F(r)iction in Machines: Accounting Hallucinations of Large Language Models[*]

Albert Di Wang
The University of Texas at Austin
albert.di.wang@utexas.edu

Abstract

Large language models (LLMs) lower information processing costs but introduce a new friction when generating outputs inconsistent with facts, i.e., hallucinations. In accounting information acquisition, these hallucinations take two forms: deviating from existing financials and fabricating nonexistent financials. I develop a framework to quantify both types of accounting hallucinations and examine whether the production and dissemination of accounting information contribute to their occurrence. By querying LLMs for pre- and post-IPO financials of SEC-reporting firms, I find that both types of hallucinations are prevalent, with the severity varying across prompt designs, queried items, years, and firms. Using correlation analysis and two exogenous shocks (EDGAR adoption and LLM knowledge cutoff), I show that *greater* accounting information production and dissemination lead LLMs to produce *fewer* deviations but *more* fabrications. In addition, I examine popular remedies such as web search and file upload and find them inadequate for addressing the issue. This paper offers an initial assessment of accounting hallucinations in LLMs and highlights this emerging form of information friction in the era of machine-assisted information processing.

[*] I am deeply grateful to the members of my dissertation committee, Shuping Chen (chair), Urooj Khan, John McInnis, Yong Yu, and Mingyuan Zhou (external, Google DeepMind), for their invaluable guidance and support throughout this project. I appreciate the helpful comments and suggestions from Jade Chen, Christian Hutzler, Nick Hallman, Nathan Herrmann, Matt Kubic, Steve Kachelmeier, Volker Laux, Lillian Mills, Sara Toynbee, Ronghuo Zheng, and workshop participants at the University of Texas at Austin. I also thank Enrique Gomez and Shijie Yang for generously sharing the EDGAR adoption data. I gratefully acknowledge Shuping Chen, the Department of Accounting, and the Eugene and Dora Bonham Memorial Fund in Business for their generous financial support.

1. INTRODUCTION

Understanding when and why large language models (LLMs) may provide inaccurate accounting information is critical for their responsible and effective use in financial decision-making. Answering these two fundamental questions is challenging due to the lack of a measurement framework and an identification strategy. In this paper, by addressing these two hurdles, I provide the first systematic assessment of accounting hallucinations in LLMs and study whether accounting information production and dissemination contribute to the rise of such hallucinations.

LLMs have been widely adopted by retail investors for all stages of information processing: awareness, acquisition, and integration (Blankespoor, Croom, and Grant 2024; Ecker et al. 2025). However, while they help reduce information frictions by facilitating processing, they also introduce a new form of friction when generating outputs inconsistent with facts, i.e., hallucinations. In the context of accounting information acquisition, hallucinations typically manifest in two forms: deviating from existing financials (*accounting deviations*) and fabricating nonexistent financials (*accounting fabrications*).[1]

I expect accounting information production and dissemination to influence accounting hallucinations in LLMs through two sequential and necessary steps. First, public companies produce accounting information and disclose it to the public domain, where information intermediaries such as analysts and journalists further process and disseminate it. Second, raw or processed accounting information is scraped, cleaned, and compiled by data providers into large web datasets, which are then used by LLM developers for model training.

[1] A deviation occurs when an LLM has encountered a piece of information in its training data but generates an incorrect value when queried, whereas a fabrication refers to information the LLM has never seen in training yet still produces a value when asked. These concepts are formally discussed in Section 2.3 and operationalized in Section 3.

However, having more accounting information is not unambiguously beneficial for LLMs. I predict that *greater* information production and dissemination lead LLMs to produce *fewer* deviations but *more* fabrications. An LLM is typically a transformer model trained to predict the next word in a sequence (Vaswani et al. 2017). Given an incomplete sentence, the model predicts the conditional probability of each word in the vocabulary based on the preceding context and selects the most probable word as the next. LLM training is essentially fitting the model to text data to learn this probability prediction function (Brown et al. 2020). Consequently, the more frequently a certain co-occurrence (e.g., firm-year-income and $100B) appears in the training data, the higher the probability the model learns to assign to that co-occurrence (Elazar et al. 2022).

Given how LLMs work and are trained, greater accounting information production and dissemination increase the likelihood that correct co-occurrences are included in the training data. Therefore, when asked about a firm's financials in a specific year, the model is more likely to have encountered the correct firm-year-item-value combination during training and to generate the correct value. This leads to *fewer* accounting deviations (H1). In contrast, if the requested financial information does not exist (e.g., a firm's income before it was founded), greater information production and dissemination cannot add the correct co-occurrence to the training data. More problematically, increased information production and dissemination may amplify the inclusion of incorrect co-occurrences (e.g., firm-year-item with an irrelevant value), leading the model to assign higher probabilities to any combination involving that firm-year-item identifier and, consequently, to generate *more* accounting fabrications (H2).

Before testing my predictions, I first develop a framework to quantify both types of accounting hallucinations. I use four financial statement items (total assets, total liabilities,

revenue, and net income) to represent the information retail investors typically seek using LLMs. The goal is not to assess hallucinations across all possible accounting information but to focus on salient items that LLMs are most likely to retrieve accurately. I design four prompting strategies (baseline, context-rich, chain-of-thought, and few-shot prompting) to proxy how investors request such information. Similarly, the goal is not to exhaust all possible prompting strategies but to reasonably approximate how retail investors craft their queries in practice.[2] By querying ChatGPT-4.1 (released April 2025, the latest model when this study was conducted) for the value of a financial item in a given firm-year using a specific prompting strategy, I define two variables based on its response to identify accounting hallucinations: *LLM_Refusal*, an indicator of whether the model refuses to answer, and *LLM_Deviation*, the absolute percentage error between the returned value and the true value in corporate disclosures.

To measure accounting deviations, I apply this framework to query financials for all the United States Securities and Exchange Commission (U.S. SEC) reporting companies from 1990 to 2024 (the accounting deviation sample). I find that such deviations are highly prevalent and vary substantially by prompt design. Under the best prompting strategy, the average deviation from the true values is 48%. To measure accounting fabrications, I apply the framework to query pre-initial-public-offering (pre-IPO) financials of all U.S. IPOs after 2000 (the accounting fabrication sample). To the extent that the model has never encountered pre-IPO financials in its training data, I classify a response as a fabrication if the model does not refuse to answer (i.e., *LLM_Refusal* = 0). Similar to deviations, fabrications are widespread and vary by prompt design. Under the best prompting strategy, the model fabricates accounting numbers 36% of the time. I

[2] See Section 2.1 for the survey and field evidence from Blankespoor, Croom, and Grant (2024) and Ecker et al. (2025). Information acquisition is a major use case of LLMs among retail investors, and many of their queries resemble the task I used in this study.

acknowledge that non-refusals for pre-IPO financials may not reflect fabrications, as such information could exist publicly through private disclosures, leaks, or IPO filings. Robustness checks on firms with minimal pre-IPO disclosures (e.g., special purpose acquisition companies, SPACs) and on financials beyond the model's knowledge cutoff confirm that fabrications remain prevalent.

To test my hypotheses, I first conduct a correlation analysis. I use three proxies for accounting information production and dissemination: firm size, the frequency of earnings guidance, and the number of analysts following. Regressing *LLM_Deviation* on these variables using the accounting deviation sample, I find that greater information production and dissemination are associated with fewer accounting deviations (H1). In contrast, using the accounting fabrication sample, I find that these variables measured *at IPO* are negatively associated with *LLM_Refusal* when querying *pre-IPO* financials, suggesting that greater information production and dissemination correspond to more accounting fabrications (H2). The findings hold across alternative model specifications, prompts and measures, suggesting that the observed correlations reflect fundamental relationships invariant to how retail investors craft queries or how accounting hallucinations are measured.

Accounting information can influence an LLM's accounting hallucinations only if it is produced *and* incorporated into the model's training data. The lack of transparency regarding commercial LLMs' training corpora poses a major challenge to establishing this causal link. In this paper, I address this challenge using two exogenous shocks, each corresponding to one of the two necessary steps: the SEC's mandatory adoption of the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system as a shock to accounting information production and

dissemination, and LLMs' training knowledge cutoffs as a shock to whether accounting information is included in training.

The EDGAR adoption design leverages the transition from physical to electronic filings with the SEC through the EDGAR system. The SEC assigned firms into ten groups and mandated staggered adoption of the EDAGR system between 1993 and 1996. This transition made corporate filings digitally accessible to the public, improving the dissemination of accounting information. The reduced acquisition costs further encouraged information production by investors and analysts (Gao and Huang 2020; Gomez 2024). To the extent that the disseminated and newly produced information entered LLM training corpora, I predict that querying LLMs for adopters' post-adoption financials should yield smaller accounting deviations relative to pre-adoption years (H1). Using a stacked-cohort difference-in-differences (DiD) design, I find results consistent with this prediction.

The knowledge cutoff design leverages the fact that GPT-3.5 Turbo was trained on information available up to August 31, 2021 (its knowledge cutoff), while GPT-4.1 was trained on information up to May 31, 2024. Because a model's knowledge cutoff largely reflects its development timing and is independent of firms' information production and dissemination decisions, it provides an exogenous reduction in the inclusion of accounting information in GPT-3.5 Turbo's training at its cutoff date (event time). For accounting deviations, I predict that querying GPT-3.5 Turbo for post-event financials yields larger deviations than querying for pre-event financials, relative to GPT-4.1 queries for the same periods (H1). For accounting fabrications, I predict that GPT-3.5 Turbo fabricates more frequently (i.e., exhibits lower *LLM_Refusal*) when queried for *post-event* financials of firms with greater information

production and dissemination *prior to the cutoff* (H2). I find results consistent with both predictions.

I acknowledge that my research designs have several limitations. First, there is measurement error in my accounting hallucination measures. Because the training data of GPT-4.1 are undisclosed, it is uncertain whether the model has encountered specific financials during training. For example, if a pre-IPO financial is included in training, the fabrication measure may reflect deviation instead. Second, LLM-generated measures violate the no-leakage condition (Ludwig et al. 2025), creating anticipation effects in the outcome variable (Roth et al. 2023). This anticipation likely biases estimation in the two natural experiment designs. Third, the knowledge cutoff date may not precisely represent the actual timing of accounting information exclusion. LLMs are often further fine-tuned using human feedback after training, introducing post-cutoff information (Levy 2024). However, these limitations are expected to bias but attenuate the estimated effects.[3] Because my goal is to establish causality rather than estimate the effect magnitudes, the research designs collectively achieve this purpose, and I explicitly caution the interpretation of effect sizes.

Lastly, I examine two increasingly popular approaches to mitigating hallucinations: web search, where the LLM is used only to locate and extract information from search engine results, and file upload, where the user provides the document containing the correct information for the model to parse. Both are variants of retrieval-augmented generation (RAG), which shifts reliance from the LLM's internal knowledge to externally provided sources (Lewis et al. 2020). Because file upload requires users to identify the correct document, it precludes accounting fabrications. I therefore test both methods using the accounting deviation sample.

[3] I discuss all limitations in detail in Section 3.4 and Section 4.4.

I find that average deviation remains high under web search (22% from the true values), and the monetary cost is nearly 200 times greater than standard queries. Although file upload yields a lower average deviation (7%), it remains economically significant, and the need for users to identify, download, and upload the correct document undermines the purpose of using LLMs for information acquisition. More interestingly, preliminary analysis shows that under RAG methods, accounting hallucinations exhibit weaker or no association with information production and dissemination. Instead, search quality (web search) and file complexity (file upload) become significant correlates, suggesting that these approaches expose investors to new forms of frictions in machines.

Implications and Contributions

My paper has two implications to LLM users. First, it provides a measurement framework for accounting hallucinations. As LLMs become everyday tools for retail investors in financial decision-making, the need for such measurement grows increasingly urgent. I introduce a framework that is comparable across firm-years and models and adaptable to different information types. Although future advances in LLMs may reduce accounting hallucinations, the opacity of their training data means that capital market participants still require a consistent method to assess model reliability before relying on them for financial decisions. In this regard, my framework provides a benchmark test for accounting information acquisition that can serve as a standard for evaluating future LLMs.

Second, my findings suggest that LLMs exhibit systematically biased knowledge across firms and over time. This cautions their use for information processing, both in financial decision-making and in research applications. For example, investment decisions that explicitly or implicitly rely on LLMs' uneven knowledge may distort resource allocation in capital

markets. Likewise, researchers use LLMs to read lengthy corporate disclosures and extract or summarize information. If LLMs possess uneven knowledge across firms or periods, their task performance will vary accordingly, introducing systematic measurement error. Such error raises a serious construct validity concern, as effects identified using LLM-generated variables may reflect variation in LLMs' knowledge rather than the construct under study.

My paper makes two contributions to the literature. First, it adds to the nascent literature on the limitations of Generative Artificial Intelligence (GenAI) in financial decision-making. While a growing body of research examines the use of LLMs by market participants (Bertomeu et al. 2025; Blankespoor, Croom, and Grant 2024; Bradshaw et al. 2025; Cheng, Lin, and Zhao 2024; Ecker et al. 2025) and their ability to process information (Bai et al. 2023; Chang et al. 2025; Chen et al. 2023; Kim and Nikolaev 2024; Lopez-Lira and Tang 2023), far less is known about their limitations in this context. To my knowledge, only three contemporaneous studies explore this question: Levy (2024), who finds that LLMs suffer from look-ahead bias, leading to illusory predictability in accounting and finance reasoning; Cao, Wang, and Yi (2025), who show that LLMs exhibit foreign bias in financial forecasts; and Motoki and Pinto (2025), who document LLMs' country bias in fraud detection. Complementing these papers, I study both the measurement and the cause of a fundamental limitation in using LLMs for information acquisition, accounting hallucinations. Together, we provide the foundation to examine the real effects of these limitations on capital markets.

Second, my paper contributes to the literature on information processing costs. A large and growing body of research examines how these costs shape investors' information choices, trading behavior, and market outcomes (see Blankespoor, deHaan, and Marinovic 2020 for a review). Traditionally, information processing costs refer to the effort and expense of monitoring,

acquiring, and analyzing firm disclosures—frictions that apply primarily to humans. With the rise of technologies such as social media (Blankespoor, Miller, and White 2014), machine-readable reporting (Blankespoor 2019), and robo-journalism (Blankespoor, deHaan, and Zhu 2018), many of these costs have been substantially reduced. While these tools enhance *human information processing*, we are entering an era of *machine information processing* in which machines are the primary consumers of information. Machines determine what to present and how to interpret content through functions such as recommendation, summarization, and extraction. The frictions they face are likely distinct from those faced by humans and deserve greater attention. My findings offer an initial step toward understanding the frictions in machines.

Paper Organization

The remainder of this paper is organized as follows. Section 2 presents the institutional background on retail investors' adoption of LLMs for information processing, the technological background on the working mechanism, training, and evaluation of LLMs, and the hypothesis development. In Section 3, I develop the measurement framework for accounting hallucinations and construct the samples for measuring these hallucinations. Section 4 introduces the research design used to establish causality and presents sample construction for the design. In Section 5, I show the prevalence of accounting hallucinations and analyze their severity across dimensions such as how investors craft queries and what information they seek. I also provide empirical evidence on the cause of accounting hallucinations. Section 6 examines potential solutions and Section 7 concludes the paper.

2. BACKGROUND AND THEORY

2.1 Retail Investors' Adoption of LLMs

This paper focuses on information acquisition by retail investors for two reasons. First, retail investors face high information processing costs and LLMs are an accessible tool to help alleviate these limitations (Chang et al. 2025). Second, acquisition is an indispensable step in information processing for financial decision-making (Blankespoor, deHaan, and Marinovic 2020).

Survey and field evidence indicate widespread adoption by retail investors. Blankespoor, Croom, and Grant (2024) survey 2,175 retail investors and find that 47% already use GenAI to process financial information or support investment decisions. Using data from a leading conversational GenAI platform with over one million daily active users, Ecker et al. (2025) identify 1.75 million queries related to listed companies during the first half of 2024, averaging about 1.75 queries per user per day.

Information acquisition is the major use case among retail investors. According to Blankespoor, Croom, and Grant (2024), searching (29%), retrieving (28%), and identifying topics in financial documents (18%) are among the most common tasks (p. 39). Other tasks like screening for securities (27.8%) also implicitly involve information acquisition (e.g., "show me some stocks at 52-week highs," p. 45). Similarly, Ecker et al. (2025) show that information acquisition tasks account for 44.5% of all queries (e.g., "what was the return on assets of Bank of China Limited and Postal Savings Bank of China Co. Ltd in 2023," p. 42).

Hallucination is the leading concern among retail investors. In the survey, 54% identify reliability and accuracy as its biggest limitation, while 43% note a lack of source transparency in LLMs' responses.

2.2 Working Mechanism, Training, and Evaluation of LLMs

An LLM is a function that maps textual input (a prompt or query) to textual output (a completion, response, or generation).[4] It first represents the input numerically using the *word2vec* technique (Mikolov et al. 2013). Essentially, the input is a sequence of words $w_1, w_2, \dots, w_i$ drawn from a vocabulary V, and *word2vec* works like a dictionary that converts each word to a vector. Then, the LLM uses the sequence of vectors to predict the probability distribution of the next word, as in Equation (1). Lastly, it selects the word from the vocabulary with the highest predicted probability as the next word (i.e., the first word of the output), as in Equation (2). Iterating the last two steps generates the full output sequence.

$$\hat{p}(w_{i+1} = v \mid w_1, w_2, \dots, w_i) = LLM_\theta(w_1, w_2, \dots, w_i), \text{ where } v \in V \tag{1}$$

$$w_{i+1} = \arg\max_{v \in V} \ \hat{p}(v | w_1, w_2, \dots, w_i) \tag{2}$$

Predicting the probability distribution of the next word using Equation (1) is analogous to predicting the value of the dependent variable using a linear regression model. The differences are twofold. First, unlike a linear regression model that has only a few parameters (e.g., one β in simple liner regression), an LLM contains billions of parameters (denoted by θ) due to its highly complex architecture.[5] Second, whereas a linear regression model usually maps the input to a predicted value ($\hat{y}$), an LLM maps it to a predicted probability distribution ($\hat{p}$) over the vocabulary (i.e., N predicted values where N is the size of the vocabulary). Applying Equation (2) to the distribution selects the word with the highest predicted probability as the next word.

[4] I focus on decoder-only LLMs (e.g., ChatGPT, Llama, Gemini, Grok, Claude, and DeepSeek) and simplify the technical details. See Naveed et al. (2025) and Zhao et al. (2023) for a comprehensive introduction to LLMs.
[5] For example, GPT-1, 2, and 3 contain 117 million, 1.5 billion, 175 billion parameters, respectively (Radford et al. 2018; Radford et al. 2019; Brown et al. 2020).

Training an LLM is the process of fitting textual data to estimate the parameters θ in $LLM_\theta(\cdot)$.[6] As in linear regression, this requires collecting training data and defining an objective function. The training data for commercial LLMs, though rarely disclosed, is widely believed to encompass most publicly available web content.[7] The objective function of LLMs, unlike the sum of squared residuals in linear regression, is the likelihood of observing a given text sequence, $P(w_1, w_2, \dots, w_n)$. Decomposing it into the product of conditional probabilities, $p(w_1), p(w_2|w_1), \cdots, p(w_n|w_1, w_2, \dots, w_{n-1})$, and replacing them with $LLM_\theta(\cdot)$ from Equation (1) provide a way to parameterize it as in Equation (3). Solving the maximization problem for Equation (3) yields the parameters θ. In this way, $LLM_\theta(\cdot)$ "learns" to assign the highest probability to the most plausible next word.[8]

$$\hat{P}(w_1, w_2, \dots, w_n) = LLM_\theta(w_1) \times LLM_\theta(w_2|w_1) \times \cdots \times LLM_\theta(w_n|w_1, w_2, \dots, w_{n-1}). \quad (3)$$

Evaluating the performance of an LLM cannot rely on individual examples due to its nondeterministic nature.[9] Generating the next word using Equation (2) often produces repetitive or low-quality text because selecting the most probable next word does not ensure an optimal overall output. To mitigate this, top-*k* sampling (Radford et al. 2019) and temperature scaling (Maddison, Mnih, and Teh 2016) are commonly used, introducing stochasticity into LLM generations. This means that the same prompt on the same LLM can yield different completions across runs. And this inherent randomness is so far inevitable (de Kok 2025; Levy 2024).

---

[6] I focus on the pre-training stage of LLMs, which enables them to predict the next word.
[7] Many LLMs are trained on large web corpora such as CommonCrawl. For example, GPT-2's training corpus, WebText, was scraped from about 45 million links and includes over 8 million documents (Radford et al. 2019). GPT-3's corpus, a cleansed version of CommonCrawl, spans more than 300 billion webpages and approximately 95 to 102 billion English words, supplemented with an expanded WebText, two internet-based book corpora, and English-language Wikipedia (Brown et al. 2020).
[8] This learning process can take weeks or even months because, unlike linear regression, there are no closed-form estimators for this optimization problem. Estimating θ requires numerical methods such as stochastic gradient descent and backpropagation (Rumelhart, Hinton, and Williams 1986).
[9] OpenAI acknowledges this in its document.

2.3 Hypothesis Development

LLM hallucinations stem inherently from the training and evaluation procedures used in the current development paradigm (Kalai et al. 2025). Learning theory shows that under this paradigm, hallucinations can be mitigated but not eliminated (Xu, Jain, and Kankanhalli 2024).[10] Thus, quantifying their severity in information acquisition is essential for financial decision-making of retail investors (Section 3). Equally important is understanding their causes. What is especially relevant for financial applications is whether accounting information plays a role in causing accounting hallucinations of LLMs.[11]

Accounting information can affect accounting hallucinations through two consecutive and necessary steps. First, public companies produce the information internally and disclose it to the public domain. Information intermediaries such as analysts and journalists further process and disseminate these disclosures. Second, raw or processed accounting information is scraped, cleaned, and compiled by data providers into large web datasets that are then provided to LLM developers for model training.

Following the taxonomy of LLM hallucinations (Huang et al. 2025), I classify accounting hallucinations into two types: deviating from existing financials (accounting deviations) and fabricating nonexistent financials (accounting fabrications).[12] Consistent with the two-step process through which accounting information influences accounting hallucinations, I define "existing" as information publicly disclosed *and* included in training data (i.e, the model has seen

[10] LLMs learn a vast set of functions, each performing a specific task such as extraction, summarization, or prediction. Prompts allow humans to use natural language to invoke a particular function and supply its input. This paper mathematically proves that LLMs cannot learn all computable functions and will therefore inevitably hallucinate when used as general problem solvers.
[11] Model architecture and training methods also contribute to LLM hallucination (Li et al. 2023; Yang et al. 2024).
[12] These are referred to as factuality hallucinations, which directly harm investors. Another major type is faithfulness hallucinations, where the output diverges from the user's request. As shown in Section 5.1, faithfulness hallucinations are very rare in the information acquisition setting when prompts are clearly crafted.

the data before), whereas "nonexistent" refers to information either undisclosed *or* excluded from training data (i.e., the model has never seen the data before).

I predict that greater accounting information production and dissemination reduce accounting deviations. As shown in Equation (1), an LLM models the conditional probability of the next word (the requested information, e.g., the value of total assets) given the preceding words (the query, e.g., the total assets of Meta Platforms, Inc. in 2024). The more frequently the correct combination (e.g., Meta-2024-Assets-$276B) is produced and disseminated in the public domain and included in the training data, the more likely the model learns through Equation (3) to assign the highest probability to the correct value and output it via Equation (2). Therefore, to the extent that information production and dissemination increase the likelihood of accounting information being included in LLM training corpora, the LLM-generated accounting data should more likely reflect the correct value of the requested information.

H1: *Greater accounting information production and dissemination reduce accounting deviations of LLMs, ceteris paribus.*

By contrast, I predict that greater accounting information production and dissemination increase accounting fabrications. Fabrications arise when the requested information in the query does not exist. Thus, more information production and dissemination increase the likelihood of incorrect combinations being included (e.g., Meta-2001-Assets-Number, although Meta Platforms, Inc. was founded in 2004 and was called Facebook, Inc.). These co-occurrences strengthen statistical associations between firm-year-item identifiers and irrelevant numbers. Because LLMs are trained to prioritize fluency rather than factual accuracy, as shown in Equation (3), they tend to assign higher conditional probabilities to generations involving firms that appear more frequently in training, and thus output fabricated financials.

H2: *Greater accounting information production and dissemination increase accounting fabrications of LLMs, ceteris paribus.*

It is important to note that these predictions are not mere extensions of existing findings in the LLM literature to the context of accounting information acquisition. To the best of my knowledge, there is no existing evidence on how accounting information shapes accounting hallucinations, whether in terms of association or causation. Empirical findings in the LLM literature suggest that correct/incorrect co-occurrences decrease/increase deviation, but no prior work studies whether co-occurrences induce fabrication.[13]

3. MEASUREMENT OF ACCOUNTING HALLUCINATIONS

3.1 Accounting Information Retrieval Using LLMs

I formalize the task of accounting information retrieval using LLMs as feeding in a prompt tuple (*Firm, Year, Item, Extra_Input*) and getting back a completion tuple (*Retrieved_Value, Extra_Output*). In this setup, *Firm* refers to the company's historical name as of the queried *Year*, and *Item* denotes the specific financial information being requested. *Extra_Input* includes any auxiliary information used in different prompt designs to help the model retrieve the correct value. *Retrieved_Value* is the model's output for the queried item, which is later compared with *True_Value* to quantify LLM hallucinations. *Extra_Output* captures any additional content generated as required by *Extra_Input*.

[13] Specifically, Kaplan et al. (2020) identify the Scaling Laws governing LLM performance, showing that training data size affects model performance. Kandpal et al. (2023) and Mallen et al. (2022) further find that question-answering accuracy improves with the number of query-relevant documents and the popularity of the queried entity in the training corpus. Elazar et al. (2022) and Zhang, Li, and Wu (2024) show that LLMs learn co-occurrences of identifier-value pairs rather than factual knowledge. Kang and Choi (2023) show that although correct information exists in the training data, frequent co-occurrences of incorrect pairs are associated with LLMs generating outputs that deviate from the correct information.

To represent the information retail investors seek using LLMs, I use four financial statement items (total assets, total liabilities, revenues, and net income) from 10-K filings of SEC-reporting firms. I acknowledge that retail investors use LLMs to acquire a wide range of information beyond these accounting items. However, my goal is not to evaluate hallucinations across all possible information requests but to focus on salient items that LLMs have the best chance of retrieving accurately. Failure to do so would cast serious doubt on their ability to perform more sophisticated acquisition tasks. Moreover, these items are well-defined, mandatory, broadly available across firm-years, and numerical, making hallucination measures easy to compute and available and comparable across firms and years.

To proxy how retail investors request information using LLMs, I design four prompting strategies ranging from **baseline prompting**, where a direct question is posed to the model, to more sophisticated approaches: **context-rich prompting**, which provides definitions and data source cues; **chain-of-thought prompting** (CoT), which guides the model's reasoning; and **few-shot prompting**, which supplements the query with illustrative examples of inputs and expected outputs. I acknowledge that retail investors may craft prompts in diverse ways beyond these strategies. Again, the goal is not to exhaust all possible prompting strategies or identify the optimal one for this task but to reasonably approximate how retail investors might use LLMs. I provide details and examples for the four prompting strategies in Appendix A.[14]

3.2 Accounting Deviation and Fabrication Measures

I measure accounting **deviations** using *LLM_Deviation*, defined as the absolute percentage error, i.e., the absolute difference between *Retrieved_Value* and *True_Value*, scaled by

[14] Using these strategies to retrieve information from the internet and uploaded files is discussed in Section 6.

True_Value.[15] A larger *LLM_Deviation* indicates a more severe hallucination. I also construct a binary measure, *LLM_Material*, equal to one if the deviation exceeds 5% and zero otherwise.

I measure accounting fabrications using *LLM_Refusal*, defined as whether an LLM explicitly acknowledges a lack of information and returns "NA" as instructed in the prompt. If the queried information exists in the public domain and is included in training but the model refuses to answer, this reflects a **knowledge deficit** (i.e., failure to learn). Conversely, if the information does not exist in the public domain or is not included in training (i.e., the model has never seen it before) but the model still provides an answer, it indicates a **fabrication**.

I also construct an adjusted *LLM_Refusal* to more accurately reflect fabrication. LLMs have an inherent tendency of refusal when their confidence is not sufficiently high, even when the quired information exists and is included in their training data (Yin et al. 2023). The raw *LLM_Refusal* may capture this tendency and overestimate the refusal rate (i.e., underestimate fabrication). To control for this tendency, I subtract two components from the raw *LLM_Refusal* in a firm-year: the firm's average *LLM_Refusal* in other years of which the queried information exists (removing within-firm refusal tendency), and the year's average *LLM_Refusal* in other firms of which the queried information exists (removing within-year refusal tendency).

3.3 Sample Construction for Measuring Accounting Hallucinations

I employ ChatGPT-4.1 to measure accounting hallucinations because OpenAI's ChatGPT is the most widely used model among retail investors, and this version was the latest available when the study was conducted (released April 2025)[16].

[15] I use the absolute value because I have no prior about the direction of the deviation.
[16] According to Blankespoor, Croom, and Grant (2024), OpenAI ChatGPT is the most widely used model (75%), followed by Google Gemini (37%) and Microsoft Copilot (36%).

To construct the deviation sample, I begin with firm-year observations in Compustat Fundamentals Annual from 1990 to 2024. I restrict the sample to firms that are incorporated and headquartered in the U.S., listed on the NYSE, AMEX, or NASDAQ, and not in the Finance or Utility industries. I exclude observations with missing firm identifiers or characteristics from CRSP, I/B/E/S, Thomson Reuters, and SRAF 10-X Header Dataset. I further drop firm-years having fiscal-year-end dates after or within 90 days before ChatGPT-4.1's knowledge cutoff date (May 31, 2024, for the model version I use).[17] I then query the model for the four financial statement items for each firm-year observation to calculate accounting deviations. The final sample consists of 349,720 firm-year-item observations, with 87,430 firm-years of 7,593 unique firms. The process is summarized in Panel A of Table 1.

To construct the fabrication sample, I obtain all U.S. IPOs from Audit Analytics and acquire their firm-year observations from the deviation sample. I require these IPOs to be first time appearance of the firms in the market by dropping firms with multiple IPO dates and firms with financials in Compustat even before the IPO date.[18] I further drop firms that ever-switched fiscal-year-end dates (2%). I supplement the sample with either public (early IPOs) or private firms (Form D filers) as control firms for *LLM_Refusal* adjustment (Section 3.2). For the IPO firms, I create and only keep five years of observations prior to their IPO. This results in 4,855 pre-IPO firm-years for 971 IPOs. I then query the model for the four financial statement items for each pre-IPO firm-year to identify potential accounting fabrications. The final sample thus consists of 19,420 firm-year-item observations. The process is summarized in Panel B of Table 1.

[17] The knowledge cutoff date is the point in time after which the model has not been trained on new data.
[18] Compustat does not backfill pre-IPO financials even for the financials in registration statements.

3.4 Measurement Error

As conceptually defined in Section 2.3, an accounting deviation occurs if the queried information **exists in the public domain** *and* is **included in the LLM training** corpus and an accounting fabrication occurs if the queried information either does **not exist in the public domain** *or* is **not included in the LLM training** corpus.

Therefore, for using financials from Compustat to identify accounting deviations, I assume that these financials meet both conditions. This assumption is reasonable because these financials are mandatorily disclosed, and ChatGPT-4.1's training corpus, though undisclosed, is believed to encompass most publicly available web content (Section 2.2). To the extent either condition is not met, *LLM_Deviation* may capture accounting fabrications rather than deviations.

Similarly, for using pre-IPO financials of IPO firms to identify accounting fabrications, I assume that these financials are likely to meet at least one of the two conditions. However, pre-IPO financials may meet neither because (1) private firms may release their financials for private placement purposes or such information may be leaked into the public domain, and (2) IPO firms are required to disclose two to three years of pre-IPO financials in their registration statements, depending on firm size. In this case, *LLM_Refusal* may partly capture knowledge deficits rather than fabrications. While my prompt design, which requests financials under GAAP, should mitigate this measurement error, I later address it using a subsample of firms and an alternative setting that is less, or not at all, subject to this issue in Sections 5.1 and 5.4, respectively.

4. RESEARCH DESIGN AND SAMPLE CONSTRUCTION

4.1 Correlation Analysis

I use firm size (*MarketCap*) to proxy for the level of accounting information production, the number of management earnings forecasts (*FreqEarn*) to capture managerial dissemination efforts, and the number of analysts following the firm (*Analyst*) to capture dissemination by information intermediaries. According to **H1**, I predict that these proxies are negatively correlated with accounting deviations of LLMs (*LLM_Deviation*). I estimate the following linear regression to test the prediction.

$$LLM_Deviation_{i,t,k} = \beta_1 MarketCap_{i,t} + \beta_2 FreqEarn_{i,t} + \beta_3 Analyst_{i,t} + \gamma' X_{i,t} + \alpha_i + \delta_t + \theta_k + \epsilon_{i,t,k} \quad (4)$$

where i, t, and k is the subscript for firm, year, and financial statement item, respectively. X is a vector of time-varying control variables including leverage, profitability, growth opportunity, sales growth, institutional ownership, and stock volatility. α_i, δ_t, and θ_k stands for firm, year, and financial statement item fixed effects, respectively.

In contrast, according to **H2**, I predict that these proxies for information production and dissemination are positively correlated with accounting fabrications of LLMs. Specifically, I regress *LLM_Refusal* of *pre-IPO* firm-year-items on the three proxies measured *at the IPO* to test this prediction. Because *LLM_Refusal* is an inverse measure of accounting fabrications, I expect the coefficients (βs) in the following linear regression model to be negative.

$$LLM_Refusal_{i,t,k} = \beta_1 MarketCap_i + \beta_2 FreqEarn_i + \beta_3 Analyst_i + \gamma' X_i + \alpha_l + \delta_t + \theta_k + \epsilon_{i,t,k} \quad (5)$$

where l is the subscript for industry and α_l stands for industry fixed effects. Other variables are the same as in Equation (4). I do not include firm fixed effects here because they will absorb the information production and dissemination proxies.

4.2 Mandatory Adoption of EDGAR

The SEC required reporting companies to transition from physical to digital filings using the EDGAR system. For firms, this mandate did not necessitate producing additional information but rather digitizing existing reports, thereby mainly facilitating the dissemination of accounting information. For investors and analysts, EDGAR adoption substantially reduced the cost of accessing, monitoring, and disseminating corporate disclosures. Prior research shows that these lowered acquisition costs encouraged further information production by market participants (Gao and Huang 2020; Gomez 2024). To the extent that the disseminated and the additionally produced information made their way into LLM training corpora, the adoption provides a good setting to test **H1**.

I predict that querying LLMs for adopters' post-adoption financials should yield lower refusal rates (i.e., fewer knowledge deficits) and smaller accounting deviations relative to pre-adoption years, compared with a control group. To test this prediction, I leverage the mandatory adoption between 1993 and 1996 across ten groups of firms. I adopt a stacked-cohort DiD design given the staggered nature of the adoption (Baker, Larcker, and Wang 2022; Barrios 2021). To construct the cohorts, I begin with firms that adopted EDGAR (treated) in a given year (e.g., 1996) and define an event window around the adoption year (e.g., t-2 through t+1, or 1994-1997). The cohort includes all firm-years of treated firms within this window. Control firm-years are drawn from the same period but consist of firms that adopted *before* the event window (e.g., in 1993) and are assigned a pseudo-adoption year matching the treated firms' actual adoption year.[19] Finally, I stack the resulting cohorts to form the analysis sample and estimate treatment effects using a DiD model (Cengiz et al. 2019; Deshpande and Li 2019; Goodman-Bacon 2021).

[19] I use early adopters, rather than late adopters, as the control group due to anticipation effects inherent in using LLMs to generate measures (Ludwig, Mullainathan, and Rambachan 2025). See Section 4.4 for further discussion.

$$Outcome_{c,i,t,k} = \beta_1 Treated_{c,i} * Post_{c,t} + \gamma' X_{c,i,t} + \alpha_{c,i} + \delta_{c,t} + \theta_{c,k} + \epsilon_{c,i,t,k} \quad (6)$$

where c is the subscript for cohort, $Outcome$ denotes either *LLM_Refusal* (measuring knowledge deficits here) or *LLM_Deviation*, X is a vector of time-varying control variables as in Equation (4) plus firm size, number of earnings forecasts, and number of analysts following, $\alpha_{c,i}$, $\delta_{c,t}$, and $\theta_{c,k}$ stands for cohort-firm, cohort-year, and cohort-financial statement item fixed effects, respectively. Other variables are the same as in Equation (4). I expect the coefficient (β_1) to be negative.

4.3 Knowledge Cutoffs across LLMs

LLMs typically have a knowledge cutoff date, meaning their training data include only information available up to that point in time.[20] I exploit variation in knowledge cutoff dates across two different LLMs as an exogenous shock to the inclusion of accounting information in training data.[21] I refer to the model with the earlier cutoff date as the early model, the one with the later cutoff date as the late model, and the earlier cutoff date as the event time.

To the extent that accounting information produced and disseminated after the early model's cutoff date was incorporated into the late model's training data, this setting provides a natural experiment to test **H1**. I predict that querying the early model for post-event financials should yield larger accounting deviations relative to querying for pre-event financials, compared with querying the late model for pre- and post-event financials. To test this prediction, I use GPT-3.5 Turbo (cutoff date August 31, 2021) as the early model and GPT-4.1 (cutoff date May 31,

An alternative would be to use non-adopters; however, in the EDGAR setting, all firms were mandated to adopt, leaving no such group available.

[20] Levy (2024) points out that the knowledge cutoff is not an accurate cutoff for the temporal boundary of information included in model training, as post-training processes may inject additional information into the model. I discuss the implications of this limitation for my inference in Section 4.4.

[21] As discussed in Section 2.3, this is a necessary condition for accounting information to influence accounting hallucinations.

2024) as the late model. The retrieved financials from GPT-3.5 Turbo are assigned as the treatment group, while those from GPT-4.1 serve as the control group, with the GPT-3.5 Turbo cutoff as the event time. I use the DiD design with the following specification.

$$LLM_Deviation_{m,i,t,k} = \beta_1 Treated_m * Post_t + \gamma' X_{i,t} + \alpha_{m,i} + \delta_{m,t} + \theta_{m,k} + \epsilon_{m,i,t,k} \quad (7)$$

where m is the subscript for model version, X is a vector of time-varying control variables as in Equation (4) plus firm size, number of earnings forecasts, and number of analysts following, $\alpha_{m,i}$, $\delta_{m,t}$, and $\theta_{m,k}$ stands for model-firm, model-year, and model-financial statement item fixed effects, respectively. Other variables are the same as in Equation (4). I expect the coefficient (β_1) to be positive.

This setting also provides a natural experiment to test **H2**. Because no accounting data after August 31, 2021 were used to train GPT-3.5 Turbo, any instances in which the model provides post-event financials rather than refusing to answer can be confidently classified as fabrications. Based on **H2**, I predict that GPT-3.5 Turbo will fabricate more frequently (i.e., exhibit lower *LLM_Refusal*) when queried for *post-event financials* of firms with greater accounting information production and dissemination *prior to or at the event*.

H2 can also be tested by replacing the dependent variable in Equation (7) with *LLM_Refusal* and extending it to a triple-DiD framework. Specifically, the coefficient (β_1) captures GPT-3.5 Turbo's post-event *LLM_Refusal* relative to its pre-event level, benchmarked against *LLM_Refusal* from GPT-4.1. This setup directly corresponds to the adjustment described in Section 3.2. In other words, β_1 captures adjusted *LLM_Refusal*. Therefore, according to **H2**, I predict that β_1 will be significantly smaller for firms with greater information production and dissemination. To test this prediction, I split the sample into two based on the proxy for information production and dissemination (*MarketCap*, *FreqEarn*, or *Analyst*) and run Equation

(7) using *LLM_Refusal* as the dependent variable for the two subsamples. I then test the difference between the estimated coefficients on *Treated * Post* across the two subsamples (i.e., the triple-difference estimator).

4.4 Strengths and Weaknesses of the Research Design

The first weakness concerns the impact of measurement error for my inference. As discussed in Section 3.4, my deviation and fabrication measures may partially capture each other because it is uncertain whether SEC-reporting firms' post-IPO and pre-IPO financials are included in model training. Given my prediction that accounting information production and dissemination have opposite effects on different types of accounting hallucinations, such measurement error likely biases against finding results in the correlation analysis.

The second weakness concerns potential confounding from unobservable variables in the correlation analysis. For example, LLMs may not refuse to provide pre-IPO financials because private firms disclose such information and larger ones are more likely to do so, rather than larger ones producing more information that causes fabrication. However, my prediction of opposite effects across the two types of hallucinations sets a high bar for confounding, as any such factor would need to cause an increase in one type but a reduction in the other. In addition, I have two exogenous shocks to further strengthen my inference.

The EDGAR design has two limitations. First, it assumes that EDGAR-induced changes in information production and dissemination affect the composition of LLM training data, which is unknown. Therefore, I use the knowledge cutoff design to complement the EDGAR design. As discussed in Section 2.3, accounting information affects accounting hallucinations through two steps: production and dissemination, and inclusion in training. EDGAR adoption serves as an exogenous shock to the first step, while the knowledge cutoff determines the second.

Second, the DiD estimate is likely biased. Because GPT-4.1 is trained on historical corpora up to its knowledge cutoff, both pre- and post-adoption information are likely included in its training, violating the no-leakage condition (Ludwig et al. 2025). As a result, LLM-retrieved pre-adoption financials may be influenced by post-adoption data, creating anticipation effects in the outcome variable (Roth et al. 2023). Anticipation in the treatment group should attenuate the estimated effect size. Using early adopters as control firms mitigates anticipation in the control group but may introduce new bias if the effects of EDGAR adoption is time-varying (Baker, Larcker, and Wang 2022). However, this should further attenuate the estimated effect size.

The limitation of the knowledge cutoff design concerns the accuracy of the early model's cutoff timing. LLMs are often further fine-tuned using human feedback after its training to align their behavior with human instructions, which may introduce information created after the stated cutoff (Levy 2024). To the extent that post-cutoff knowledge is incorporated into the early model, the estimated effect size in Equation (7) for testing H1 will be biased (attenuated), as the shock to accounting information inclusion becomes weaker. However, this issue is unlikely to affect the test of H2, since the triple-DiD design differences out the bias across both high and low information production and dissemination groups (Olden and Møen 2022).

The knowledge cutoff design have two strengths. First, in testing H1, it compares accounting hallucinations for the same firm and year across different models. This within-firm-year comparison helps eliminate the possibility that time-varying firm characteristics are potential confounders. Second, in testing H2, the choice of knowledge cutoff largely reflects the model's development timing and thus is independent from firms' information production and dissemination decisions before the cutoff.

4.5 Sample Construction for Identifying the Cause of Accounting Hallucinations

For the EDGAR adoption design, I construct the cohort-firm-year sample described in Section 4.2 as in Panel C of Table 1. Each cohort-firm-year observation requires querying the LLM for four financial statement items (13,045 cohort-firm-years * 4 items = 52,180 requests). For the knowledge cutoff design, I construct the model-firm-year sample described in Section 4.3 as in Panel D of Table 1. Each model-firm-year observation requires querying the LLM for four financial statement items (54,400 model-firm-years * 4 items = 217,600 requests). In all regressions, I further limit the sample to requests that yield valid outputs to measure *LLM_Refusal*, and are not refused by the model to measure *LLM_Deviation*.[22] Detailed definitions of all variables are provided in Appendix B.

5. EMPIRICAL RESULTS

5.1 Accounting Deviations and Fabrications

Table 2 and Figure OA1 show that accounting deviations vary substantially by prompt design and information requested (firm's industry, calendar year, and item). This poses a risk for retail investors as they are likely unaware of the uneven response quality across different queries. Specifically, Panel A of Table 2 presents the variation across prompting strategies. All strategies perform effectively: no requests fail due to unsupported input, and fewer than 0.08% generate unparsable responses. CoT prompting yields the lowest average deviation (31%) but a prohibitively high refusal rate (61%). Few-shot prompting produces a much lower refusal rate (8%) and the second-lowest average deviation (48%). I therefore use the data generated under

[22] As shown in Section 5.1, the incidence of invalid outputs is rare. They may arise for two reasons. First, a prompt may contain content flagged by the model as harmful, or a file upload may be of an unsupported type (see OpenAI's content moderation guidance). Second, the model may fail to follow instructions and produce output that are not parsable (i.e., faithfulness hallucination).

few-shot prompting as my main sample in subsequent analyses. Panels B and C report summary statistics for *LLM_Deviation* and firm characteristics, respectively. Figure OA1 presents the variation of accounting deviations across queried firm's industry, calendar year, and item.

Similarly, Table 3 and Figure OA2 show that accounting fabrications vary substantially by prompt design and information requested (firm's industry, event year, and item), posing the same risk of uneven response quality across different queries. Specifically, Panel A of Table 3 presents the variation across prompting strategies, and all strategies perform effectively with no unsupported input and very few unparsable output. CoT prompting achieves the highest absolute *LLM_Refusal*. However, if I adjust *LLM_Refusal* by benchmarking it against either public or private firms, it performs as the worst. By contrast, few-shot prompting yields a relatively high *LLM_Refusal* under all measurement approaches. But still, the model fabricates numbers 36% to 45% of the time when pre-IPO financials are queried. I use the data generated under few-shot prompting as my main sample in subsequent analyses. Panels B reports summary statistics for firm characteristics and Figure OA2 presents the variation of accounting fabrications across queried firm's industry, event year, and item.

As discussed in Section 3.4, LLM-retrieved pre-IPO financials may not represent fabrications if such data were disclosed or leaked and included in training. I provide two analyses supporting that pre-IPO financials are likely fabrications. First, I summarize *LLM_Refusal* for information-scarce firms—emerging growth companies (EGCs) and special purpose acquisition companies (SPACs)—which are either sensitive to information leakage or have little pre-IPO information to disclose (Barth, Landsman, and Taylor 2017; Blankespoor et al. 2022). As shown in Table OA1, the model still retrieves pre-IPO financials 33%-43% of the time for EGCs and 6%-33% for SPACs.

Second, if non-refusals when querying pre-IPO financials are driven by leaked or disclosed information already included in LLM training, then pre- and post-IPO *LLM_Deviation* should be comparable, as both periods' financials would exist in the training data. Using two years of pre-IPO financials obtained from XBRL tags of Form S-1 filings, I plot pre- and post-IPO deviation levels in Figure OA3. I also estimate the pre- versus post-IPO difference in *LLM_Deviation*, benchmarked against already public firms in Table OA2. As shown, pre-IPO deviations are significantly higher than those in the post-IPO period.

These results suggest that pre-IPO non-refusals more likely represent accounting fabrications. In Section 5.4, I further show that the LLM generates financials even for periods after its knowledge cutoff, which are clearly accounting fabrications.

5.2 Results for Correlation Analysis

The results in Table 4 support **H1**. All three proxies for accounting information production and dissemination are negatively correlated with accounting deviations of LLMs, and this relationship remains robust across model specifications. Column (1) includes only the three proxies and fixed effects, while Column (2) adds time-varying firm characteristics. Column (3) switches from firm clustering to two-way clustering of standard errors by firm and year. In Table OA3, I further show that the findings hold when using alternative prompting strategies and deviation measures, suggesting that the observed correlation reflects a fundamental relationship that is invariant to how users craft queries or how deviations are measured.

The results in Table 5 support **H2**. Two of the three proxies are negatively correlated with *LLM_Refusal* (the inverse of accounting fabrications), and this relationship remains robust across model specifications. Column (1) includes only the three proxies and fixed effects. Columns (2) and (3) additionally control for time-varying firm characteristics and two IPO-related variables

that capture information production during the offering, respectively. I do not apply two-way clustering by industry and year because the fabrication sample covers only five years of observations. In Table OA4, I further show that the findings hold under alternative prompting strategies and adjusted fabrication measures, suggesting that the observed correlation reflects a fundamental relationship that is invariant to how users craft queries or how fabrications are measured.

5.3 Results for the EDGAR Adoption Design

The results in Table 6 support **H1**. As shown in Columns (1) and (3), EDGAR adoption significantly reduces *LLM_Refusal* (measuring knowledge deficits in this setting) and *LLM_Deviation*. When controlling for firms' time-varying information environments and other characteristics, the effect remains significant for knowledge deficits (Column (2)) but diminishes for accounting deviations (Column (4)). This suggests that EDGAR adoption primarily reduces hallucinations by encouraging information production and dissemination that enhance firms' information environments, rather than by incorporating digitalized filings directly into training. This finding aligns with Wang and Levy (2024), who show that EDGAR data has little direct overlap with typical LLM training sources. The dynamic DiD plots in Figure 1 further illustrate that reductions in knowledge deficits and accounting deviations occur only after adoption.[23]

5.4 Results for the Knowledge Cutoff Design

The results in Table 7 support **H1**. The exclusion of accounting information from the training data of GPT-3.5 Turbo due to its knowledge cutoff substantially increases the model's accounting deviations. Figure 2 plots both the unconditional time series of the outcome variable

[23] Due to the short adoption window of EDGAR (1993-1996), the stacked DiD analysis is limited to a four-year event window. This may reduce the power of the pre-trend test (Hribar, McInnis, and Wang 2025). I use a longer pre-event window of ten years in the knowledge cutoff design.

and the event-study dynamics over a ten-year pre-cutoff period. Both plots show no evidence of a pre-trend. Heterogeneity analysis in Table OA5 further indicates that firms with weaker information environments experience the greatest deterioration in the information quality of LLM-generated outputs when their data are excluded from training.

Figure 3 and the results in Table 8 support **H2**. Panel A of Figure 3 shows that LLMs generate financials even for periods beyond their knowledge cutoff, as evidenced by GPT-3.5 Turbo's non-refusals after year 0. Panel B further compares *LLM_Refusal* across firms with more versus less information production and dissemination prior to the cutoff. Firms with more *pre-cutoff* information production and dissemination exhibit lower *post-cutoff LLM_Refusal* (i.e., more fabrications). Results in Table 8 reinforce this finding using the triple-DiD analysis. Under two of the three proxies, firms with less information production and dissemination experience a greater increase in *LLM_Refusal* after the cutoff, indicating fewer accounting fabrications, whereas firms with more show the opposite pattern.

6. POTENTIAL SOLUTIONS TO ACCOUNTING HALLUCINATIONS

I examine two increasingly popular approaches to mitigating hallucinations: web search, where the LLM is used only to locate and extract information from search engine results, and file upload, where the user provides the document containing the correct answer for the LLM to parse. Both methods are variants of RAG, which shifts reliance from the LLM's formed knowledge to externally provided sources (Lewis et al. 2020). In the ChatGPT interface, RAG features are not enabled by default. Web search is either automatically triggered at OpenAI's

discretion or must be manually activated by the user.[24] File upload function operates only when the user manually provides a document.

I focus on accounting deviations of LLMs when testing these two potential solutions to ensure comparability. Because file upload requires the user to locate and provide the document containing the requested information, it eliminates the possibility of non-existent information being asked and thus fabrication. I use a 10% random subsample from Panel A of Table 1 to keep the analysis manageable in terms of time and cost. I then exclude firm-year observations without a Form 10-K filing in Compustat's EDGAR database. The final sample consists of 6,351 firm-year observations from 697 unique firms. As each firm-year requires querying the LLM for four financial statement items, the resulting dataset contains 25,404 firm-year-item observations.

RAG methods mitigate accounting deviations of LLMs, but the deviations remain large. As shown in Panel A of Table 9, average *LLM_Deviation* decreases from 47% under few-shot prompting to 22% with web search and 7% with file upload. Although these improvements are meaningful, the remaining deviations are still economically significant. For instance, the best-performing method (RAG with file upload) produces accounting values that deviate from actual reports by more than 5% in 16% of cases.

The implementation costs of RAG methods are prohibitively high. As shown in Panel B of Table 9, the cost of web search is nearly 200 times greater than that of few-shot prompting. Moreover, file upload requires investors to manually identify, download, and upload the correct filings, which defeats the purpose of using LLMs to reduce information acquisition costs.

More importantly, using RAG methods appears to expose retail investors to information frictions in the external data sources (in this case, the search engine and uploaded documents).

[24] https://openai.com/index/introducing-chatgpt-search/.

As a preliminary analysis, I conduct the correlation analysis in Equation (4) to the samples used for testing RAG methods. The regression results in Table 10 confirm that, even within this 10% subsample, firms' information production and dissemination remain significantly correlated with accounting deviations (Column (1)). However, accounting deviations under RAG methods show weaker or no association with information production and dissemination (Columns (2) and (4)). Instead, deviations become correlated with whether the web search successfully locates a link from the SEC's official website (Column (3)) and whether the uploaded file contains an excerpt highly similar to the query (Column (5)). These preliminary results suggest that using such RAG methods introduces new forms of machine frictions to retail investors.

7. CONCLUSION

This paper provides the first systematic assessment of accounting hallucinations in LLMs and investigates whether accounting information production and dissemination cause their emergence. I develop a measurement framework for two distinct types of accounting hallucinations: deviations from existing financials and fabrications of nonexistent financials. I find both are pervasive and systematic, even under advanced prompting strategies or RAG methods like web search and file upload. These results underscore that, despite their sophistication, LLMs frequently misstate accounting information and that these errors are not random but vary predictably with firms' information production and dissemination. Specifically, I find that greater information production and dissemination causes fewer deviations but more fabrications. These findings highlight that information production and dissemination can both enhance and distort LLM generation of accounting information.

These empirical findings suggest that LLMs exhibit uneven knowledge across firms and over time, which raises important implications for their use in financial decision-making and research. For investors, reliance on LLMs' inconsistent knowledge can distort resource allocation and misinform trading decisions. For researchers, dependence on LLM-generated variables introduces systematic measurement errors that threaten construct validity, as observed effects may be caused by model knowledge rather than the construct under study.

This study contributes to the literature on limitations of GenAI in financial decision-making and the literature on information processing costs. As machines become the major consumer of information in capital markets, the frictions they face (e.g., stemming from incompleteness or bias in data) require more attention. This paper is an initial attempt toward understanding the frictions in machines.

REFERENCE

Bai, J.J., Boyson, N.M., Cao, Y., Liu, M. and Wan, C., 2023. Executives vs. chatbots: Unmasking insights through human-AI differences in earnings conference Q&A. *Northeastern U. D'Amore-McKim School of Business Research Paper*, (4480056).

Baker, A.C., Larcker, D.F. and Wang, C.C., 2022. How much should we trust staggered difference-in-differences estimates?. *Journal of Financial Economics*, *144*(2), pp.370-395.

Barrios, J.M., 2022. Staggeringly problematic: A primer on staggered DiD for accounting researchers. *Available at SSRN 3794859*.

Barth, M.E., Landsman, W.R. and Taylor, D.J., 2017. The JOBS Act and information uncertainty in IPO firms. *The Accounting Review*, *92*(6), pp.25-47.

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

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

Blankespoor, E., Croom, J. and Grant, S.M., 2024. Generative AI and investor processing of financial information. *Available at SSRN 5053905*.

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

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

Blankespoor, E., Hendricks, B.E., Miller, G.S. and Stockbridge Jr, D.R., 2022. A hard look at SPAC projections. *Management Science*, *68*(6), pp.4742-4753.

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

Bond, S.A., Klok, H. and Zhu, M., 2023. Large language models and financial market sentiment. *Available at SSRN 4584928*.

Bradshaw, M.T., Ma, C., Yost, B. and Zou, Y., 2025. Generative AI Use by Capital Market Information Intermediaries: Evidence from Seeking Alpha. *Available at SSRN*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. *Advances in neural information processing systems*, *33*, pp.1877-1901.

Cengiz, D., Dube, A., Lindner, A. and Zipperer, B., 2019. The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, *134*(3), pp.1405-1454.

Chang, A., Dong, X., Martin, X. and Zhou, C., 2023. AI democratization, return predictability, and trading inequality. *Available at SSRN 4543999*.

Chen, J., Tang, G., Zhou, G. and Zhu, W., 2025. ChatGPT and DeepSeek: Can they predict the stock market and macroeconomy?. *arXiv preprint arXiv:2502.10008*.

Cheng, Q., Lin, P. and Zhao, Y., 2024. Does Generative AI Facilitate Investor Trading? Evidence from ChatGPT Outages. *Evidence from ChatGPT Outages (June 21, 2024). Singapore Management University School of Accountancy Research Paper*, (2025-186).

Christ, M.H., Kim, M.J. and Yip, M.A., 2024. Survey Evidence on the Determinants and Consequences of Artificial Intelligence Use in Sell-Side Equity Research. *Available at SSRN 5071836*.

de Kok, T., 2025. ChatGPT for textual analysis? How to use generative LLMs in accounting research. *Management Science*.

Deshpande, M. and Li, Y., 2019. Who is screened out? Application costs and the targeting of disability programs. *American Economic Journal: Economic Policy*, *11*(4), pp.213-248.

Ecker, F., Li, X., Li, Y. and Wu, F., 2025. How Stock Market Participants Use Generative Artificial Intelligence: Evidence from User-Platform Interaction Data. *Available at SSRN*.

Elazar, Y., Kassner, N., Ravfogel, S., Feder, A., Ravichander, A., Mosbach, M., Belinkov, Y., Schütze, H. and Goldberg, Y., 2022. Measuring causal effects of data statistics on language model'sfactual'predictions. *arXiv preprint arXiv:2207.14251*.

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

Gomez, E.A., 2024. The effect of mandatory disclosure dissemination on information asymmetry among investors: Evidence from the implementation of the EDGAR system. *The Accounting Review*, *99*(1), pp.235-257.

Goodman-Bacon, A., 2021. Difference-in-differences with variation in treatment timing. *Journal of econometrics*, *225*(2), pp.254-277.

Hribar, P., McInnis, J.M. and Wang, A.D., 2025. Trends in Accounting Research. *Available at SSRN 5492586*.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. and Liu, T., 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), pp.1-55.

Kalai, A.T., Nachum, O., Vempala, S.S. and Zhang, E., 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Kandpal, N., Deng, H., Roberts, A., Wallace, E. and Raffel, C., 2023, July. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning* (pp. 15696-15707). PMLR.

Kang, C. and Choi, J., 2023. Impact of co-occurrence on factual knowledge of large language models. *arXiv preprint arXiv:2310.08256*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D., 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kim, A.G. and Nikolaev, V.V., 2024. Context-based interpretation of financial information. *Journal of Accounting Research*.

Kojima, T., Gu, S.S., Reid, M., Matsuo, Y. and Iwasawa, Y., 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, pp.22199-22213.

Levy, B., 2024. Caution ahead: Numerical reasoning and look-ahead bias in AI models. *Available at SSRN 5082861*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T. and Riedel, S., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, pp.9459-9474.

Li, Z., Zhang, S., Zhao, H., Yang, Y. and Yang, D., 2023. Batgpt: A bidirectional autoregessive talker from generative pre-trained transformer. *arXiv preprint arXiv:2307.00360*.

Lopez-Lira, A. and Tang, Y., 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *Return Predictability and Large Language Models (April 6, 2023)*.

Ludwig, J., Mullainathan, S. and Rambachan, A., 2025. *Large language models: An applied econometric framework* (No. w33344). National Bureau of Economic Research.

Maddison, C.J., Mnih, A. and Teh, Y.W., 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D. and Hajishirzi, H., 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. and Mian, A., 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, *16*(5), pp.1-72.

Olden, A. and Møen, J., 2022. The triple difference estimator. *The Econometrics Journal*, *25*(3), pp.531-553.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), p.9.

Roth, J., Sant'Anna, P.H., Bilinski, A. and Poe, J., 2023. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, *235*(2), pp.2218-2244.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J., 1986. Learning representations by back-propagating errors. *nature*, *323*(6088), pp.533-536.

Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S. and Chadha, A., 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, S. and Levy, B., 2024. BeanCounter: A low-toxicity, large-scale, and open dataset of business-oriented text. *Advances in Neural Information Processing Systems*, *37*, pp.91653-91690.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V. and Zhou, D., 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, pp.24824-24837.

Xu, Z., Jain, S. and Kankanhalli, M., 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Yang, Y., Chern, E., Qiu, X., Neubig, G. and Liu, P., 2024. Alignment for honesty. *Advances in Neural Information Processing Systems*, *37*, pp.63565-63598.

Yin, Z., Sun, Q., Guo, Q., Wu, J., Qiu, X. and Huang, X., 2023. Do large language models know what they don't know?. *arXiv preprint arXiv:2305.18153*.

Yoo, M., 2024. How Much Should We Trust Large Language Model-Based Measures for Accounting and Finance Research? *Available at SSRN*.

Zhang, X., Li, M. and Wu, J., 2024. Co-occurrence is not factual association in language models. *Advances in Neural Information Processing Systems*, *37*, pp.64889-64914.

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z. and Du, Y., 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, *1*(2).

Appendix A. Prompting Strategies

Please note that the prompts and completions presented here are for illustrative purposes only. As explained in Section 2.2, using the same prompt on the same LLM can yield different completions across runs due to LLMs' nondeterministic nature. Therefore, it is meaningless to evaluate LLMs' performance based on individual examples.

Baseline Prompting

This prompting strategy reflects how a retail investor would typically ask such a question. Following conventions in finance and accounting literature (Bond, Klok, and Zhu 2023; Yoo 2024), I suppress explanatory text and impose an output format to facilitate response parsing. The baseline prompt and its completion are as follows:

```
Prompt: What is the Total Assets under GAAP of Aar Corp for the fiscal year 1990? Do
not include any explanatory text. Return the value in millions of USD in the
following JSON format: {'value': float}. If you do not have it, return NA for the
value.
Completion: {'value': 233.6}
```

Context-rich Prompting

A commonly used technique that has been shown to improve response quality is prompt engineering with contextual information (Sahoo et al. 2024). Following prompt design guidance from OpenAI and de Kok (2025), I augment the baseline prompt with additional information, such as the definition of the queried item (e.g., total assets), where it is typically disclosed (e.g., balance sheet in annual reports), and where such reports can be found (e.g., EDGAR).[25] This strategy mirrors the efforts a retail investor might undertake to craft a more sophisticated query. The context-rich prompt and its completion are as follows:

```
Prompt: What is the Total Assets under GAAP of Aar Corp for the fiscal year 1990?
Total Assets is the item representing the total value of assets reported on the
Balance Sheet. Public firms report their Total Assets in the Balance Sheet of annual
reports. And annual reports can be accessed through firm websites or Form 10-K
filings in EDGAR. Do not include any explanatory text. Return the value in millions
of USD in the following JSON format: {'value': float}. If you do not have it, return
NA for the value.
Completion: {'value': 233.6}
```

Chain-of-Thought Prompting

Another widely used prompt engineering technique is CoT that instructs the model to generate a sequence of intermediate reasoning steps before arriving at the final answer (Kojima et al. 2022; Wei et al. 2022). This approach has been shown to significantly enhance LLM performance in reasoning tasks. The CoT prompt and its completion are as follows:

```
Prompt: What is the Total Assets under GAAP of Aar Corp for the fiscal year 1990?
Explain step by step how you find the answer. Return the brief explanation and the
value in millions of USD in the following JSON format: {'explanation': string,
'value': float}. If you do not have it, return NA for the value.
Completion: {'explanation': "To find the Total Assets under GAAP for AAR Corp for the
fiscal year 1990, I would need to locate the company's balance sheet for that year,
typically found in their annual report or 10-K filing. The Total Assets figure is
usually listed as the last line in the 'Assets' section of the balance sheet.
However, since I do not have access to the specific financial statements or data for
```

[25] Prompt engineering guidance from OpenAI: https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api.

```
AAR Corp for 1990 in the provided context, I cannot provide the exact value.",
'value': 'NA'}
```

Few-shot Prompting

Another widely adopted prompt engineering strategy is in-context learning, which enables large language models to adapt to a given task by providing examples within the prompt (Brown et al. 2020). A common implementation of this approach is few-shot prompting, where the user includes a small number of input-output examples to guide the model's behavior. The few-shot prompt and its completion are as follows:

```
Prompt: What is the Total Assets under GAAP of Aar Corp for the fiscal year 1990? Do
not include any explanatory text. Return the value in millions of USD in the
following JSON format: {'value': float}. If you do not have it, return NA for the
value. Here are three examples:

Prompt 1: What is the Total Assets under GAAP of Tuboscope Vetco International Corp
for the fiscal year 1995?
Completion 1: {'value': 306.679}
Prompt 2: What is the Total Assets under GAAP of Spectrum Control Inc for the fiscal
year 2002?
Completion 2: {'value': 81.824}
Prompt 3: What is the Total Assets under GAAP of Adtran Holdings Inc for the fiscal
year 2008?
Completion 3: {'value': 473.615}

Completion: {'value': 233.894}
```

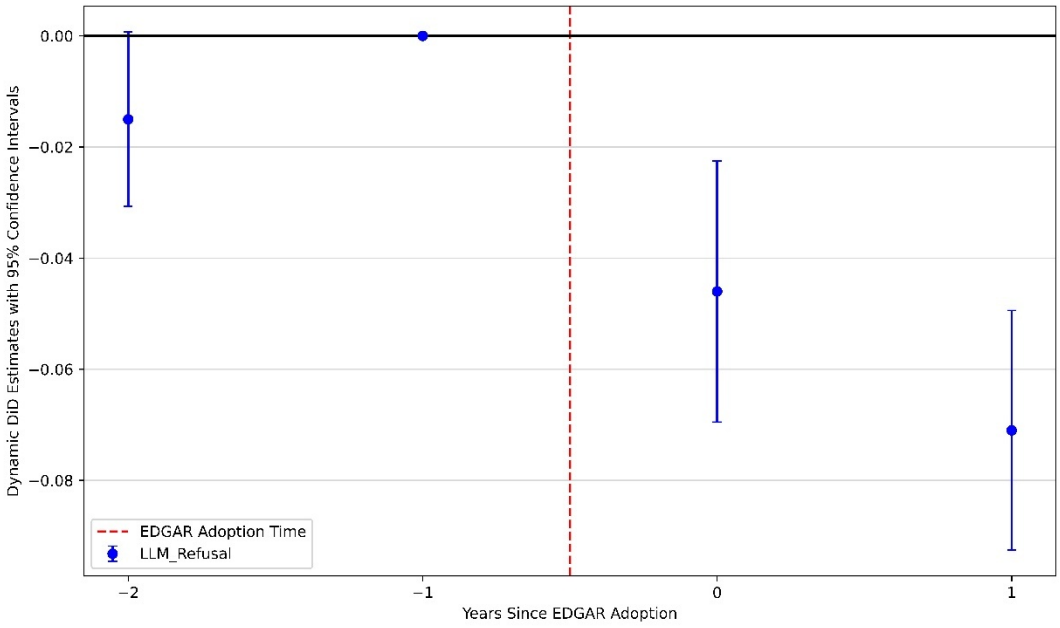
Appendix B. Variable Definitions

Variable	Definition	Source
Hallucination-related Variables		
True_Value	The true value of a financial statement item (total assets, total liabilities, revenues, or net income) for a firm in a year.	Compustat
Retrieved_Value	The retrieved value from the LLM when it is queried about a financial statement item (total assets, total liabilities, revenues, or net income) for a firm in a year.	OpenAI
LLM_Refusal	An indicator equal to one if the LLM acknowledges a lack of knowledge regarding the queried information, and zero otherwise.	Compustat, OpenAI
LLM_Deviation	The absolute difference between *True_Value* and *Retrieved_Value*, scaled by *True_Value*, for a firm in a year.	Compustat, OpenAI
LLM_Material	An indicator equal to one if *LLM_Deviation* exceeds 0.05, and zero otherwise.	Compustat, OpenAI
Firm Characteristics		
MarketCap	The natural logarithm of a firm's market value of equity (MKVALT or PRCC_F * CSHO) in a year.	Compustat
FreqEarn	The natural logarithm of one plus the number of earnings forecasts issued by a firm in a year.	I/B/E/S
Analyst	The natural logarithm of one plus the number of analysts following a firm in a year.	I/B/E/S
Leverage	The ratio of a firm's total liabilities (LT) to its total assets (AT) in a year.	Compustat
ROA	The ratio of a firm's net income (NI) to its total assets (AT) in a year.	Compustat
MTB	The ratio of a firm's market value of equity (MKVALT or PRCC_F * CSHO) to its book value of equity (CEQ) in a year.	Compustat
SalesGrowth	The difference between a firm's sales in a year and its sales in the last year, divided by its sales in the last year.	Compustat
Institution	The average of quarterly institutional ownership ratios for a firm in a year. Quarterly institutional ownership ratios are measured by the total number of shares owned by institutions divided by the number of shares outstanding.	Thomson/Refinitiv
ReturnVol	The standard deviation of a firm's daily stock return in a year.	CRSP
EGC	An indicator equal to one if the IPO firm is an emerging growth company, and zero otherwise.	Audit Analytics
SPAC	An indicator equal to one if the IPO firm is a special purpose acquisition company, and zero otherwise.	Audit Analytics
Other Variables		
SECWeb	An indicator equal to one if the RAG with Web Search approach sources the data from the SEC's website (e.g., EDGAR).	OpenAI
FileMatchScore	A continuous number between 0 and 1 returned by OpenAI when using RAG with File Upload approach. It denotes the highest similarity score any excerpt in the file has with the query. The model then retrieves the value of the queried item from the excerpt with the highest similarity score.	OpenAI

Figure 1. Effect Dynamics of EDGAR Adoption on Accounting Deviations

Notes: This figure illustrates the effect dynamics of EDGAR Adoption on the LLM's knowledge deficits (Panel A) and accounting deviations (Panel B), along with the 95% confidence intervals. Because EDGAR adoption merely digitalizes existing accounting information, *LLM_Refusal* in this setting captures knowledge deficits (i.e., failures to learn) rather than accounting fabrications.

Panel A. Knowledge Deficits



Panel B. Accounting Deviations

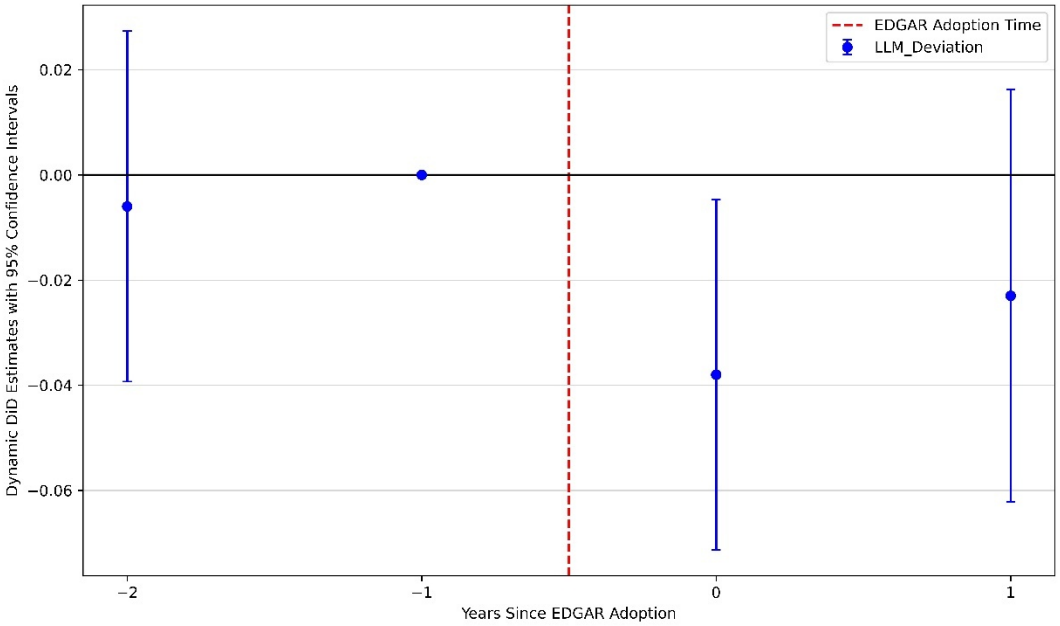
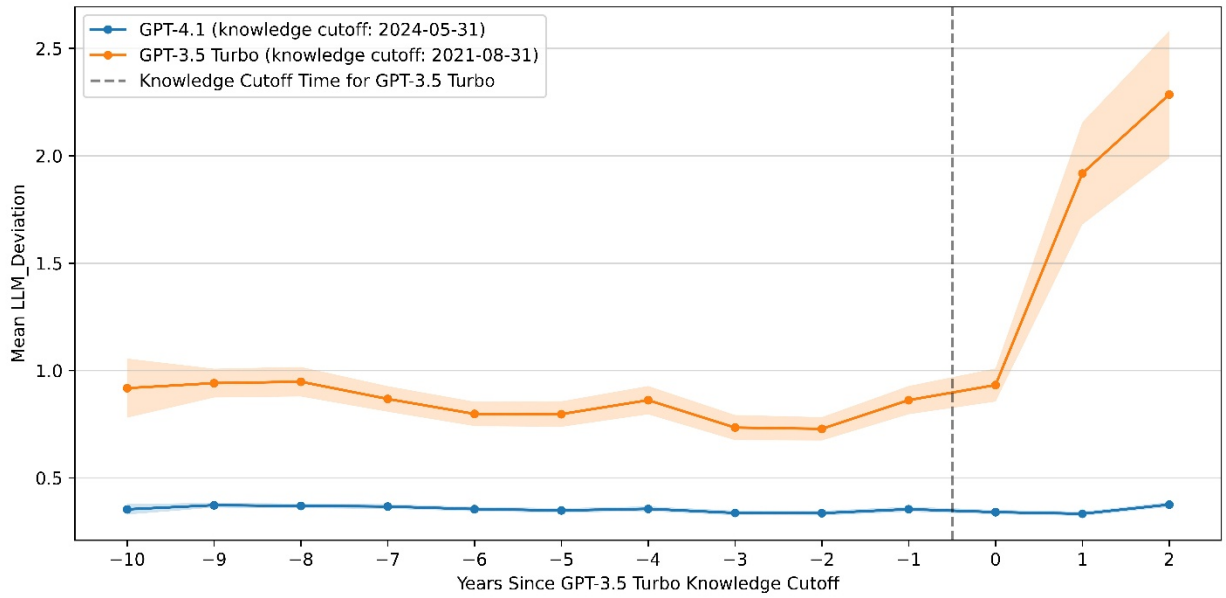


Figure 2. Effect Dynamics of LLM Knowledge Cutoff on Accounting Deviations

Notes: This figure illustrates the effect dynamics of LLM knowledge cutoff on accounting deviations. Panel A plots the unconditional mean *LLM_Deviation* around GPT-3.5 Turbo's knowledge cutoff, and Panel B plots the estimates from the event study, along with the 95% confidence intervals.

Panel A. Unconditional Plot



Panel B. Event Study

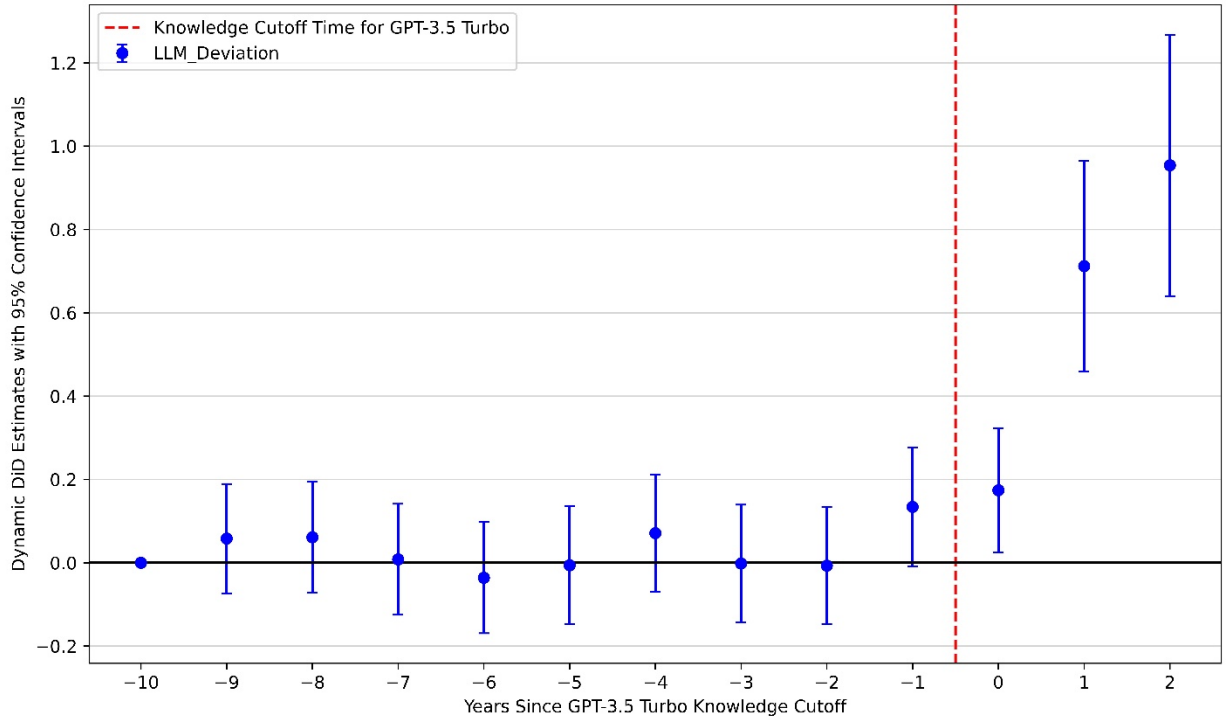
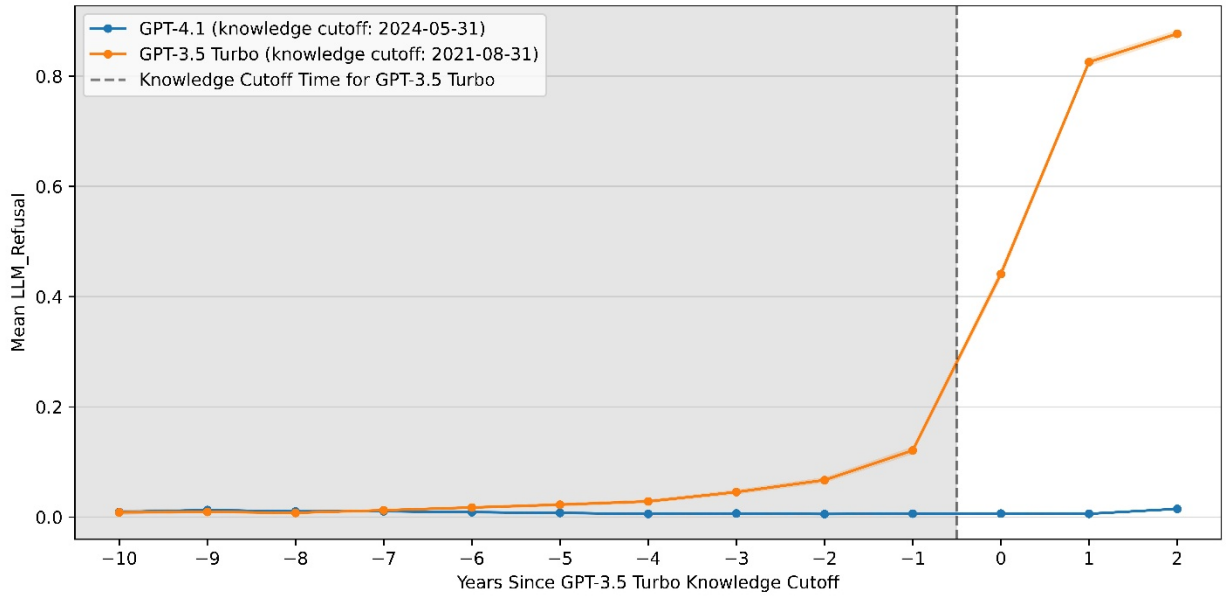


Figure 3. Accounting Fabrications after LLM Knowledge Cutoff

Notes: This figure illustrates the accounting fabrications when querying the LLM for post-knowledge-cutoff financials, along with the 95% confidence intervals. The orange line in Panel A (the green and red lines in Panel B) plots the mean *LLM_Refusal* when querying GPT-3.5 Turbo for financials in years around the model's knowledge cutoff date. Since post-cutoff accounting data is not included in training, no-refusal (i.e., mean *LLM_Refusal*<1.0) on the right-hand side (non-grey shaded area) indicates accounting fabrications. The blue line in Panel A (the blue and orange lines in Panel B) represents the mean *LLM_Refusal* from querying GPT-4.1.

Panel A. Post-Knowledge-Cutoff Accounting Fabrications



Panel B. Accounting Fabrications by Firm Groups

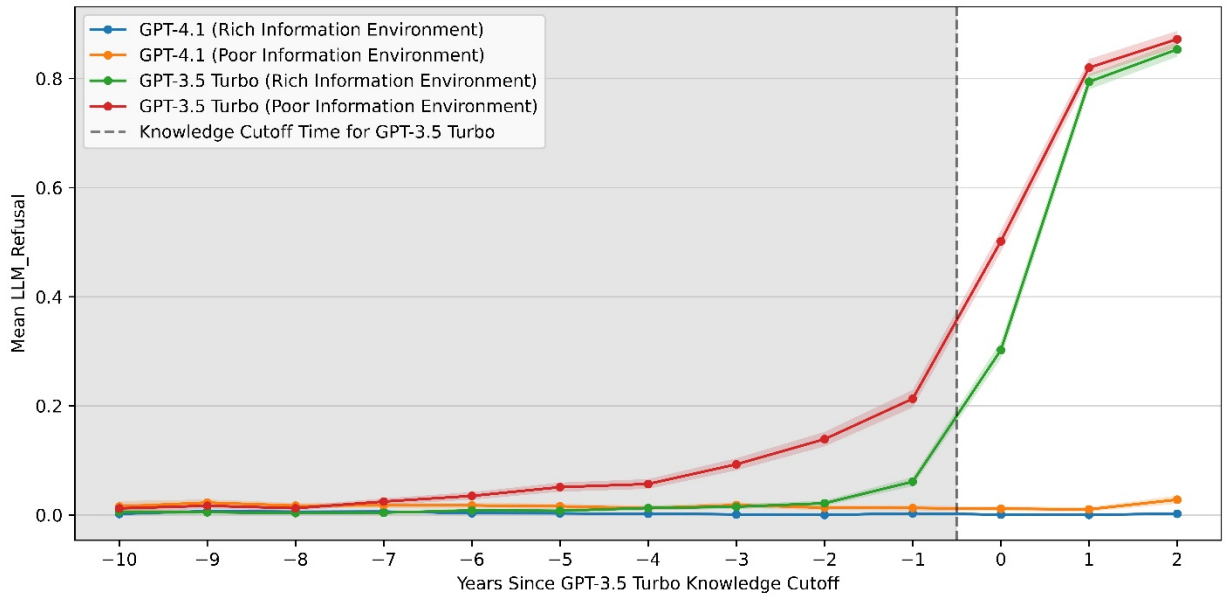


Table 1. Sample Construction

Notes: This table summarizes the construction of the samples used throughout the paper. Panels A-D outline the construction of samples for accounting deviations, accounting fabrications, EDGAR adoption, and LLM knowledge cutoff, respectively.

Panel A. Accounting Deviations

#	Steps	# Firm-year Left	# Firm Left
1	Select firm-years from Compustat that have fiscal-year-end dates between 1990 and 2024, are incorporated and headquartered in the U.S., listed on NYSE/AMEX/NASDAQ, and not in Finance (SIC: 6000-6999) or Utility industry (SIC: 4900-4999)	110,893	8,996
2	Drop observations with missing firm identifier PERMNO, TICKER, CIK, or COMPANYID	94,109	7,975
3	Drop observations with missing major variables in Compustat, CRSP, IBES, Thomson, and SRAF 10-X Header	89,777	7,687
4	Drop observations with fiscal-year-end dates after or within 90 days before the LLM's knowledge cutoff date	87,430	7,593

Panel B. Accounting Fabrications

#	Steps	# Firm-year Left (# Cohort-firm-year)	# Firm Left
1	Following Panel A: Drop observations of firms missing IPO records in Audit Analytics	17,028	2,340
2	Drop observations of firms matched with multiple IPO dates	16,956	2,334
3	Drop observations of firms that have observations before IPO dates	15,186	2,237
4	Drop observations of firms ever switched fiscal-year-end dates	14,747	2,195
5a	Create a stacked cohort sample with a 10-year window and public firms as control firms	16,866 (46,290)	1,185
5b	Create a stacked cohort sample with a 10-year window and private firms as control firms	12,333 (21,540)	1,131
5c	Stack 5a and 5b but only keep IPO firms' pre-IPO unique observations	4,855	971

Panel C. EDAGR Adoption

#	Steps	# Firm-year Left (# Cohort-firm-year)	# Firm Left
1	Following Panel A: Drop observations of firms that are not in the SEC EDGAR adoption list	40,498	2,374
2	Drop observations of firms ever switched fiscal-year-end dates	34,787	2,081
3	Create a stacked cohort sample with a 4-year window and already-adopted firms as control firms	8,603 (13,045)	1,663

Panel D. LLM Knowledge Cutoff

#	Steps	# Firm-year Left (# Model-firm-year)	# Firm Left
1	Following Panel A: Duplicate the sample for GPT-3.5 Turbo	87,430 (174,860)	7,593
2	Drop observations of year before 2012 to have 10-year pre-period	28,075 (56,150)	3,933
3	Drop observations of firms with fiscal-year-end dates after or within 90 days before the corresponding model's knowledge cutoff	27,937 (55,874)	3,932
4	Drop observations of firms ever switched fiscal-year-end dates	27,200 (54,400)	3,847

Table 2. Summary Statistics for the Sample of Accounting Deviations

Notes: This table reports summary statistics for the sample of accounting deviations. For Panel A, the data used here is the Compustat firm-year sample created in Panel A of Table 1. Each firm-year observation requires querying the LLM for four financial statement items (87,430 firm-years * 4 items = 349,720 requests). Panel A summarizes LLM queries and responses across different prompting strategies. "Unsupported Input" and "Unparsable Output" reflect the share of requests either rejected by the LLM due to unsupported inputs or producing outputs that do not conform to the expected format. The mean *LLM_Refusal* is the average *LLM_Refusal* across all requests passing the input and output checks. The mean *LLM_Deviation* and the mean *LLM_Material* are the average *LLM_Deviation* and *LLM_Material* across all requests that pass the input and output checks and are not refused by the LLM (i.e., conditional on *LLM_Refusal*=0), respectively. In all cases, higher values of *LLM_Deviation* suggest a higher degree of deviation in the LLM's output. For Panel B, the data used here is from Panel A but being restricted to requests that pass the input and output checks and are not refused by the LLM. Panel B presents summary statistics of *LLM_Deviation* across different prompting strategies. For Panel C, the data used here is the observations from Panel B under few-shot prompting. Panel C presents summary statistics on firm characteristics of the sample firms. All continuous variables are winsorized at 1% and 99%. Detailed definitions of all variables are provided in Appendix B.

Panel A. LLM Queries for Accounting Deviations

Prompting Strategy	# LLM Request	Unsupported Input	Unparsable Output	Mean *LLM_Refusal*	Mean *LLM_Deviation*	Mean *LLM_Material*
Baseline	349,720	0.00%	0.03%	0.04	0.66	0.84
Context-rich	349,720	0.00%	0.04%	0.11	0.54	0.83
Chain-of-Thought	349,720	0.00%	0.01%	0.61	0.31	0.70
Few-shot	349,720	0.00%	0.08%	0.08	0.48	0.84

Panel B. Summary Statistics for *LLM_Deviation*

Prompting Strategy	Count	Mean	STD	25%	Median	75%
Baseline	336,132	0.66	1.13	0.10	0.30	0.63
Context-rich	309,844	0.54	0.78	0.09	0.28	0.59
Chain-of-Thought	136,243	0.31	0.40	0.03	0.15	0.38
Few-shot	319,791	0.48	0.61	0.10	0.29	0.58

Panel C. Firm Characteristics

Variable	Count	Mean	STD	25%	Median	75%
MarketCap	319,791	6.26	2.07	4.77	6.22	7.64
FreqEarn	319,791	0.71	0.99	0.00	0.00	1.61
Analyst	319,791	1.83	1.03	1.10	1.95	2.64
Leverage	319,791	0.50	0.26	0.30	0.49	0.65
ROA	319,791	-0.02	0.22	-0.03	0.04	0.08
MTB	319,791	3.34	5.40	1.29	2.21	3.93
SalesGrowth	319,791	0.22	0.66	-0.01	0.09	0.24
Institution	319,791	0.53	0.32	0.26	0.56	0.81
ReturnVol	319,791	0.04	0.02	0.02	0.03	0.04

Table 3. Summary Statistics for the Sample of Accounting Fabrications

Notes: This table reports summary statistics for the sample of accounting fabrications. For Panel A, the data used here is the pre-IPO sample in step 5c in Panel B of Table 1. Each observation requires querying the LLM for four financial statement items (4,855 firms * 4 items = 19,420 requests). Panel A summarizes LLM queries and responses across different prompting strategies. "Unsupported Input" and "Unparsable Output" reflect the share of requests either rejected by the LLM due to unsupported inputs or producing outputs that do not conform to the expected format. The mean *LLM_Refusal* (absolute value) is the average *LLM_Refusal* across all requests passing the input and output checks. The mean *LLM_Refusal* (relative to public firms) is an adjusted measure specified in Section 3.2. It is calculated as the average of (pre-IPO refusal of IPO firms - post-IPO refusal of IPO firms) - (pre-IPO refusal of public firms - post-IPO refusal of public firms) across requests. The mean *LLM_Refusal* (relative to private firms) is constructed analogously, using private firms as the benchmark group. In all cases, higher values of *LLM_Refusal* suggest a lower degree of fabrication in the LLM's output. For Panel B, the data used here is the observations from Panel A under few-shot prompting, further restricting to requests with valid input and output and firms with non-missing first two years' observations post-IPO. Panel B presents summary statistics of the sample firms. All continuous variables are winsorized at 1% and 99%. Detailed definitions of all variables are provided in Appendix B.

Panel A. LLM Queries for Accounting Fabrications

Prompting Strategy	# LLM Request	Unsupported Input	Unparsable Output	Mean *LLM_Refusal* (absolute value)	Mean *LLM_Refusal* (relative to public firms)	Mean *LLM_Refusal* (relative to private firms)
Baseline	19,420	0.00%	0.01%	0.45	0.41	0.58
Context-rich	19,420	0.00%	0.00%	0.60	0.54	0.61
Chain-of-Thought	19,420	0.00%	0.02%	0.88	0.38	0.52
Few-shot	19,420	0.00%	0.01%	0.61	0.55	0.64

Panel B. Firm Characteristics

Variable	Count	Mean	STD	25%	Median	75%
MarketCap	16,178	6.49	1.35	5.68	6.50	7.39
FreqEarn	16,178	0.40	0.64	0.00	0.00	0.81
Analyst	16,178	1.86	0.68	1.50	1.94	2.28
Leverage	16,178	0.47	0.28	0.24	0.42	0.65
ROA	16,178	-0.09	0.23	-0.19	0.00	0.05
MTB	16,178	4.86	7.44	2.01	3.65	6.90
SalesGrowth	16,178	0.46	0.79	0.09	0.26	0.53
Institution	16,178	0.44	0.29	0.23	0.41	0.64
ReturnVol	16,178	0.04	0.01	0.03	0.03	0.05

Table 4. Correlation between Accounting Deviations and Information Production & Dissemination

Notes: This table presents regression results for the correlation between accounting deviations and the production and dissemination of accounting information. These regressions use the firm-year panel from Panel A of Table 1, where each firm-year observation requires querying the LLM with few-shot prompting for four financial statement items. Accounting deviations are measured by *LLM_Deviation*. Higher values of *LLM_ Deviation* indicate a higher degree of deviation. *MarketCap*, *FreqEarn*, and *Analysts* serve as proxies for the production and dissemination of accounting information. All regressions include accounting item, firm, and calendar year fixed effects. Detailed definitions of all variables are provided in Appendix B. *t*-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

D.V. = *LLM_Deviation*	**(1)**	**(2)**	**(3)**
MarketCap	-0.042***	-0.053***	-0.053***
	(-17.445)	(-18.542)	(-11.217)
FreqEarn	-0.007***	-0.005**	-0.005**
	(-3.221)	(-2.327)	(-2.041)
Analysts	-0.046***	-0.038***	-0.038***
	(-13.736)	(-10.811)	(-7.981)
Leverage		-0.081***	-0.081***
		(-6.640)	(-4.912)
ROA		0.129***	0.129***
		(10.563)	(9.293)
MTB		0.002***	0.002***
		(5.776)	(4.222)
SalesGrowth		-0.003	-0.003
		(-1.554)	(-1.410)
Institution		-0.036***	-0.036***
		(-3.805)	(-3.768)
ReturnVol		0.774***	0.774***
		(6.238)	(3.977)
Item FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Clustering	Firm	Firm	Firm, Year
Observations	319,706	319,706	319,706
Adj. R-squared	0.291	0.293	0.293

Table 5. Correlation between Accounting Fabrications and Information Production & Dissemination

Notes: This table presents regression results for the correlation between accounting fabrications and the production and dissemination of accounting information. These regressions use the firm-year-item panel from step 5c in Panel B of Table 1 and few-shot prompting. Accounting fabrications are measured by *LLM_Refusal*. Higher values of *LLM_Refusal* indicate a lower degree of fabrication. Firm characteristics are measured as the average over the first two years following the IPO. *MarketCap*, *FreqEarn*, and *Analysts* serve as proxies for the production and dissemination of accounting information. All regressions include accounting item, 4-digit SIC industry, and event year fixed effects. Detailed definitions of all variables are provided in Appendix B. *t*-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

D.V. = *LLM_Refusal*	**(1)**	**(2)**	**(3)**
MarketCap	0.008	0.013	0.012
	(0.997)	(1.271)	(1.103)
FreqEarn	-0.084***	-0.078***	-0.077***
	(-5.320)	(-5.070)	(-5.068)
Analysts	-0.060***	-0.068***	-0.062***
	(-2.739)	(-3.266)	(-2.679)
Leverage		-0.031	-0.026
		(-0.605)	(-0.497)
ROA		-0.085	-0.078
		(-1.224)	(-1.116)
MTB		0.001	0.001
		(0.497)	(0.578)
SalesGrowth		0.022*	0.023*
		(1.668)	(1.660)
Institution		0.032	0.029
		(0.986)	(0.841)
ReturnVol		0.389	0.504
		(0.348)	(0.451)
EGC			0.020
			(0.368)
SPAC			0.124
			(1.337)
Item FE	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Clustering	Industry	Industry	Industry
Observations	16,178	16,178	16,178
Adj. R-squared	0.386	0.388	0.388

Table 6. Effect of EDGAR Adoption on Accounting Deviations

Notes: This table presents the effect of EDGAR adoption on knowledge deficits (*LLM_Refusal*) and accounting deviations (*LLM_Deviation*). The data used here is the cohort-firm-year sample from Panel C of Table 1, where each observation requires querying the LLM for four financial statement items (13,045 cohort-firm-years * 4 items = 52,180 requests). The requests are further restricted to those that pass the input and output checks (column (1) and (2)) and are not refused by LLM (column (3) and (4)). Detailed variable definitions are provided in Appendix B. *t*-statistics are reported in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

	(1)	**(2)**	**(3)**	**(4)**
Dependent Variable =	*LLM_Refusal*		*LLM_Deviation*	
*Treated * Post*	-0.051***	-0.050***	-0.028*	-0.012
	(-4.640)	(-4.520)	(-1.722)	(-0.774)
MarketCap		0.002		-0.088***
		(0.270)		(-7.449)
FreqEarn		-0.001		0.001
		(-0.170)		(0.113)
Analysts		-0.010*		-0.045***
		(-1.830)		(-4.150)
Leverage		-0.027		-0.149***
		(-1.175)		(-2.928)
ROA		-0.011		0.006
		(-0.392)		(0.116)
MTB		-0.001		0.005***
		(-1.505)		(3.714)
SalesGrowth		-0.003		-0.036***
		(-0.653)		(-3.036)
Institution		0.005		-0.032
		(0.221)		(-0.732)
ReturnVol		0.120		0.809
		(0.370)		(1.550)
Cohort-Item FE	Yes	Yes	Yes	Yes
Cohort-Firm FE	Yes	Yes	Yes	Yes
Cohort-Year FE	Yes	Yes	Yes	Yes
Clustering	Cohort-Firm	Cohort-Firm	Cohort-Firm	Cohort-Firm
Observations	52,106	52,106	46,164	46,164
Adj. R-squared	0.566	0.566	0.359	0.363

Table 7. Effect of LLM Knowledge Cutoff on Accounting Deviations

Notes: This table presents the effect of LLM knowledge cutoff on accounting deviations, leveraging the knowledge cutoff difference between GPT-4.1 and GPT-3.5 Turbo. The data used here is the model-firm-year sample from Panel D of Table 1, where each observation requires querying the LLM for four financial statement items (54,400 model-firm-years * 4 items = 217,600 requests). The requests are further restricted to those that pass the input and output checks for and are not refused by both LLMs. Detailed variable definitions are provided in Appendix B. *t*-statistics are reported in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

D.V. = *LLM_Deviation*	**(1)**	**(2)**	**(3)**
*Treated * Post*	0.585***	0.376***	0.385***
	(12.548)	(8.197)	(8.413)
MarketCap			-0.102***
			(-7.922)
FreqEarn			-0.031***
			(-3.567)
Analysts			-0.086***
			(-4.703)
Leverage			-0.192***
			(-3.925)
ROA			0.245***
			(4.534)
MTB			-0.000
			(-0.417)
SalesGrowth			-0.012
			(-1.114)
Institution			-0.057
			(-1.426)
ReturnVol			0.064
			(0.123)
Model-Firm FE	Yes	Yes	Yes
Model-Year FE	Yes	Yes	Yes
Model-Item FE	No	Yes	Yes
Clustering	Model-Firm	Model-Firm	Model-Firm
Observations	192,891	192,891	192,891
Adj. R-squared	0.178	0.199	0.200

Table 8. Triple-Differences Analysis for the Effect of Accounting Information Production and Dissemination on Accounting Fabrications

Notes: This table presents the triple-differences analysis for the effect of accounting information production and dissemination on accounting fabrications. The data used here is the model-firm-year sample from Panel D of Table 1, where each observation requires querying the LLM for four financial statement items (54,400 model-firm-years * 4 items = 217,600 requests). The requests are further restricted to those that pass the input and output checks for both LLMs. I further divide the sample into high and low subsamples if pre-cutoff *MarketCap*, *FreqEarn*, or *Analysts* is in the top and bottom 33%, respectively. The analysis for each subsample employs a generalized DiD regression. For triple-DiD estimate, I use Wald test to test the statistical significance of the cross-equation parameter equality for *Treated * Post* and present the differences and their corresponding significance in the last row. Detailed variable definitions are provided in Appendix B. *t*-statistics are reported in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

D.V. = *LLM_Refusal*	**(1)** High *MarketCap*	**(2)** Low *MarketCap*	**(3)** High *FreqEarn*	**(4)** Low *FreqEarn*	**(5)** High *Analysts*	**(6)** Low *Analysts*
*Treated * Post*	0.605***	0.617***	0.590***	0.608***	0.609***	0.607***
	(111.304)	(85.187)	(101.791)	(97.874)	(106.807)	(85.538)
MarketCap	-0.008***	-0.005**	-0.008***	-0.003	-0.008***	-0.001
	(-3.332)	(-2.020)	(-3.432)	(-1.467)	(-3.744)	(-0.486)
FreqEarn	0.001	0.003	-0.001	0.009*	0.001	0.006**
	(0.813)	(1.162)	(-0.525)	(1.875)	(0.482)	(2.070)
Analysts	-0.009**	-0.005	-0.001	-0.009**	-0.006	-0.006*
	(-2.228)	(-1.304)	(-0.225)	(-2.421)	(-1.474)	(-1.861)
Leverage	0.013	-0.026***	0.004	-0.008	0.015*	-0.019*
	(1.450)	(-2.956)	(0.392)	(-1.036)	(1.954)	(-1.808)
ROA	0.038***	0.003	0.033***	0.015**	0.030***	0.007
	(2.775)	(0.377)	(2.631)	(2.312)	(3.088)	(0.886)
MTB	-0.000	0.000	-0.000	0.000	0.000	-0.000
	(-0.243)	(0.926)	(-0.541)	(0.281)	(0.096)	(-0.018)
SalesGrowth	0.002	0.001	-0.001	0.001	-0.002	0.003
	(0.883)	(0.986)	(-0.398)	(0.690)	(-1.087)	(1.620)
Institution	0.008	-0.010	0.002	-0.009	0.016*	-0.014
	(0.770)	(-0.865)	(0.209)	(-1.001)	(1.764)	(-1.108)
ReturnVol	0.014	-0.007	0.045	-0.019	0.007	-0.044
	(0.088)	(-0.068)	(0.305)	(-0.202)	(0.048)	(-0.418)
Model-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Model-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Model-Item FE	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Model-Firm	Model-Firm	Model-Firm	Model-Firm	Model-Firm	Model-Firm
Observations	76,852	61,172	76,160	78,170	74,594	64,594
Adj. R-squared	0.644	0.616	0.602	0.639	0.641	0.621
Triple-DiD Estimate: *High-Low*	**-0.012***		**-0.018***		0.002	

Table 9. Accounting Hallucinations and Implementation Cost of Potential Solutions

Notes: This table reports the accounting hallucinations (Panel A) and the implementation cost (Panel B) of potential solutions. The data used here is from Panel A of Table 1, but I randomly select 10% of firms and filter out firm-year observations without an SEC filing raw document in Compustat's EDGAR filing database. The final sample contains 6,351 firm-year observations from 697 unique firms. Each firm-year observation requires querying the LLM for four financial statement items (6,351 firm-years * 4 items = 25,404 requests). Panel A summarizes LLM queries and responses across different prompting strategies. "Unsupported Input" and "Unparsable Output" reflect the share of requests either rejected by the LLM due to unsupported inputs or producing outputs that do not conform to the expected format. The mean *LLM_Refusal* is the average *LLM_Refusal* across all requests passing the input and output checks. The mean *LLM_Deviation* and the mean *LLM_Material* are the average *LLM_Deviation* and *LLM_Material* across all requests that pass the input and output checks and are not refused by the LLM (i.e., conditional on *LLM_Refusal*=0), respectively. In all cases, higher values of *LLM_Deviation* suggest a higher degree of deviation in the LLM's output. Panel B reports the cost of implementing different prompting strategies on the 10% subsample. Total cost consists of four components: the number of API requests sent to OpenAI (Request Cost), the total length of all queries (Input Cost), the total length of all responses (Output Cost), and the non-monetary effort required to implement the methodology (Labor Cost). For the RAG with File Upload approach, I first locate the correct filing in EDGAR for the requested financial statement item (i.e., the 10-K or 10-K/A for the queried firm-year), download it, and remove images to reduce file size. I then manually upload the cleaned filing to OpenAI, instruct the model to answer questions using only the uploaded file, and delete the file after completion. This process is repeated for each firm-year-item request to ensure the model has access only to the correct filing for the query.

Panel A. Accounting Hallucinations

Prompting Strategy	# LLM Request	Unsupported Input	Unparsable Output	Mean *LLM_Refusal*	Mean *LLM_Deviation*	Mean *LLM_Material*
Baseline	25,404	0.00%	0.04%	0.01	0.58	0.83
Context-rich	25,404	0.00%	0.06%	0.05	0.52	0.83
Chain-of-Thought	25,404	0.00%	0.00%	0.58	0.30	0.70
Few-shot	25,404	0.00%	0.05%	0.03	0.47	0.84
RAG with Web Search	25,404	0.00%	0.89%	0.12	0.22	0.41
RAG with File Upload	25,404	5.50%	0.01%	0.02	0.07	0.16

Panel B. Implementation Cost

Prompting Strategy	# LLM Request	#M Input Token	#M Output Token	Request Cost	Input Cost	Output Cost	Labor Cost	Total Cost
Baseline	25,404	1.78	0.23	0.00	1.78	0.93	-	2.71
Context-rich	25,404	3.15	0.23	0.00	3.15	0.92	-	4.07
Chain-of-Thought	25,404	2.11	2.80	0.00	2.11	11.18	-	13.29
Few-shot	25,404	4.77	0.23	0.00	4.77	0.92	-	5.70
RAG with Web Search	25,404	26.06	1.61	889.14	52.11	12.91	-	954.17
RAG with File Upload	25,404	83.66	1.29	63.51	83.66	5.16	**Intense**	152.33+**Labor**

Table 10. Determinants of Accounting Hallucinations under Potential Solutions

Notes: This table presents regression results for the correlation between accounting deviations and their potential determinants, under RAG methods. The data used here is from Panel A of Table 1, but I randomly select 10% of firms and filter out firm-year observations without an SEC filing raw document in Compustat's EDGAR filing database. The final sample contains 6,351 firm-year observations from 697 unique firms. Each firm-year observation requires querying the LLM for four financial statement items (6,351 firm-years * 4 items = 25,404 requests). For regressions, I further filer out requests that do not pass the input and output checks or are refused by the LLM under a certain prompting strategy. All regressions include accounting item, firm, and calendar year fixed effects. Detailed definitions of all variables are provided in Appendix B. *t*-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

D.V. = *LLM_Deviation*	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Prompting Strategy	Few-shot	RAG with Web Search		RAG with File Upload	
MarketCap	-0.045***	0.003	0.003	0.003	0.003
	(-4.358)	(0.554)	(0.505)	(1.109)	(1.114)
FreqEarn	-0.017**	-0.004	-0.004	-0.001	-0.001
	(-2.184)	(-0.723)	(-0.725)	(-0.532)	(-0.501)
Analysts	-0.022*	-0.018*	-0.017*	0.000	-0.000
	(-1.772)	(-1.799)	(-1.787)	(0.051)	(-0.003)
SECWeb			0.015**		
			(2.002)		
FileMatchScore					-0.026***
					(-5.288)
Leverage	-0.040	-0.006	-0.006	-0.001	-0.001
	(-1.007)	(-0.230)	(-0.232)	(-0.060)	(-0.084)
ROA	0.070	-0.061**	-0.060**	-0.035***	-0.035***
	(1.199)	(-2.453)	(-2.426)	(-3.713)	(-3.734)
MTB	0.001	0.000	0.000	-0.001*	-0.001*
	(0.612)	(0.167)	(0.173)	(-1.899)	(-1.887)
SalesGrowth	0.006	0.008	0.008	-0.002	-0.002
	(0.751)	(1.286)	(1.281)	(-1.023)	(-1.013)
Institution	0.019	-0.029	-0.029	-0.014	-0.014
	(0.578)	(-1.012)	(-1.018)	(-0.930)	(-0.931)
ReturnVol	0.592	0.262	0.252	0.146	0.137
	(1.518)	(0.724)	(0.697)	(1.120)	(1.049)
Item FE	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Clustering	Firm	Firm	Firm	Firm	Firm
Observations	24,471	22,183	22,183	23,504	23,504
Adj. R-squared	0.265	0.226	0.226	0.057	0.059

Online Appendix for

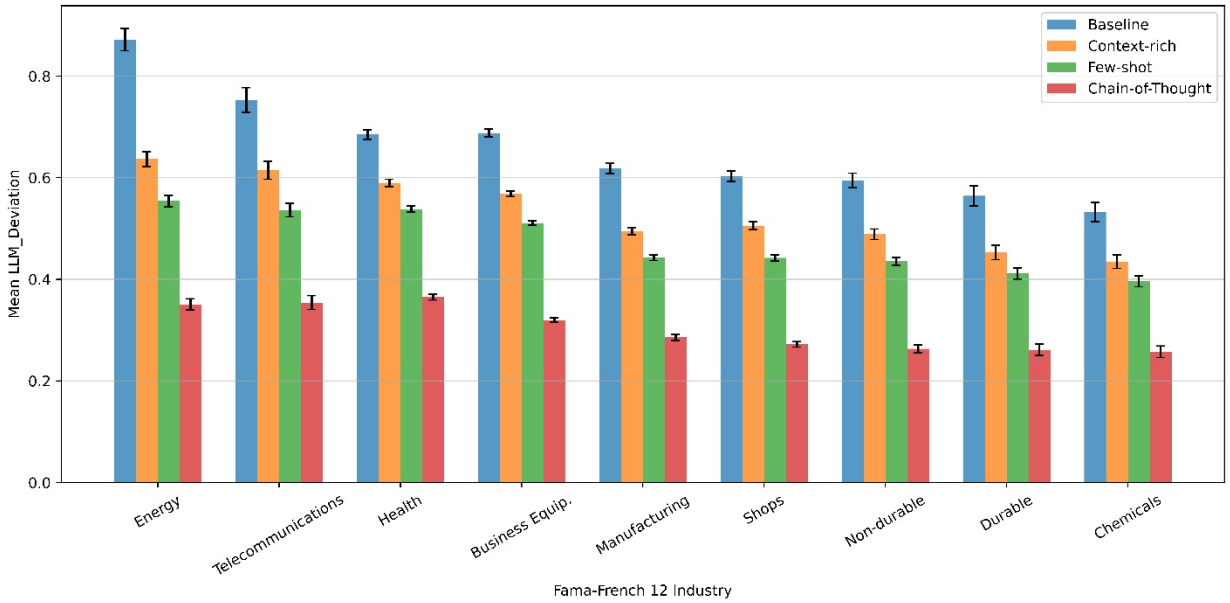
"F(r)iction in Machines: Accounting Hallucinations of Large Language Models"

Albert Di Wang
The University of Texas at Austin
albert.di.wang@utexas.edu

Figure OA1. Accounting Deviations across Industries, Calendar Years, and Financial Statement Items

Notes: This figure presents accounting deviations across Fama-French 12 industries (Panel A), calendar years (Panel B), and financial statement items (Panel C). The bars below are the average *LLM_Deviation* across requests that pass the input and output checks and are not refused by the model under the corresponding condition, along with their 95% confidence intervals. Higher values of mean *LLM_Deviation* indicate a higher degree of deviation.

Panel A. Accounting Deviations by Fama-French 12 Industries



Panel B. Accounting Deviations by Calendar Years

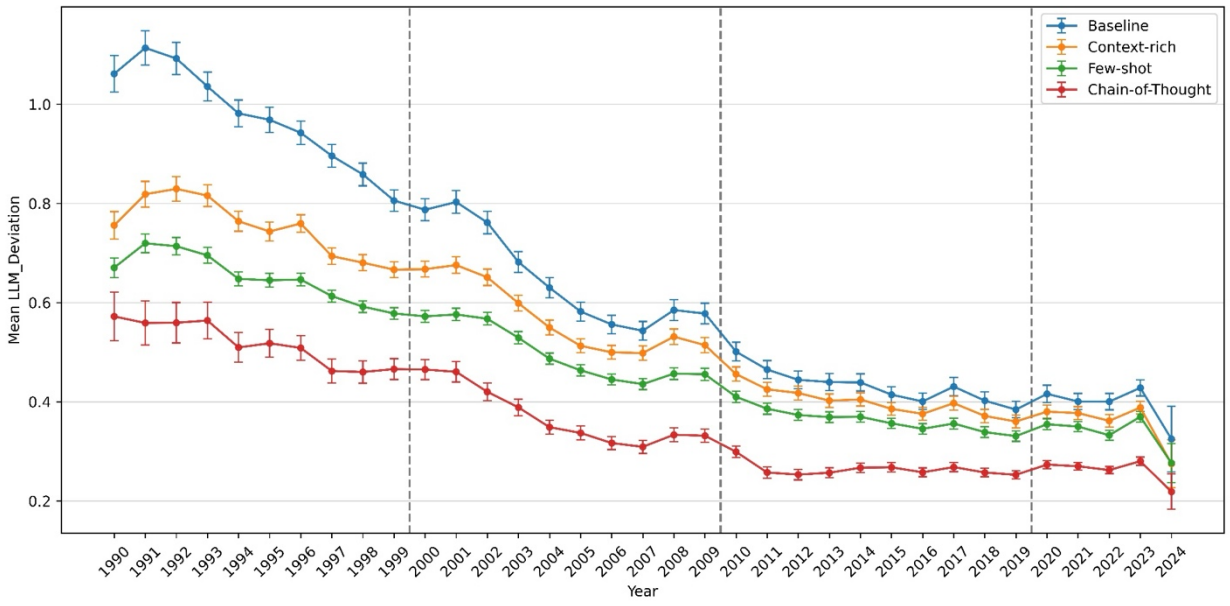


Figure OA1. Accounting Deviations across Industries, Calendar Years, and Financial Statement Items (continued)

Panel C. Accounting Deviations by Financial Statement Items

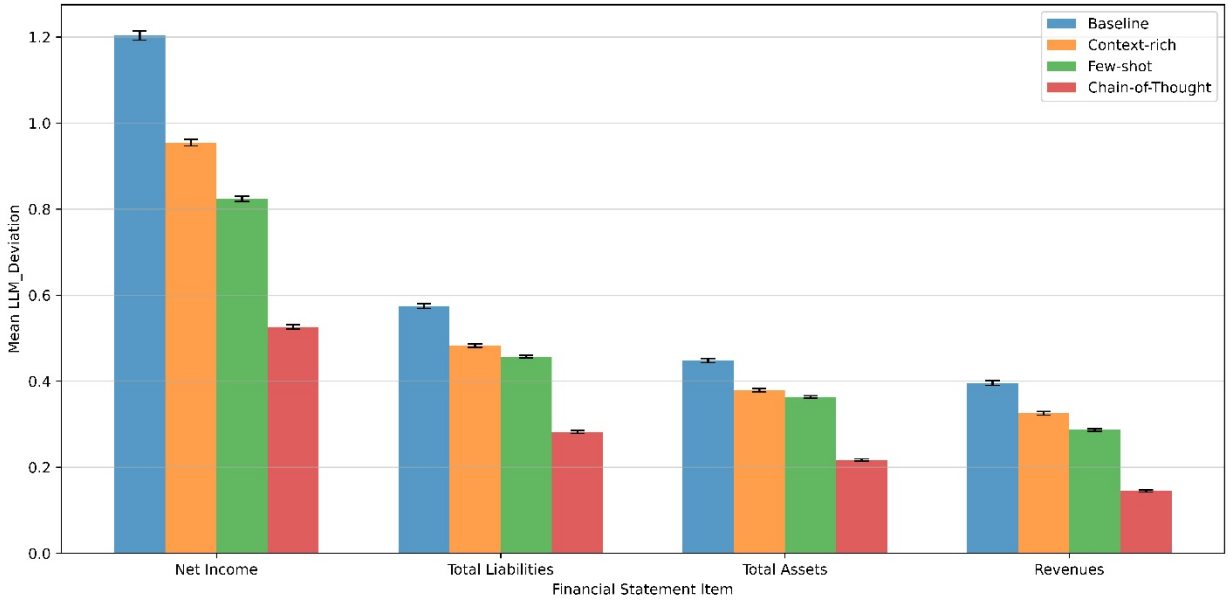
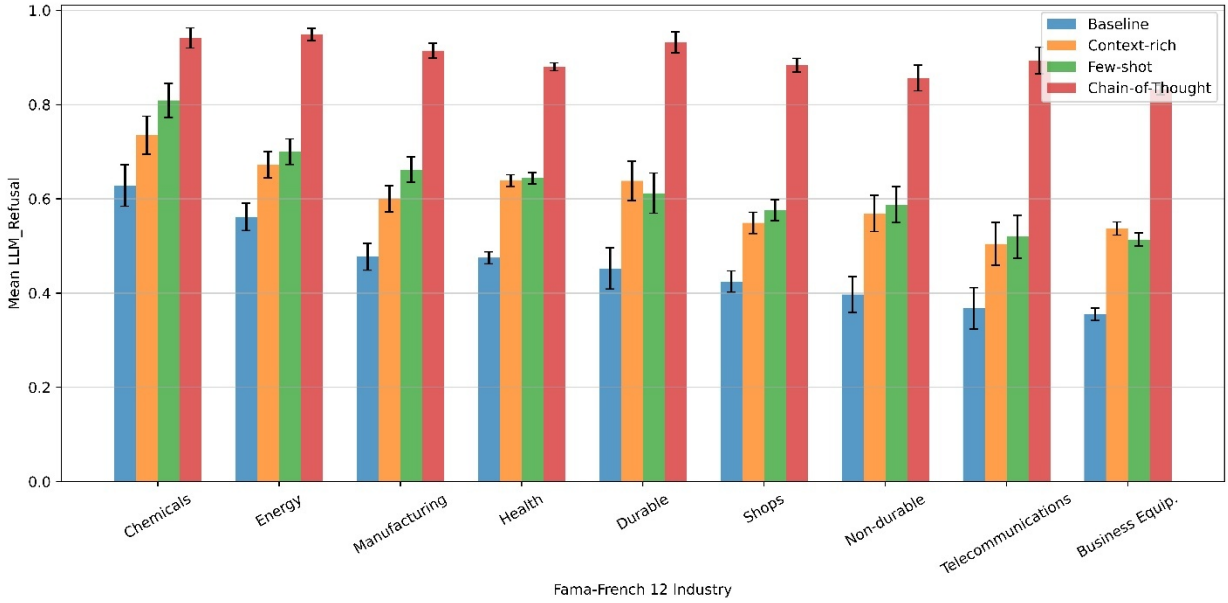


Figure OA2. Accounting Fabrications across Industries, Event Years, and Financial Statement Items

Notes: This figure presents accounting fabrications across Fama-French 12 industries (Panel A), event years (Panel B), and financial statement items (Panel C). The bars below are the average *LLM_Refusal* across requests under the corresponding condition, along with their 95% confidence intervals. Higher values of mean *LLM_Refusal* indicate a lower degree of fabrication. In Panel B, no-refusals (i.e., mean *LLM_Refusal*<1.0) on the left-hand side (non-grey shaded area) indicate accounting fabrications. The green and orange lines represent the mean *LLM_Refusal* for private firms (those that have never gone public) and public firms (those that completed IPOs prior to the observation window), respectively, and serve as benchmarks.

Panel A. Accounting Fabrications by Fama-French 12 Industries



Panel B. Accounting Fabrications by Event Year

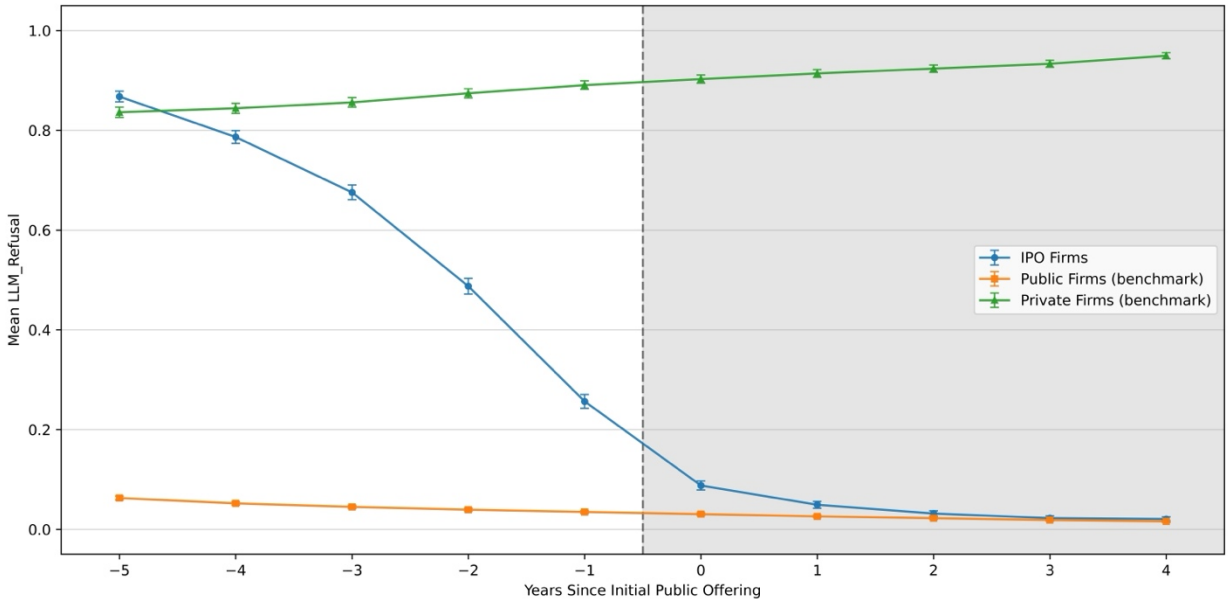


Figure OA2. Accounting Fabrications across Industries, Event Years, and Financial Statement Items (continued)

Panel C. Accounting Fabrication by Financial Statement Items

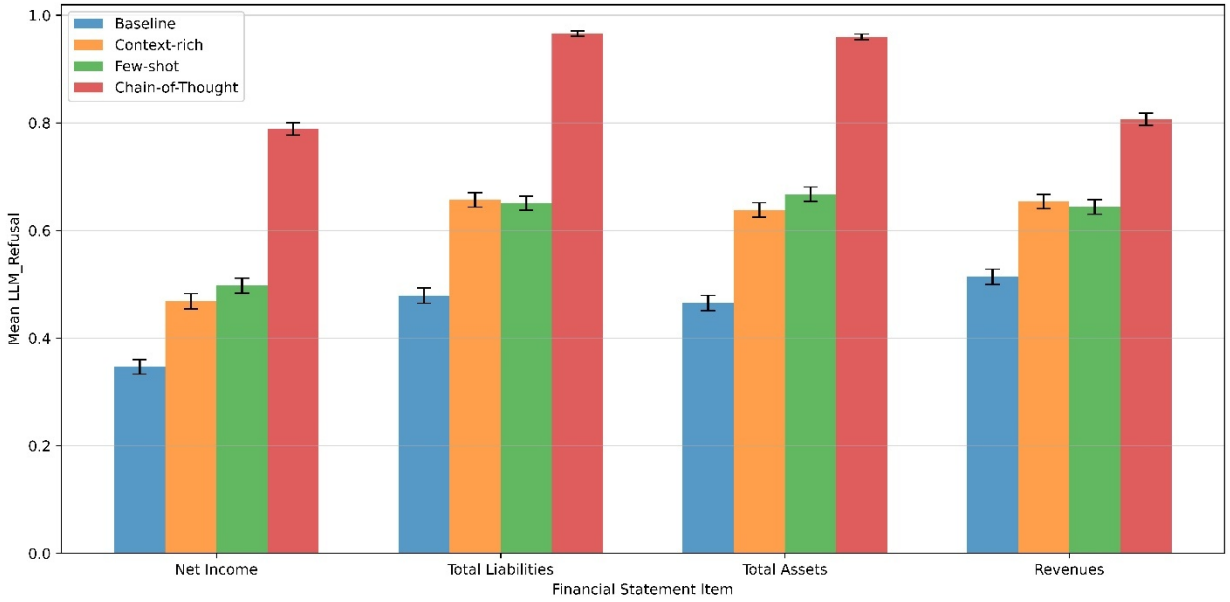
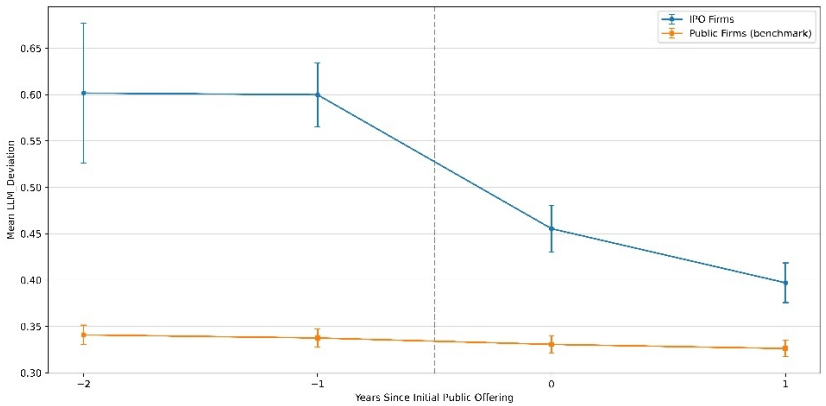


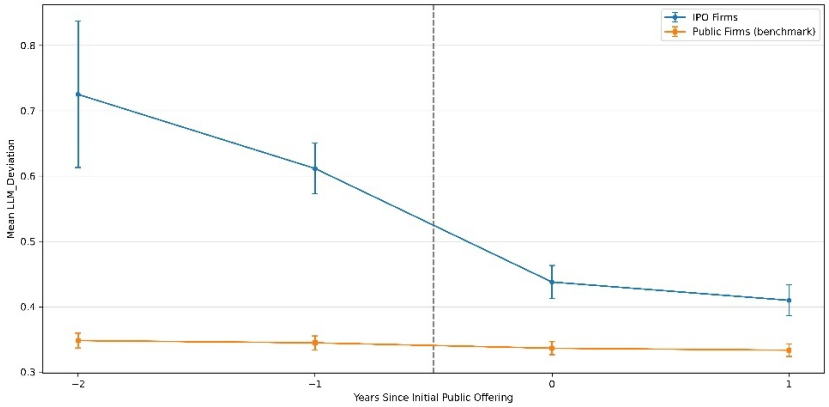
Figure OA3. Account Deviations Around Initial Public Offering

Notes: This figure presents the level of accounting deviations around IPO events, with 95% confidence intervals. In each panel, the blue line plots the mean *LLM_Deviation* across queries. The orange line represents the mean *LLM_Deviation* for public firms, serving as a benchmark. Private firms (those that have never gone public) are not used in this analysis, as the absence of true values prevents computation of accounting deviations. The analysis window spans four years centered on IPOs because only two years of pre-IPO accounting information are mandatorily disclosed for all firms. Both pre- and post-IPO *LLM_Deviation* are computed using true values parsed from XBRL tags to ensure comparability.

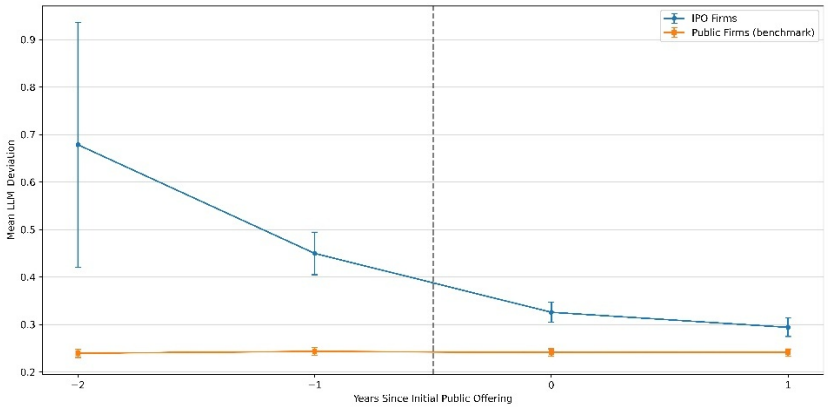
Panel A. Baseline Prompting



Panel B. Context-rich Prompting



Panel C. Chain-of-Thought Prompting



Panel D. Few-shot Prompting

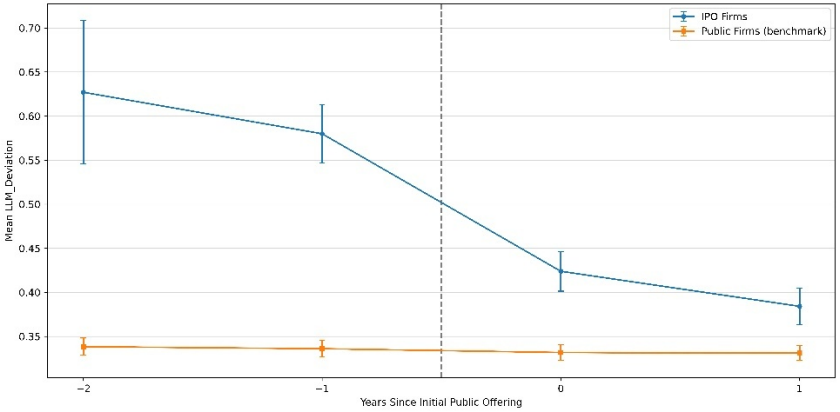


Table OA1. Summary Statistics of LLM Fabrication in the IPO Setting for Information-scarce Companies

Notes: This table reports summary statistics for the sample of accounting fabrications, focusing on information-scarce firms. The data used here is the pre-IPO sample in step 5c in Panel B of Table 1, but further restricted to emerging growth companies (EGCs) and special purpose acquisition companies (SPACs) in Panel A and B, respectively. Each observation requires querying the LLM for four financial statement items. "Unsupported Input" and "Unparsable Output" reflect the share of requests either rejected by the LLM due to unsupported inputs or producing outputs that do not conform to the expected format. The mean *LLM_Refusal* (absolute value) is the average *LLM_Refusal* across all requests passing the input and output checks. The mean *LLM_Refusal* (relative to public firms) is an adjusted measure specified in Section 3.2. It is calculated as the average of (pre-IPO refusal of IPO firms - post-IPO refusal of IPO firms) - (pre-IPO refusal of public firms - post-IPO refusal of public firms) across requests. The mean *LLM_Refusal* (relative to private firms) is constructed analogously, using private firms as the benchmark group. In all cases, higher values of *LLM_Refusal* suggest a lower degree of fabrication in the LLM's output.

Panel A. Accounting Fabrications Among EGCs

Prompting Strategy	**# LLM Request**	**Unsupported Input**	**Unparsable Output**	**Mean** *LLM_Refusal* (absolute value)	**Mean** *LLM_Refusal* (relative to public firms)	**Mean** *LLM_Refusal* (relative to private firms)
Baseline	460	0.00%	0.00%	0.51	0.47	0.64
Context-rich	460	0.00%	0.00%	0.68	0.58	0.65
Chain-of-Thought	460	0.00%	0.00%	0.95	0.36	0.50
Few-shot	460	0.00%	0.00%	0.67	0.57	0.66

Panel B. Accounting Fabrications Among SPACs

Prompting Strategy	**# LLM Request**	**Unsupported Input**	**Unparsable Output**	**Mean** *LLM_Refusal* (absolute value)	**Mean** *LLM_Refusal* (relative to public firms)	**Mean** *LLM_Refusal* (relative to private firms)
Baseline	1,080	0.00%	0.00%	0.79	0.65	0.82
Context-rich	1,080	0.00%	0.00%	0.92	0.68	0.76
Chain-of-Thought	1,080	0.00%	0.00%	0.99	0.38	0.53
Few-shot	1,080	0.00%	0.00%	0.94	0.67	0.76

Table OA2. Effect of IPOs on Accounting Deviations

Notes: This table presents the effect of IPO events on accounting deviations. The data used here is the cohort-firm-year sample from step 5a in Panel B of Table 1 where each observation requires querying the LLM for four financial statement items (46,290 cohort-firm-years * 4 items = 185,160 requests). The requests are further restricted to those that (1) a four-year window centered on IPOs, (2) pass the input and output checks, (3) are not refused by LLM, and (4) the corresponding XBRL value is not missing. The four-year window is chosen to align with the mandatory disclosure of two years of pre-IPO accounting information in registration statements for all firms. Both pre- and post-IPO *LLM_Deviation* measures are computed using true values parsed from XBRL filings to ensure comparability. Detailed variable definitions are provided in Appendix B. *t*-statistics are reported in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively

D.V. = *LLM_Deviation*	**(1)**	**(2)**	**(3)**	**(4)**
Prompting Strategy	Baseline	Context-rich	Chain-of-Thought	Few-shot
*Treated * Post*	-0.179***	-0.206***	-0.138***	-0.195***
	(-9.584)	(-9.691)	(-5.664)	(-11.306)
Cohort-Item FE	Yes	Yes	Yes	Yes
Cohort-Firm FE	Yes	Yes	Yes	Yes
Cohort-Year FE	Yes	Yes	Yes	Yes
Clustering	Cohort-Firm	Cohort-Firm	Cohort-Firm	Cohort-Firm
Observations	20,515	20,205	10,534	20,032
Adj. R-squared	0.356	0.351	0.335	0.336

Table OA3. Correlation between Accounting Deviations and Information Production & Dissemination: Prompting Strategies, Material Deviations, and Knowledge Deficits

Notes: This table presents regression results for the correlation between accounting deviations and the production and dissemination of accounting information. Detailed definitions of all variables are provided in Appendix B. *t*-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A. Prompting Strategies

D.V. = *LLM_Refusal*	**(1)**	**(2)**	**(3)**
Prompting Strategy	Baseline	Context-rich	Chain-of-Thought
MarketCap	-0.097***	-0.070***	-0.031***
	(-9.769)	(-10.957)	(-10.885)
FreqEarn	-0.002	-0.007*	-0.004*
	(-0.448)	(-2.031)	(-1.825)
Analysts	-0.078***	-0.047***	-0.021***
	(-7.890)	(-7.357)	(-6.422)
Controls, FEs, Clustering	All, Yes, Yes	All, Yes, Yes	All, Yes, Yes
Observations	336,112	309,709	135,719
Adj. R-squared	0.316	0.279	0.299

Panel B. Material Deviations

D.V. = *LLM_Material*	**(1)**	**(2)**	**(3)**	**(4)**
Prompting Strategy	Baseline	Context-rich	Chain-of-Thought	Few-shot
MarketCap	-0.013***	-0.014***	-0.024***	-0.010***
	(-7.351)	(-8.149)	(-10.685)	(-6.078)
FreqEarn	-0.010***	-0.009***	-0.009***	-0.009***
	(-4.892)	(-4.868)	(-3.827)	(-4.913)
Analysts	-0.000	0.001	-0.006	0.001
	(-0.186)	(0.599)	(-1.573)	(0.537)
Controls, FEs, Clustering	All, Yes, Yes	All, Yes, Yes	All, Yes, Yes	All, Yes, Yes
Observations	336,112	309,709	135,719	319,706
Adj. R-squared	0.203	0.206	0.302	0.201

Panel C. Knowledge Deficits

D.V. = *LLM_Refusal*	**(1)**	**(2)**	**(3)**	**(4)**
Prompting Strategy	Baseline	Context-rich	Chain-of-Thought	Few-shot
MarketCap	-0.006***	-0.009***	-0.032***	-0.009***
	(-3.163)	(-3.855)	(-10.094)	(-3.935)
FreqEarn	0.005**	0.013***	-0.011***	0.009***
	(2.251)	(3.454)	(-4.627)	(3.185)
Analysts	-0.017***	-0.031***	-0.006*	-0.023***
	(-3.792)	(-6.529)	(-1.874)	(-5.087)
Controls, FEs, Clustering	All, Yes, Yes	All, Yes, Yes	All, Yes, Yes	All, Yes, Yes
Observations	349,598	349,573	349,687	349,446
Adj. R-squared	0.358	0.558	0.601	0.505

Table OA4. Correlation between Accounting Fabrications and Information Production & Dissemination: Prompting Strategies and Adjusted Fabrications

Notes: This table presents regression results for the correlation between accounting fabrications and the production and dissemination of accounting information. Detailed definitions of all variables are provided in Appendix B. *t*-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. *LLM_RelativeRefusal* is the mean *LLM_Refusal* relative to public firms. It is an adjusted measure specified in Section 3.2 and is calculated as the average of (pre-IPO refusal of IPO firms - post-IPO refusal of IPO firms) - (pre-IPO refusal of public firms - post-IPO refusal of public firms) across requests.

Panel A. Prompting Strategies

D.V. = *LLM_Refusal*	**(1)**	**(2)**	**(3)**
Prompting Strategy	Baseline	Context-rich	Chain-of-Thought
MarketCap	0.021	0.004	-0.034***
	(1.572)	(0.419)	(-3.582)
FreqEarn	-0.044***	-0.050***	-0.001
	(-2.603)	(-3.065)	(-0.100)
Analysts	-0.065**	-0.049**	-0.002
	(-2.566)	(-2.362)	(-0.188)
Controls	All	All	All
Item FE	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Clustering	Industry	Industry	Industry
Observations	16,179	16,180	16,176
Adj. R-squared	0.357	0.385	0.277

Panel B. Adjusted Fabrications

D.V. = *LLM_RelativeRefusal*	**(1)**	**(2)**	**(3)**	**(4)**
Prompting Strategy	Baseline	Context-rich	Chain-of-Thought	Few-shot
MarketCap	0.017	0.004	0.074***	0.009
	(1.270)	(0.398)	(5.257)	(0.822)
FreqEarn	-0.046***	-0.050***	-0.043**	-0.074***
	(-2.710)	(-3.013)	(-2.038)	(-4.857)
Analysts	-0.046*	-0.024	0.010	-0.034
	(-1.709)	(-1.012)	(0.435)	(-1.217)
Controls	All	All	All	All
Item FE	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Clustering	Industry	Industry	Industry	Industry
Observations	16,179	16,180	16,176	16,178
Adj. R-squared	0.348	0.369	0.250	0.371

Table OA5. Effect Heterogeneities of LLM Knowledge Cutoff on Accounting Deviations

Notes: This table presents the heterogeneity of the effect of LLM knowledge cutoff on accounting deviations. The data used here is the model-firm-year sample from Panel D of Table 1, where each observation requires querying the LLM for four financial statement items (54,400 model-firm-years * 4 items = 217,600 requests). The requests are further restricted to those that pass the input and output checks for and are not refused by both LLMs. I further divide the sample into high and low subsamples if pre-cutoff *MarketCap*, *FreqEarn*, or *Analysts* is in the top and bottom 33%, respectively. The analysis for each subsample employs a generalized DiD regression. For triple-DiD estimate, I use Wald test to test the statistical significance of the cross-equation parameter equality for *Treated * Post* and present the differences and their corresponding significance in the last row. Detailed variable definitions are provided in Appendix B. *t*-statistics are reported in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

D.V. = *LLM_Deviation*	**(1)** High *MarketCap*	**(2)** Low *MarketCap*	**(3)** High *FreqEarn*	**(4)** Low *FreqEarn*	**(5)** High *Analysts*	**(6)** Low *Analysts*
*Treated * Post*	0.249***	0.653***	0.348***	0.426***	0.277***	0.502***
	(4.801)	(5.364)	(5.432)	(4.880)	(4.769)	(4.497)
MarketCap	-0.111***	-0.088***	-0.110***	-0.097***	-0.107***	-0.119***
	(-4.150)	(-3.754)	(-4.787)	(-4.782)	(-4.485)	(-4.566)
FreqEarn	-0.036***	-0.024	-0.041***	0.009	-0.040***	-0.050**
	(-2.866)	(-1.065)	(-3.089)	(0.259)	(-3.119)	(-2.112)
Analysts	-0.101**	-0.068**	-0.024	-0.086***	-0.120**	-0.081***
	(-2.183)	(-2.370)	(-0.759)	(-2.967)	(-2.441)	(-3.158)
Leverage	-0.139	-0.161**	-0.044	-0.176**	-0.221**	-0.248**
	(-1.557)	(-2.020)	(-0.565)	(-2.453)	(-2.334)	(-2.578)
ROA	-0.518***	0.384***	-0.047	0.284***	-0.191*	0.366***
	(-4.834)	(4.915)	(-0.475)	(3.583)	(-1.901)	(3.830)
MTB	0.000	-0.000	-0.001	-0.001	0.000	-0.002
	(0.325)	(-0.217)	(-0.748)	(-0.345)	(0.259)	(-0.616)
SalesGrowth	0.028	-0.032**	-0.007	-0.011	-0.000	-0.024
	(0.783)	(-1.986)	(-0.367)	(-0.808)	(-0.002)	(-1.176)
Institution	-0.100	-0.127	-0.097	-0.047	-0.028	-0.069
	(-1.513)	(-1.295)	(-1.509)	(-0.615)	(-0.423)	(-0.828)
ReturnVol	0.707	-0.054	-0.870	0.768	-1.762*	0.518
	(0.618)	(-0.070)	(-0.773)	(0.994)	(-1.773)	(0.634)
Model-Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Model-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Model-Item FE	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Model-Firm	Model-Firm	Model-Firm	Model-Firm	Model-Firm	Model-Firm
Observations	70,428	54,045	70,992	68,059	69,003	56,748
Adj. R-squared	0.091	0.234	0.162	0.205	0.107	0.260
Triple-DiD Estimate: *High-Low*	**-0.404*** **		-0.078		**-0.225***	