \in (1, \bar{a})$. Setting $\bar{C}^2 \equiv h(a_0)$ and using the strict monotonicity of h,

$$\operatorname{sign}\left(\frac{da_p^*}{dC}\right) = \begin{cases} -1, & C \in (C^*, \bar{C}), \\ 0, & C = \bar{C}, \\ +1, & C \in (\bar{C}, \infty), \end{cases}$$

so a_p^* attains its unique minimum over $C > C^*$ at $C = \bar{C}$. ■

Proof of Proposition 2. Regime S uniformly dominates Regime D+AI for all $C > 1$ if and only if $\Delta_S(C) > \Delta^*_{D+AI}(C)$ for every $C > 1$, where

$$\Delta_S(C) = \frac{A^2(C^2-1)}{C^2(A+\sigma_\gamma^2)}, \qquad \Delta^*_{D+AI}(C) = \frac{A^2}{C^2(A+B)}\,\Psi^*(C),$$

with $\Psi^*(C) = \max_{a_d \in [1, C^2]} \Psi(a_d)$ as in (A.15). Canceling $A^2/C^2 > 0$, the dominance condition reduces to

$$\frac{\Psi^*(C)}{C^2-1} < \frac{A+B}{A+\sigma_\gamma^2}, \qquad \forall\, C > 1. \tag{A.39}$$

By Plackett's identity, $\Phi_2(d, d; \rho)$ is strictly increasing in ρ, and the Gaussian copula satisfies $\lim_{\rho \to 1} \Phi_2(d, d; \rho) = \Phi(d)$. Since $\rho(a_d) < 1$ for finite a_d,

$$f(a_d) < \Phi(d), \qquad \forall\, a_d < \infty. \tag{A.40}$$

For $a_d = 1$: $(C^2-1)\, f(1) = (C^2-1)\, \Phi(d)^2 < (C^2-1)\, \Phi(d)$ since $\Phi(d) \in (0,1)$. For $a_d > 1$: $C^2 - a_d < C^2 - 1$ and $f(a_d) < \Phi(d)$. In both cases $(C^2 - a_d)\, f(a_d) < (C^2-1)\, \Phi(d)$. Taking the maximum preserves the strict inequality:

$$\Psi^*(C) < (C^2-1)\, \Phi(d). \tag{A.41}$$

From (A.41), $\Psi^*(C)/(C^2-1) \leq \Phi(d)$. For the matching lower bound, set $a_d = C$ (feasible

for $C > 1$):

$$\frac{\Psi^*(C)}{C^2-1} \geq \frac{(C^2-C)\,f(C)}{C^2-1} = \frac{C}{C+1}\,f(C).$$

As $C \to \infty$, $C/(C+1) \to 1$ and $f(C) \to \Phi(d)$. Therefore

$$\lim_{C\to\infty} \frac{\Psi^*(C)}{C^2-1} = \Phi(d), \qquad \sup_{C>1} \frac{\Psi^*(C)}{C^2-1} = \Phi(d). \tag{A.42}$$

Sufficiency. Suppose $\sigma_\gamma^2 \leq T \equiv (A+B)/\Phi(d) - A$, i.e. $(A+B)/(A+\sigma_\gamma^2) \geq \Phi(d)$. By (A.41), for every finite $C > 1$,

$$\frac{\Psi^*(C)}{C^2-1} < \Phi(d) \leq \frac{A+B}{A+\sigma_\gamma^2},$$

which is exactly (A.39).

Necessity (contrapositive). Suppose $\sigma_\gamma^2 > T$, so $(A+B)/(A+\sigma_\gamma^2) < \Phi(d)$. Let $\epsilon = \Phi(d) - (A+B)/(A+\sigma_\gamma^2) > 0$. By (A.42), there exists C_0 such that for all $C > C_0$,

$$\frac{\Psi^*(C)}{C^2-1} > \Phi(d) - \epsilon = \frac{A+B}{A+\sigma_\gamma^2},$$

so (A.39) is violated.

Comparative statics. Since $\Phi(d) = 1-\lambda$, the threshold is $T = (A+B)/(1-\lambda) - A$, and

$$\frac{\partial T}{\partial \lambda} = \frac{A+B}{(1-\lambda)^2} > 0. \tag{A.43}$$

A higher λ raises T, expanding the region where Regime S dominates. ■

C Experiment Screenshots

This appendix presents screenshots from the online experiment interface. The experiment is accessible at https://singular-truffle-59ec1c.netlify.app. Participants were paid $2.50 for approximately 15 minutes of participation (equivalent to $10 per hour) and received a bonus of $0.10 for each correct answer.

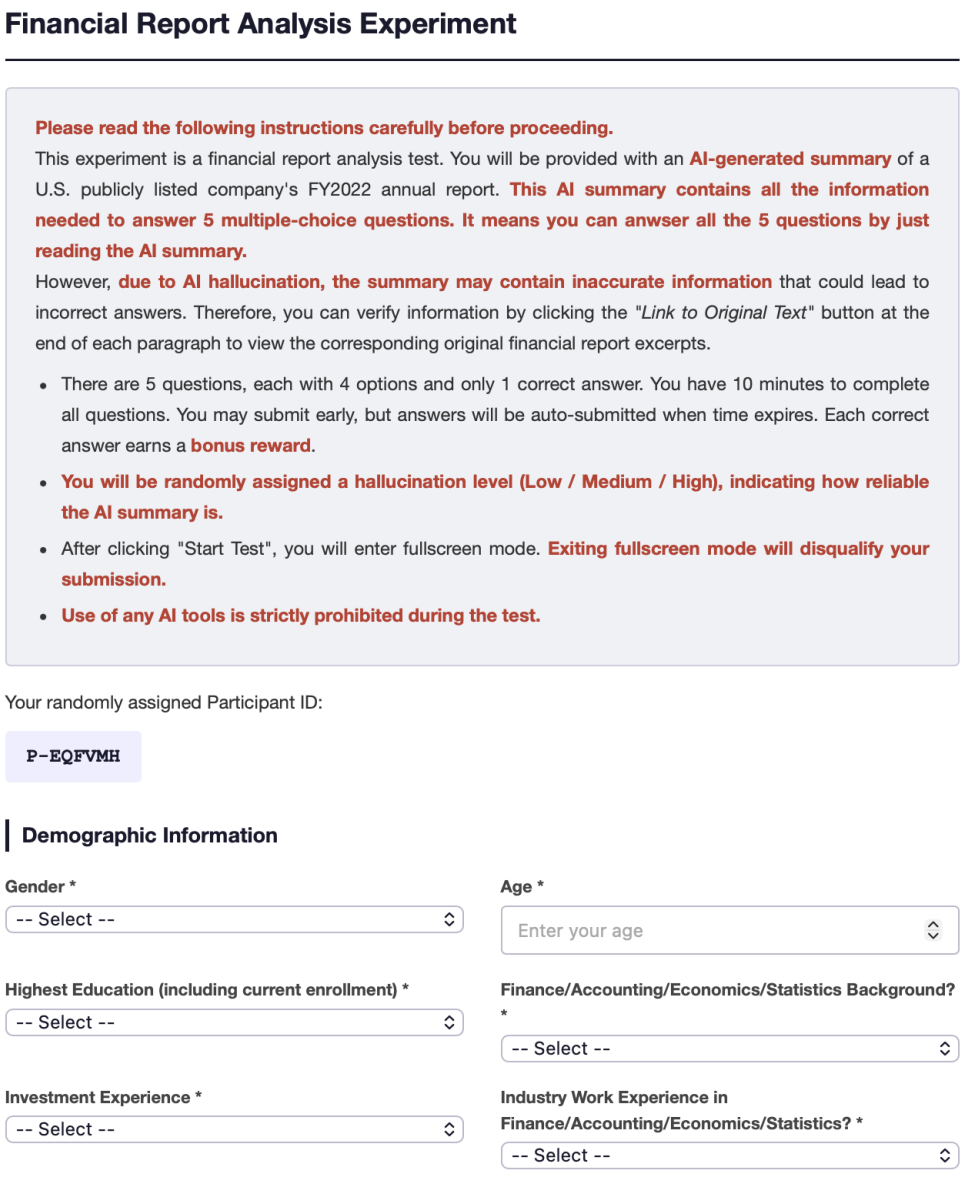

Figure 1: Experiment Landing Page: Instructions and Demographic Information

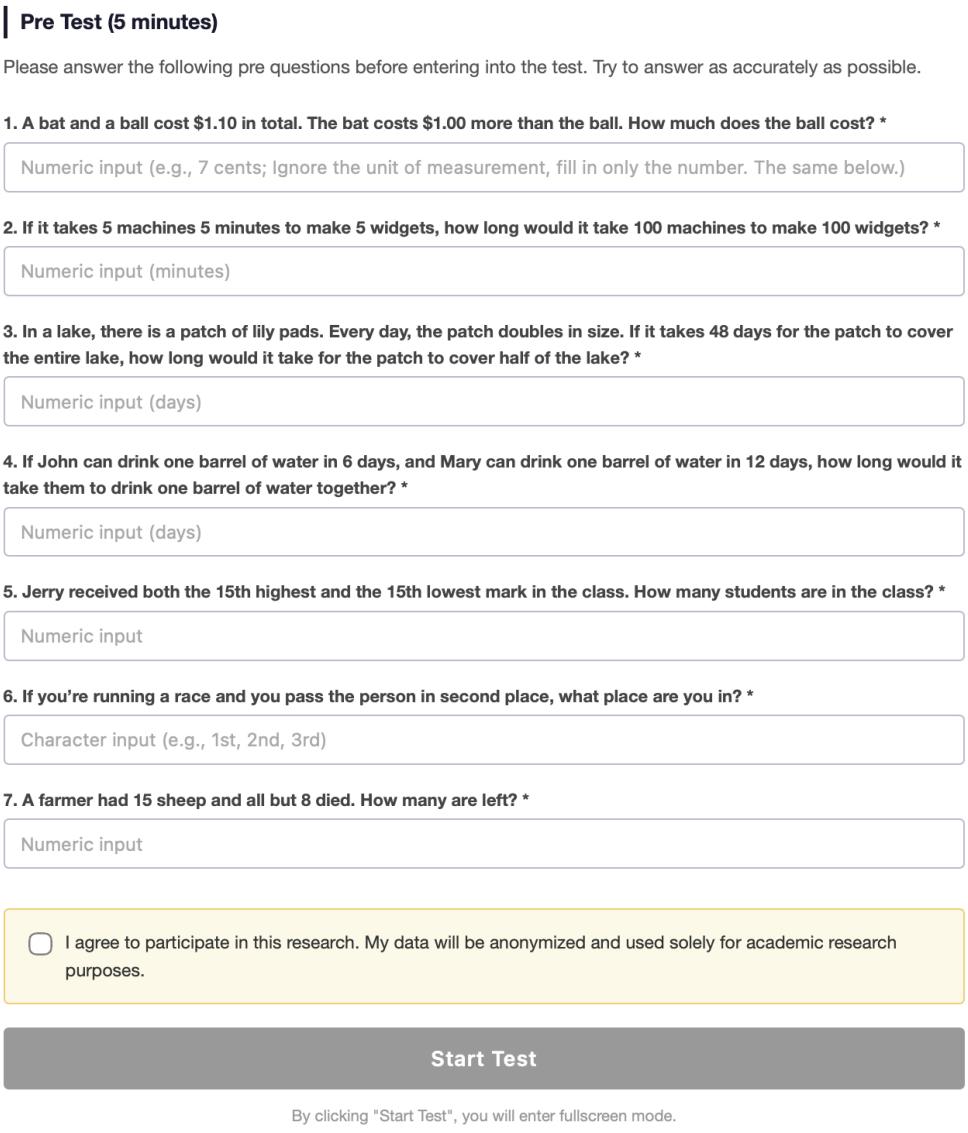

Figure 2: Pre-Test: Cognitive Reflection Test (CRT) and Consent

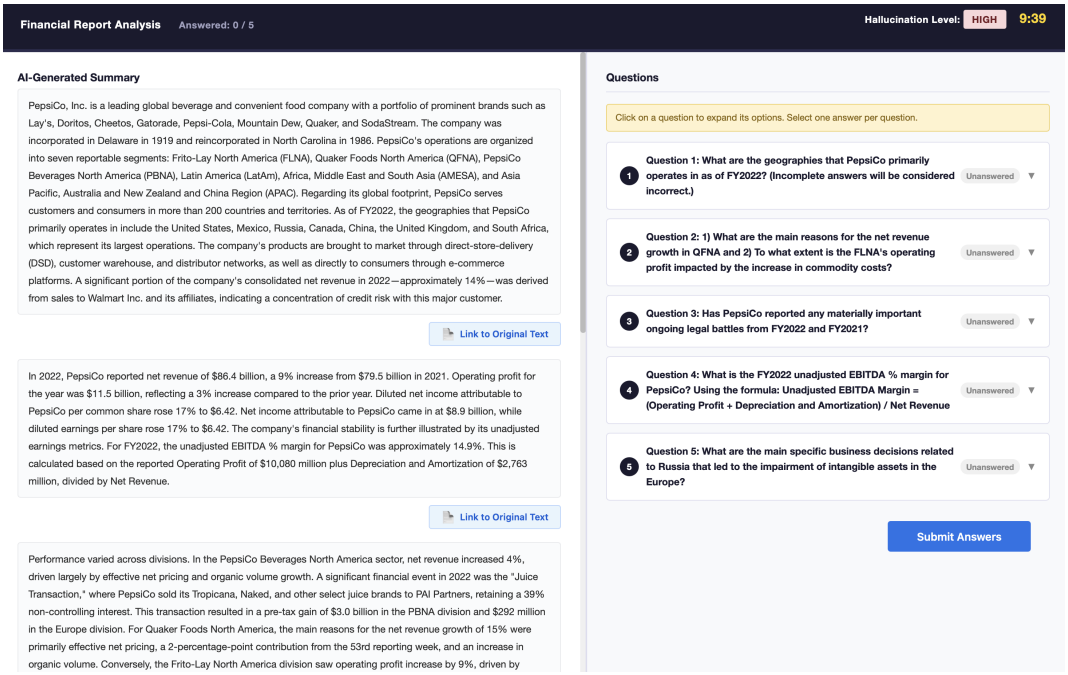

Figure 3: Main Test Interface: AI-Generated Summary (Left Panel) and Questions (Right Panel)

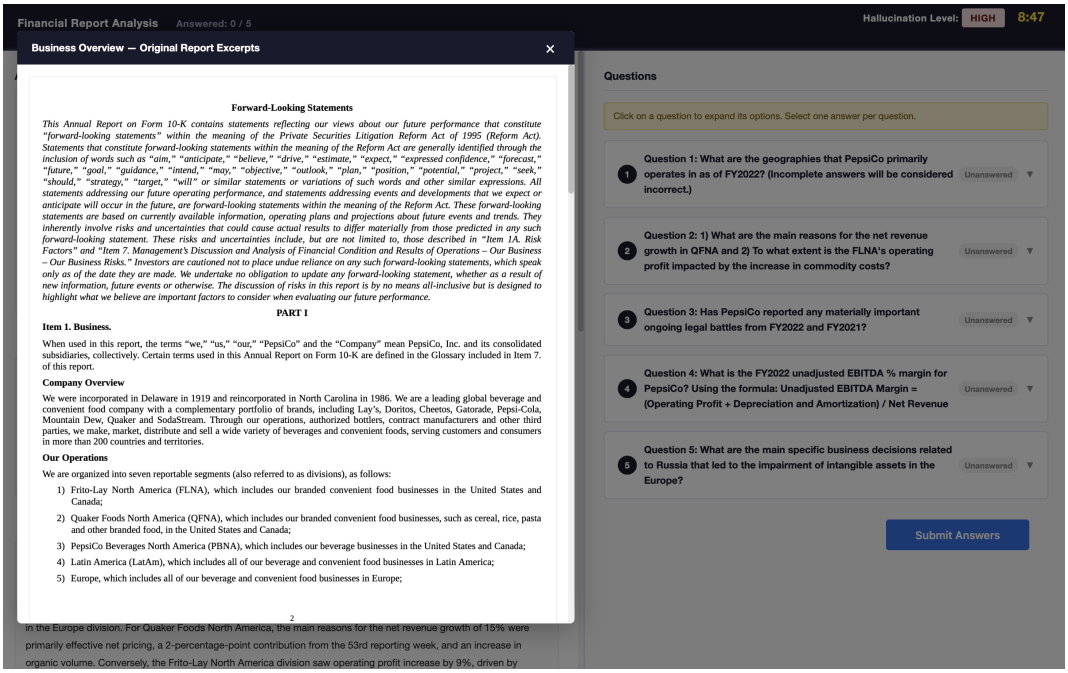

Figure 4: Diagnostic Channel: Original Report Excerpt Modal Overlay

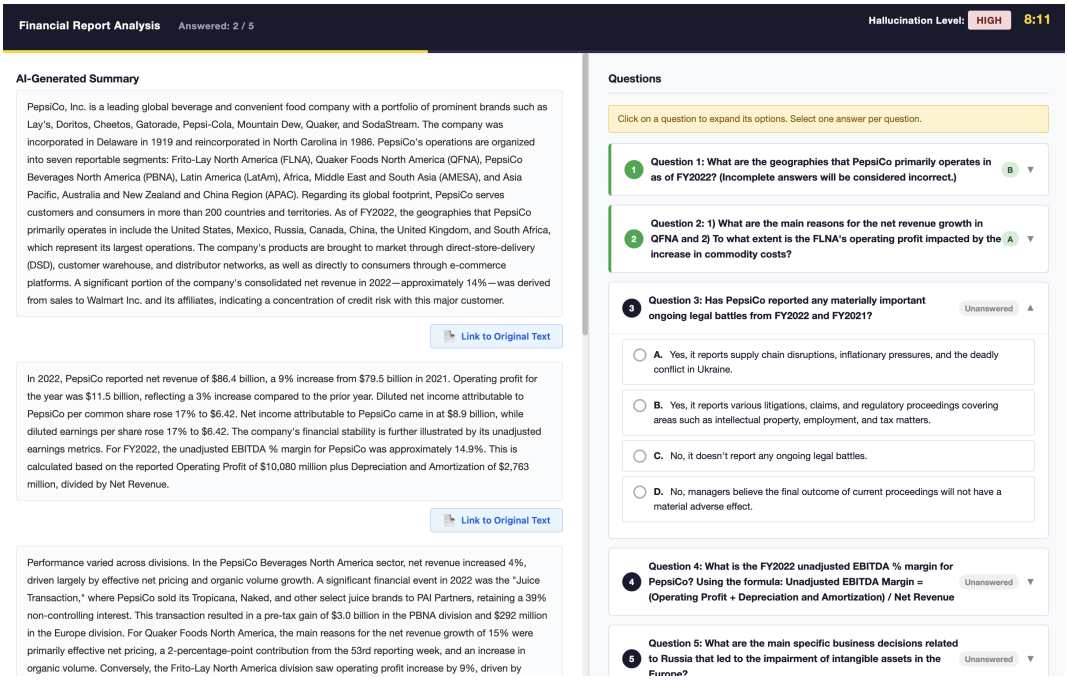

Figure 5: Main Test Interface: Question Expansion and Answer Selection

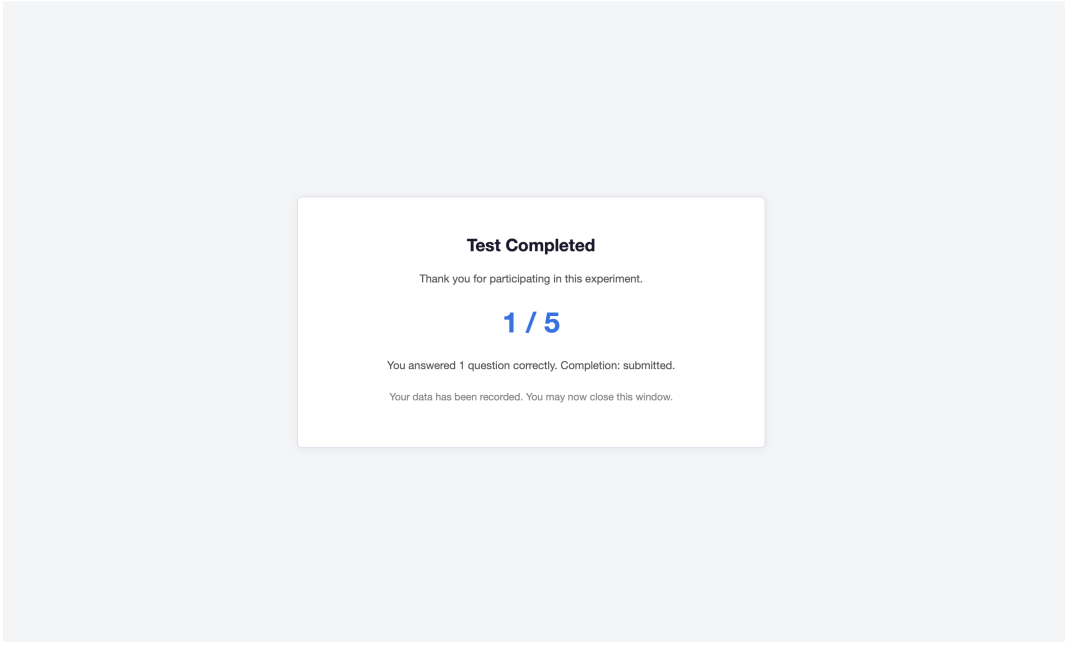

Figure 6: Test Completion: Score Feedback and Submission Confirmation