AI in Finance and Information Overload

Attila Balogh Antoine Didisheim Luciano A. Somoza Hanqing Tian

March 8, 2025

Abstract

Artificial intelligence is reshaping financial markets, yet the limits to its rationality remain underexplored. This paper documents information overload in Large Language Models applied to financial analysis. Using earnings forecasts from corporate calls and market reaction predictions from news, we show that predictive accuracy follows an inverted U-shaped pattern, where excessive context degrades performance. Larger LLMs mitigate this effect, increasing the optimal context length. Our findings highlight a fundamental limitation of AI-driven finance: more data is not always better, emphasizing the need for empirical calibration to determine the appropriate amount of context for each application.

JEL classification: G14, G17, C45, C55
Keywords: Large Language Models, AI in Finance, Information Overload

---

Balogh, Didisheim, Tian: Department of Finance, 198 Berkeley Street, University of Melbourne, Melbourne, VIC 3010, Australia. Somoza: ESSEC Business School, 10 Bd de l'Oise, 95000 Cergy, France. Send correspondence to balogh@unimelb.edu.au. This research was supported by the University of Melbourne's Research Computing Services and resources from the Australian Government's National Computational Infrastructure. All errors are our own.

1. Introduction

In his Nobel Prize lecture, Thaler argued that *"if errors are predictable, then departures from rational choice models can also be predictable. [...] It implies that, at least in principle, it would be possible to improve the explanatory power of economics by adding psychological realism"* (Thaler, 2018). He contended that the classical picture of economic rationality (Muth, 1961) must be tempered with the study of humans' cognitive biases and limitations of the agents who collectively shape equilibrium outcomes. However, as financial markets are rapidly integrating innovation in artificial intelligence (AI) (Aldasoro, Gambacorta, Korinek, Shreeti, and Stein, 2024), impacting trading (Cheng, Lin, and Zhao, 2024), innovation (Babina, Fedyk, He, and Hodson, 2024), and information processing (Abis and Veldkamp, 2024), biases embedded in AI systems are becoming increasingly relevant alongside human biases. Thus, economists have to expand their understanding of the limits to rationality limitations beyond those of humans.

We contribute to this effort by identifying a specific limitation in the rationality of Large Language Models (LLMs). When using an LLM for financial analysis, one has to provide the model with a prompt containing instructions and information. In the absence of cognitive biases, the rational choice is to provide as much information as the model can process. In this paper, we demonstrate that frictions in the algorithm can degrade performance due to information overload. We establish this by performing two tests: one in corporate finance focusing on earnings forecasts from earnings conference calls, and one in asset pricing focusing on the market reaction to the news. In both experiments, we reveal that beyond a certain threshold, additional context degrades predictive accuracy rather than improving it. The relationship has an inverted U-shape, with an optimal amount of context for each problem. We also find that larger LLM models can alleviate the adverse effects of excessive context, thus increasing the optimal context quantity.

The fact that humans are subject to information overload is well established in the literature (Schroder, Driver, and Streufert, 1967). Excessive information can lead to stress, anxiety, and confusion, ultimately impairing performance (Chernev, Böckenholt, and Goodman, 2015). Clearly, this channel is not relevant for LLMs, as they do not experience emotions. Instead, we propose two possible sources of friction that can explain the observed bounded rationality patterns.

The first stems from the layered architecture of neural networks shared by all LLMs. LLMs rely on fixed-dimensional representations, meaning that each layer must encode and transmit information within pre-set constraints. As context expands, encoding imperfections compound, limiting the model's ability to retain and process relevant information.

The second friction arises from LLM training procedures. The maximum input size of LLMs, known as the context window, is iteratively increased during training (Grattafiori, Dubey, Jauhri, Pandey, Kadian, Al-Dahle, Letman, Mathur, Schelten, Vaughan, et al., 2024; Hou, Zhao, Liu, Yang, Wang, Li, Luo, Lo, Grundy, and Wang, 2024). This procedure does not guarantee equal performance for short or long contexts within the context window. In practice, standard training protocols focus on ensuring retrieval within a large context window rather than optimizing the filtration of redundant or irrelevant information. As a result, the inverted-U shape shown in this paper may, in part, be caused by the way LLMs are trained to expand their context window.

If LLMs processed information without friction, incorporating additional earnings call transcript data should either improve forecasting accuracy or have no effect. Instead, we observe a clear inverted U-shaped relationship between context size and performance. Initially, more context enhances predictive accuracy. The forecasting accuracy starts at 52% when the model processes only the first 10% of the most recent earnings call, while the F1-score begins at 57%. Accuracy peaks at 55%, and the F1-score reaches 65%, when half of the latest call's transcript is included. Beyond this threshold, additional context becomes detrimental: accuracy and the F1-score decline when the entire transcript is provided, and both metrics drop further when the model receives all three prior earnings calls in their entirety. Regressions find that patterns in accuracy are consistent with basic economic intuition and the earlier findings by Kim, Muhn, and Nikolaev (2025) prior to the withdrawal of their working paper.

Next, we turn to asset pricing and use LLMs to predict market reactions to stock-specific news. To minimize noise, we focus on same-day news and stock returns in a nowcasting setup. We progressively increase the amount of contextual information by appending up to forty stock-specific news items in reverse chronological order, adding one older item at a time. As in the earnings call experiment, expanding the context initially enhances performance but eventually hinders it. While incorporating up to ten news items improves predictive accuracy, performance declines as more context is added. By the time forty news items are included, the model's predictions become no better than random chance.

Finally, we construct two long-short portfolios using a large dataset of Reuters news: one based solely on short, time-sensitive reports labeled as "Alerts", and another that supplements each alert with the most recent previous full news article about the same firm as additional context. To ensure computational feasibility, we follow Chen, Kelly, and Xiu (2024b) and use LLMs to embed the news—both with and without context—into vector representations. These embeddings serve as inputs to a predictive model of firm returns. The forecasted returns are then used as signals to construct equally weighted and value-weighted

long-short portfolios, with and without contextual information.

In line with the information overload hypothesis, the context-based strategy does not consistently outperform the no-context strategy. Instead, both exhibit similar out-of-sample Sharpe ratios, consistent with Chen et al. (2024b), ranging from 4.95 for the smallest LLM used in our experiment to 5.84 for the largest one. Notably, across all configurations, we find that as the size and quality of the LLM used for embedding increases, the relative performance of the context-based portfolio improves. The alpha of the context strategy relative to the alert-only strategy rises with model size, both in magnitude and statistical significance. Similarly, in a tangency portfolio constructed from the two strategies, the weight assigned to the context-based portfolio increases as the model size grows. These results highlight the significance of information overload when using LLMs for key asset pricing applications such as portfolio formation. They also demonstrate that the size and quality of the LLM play a crucial role in mitigating its negative effects. When the model is sufficiently advanced and the task relatively simple, providing the LLM with as much context as possible tends to be beneficial. However, ex-ante, an economic agent cannot determine whether the model is robust enough to handle extensive context effectively. Thus, our findings suggest that identifying the optimal amount of context requires empirical experimentation.

The literature on the applications of LLMs in financial economics is recent but expanding rapidly (see, e.g., Korinek, 2023, 2024; Eisfeldt and Schubert, 2024). We contribute to three main strands: the one using LLMs to forecast financial variables, the one trying to simulate human behavior, and the one investigating biases in LLMs.

There is extensive evidence on the value of artificial intelligence for forecasting in a financial setting (see, e.g., Giglio, Kelly, and Xiu, 2022; Chen et al., 2024b). Cao, Jiang, Wang, and Yang (2024) find that machine learning algorithms can surpass human analysts in predicting returns. Similarly, Lopez-Lira and Tang (2024) find that feeding news to ChatGPT and asking it to predict the next day's stock price return yields significant returns. Using accounting data, Kim et al. (2025) report that large language models outperform conventional analysts in forecasting earnings. We contribute to this literature by showing how both corporate finance and asset pricing LLM-based forecasts have an optimal amount of context beyond which their performance declines.

LLMs are also increasingly used in simulating agent behavior and expectations, with papers exploring LLM's human-like patterns and their deviations from human reasoning. Ashokkumar, Hewitt, Ghezae, and Willer (2024) test the predictive capabilities of LLMs concerning the outcomes of social science experiments, finding strong results. Conversely, Ludwig and Mullainathan (2024) exploit the difference between human cognition and LLMs to propose a systematic procedure for generating novel hypotheses about human behavior.

In this context, we emphasize that the rationality of LLMs is inherently bounded due to their susceptibility to information overload. This limitation is particularly relevant for those employing LLMs to model human behavior.

Other papers in the literature highlight the presence of systematic biases in LLMs. Bybee (2023) provides evidence that AI-generated economic forecasts reflect biases akin to those found in expert survey predictions. Fedyk, Kakhbod, Li, and Malmendier (2024) find that AI-driven economic analyses and projections tend to align with the perspectives of young, high-income males. Chen, Green, Gulen, and Zhou (2024a) show that LLM-generated forecasts overemphasize recent trends, much like human forecasters. In each of these instances, biases inherent in the training data provide a plausible explanation because LLM outputs largely mirror the patterns found in their training data. If the majority of financial content online is created by young men, this bias is likely to permeate LLM outputs. In theory, such bias could be mitigated by adjusting the LLM's training sample to offer a more balanced perspective on finance. By contrast, the bias documented in this paper is not easily attributed to the training sample. If, as we argue, the friction arises from imperfect layered encoding that leads to information overload, this LLM-induced bias becomes both inevitable and predictable. In this respect, our paper is most closely related to Chen, Didisheim, and Somoza (2025), who document that the conditional probabilities estimated by the models underlying LLMs contain both valuable economic information and systematic biases. Given AI's increasing integration into financial systems, this predictability makes LLM biases particularly relevant for economists.

2. LLM's Bounded Rationality

Neural networks are universal approximators (Hornik, Stinchcombe, and White, 1989): with sufficient data, a highly parameterized network can approximate any function arbitrarily well. This might suggest that neural networks, and LLMs in particular, have a near-perfect ability to process and filter out redundant or irrelevant information, implying that too much information should almost always be preferable to too little information. This intuition is wrong.

This section examines the tradeoff LLMs face between receiving too little and too much information, linking this pattern to the economic concept of bounded rationality (Simon, 1990). We then explore two key sources of this limitation. First, we discuss how informational bottlenecks within LLMs, shaped by imperfect encoding processes, constrain their ability to process expanding context effectively. Second, we consider how the method used to train LLMs to handle large inputs, particularly the expansion of the context window, may

contribute to the observed performance patterns. We stress that the mechanisms presented in Sections 2.2 and 2.3 propose elements of an answer to help rationalize and understand our empirical results. However, given the opacity and complexity of state-of-the-art LLMs, formally establishing that these mechanisms alone drive our empirical results is beyond the scope of this paper.

2.1. Economic Intuition

Consider first a rational agent whose task is to analyze the price impact news about a firm. She can either access the news alone, or the news and an arbitrarily large amount of context: older news about the same firm, regulatory filings, or relevant macroeconomic data. How would the expected accuracy of her forecast change as a function of the amount of context? Under the assumption of classical complete economic rationality, the marginal impact of additional context should never be negative and converge to zero, as shown by the solid line in Figure 1. As the agent receives additional context, most of it will be redundant, outdated, or provide only a marginal value beyond the information already available.

However, in practice, this assumption of classical economic rationality does not hold. Economic research has long understood that economic rationality is best understood as bounded by cognitive limitations and biases (see, e.g., Simon, 1955; Stigler, 1961; Radner, 1972; Kahneman and Tversky, 1979).

Specifically, behavioral economics and psychological research highlight that as the volume of available information increases, decision-making does not necessarily improve monotonically but may instead deteriorate due to informational overload, as shown in Figure A1 (Schroder et al., 1967; Eppler and Mengis, 2004). This bound creates an inverted U-shaped relationship between context and performance: an initial benefit from adding context is followed by a decline once processing constraints are exceeded, as shown by the dashed green line in Figure 1.

In Sections 4, 5, and 5.1, we show empirical applications in corporate finance and asset pricing where the LLM's relationship to context is reminiscent of the green-dotted line in Figure 1. However, while empirical patterns of the LLM's output may appear similar to human psychological biases, the analogy has limitations.

With human economic agents, information overload can be explained from an emotional and cognitive perspective. Agents overwhelmed by surplus information experience impaired performance due to stress, anxiety, and confusion (Inbar, Pizarro, and Bloom, 2009; Chernev et al., 2015). Neurological studies confirm that excessive information activates brain regions responsible for decision-making and emotional regulation, exacerbating the effects of overload

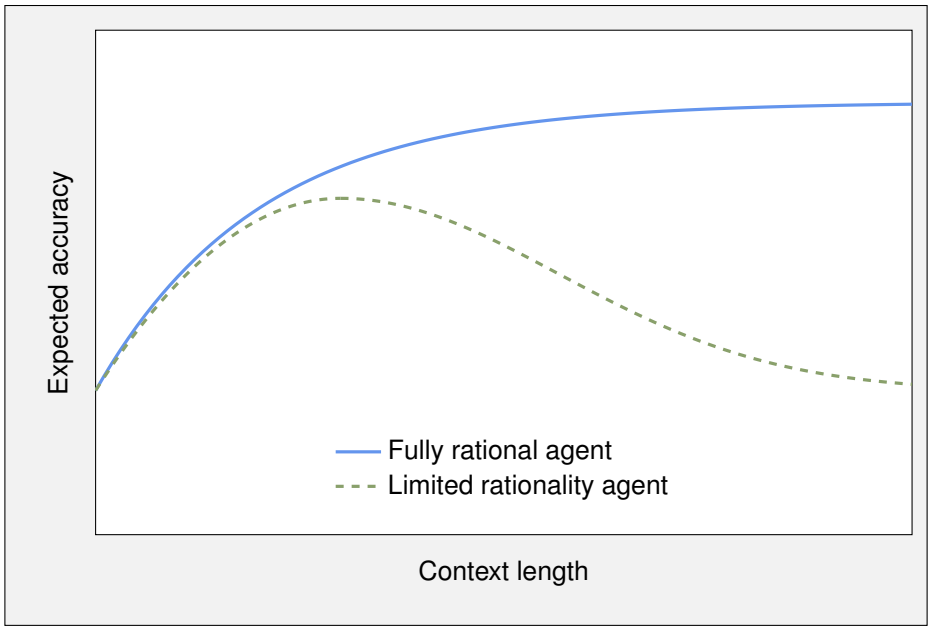


Fig. 1. Performance as a Function of Context

This figure illustrates the relationship between the amount of context and expected accuracy for two types of decision-makers. The solid blue curve represents a fully rational agent, whose performance improves monotonically as additional context is provided. The dashed green curve depicts a limited rationality agent, whose performance initially improves with more context but eventually declines due to information overload. The divergence in these curves highlights the impact of cognitive and informational constraints on decision-making, which apply not only to human agents but also to LLMs constrained by encoding limitations and processing inefficiencies.

(Dimoka, 2012). Unlike human decision-makers, LLMs are stochastic mathematical models. Their limitations do not stem from imperfect emotional regulation and, in general, can't be understood within the behavioral-economic framework. These biases are best understood as pseudo-behavioral biases—patterns that resemble human cognitive errors but arise from entirely different underlying mechanisms. While LLMs are trained to mimic human reasoning well enough to make psychological comparisons tempting, their actual decision-making processes are fundamentally different. Many of their biases and mistakes are distinctly non-human, driven by computational and statistical artifacts rather than cognitive limitations. A prime example is hallucination, where LLMs generate entirely false yet confidently stated responses—an error that has no direct counterpart in human cognition. Thus, to accurately diagnose the limitations of LLMs for applications in economics and finance, we must move beyond behavioral analogies and instead examine the mathematical and computational

constraints that define their performance. Unlike humans, who struggle with information overload due to emotional and cognitive fatigue, LLMs face bottlenecks imposed by their architecture, encoding methods, and training procedures.

Other recent research has identified pseudo-behavioral biases in AI-driven economic forecasting. Bybee (2023) finds that AI-generated forecasts exhibit errors similar to those found in expert surveys. Fedyk et al. (2024) show that AI economic analyses tend to reflect the perspectives of young, high-income males, mirroring demographic biases in financial literature. Chen et al. (2024a) demonstrate that LLM forecasts, much like human analysts, overweight recent trends. In these cases, the source of bias is relatively clear: it stems from patterns embedded in the training data.

By contrast, the information overload we document in this paper does not have an obvious link to training data. Instead, we argue that it arises from fundamental computational constraints within LLMs. To explain this phenomenon, we propose two key mechanisms. First, the limited size and quality of information encoding restricts an LLM's ability to process increasing amounts of context effectively. Second, the methods used to train LLMs to handle large context windows create structural inefficiencies that lead to declining performance at high context lengths. We examine both mechanisms next.

2.2. Information Bottlenecks in Neural Networks

To address the engineering challenges across different subfields of machine learning, such asvision, audio, text, engineers have developed and refined a diverse set of techniques. As a result, neural networks differ significantly in architecture, training procedures, and parametrization. This growing complexity has increased both the barrier to entry and the overall opacity of the those models.

Despite this diversity, a common feature of most neural networks is their layer structure. Each layer, except for the first, receives input from one or more preceding layers. Symmetrically, each layer, except for the last, passes as output some or all the input to the next layer. Transformers models used in most state-of-the-arts LLMs are no exception. And with transformers just as with every other models, each layer is of some large but fixed dimension. Hence, each layer functions as an implicit informational bottleneck.

To illustrate this constraint, consider a simplified feedforward neural network. While the architecture of state-of-the-art LLMs is orders of magnitude more complex, the core intuition of information processing through successive layers remains applicable. Assume a network

with three hidden layers, defined as follows:

$$\begin{aligned}
\text{Input:} \quad & \mathbf{x} \in \mathbb{R}^{d_{\text{in}}} \\
\text{Hidden Layer 1:} \quad & \mathbf{h}_1 = \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1), \quad \mathbf{W}_1 \in \mathbb{R}^{k \times d_{\text{in}}}, \mathbf{b}_1 \in \mathbb{R}^k \\
\text{Hidden Layer 2:} \quad & \mathbf{h}_2 = \sigma(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2), \quad \mathbf{W}_2 \in \mathbb{R}^{m \times k}, \mathbf{b}_2 \in \mathbb{R}^m \\
\text{Hidden Layer 3:} \quad & \mathbf{h}_3 = \sigma(\mathbf{W}_3\mathbf{h}_2 + \mathbf{b}_3), \quad \mathbf{W}_3 \in \mathbb{R}^{n \times m}, \mathbf{b}_3 \in \mathbb{R}^n \\
\text{Output:} \quad & \hat{\mathbf{y}} = \mathbf{W}_4\mathbf{h}_3 + \mathbf{b}_4, \quad \mathbf{W}_4 \in \mathbb{R}^{d_{\text{out}} \times n}, \mathbf{b}_4 \in \mathbb{R}^{d_{\text{out}}}
\end{aligned}$$

where $\mathbf{x}$ is the input vector of dimension d_{in}, $\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3$ are the output of each hidden layer, and $\hat{\mathbf{y}}$ is the output vector of dimension d_{out}. The weight matrices $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{W}_4$ and bias vectors $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{b}_4$ are the learnable parameters of the network, with dimensions specified in each layer. The activation function $\sigma(\cdot)$ is applied element-wise and is typically chosen as a nonlinear function such as the ReLU or sigmoid.

When making a prediction with this model, the original data $\mathbf{x}$ flows through the three layers to finally be projected as the output $\hat{\mathbf{y}}$ by the final matrix of trainable parameters $\mathbf{W}_4$. While going through this process, each of the hidden layers play iteratively the role of informational bottleneck while the previous parts acts as an implicit encoder.[1] In this example, we have:

1. **First informational bottleneck:** When the model makes a prediction $\hat{\mathbf{y}}$, the information flows through. When passing through layer 1, the information contained in the vector $\mathbf{x}$ is encoded into a space of dimension k. This fixed k is the current informational bottleneck, and the network's task up to this point is to use this limited informational space to keep all relevant information, discard irrelevant ones, and format it in a way that is optimal for the second layer.
   The quality of this projection depends at this stage on the trainable parameters $\mathbf{W}_1$ and $\mathbf{b}_1$ and their interaction with the activation function $\sigma(\cdot)$. Jointly, these parameters and the activation functions act as the implicit encoder.
2. **Second informational bottleneck:** When the information arrives in the second layer, the informational bottleneck is now the dimension m. Just as before, the role of the model is to preserve all relevant information, discard useless information, and process data so that it can be optimally encoded into the third hidden layer. The

[1] It's important to note that the encoder-bottleneck patterns above differ from the ones sometimes discussed in the machine learning literature. Some models have an explicit encoder-decoder structure where the first half of the network processes the data to produce a layer, which is usually smaller. While those models explicitly model the mechanism we explain in this section, we argue that this same mechanism is implicitly present in every machine learning model, including transformers.

parameters defining this encoding of information are now the trainable parameters $\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2$ and the activation functions $\sigma(\cdot)$.

3. **Third informational bottleneck:** The dimension n then bounds the encoding space. The parameters controlling the optimal encoding are now $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{b}_1, \mathbf{b}_2$ and $\mathbf{b}_3$. The only difference with previous layers is that the role of this encoding is now to prepare for the final projection into the output $\hat{\mathbf{y}} = \mathbf{W}_4\mathbf{h}_3 + \mathbf{b}_4$.

When choosing between adding context or not to help the LLMs make optimal economic decisions, practitioners and academics often take the model already trained (the parameters $\mathbf{W}_i$ and $\mathbf{b}_i$ are already estimated), hence, their choice is limited to the definition of $\mathbf{x}$. Specifically, an agent with some degree of economic expertise using an LLM to make an economic decision has the choice between using a very curated input $\underline{\mathbf{x}}$ with what she believes are the most relevant and essential bits of context, or a larger $\overline{\mathbf{x}}$. This larger $\mathbf{x}$ contains more context that may be relevant but may also be redundant or useless.[2]

There is always some degree of uncertainty on whether or not the additional information contained in $\overline{\mathbf{x}}$ and absent in $\underline{\mathbf{x}}$ has economic relevance. For example, when predicting future earnings with earnings calls, whether or not to include the minutes of the previous meeting into $\mathbf{x}$ is a non-trivial decision. On the one hand, it's obvious that most economically relevant information is contained in the latest information. On the other hand, contrasting the most recent management announcement to their previous view can be valuable to a skilled analyst. Hence, if one believes that the LLM is always capable of either ignoring or processing additional information, adding more information will always be beneficial: $\overline{\mathbf{x}} >> \underline{\mathbf{x}}$.

In practice, the decision of whether or not to include additional information is a trade-off between the expert's capacity to not curate the input $\underline{\mathbf{x}}$ and the LLM's capacity to process redundancies in $\overline{\mathbf{x}}$.

This capacity of an LLM to process redundant information will, in general, be correlated with the network's size. In our simplified three-layer network example, as the dimensions of the hidden layers k, m, and n increase, the size of the bottleneck grows at the same time as the dimensions of the parameters $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$ and $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$.[3] Hence, model sizes both increase the quality of the implicit encoder and decrease the constraints of a fixed dimension

---

[2] Numerically, the input's dimension is fixed $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$, by choosing a smaller or a larger text as input to an LLM one can understand this as making $\mathbf{x}$ more or less sparse.

[3] Machine learning models tend to exhibit a strong virtue of complexity, as demonstrated in finance by Kelly, Malamud, and Zhou (2024); Didisheim, Ke, Kelly, and Malamud (2024). Larger models tend to outperform smaller ones, even if only marginally and sometimes at the expense of high computational cost (see, e.g., DeepSeek-AI, Liu, Feng, Xue, Wang, Wu, Lu, Zhao, Deng, Zhang, Ruan, et al., 2025).

for each of the information bottlenecks.

The common intuition of "more context is always better" is therefore true only for a sufficiently good LLM, and this often translates into "large enough." The definition of "good enough" will vary significantly across economic tasks. While it's safe to assume that the difficulty of economic tasks will allow smaller or weaker LLMs to be "good enough," human intuition is often a poor guide to task difficulties for an LLM. For instance, earlier models like GPT-3.5 that were capable of processing complex ideas were not capable of identifying the number of times "R" appears in the word "strawberry." This particular failure was linked to other aspects of LLMs' nature than the one described in this section, but it highlights the unpredictability of what tasks are complex or trivial to LLMs.

Hence, an economic agent deciding between $\underline{\mathbf{x}}$ and $\overline{\mathbf{x}}$ will need to complement economic intuition with trial and error to gauge the optimal trade-off between her capacity to curate $\mathbf{x}$ and the LLM's capacity to process irrelevant or redundant information.

2.3. The Hidden Costs of Expanding Context Windows

LLMs' input sizes are bounded by what is commonly referred to as the context window. Context windows bound the maximum input measured in character units called tokens. Those tokens can be single characters like "A" or "5," or commonly used and meaningful fractions of words like "ing", or even complete words. With most current models the average English word is composed of 3 to 4 tokens.

Historically, early LLMs had a fairly small context window. GPT-3 had 2,048 tokens, early models of GPT-4 had 8,192 and 32,768 tokens. As the technology progressed, context windows expanded significantly. For instance, OpenAI's GPT-4 Turbo offers a context window of up to 128,000 tokens. Similarly, Anthropic's Claude 3 Opus, their flagship model, provides a context window of 200,000 tokens. The only exception is Alphabet's Gemini 1.5 model, which boasts an impressive 1 million token context window.

The existence of a fixed context window per model may reinforce the false intuition that LLMs are perfectly able to process context when the context fits into the context window. To help deconstruct this intuition, we need to understand how LLM's expand their context window.

Most state-of-the-art models are both commercial and closed source, leaving the details of their training procedures, such as those for OpenAI's GPT-4o or Alphabet's Gemini models, largely unknown. However, Meta has adopted an open-source approach, publishing detailed methodology papers along with the parameters of its LLaMA models. Since the largest of these models (LLaMA3 405B) performs comparably with early versions of GPT-4

Grattafiori et al. (2024), these papers offer invaluable insights into the workings of state-of-the-art commercial models.

Before discussing the specifics of LLaMA's context windows, it's important to understand that for all LLMs, the model's complexity and the size of the data prevent any global solutions to estimate the model parameters. Instead, the LLMs are "trained" through some variation of stochastic gradient descent. During this process, the LLM is "fed" comparatively small batches of data iteratively, updating its parameters every time to marginally improve the model's fit over the batch. The process is repeated until some arbitrary measure of convergence is reached. Theory provides very little guarantee of convergence toward a global optimum for the models, and it's reasonable to assume that most LLMs are at best trained until convergence to some local minimum.

When considering context windows, this process interacts with the fact that large context windows are not used throughout the whole training process. Indeed Grattafiori et al. (2024) explains that "[Meta's team] pre-train a model with 405B parameters on 15.6[trillion] tokens using a context window of 8K tokens." For reference, the total training sample is 15 trillion tokens. They expand on the training of the context window explaining that "In the final stages of pre-training, we train on long sequences to support context windows of up to 128K tokens. We do not train on long sequences earlier because the compute in self-attention layers grows quadratically in the sequence length. We increase the supported context length in increments, pre-training until the model has successfully adapted to the increased context length." Note that this quadratic expense in computational cost is a limit that is likely to affect all AI companies, and thus it's reasonable to assume that similar techniques are used by other LLM providers. Finally, they explain that they "assess successful adaptation by measuring whether (1) model performance on short-context evaluations has recovered completely and (2) the model perfectly solves needle in a haystack tasks up to that length."

The specifics of these training procedures are essential to interpreting the results presented in this paper. First, models trained to handle large context iteratively are likely to perform better on shorter contexts than extended ones. The first 8,000 tokens of the context window are always utilized throughout training, whereas the remainder is introduced only in the second half.

Second, a successful adaptation to a large context window is only defined as one that does not lose predictive power on shorter context and solves needle-in-a-haystack tasks. Those needle-in-a-haystack tasks consist of showing the LLM is able to retrieve information in the long context when prompted to ask for it. Neither of those conditions measures the model's capacity to disregard context when it's redundant or harmful, or in more general terms: to keep full rationality with larger context.

3. Data

3.1. Earning Calls

We obtain earnings call transcripts from Thomson Reuters, covering the period from 2004 to 2020. The dataset includes 70,408 earnings calls from 2,876 unique firms, with each observation corresponding to a firm's quarterly earnings announcement. To complement this dataset, we collect firms' Earnings Per Share (EPS) from Compustat, allowing us to analyze changes in earnings performance over time. Table 1 presents summary statistics for the full sample and each individual year in our dataset. The first two columns report the number of unique firms and earnings calls per year. The next five columns describe the length of earnings call transcripts, measured in number of characters, with the mean, standard deviation, minimum, median, and maximum transcript length reported. The final column provides the percentage of firm-quarters in which EPS increased relative to the previous quarter.

3.2. Financial News

We use news text data from the Thomson Reuters Real-time News Feed, covering the period from January 1996 to December 2022, with a focus on U.S. firms. The data prior to 2004 is used exclusively as the initial training sample in the analysis described in the nowcasting section. For all other experiments, we restrict the dataset to financial news published after 2004.

To ensure data quality, we apply standard preprocessing filters to remove irrelevant or extreme-length articles, following Chen et al. (2024b). Articles with fewer than 100 characters or more than 100,000 characters are excluded, as are highly redundant texts based on cosine similarity measures. To maintain relevance, we limit the sample to news items that mention a single firm, identified through ticker mentions, and for which corresponding three-day close-to-close returns are available. Reuters provides two types of news: standard articles and brief alerts. Articles include both a headline and a detailed narrative, offering comprehensive coverage of company-related developments. Alerts, in contrast, are designed for immediate updates on breaking events and contain only a headline without additional content. In this context, alerts serve as a natural baseline representing "no context" news, while articles provide a structured way to add context to each alert.

Table 2 presents summary statistics for the full sample and individual years. The first three columns report the number of unique firms, alerts, and articles. The next five columns summarize article length, reporting the mean, standard deviation, minimum, median, and

Table 1: Earnings Call Transcripts and EPS Data

This table reports summary statistics for quarterly earnings call transcripts obtained from Thomson Reuters from 2004 to 2020, complemented with Earnings Per Share (EPS) data from Compustat. The first two columns present the number of unique firms and transcripts in each year. The following columns summarize transcript length (in characters) by reporting the mean, standard deviation, minimum, median, and maximum values. The final column indicates the percentage of firm-quarters where EPS increased relative to the previous quarter. Statistics are provided for both the full sample and individual years.

			Character length					
	Firms	Transcripts	Mean	STD	Min	Median	Median	%EPS Gr.
2004	996	2,446	45,737	13,422	10,870	45,714	96,510	48.16
2005	1,153	3,834	45,804	12,948	12,630	46,155	97,529	51.28
2006	1,155	3,825	45,906	12,827	10,310	46,010	101,440	51.56
2007	1,208	3,952	45,495	12,631	12,170	46,018	93,780	50.18
2008	1,270	4,236	45,871	12,705	11,581	46,780	95,110	47.97
2009	1,247	4,357	45,499	13,389	8,350	46,010	136,320	54.30
2010	1,331	4,672	44,770	13,011	11,300	44,820	93,120	52.98
2011	1,394	4,706	44,781	13,165	10,386	45,256	101,400	49.77
2012	1,375	4,623	45,400	12,830	5,370	45,900	104,330	50.18
2013	1,279	4,451	45,451	12,683	9,160	46,030	102,421	50.91
2014	1,259	4,234	45,269	12,575	9,130	45,798	102,800	50.73
2015	1,228	4,148	45,186	12,548	12,210	45,866	101,380	50.02
2016	1,199	4,125	45,624	12,462	11,780	46,750	100,060	52.82
2017	1,318	4,437	44,608	12,934	9,870	45,160	96,490	53.19
2018	1,468	5,033	43,388	13,039	10,340	43,470	107,327	50.88
2019	1,457	4,832	43,757	12,780	10,890	44,225	121,060	48.68
2020	1,210	2,497	45,841	13,219	8,530	46,699	134,194	60.67
Total	2,876	70,408	45,123	12,899	5,370	45,650	136,320	51.26

maximum number of characters. The final columns show the percentage of observations where the firm's return is positive.

4. The Limits of Context in Earnings Forecasting with LLMs

In this section, we examine the relationship between the amount of context provided to an LLM and the quality of its predictions in a corporate finance setting. Our analysis builds on the approach of Kim et al. (2025), focusing on earnings conference calls. We use earnings call transcripts from Thomson Reuters, covering the period from 2004 to 2022, and merge them with quarterly fundamental data from Compustat to obtain corresponding Earnings Per Share figures. To assess how context affects predictive accuracy, we segment the transcripts of the three most recent earnings calls into thirty equal sections. The model, Llama Small, is then queried using a structured prompt designed to predict the direction of next quarter's EPS. The experimental setup ensures a controlled introduction of incremental context, allowing us to isolate its impact on forecasting accuracy.

System Prompt

`<|begin_of_text|><|start_header_id|>system<|end_header_id|>`

You are a financial statement analyst responsible for predicting the next quarter earnings per share (EPS) based on the given earnings call transcripts.

You should answer in a single letter (with no further explanation) `'P'` (indicating an increase in EPS next quarter) or `'N'` (indicating a decrease in EPS next quarter). `<|eot_id|>`

User Prompt

`<|start_header_id|>user<|end_header_id|>`

Below is the earnings call transcript for the current quarter: `{calls_context}`

Below are the earnings call transcripts from previous quarters starting from the most recent: `{calls_context_hist}`

Please answer `'P'` or `'N'` below in a single letter: `<|eot_id|>`

Assistant Prompt

`<|start_header_id|>assistant<|end_header_id|>`

Table 2: Summary Statistics for Financial News and Market Reactions

This table presents summary statistics for financial news articles and alerts from the Thomson Reuters Real-time News Feed, covering the period from 1996 to 2022. The first three columns report the number of unique firms, alerts, and articles in the sample. The next five columns summarize the length of articles (measured in characters), including the mean, standard deviation, minimum, median, and maximum values. The final column reports the percentage of instances where the firm's stock return was positive following the news. Statistics are provided for both the full sample and individual years.

				Article length					
	Firms	Alerts	Articles	Mean	STD	Min	Median	Median	$r_t > 0$ (%)
1996	6,241	51,223	89,991	826	700	100	606	11,211	47.67
1997	6,674	55,712	94,113	1,011	911	100	708	43,329	47.97
1998	6,564	50,355	81,552	1,375	1,156	100	1,022	31,980	47.79
1999	5,988	40,721	68,100	1,622	1,233	100	1,302	18,760	48.98
2000	4,834	37,301	59,126	1,841	1,296	101	1,451	27,648	47.69
2001	4,178	45,507	55,239	1,997	1,397	100	1,548	25,972	47.82
2002	3,555	46,470	53,279	2,038	1,390	100	1,657	23,909	47.49
2003	3,161	62,582	53,499	1,729	1,292	100	1,308	13,547	49.91
2004	4,011	81,770	62,590	1,600	1,326	100	1,155	18,788	49.61
2005	3,985	124,578	77,390	1,313	1,284	100	799	24,163	50.18
2006	4,345	186,551	89,665	1,249	1,164	100	798	17,026	50.79
2007	4,390	191,984	93,903	1,261	1,322	100	744	37,555	50.54
2008	4,235	202,239	101,181	1,139	1,279	100	562	49,845	47.34
2009	4,001	208,600	103,991	1,235	1,331	100	673	29,170	51.60
2010	3,940	224,833	98,367	1,292	1,368	100	705	22,816	51.60
2011	3,819	213,128	85,154	1,387	1,428	100	838	25,504	49.89
2012	3,711	194,956	81,654	1,347	1,767	105	653	34,736	51.27
2013	3,727	186,815	74,279	1,237	1,833	105	688	59,735	52.95
2014	3,845	225,096	68,469	1,225	1,866	106	690	59,227	51.89
2015	3,924	290,874	66,262	1,048	1,342	108	647	50,447	50.59
2016	3,833	263,969	60,815	1,054	1,374	100	639	74,163	51.02
2017	3,793	260,771	60,067	1,022	1,162	112	665	34,239	51.30
2018	3,806	251,373	64,254	1,078	1,559	100	660	54,669	51.24
2019	3,842	274,770	70,502	1,073	1,508	102	685	53,616	52.42
2020	4,000	286,400	69,486	1,052	1,334	100	693	35,552	52.48
2021	4,509	260,815	85,896	1,127	1,070	100	888	34,308	51.41
2022	4,354	247,858	81,026	1,117	1,004	100	902	32,670	49.27
Total	14,700	4,567,251	2,049,850	1,278	1,362	100	794	74,163	50.78

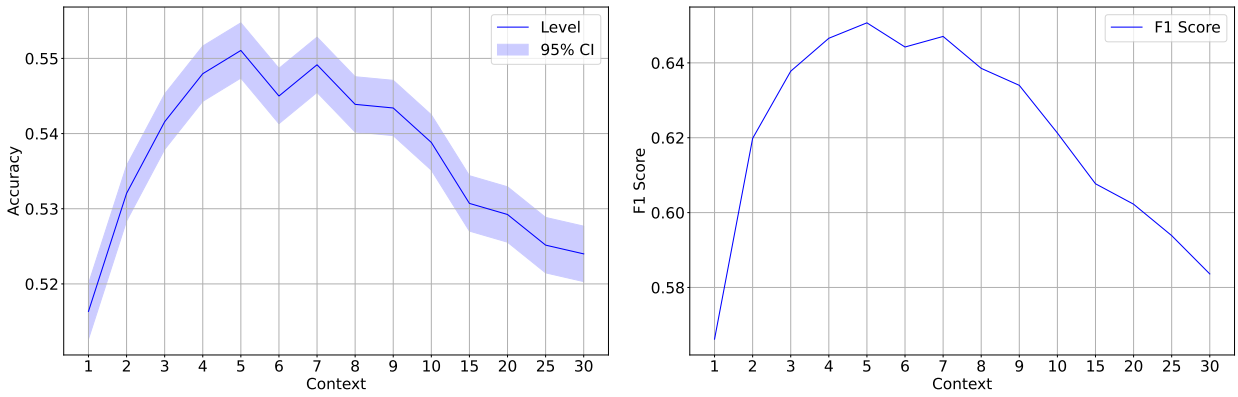


Fig. 2. EPS prediction match rate by context length

This figure presents the match rate on the left and the F1 score on the right, measuring the model's ability to predict the direction of next quarter's EPS change, defined as the sign of $\text{EPS}_{t+1} - \text{EPS}_t$. The x-axis represents the amount of contextual information provided from earnings call transcripts. Each quarter's earnings call, including the current and the two most recent past calls, is divided into ten equal sections. A context value of seven indicates that 70 percent of the current quarter's earnings call is included in the model input. A value of seventeen means the full earnings call from the current quarter is used along with 70 percent of the prior quarter's earnings call. The left panel includes a shaded area representing the 95 percent confidence interval around the Match estimates.

In the experiment, we progressively add sections from prior earnings conference calls to the model's input. The process begins with the first section and continues sequentially, ensuring that the end of one earnings call is followed by the beginning of the previous one. Based on the provided sections, the model predicts whether next quarter's EPS will increase or decrease, defined as the sign of EPS $(t+1)$ - EPS (t). To maintain consistency in style and structure, we exclude accounting data from the sample, differentiating our approach from Kim et al. (2025). This ensures that any observed effects stem from the amount of textual context rather than differences in data type. Figure 2 presents the results of the experiment. The Y-axis reports model performance, with accuracy on the left and the F1-score on the right, while the X-axis indicates the number of incremental earnings call sections included in the prompt.

The results demonstrate a clear inverted U-shaped relationship between the amount of context provided and the model's performance in predicting EPS direction. While adding a second section improves accuracy compared to using only one, further increases in context lead to diminishing and eventually negative returns in predictive performance. To test the robustness of this pattern, we replicate the analysis across different industries. Figure 3 presents the results, showing that the relationship remains consistent across sectors, suggesting that the observed effect is unlikely to be driven by firm-specific characteristics.

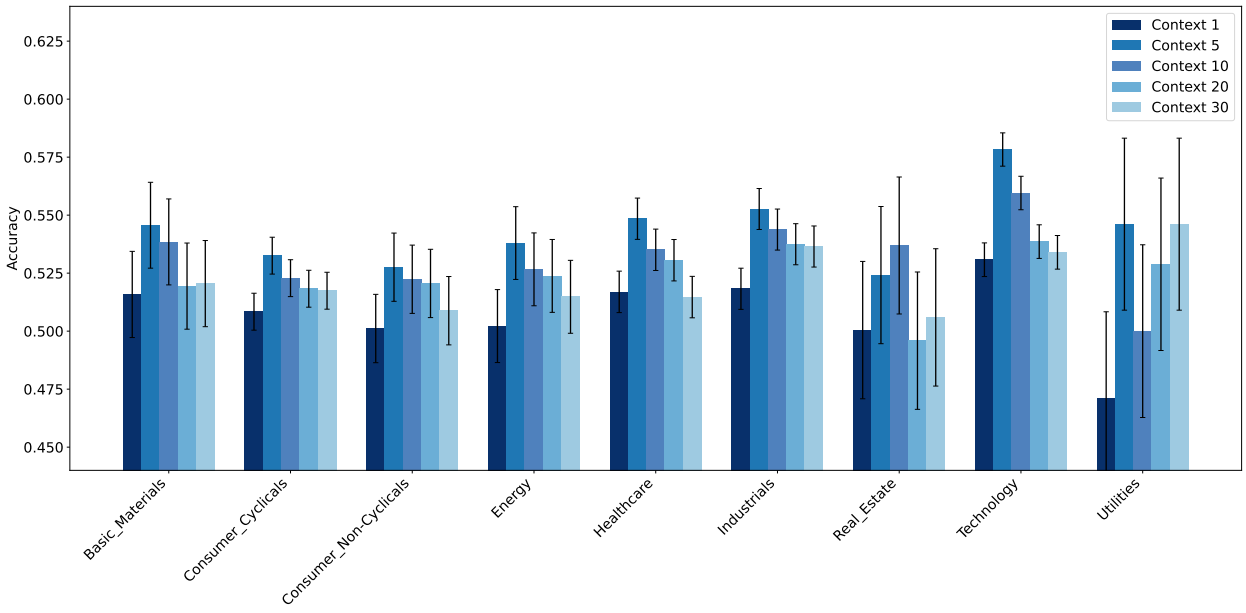


Fig. 3. Industry heterogeneity in EPS forecast match rate

This figure presents the match rate of the model's prediction of next quarter's EPS change, defined in Equation (1). Context refers to the proportion of the Q&A section included from each quarter's earnings call. Each quarter's earnings call, including the current and the two most recent past calls, is divided into ten equal sections. A context value of seven indicates that 70 percent of the current quarter's earnings call is included in the model input. A value of seventeen means the full earnings call from the current quarter is used along with 70 percent of the prior quarter's earnings call. The figure reports results separately for each industry subsample.

To further assess the robustness of our findings, we estimate the following regression, following the approach of Kim et al. (2025):

$$\mathbf{1}\Big[\hat{sgn}(\text{EPS}_{i,t+1} - \text{EPS}_{i,t}) = sgn(\text{EPS}_{i,t+1} - \text{EPS}_{i,t})\Big] = \Gamma' X_{i,t-1} + \delta_{j,t} + \varepsilon_{i,t}. \tag{1}$$

Here, $\mathbf{1}(\cdot)$ represents and indicator variable that equals 1 when the LLM's predicted sign matches the observed sign of earnings changes and 0 otherwise. We refer to this outcome variable as "match" or "match rate", where $\hat{sgn}(\text{EPS}_{i,t+1} - \text{EPS}_{i,t})$ is the LLM's prediction for EPS change for firm i from the current quarter t to the next quarter $t+1$, and $sgn(\text{EPS}_{i,t+1} - \text{EPS}_{i,t})$ is the actual observed earnings change direction. $X_{i,t}$ is a vector of firm-level control variables, $\delta_{j,t}$ is an industry-by-time fixed effect, and $\varepsilon_{i,t}$ is the error term.

Our variables of interest are Γ', which capture the relationship between firm characteristics and accurate predictions. They are consistent with Kim et al. (2025) and include firm

size, book-to-market ratio, leverage, asset tangibility, and an indicator for firms reporting a loss.

The regression results in Table 3 indicate that firm characteristics explain some variation in the model's ability to predict earnings direction, though overall explanatory power remains low. Earnings volatility is consistently negative and highly significant across all specifications, suggesting that greater uncertainty in earnings makes forecasting more difficult. Firms with unpredictable earnings histories likely provide noisier signals in their earnings calls, making it harder for the model to extract reliable patterns. The book-to-market ratio is also consistently negative and significant, indicating that the model performs better at predicting earnings changes for firms with lower book-to-market values. Growth firms, which tend to have lower book-to-market ratios, may provide clearer forward guidance or have earnings that follow more discernible trends, making their earnings calls easier for the model to interpret.

Firm size is positively associated with prediction accuracy in some specifications, though the effect is small. Larger firms may have more predictable earnings due to diversified revenue streams, greater market stability, or more standardized earnings call structures. The loss indicator is also positive and highly significant, suggesting that firms reporting losses are easier to classify in terms of future earnings direction. This likely reflects the asymmetric nature of earnings surprises. Loss-making firms tend to use more explicit language about financial distress or recovery prospects, making the model's task easier. Leverage and asset tangibility exhibit weaker and less consistent relationships with prediction accuracy. The weak negative association of leverage in some specifications may indicate that highly leveraged firms face greater uncertainty or external pressures that complicate earnings predictions. The mixed significance of asset tangibility suggests that its role in earnings predictability is less straightforward. Differences across industries in capital structure and financial reporting could explain this variation.

The inclusion of year-by-industry fixed effects improves model fit slightly, as indicated by the higher R-squared values, but overall explanatory power remains low. These findings reinforce the idea that firm-level characteristics do not fully explain the inverted U-shaped relationship between context length and predictive accuracy. Instead, the results suggest that the observed decline in performance with additional context is driven by the model's information-processing constraints rather than systematic firm-specific differences.

We also examine the stability of the observed pattern over time by evaluating the model's match rate across different years. Figure 4 presents the results, comparing predictions made using a limited context window (C5), which includes 50 percent of the current quarter's earnings call transcript, to those made using an expanded context window (C30), which

Table 3: Determinants of LLM Forecasting Accuracy in Earnings Prediction

This table presents the results of Equation (1), where the dependent variable is a dummy indicating whether Llama Smallcorrectly predicts the direction of Earnings Per Share (EPS) changes in the following quarter. All independent variables are winsorized at the 1% and 99% levels to mitigate the influence of extreme values. Standard errors, clustered at the industry level, are shown in parentheses. Variable definitions are provided in the appendix. Statistical significance is denoted by *, **, and *** at the 10%, 5%, and 1% levels.

	C 1		C 10		C 30	
	(1)	(2)	(3)	(4)	(5)	(6)
Earnings volatility	-0.132^{**}	-0.129^{***}	-0.261^{***}	-0.213^{***}	-0.254^{***}	-0.235^{***}
	(0.057)	(0.048)	(0.087)	(0.065)	(0.068)	(0.054)
Book-to-market	-0.017^{***}	-0.019^{***}	-0.031^{***}	-0.038^{***}	-0.022^{***}	-0.030^{***}
	(0.005)	(0.004)	(0.006)	(0.004)	(0.004)	(0.004)
Loss	0.031^{***}	0.032^{***}	0.101^{***}	0.107^{***}	0.068^{***}	0.073^{***}
	(0.005)	(0.005)	(0.007)	(0.007)	(0.006)	(0.006)
Size	0.004^{***}	0.006^{***}	0.003^{***}	0.005^{***}	0.001	0.001
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Leverage	-0.019^{*}	-0.016	-0.036^{***}	-0.025^{**}	-0.022^{**}	-0.019^{**}
	(0.011)	(0.010)	(0.010)	(0.011)	(0.010)	(0.009)
Asset tangibility	-0.023^{**}	-0.007	-0.009	0.016	0.015	0.029^{*}
	(0.010)	(0.014)	(0.013)	(0.018)	(0.013)	(0.015)
Constant	0.498^{***}	0.481^{***}	0.522^{***}	0.502^{***}	0.522^{***}	0.515^{***}
	(0.008)	(0.010)	(0.009)	(0.009)	(0.008)	(0.009)
Year × Industry FE	No	Yes	No	Yes	No	Yes
Observations	70,432	70,423	70,432	70,423	70,432	70,423
R-squared	0.001	0.016	0.007	0.021	0.003	0.019

includes the full transcript from the current and two prior quarters. The match rate remains higher for C5 than C30 in all years, reinforcing the finding that adding more context does not consistently improve predictive performance. The gap between the two conditions fluctuates over time, with the difference narrowing in some periods, such as 2017. This suggests that the optimal amount of context varies by year, likely due to differences in earnings informativeness and broader economic conditions. Appendix A2 further explores this variation by tracking the number of sections that yield peak performance and the impact of including transcripts from multiple past earnings calls.

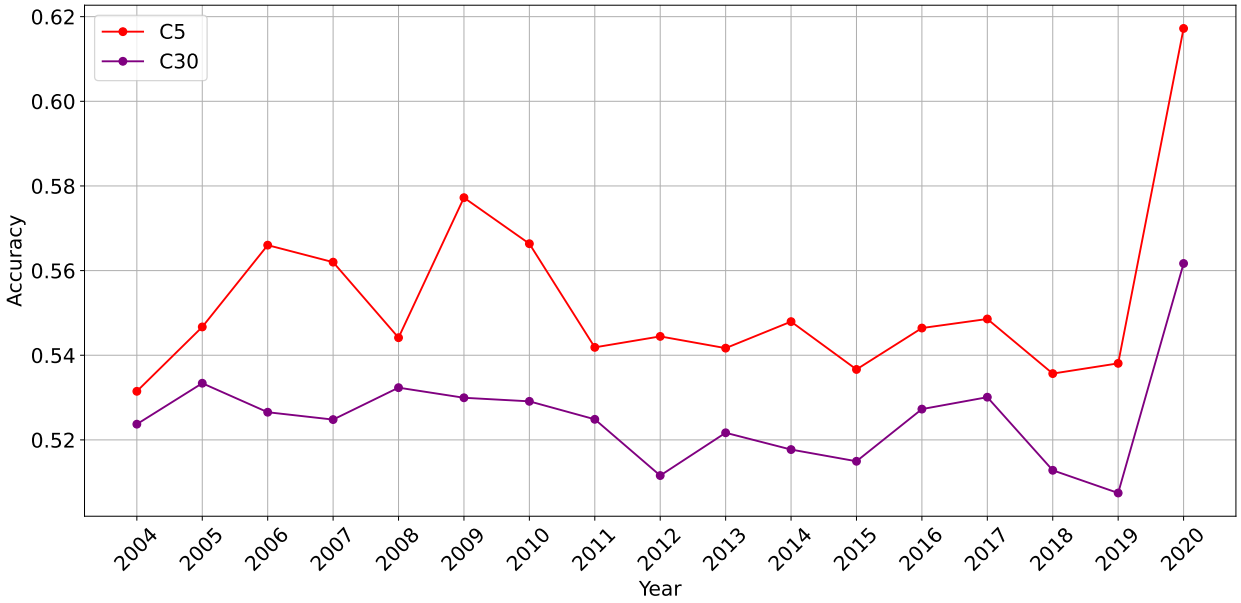


Fig. 4. Temporal variation in EPS forecasting match rate

This figure shows the match rate of the LLM's predictions over time. The red line represents predictions using a limited amount of context (C5), which includes 50 percent of the current quarter's earnings call transcript. The purple line represents predictions using all available context (C30), which includes the full transcript of the current and two prior quarters. The x-axis represents the year of prediction, while the y-axis indicates the match rate.

5. Nowcasting Stock Price Movements

This section examines the relationship between the quantity of context and the predictive performance of a large language model in an asset pricing setting. We test how additional contextual information affects the model's ability to interpret financial news and forecast immediate market reactions. Using a comprehensive dataset of news concerning U.S. publicly listed firms, we apply an LLM to classify news sentiment and assess whether its predictions

align with same-day stock price movements.

The news dataset is sourced from the Thomson Reuters Real-time News Feed and spans the period from 1996 to 2022. It includes both structured financial alerts—concise, real-time updates intended for rapid dissemination—and full-length articles that provide more extensive coverage of firm-specific developments. As in prior sections, we filter the data to remove articles that are excessively short, long, or redundant based on cosine similarity measures. To ensure relevance, we focus on firm-specific news identified through ticker mentions and retain only observations where the firm's close-to-close return on the same day is available.

For this analysis, we use Llama Small, an eight-billion-parameter LLM, to classify news sentiment as either positive or negative and compare these classifications to realized same-day stock returns. The model is queried using the following prompt:

System Prompt

`<|begin_of_text|><|start_header_id|>system<|end_header_id|>`

You are financial news analyst that answers either `P` (positive return), `N` (negative return) for the news given for a company. Please answer in a single letter. `<|eot_id|>`

User Prompt

`<|start_header_id|>user<|end_header_id|>`

Here is a piece of news today: `{headline}`.

Here are the historical news for the same company starting from most recent: `{context}`. `<|eot_id|>`

Assistant Prompt

`<|start_header_id|>assistant<|end_header_id|>`

We repeat the exercise by incrementally appending up to 30 prior stock-specific news articles to the prompt. Rather than presenting the news in strictly reverse chronological order, we introduce them in a randomized sequence while retaining each article's original timestamp, including both date and hour. This approach ensures that the model receives comparable levels of information across different samples while preserving the temporal context necessary for assessing the relevance of past news. Figure 5 presents the results.

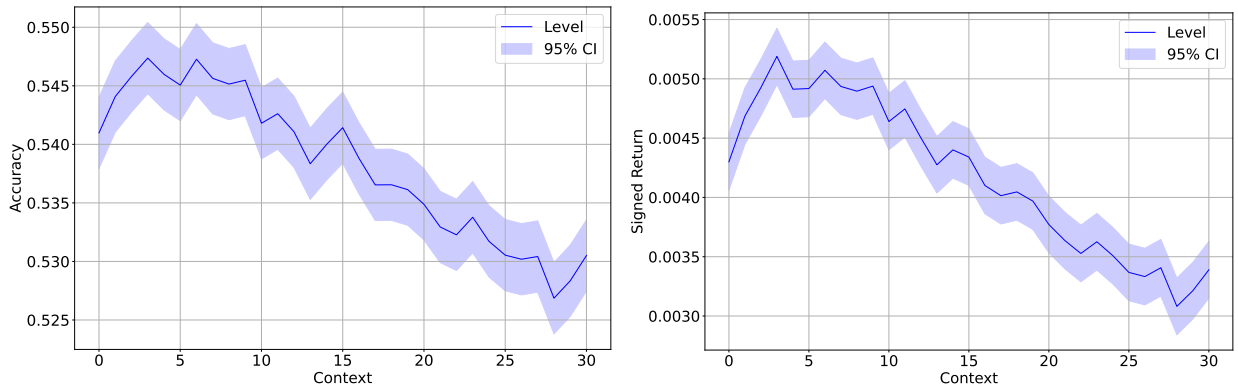


Fig. 5. The figure shows the performance in terms of accuracy (left) and Signed Return (right) of llama8B in a nowcasting exercise. We ask the model to evaluate a news as positive or negative for the market, and compare the answer to the simultaneous price reaction. We incrementally append older news in random chronological order, from one to 30.

The model's performance initially improves as additional context is provided, but after approximately ten appended news articles, its predictive accuracy begins to decline. This pattern aligns with the mechanism outlined in our bounded rationality section and the findings from our earnings call transcript experiment. While adding context enhances interpretability by allowing the model to incorporate relevant historical information, excessive context introduces cognitive overload, leading to diminishing returns and eventual degradation in predictive performance.

5.1. Long-Short Portfolios and Model Size

This section examines how information overload affects portfolio formation and whether its economic consequences vary across LLM sizes. We construct long-short portfolios based on financial alerts—short, timely news updates—with and without portfolio constraints. The results show that information overload has a measurable impact on portfolio performance, with smaller LLMs exhibiting significantly stronger effects than larger, more advanced models.

Processing a large volume of financial news poses computational challenges, particularly given the exponential growth in token processing costs within individual prompts and the linear increase in costs with the number of prompts. These constraints limit the feasibility of directly applying the previous methodologies to portfolio construction. To address this, we adopt the approach of Chen et al. (2024b), which leverages LLM-generated embeddings of financial news rather than direct textual input. These embeddings serve as inputs for a separate machine learning model trained to predict short-term stock returns.

Embedding-based forecasts offer a distinct advantage: the computational burden of running the LLM is incurred only once during the initial vectorization of the news. Subsequent portfolio construction, which involves rolling window estimation and strategy recalibration, relies on a non-LLM-based model. This approach significantly reduces computational costs and ensures that the analysis remains scalable within realistic constraints.

Hence, we construct long-short portfolios using the following steps. Additional details on the vectorization and training procedure are available in Chen et al. (2024b).

1. **Embedding Financial News:** We use LLMs to transform financial news into vector representations $v_{i,t} \in \mathbb{R}^d$, where each vector encodes the content of a news item about firm i at time t. The embedding dimension d defines the size of the resulting feature space.
2. **Return Prediction Model:** Using an 8-year rolling window, we train ridge regression models to predict open-to-open returns accumulated between t and $t+1$. The predictive model is specified as:

$$\prod_{s=t}^{t+1}(1+r_{i,s})-1=\alpha+\beta^{\top}v_{i,t}+\epsilon_{i,t+1},$$

   where $\alpha \in \mathbb{R}$ is the intercept, $\beta \in \mathbb{R}^d$ is the coefficient vector, and $\epsilon_{i,t+1}$ is the error term.
3. **Out-of-Sample Forecasting:** The trained model is then used to generate out-of-sample return forecasts for individual stocks.
4. **Long-Short Portfolio Construction:** Based on the model's predictions, we implement a long-short strategy by buying the top 30 percent of firms, which are those with the highest predicted returns, and shorting the bottom 30 percent, which are those with the lowest predicted returns, at a daily frequency.
5. **Portfolio Weighting and Return Calculation:** The weights obtained at time t are used to form a portfolio whose return is realized at time $t+1$. This differs from the training process, where news at time t is incorporated for predictive insights. For portfolio construction, weights—functions of the embeddings at time t—are applied to returns observed at time $t+1$. Unlike Chen et al. (2024b), who include $t-1$ in the training label, we exclude $t-1$ to avoid overlap with contextualizing information. The portfolio return is defined as:

$$r_{p,t+1}=\sum_{i\in\mathcal{I}_t} w_{i,t}\, r_{i,t+1},$$

where $w_{i,t}$ represents the portfolio weight assigned to firm i at time t, and $\mathcal{I}_t$ denotes the set of selected firms. We also define a value-weighted portfolio, where, within both the long and short legs, each stock is weighted by its market capitalization.

5.2. Model Size, Context, and Portfolio Performance

We use three large language models, Llama Small, Llama Medium, and Llama Large, which correspond to the Llama 3.1 8B, 70B, and 405B models, to encode two distinct information sets, generating six embedding procedures. These embeddings are then used to construct six long-short portfolios, summarized in Table 4. The first set of embeddings is based on news alerts published between 9 AM on day $t-1$ and 9 AM on day t. Alerts are concise, time-sensitive updates that often resemble article headlines rather than full reports. This subset is particularly relevant for assessing the value of additional context, as short-form news is more likely to be interpreted differently depending on surrounding information. The second set of embeddings builds on the first by incorporating prior news about the same firm as context. Instead of using alerts, we include only full articles, as they provide more comprehensive coverage and contain more fundamental information about the firm. In C_1, an article is added as context for a given firm only if it was published within the past day. Appendix A.1 provides further details on the prompt design.

Table A3 presents the annualized out-of-sample mean return, variance, and Sharpe ratio for the six equally weighted portfolios. The corresponding results for value-weighted strategies are reported in Table A4. Out-of-sample performance is strong, with Sharpe ratios ranging from 4.95 for $C_1$with Llama Small to 5.84 for C_0 with Llama Large. These findings align with Chen et al. (2024b), confirming that larger LLMs enhance predictive accuracy in portfolio construction. Both the C_0 and C_1 strategies benefit from increased model size, but the improvement is more pronounced for C_1, which incorporates additional context. This result supports the intuition outlined in our section on bounded rationality, where larger models produce higher-quality embeddings, leading to better financial performance. Context-based embeddings benefit more from this effect, as information overload diminishes with model size, allowing for more effective utilization of expanded input.

To further this intuition we estimate the tangency portfolio composed of two assets: the long short strategy with $C_1$and C_0. We estimate this tangency portfolio on the full sample for each Llamma model for both the equal-weighted and value weighted strategy.

Finally, we evaluate the performance of $r_t(C_1, M)$ relative to $r_t(C_0, M)$, where M represents one of the three LLMs: Llama Small, Llama Medium, or Llama Large. To quantify the incremental value of additional context, we estimate the following regression:

Table 4: Summary of Long-Short Portfolios by Model Size and Context

This table summarizes the six long-short portfolios constructed using embeddings from three different language models: Llama Small, Llama Medium, and Llama Large, representing the Llama 3.1 8B, 70B, and 405B models. Each model is applied to two distinct information sets. C_0 portfolios are based on embeddings of news alerts published between 9 AM on day $t-1$ and 9 AM on day t. C_1 portfolios incorporate additional context by including prior news articles about the same firm. The notation $r_t(C_i, XB)$ represents the return of the long-short portfolio at time t, where C_i denotes the embedding context ($i = 0, 1$) and XB indicates the model's parameter size in billions.

	C_0	C_1
Llama Small	$r_t(C_0, 8B)$	$r_t(C_1, 8B)$
Llama Medium	$r_t(C_0, 70B)$	$r_t(C_1, 70B)$
Llama Large	$r_t(C_0, 405B)$	$r_t(C_1, 405B)$

Table 5: Sharpe Ratios for Long-Short Portfolios

This table reports Sharpe ratios for the Long, Short, and Long-Short portfolios based on daily returns, comparing the no-context (C_0) and context (C_1) strategies. Results are presented for the three Llama models: Llama Small, Llama Medium, and Llama Large. Additional details on the mean and variance of these portfolios are provided in Tables A1 and A2.

		Equally Weighted		Value Weighted	
		C_0	C_1	C_0	C_1
Long	Llama Small	1.98	1.93	1.21	0.98
	Llama Medium	2.15	2.04	1.25	1.10
	Llama Large	2.07	2.10	1.23	1.19
Short	Llama Small	-1.49	-1.41	-0.13	0.08
	Llama Medium	-1.56	-1.53	-0.07	0.12
	Llama Large	-1.74	-1.67	-0.07	0.10
Long-Short	Llama Small	5.13	4.95	1.86	1.26
	Llama Medium	5.55	5.39	1.81	1.34
	Llama Large	5.84	5.73	1.90	1.51

Table 6: Tangency Portfolio Estimates for Long-Short Strategies Across LLM Models

This table presents the estimated tangency portfolio statistics for long-short strategies based on C_0 and C_1. The tangency portfolio is constructed using the full sample for each Llama model, with results reported separately for equal-weighted and value-weighted strategies. The table includes key performance metrics such as the annualized Sharpe ratio, annualized return, and annualized standard deviation. Additionally, it provides the tangency portfolio weights assigned to C_0 and C_1.

	Equally Weighted			Value Weighted		
	Small	Medium	Large	Small	Medium	Large
Sharpe ratio	5.18	5.61	5.92	1.87	1.84	1.95
Return	0.88	0.93	0.97	0.26	0.26	0.26
Volatility	0.17	0.17	0.16	0.14	0.14	0.13
C_0 Weight	0.68	0.66	0.60	0.86	0.79	0.72
C_1 Weight	0.32	0.34	0.40	0.14	0.21	0.28

$$r_t(C_1, M) = \alpha + \beta r_t(C_0, M) + \varepsilon_{i,t}, \tag{2}$$

The parameter α captures the portion of $r_t(C_1, M)$ that cannot be explained by $r_t(C_0, M)$, providing an estimate of the independent contribution of context. Table 7 reports the estimated values and significance levels of alpha across all configurations, including both value-weighted and equally weighted portfolios.

The results indicate a clear relationship between LLM size and the benefits of adding context. In equally weighted portfolios, both the magnitude and significance of the alpha increase with model size, suggesting that larger models extract more useful information from additional context. A similar trend is observed for value-weighted portfolios, though statistical significance is weaker. The largest model, Llama Large, is the only one for which the performance differential reaches marginal significance, with a p-value of 0.0625. Tables A1 to A4 provide further robustness checks, confirming that these results hold when controlling for risk factors, including those from the Fama-French framework.

5.3. Firm Characteristics and the Economic Value of Context

This section examines the fundamental drivers behind the differences in predictive performance when using C_0 and C_1. To assess how firm and news characteristics affect match rate improvements, we define the match rate for an individual forecast as

Table 7: Impact of Additional Context on Portfolio Performance

This table presents the regression results of $r_t(C_1, M)$ on $r_t(C_0, M)$ across different LLM sizes, reported separately for equally weighted and value-weighted portfolios. The table displays the estimated coefficient and intercept, Alpha, which represents the portion of $r_t(C_1, M)$ returns unexplained by $r_t(C_0, M)$. Standard errors are shown in parentheses. Annualized alpha is reported as a measure of the economic significance of the results. All return values, including alpha and annualized alpha, are expressed in percentage terms for ease of interpretation. Statistical significance is indicated by *, **, and *** at the 10%, 5%, and 1% levels.

	Equal-Weighted			Value-Weighted		
	Small	Medium	Large	Small	Medium	Large
Alpha	0.021***	0.022***	0.028***	0.010	0.016	0.022*
	(0.008)	(0.007)	(0.007)	(0.012)	(0.012)	(0.012)
$r_t(C_0, M)$	90.247***	90.103***	91.413***	56.485***	57.795***	59.637***
	(0.666)	(0.643)	(0.660)	(1.247)	(1.245)	(1.254)
Annualized Alpha	5.380	5.636	7.393	2.668	4.011	5.586
Observations	4,015	4,015	4,015	4,015	4,015	4,015
R-squared	0.820	0.830	0.827	0.338	0.349	0.360

$$A^c_{i,j,t}(r_{i,t}) = \begin{cases} 1 & \text{if } (P^c_{i,j,t} \geq 0.5 \text{ and } r > 0) \text{ or } (P^c_{i,j,t} < 0.5 \text{ and } r < 0), \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $P^c_{i,j,t} \in [0, 1]$ represents the out-of-sample prediction from the model with context $c \in [C_1, C_0]$ for news item n related to firm i. We estimate the following regression:

$$\mathrm{A}^{C_1}_{i,n,t}(r_{i,t}) - \mathrm{A}^{C_0}_{i,n,t}(r_{i,t}) = \alpha + \Gamma' X_{i,t} + \Lambda' Z_{i,n,t} + \delta_{j,t} + \epsilon_{i,n,t} \tag{4}$$

where $\Gamma' X_{i,t}$ represents a set of firm-level predictors, and $\Lambda' Z_{i,n,t}$ captures news-level predictors. $\delta_{j,t}$ is an industry-by-time fixed effect, and standard errors are clustered at the firm level.

The regression results in Table 8 provide insights into how firm and news characteristics influence the effectiveness of incorporating additional context in predicting stock returns. Larger firms and those with greater analyst coverage exhibit a higher match rate, suggesting that additional context is more valuable when market participants process information more efficiently (Hong, Lim, and Stein, 2000). This finding aligns with the view that well-covered

firms have more widely disseminated and scrutinized information, making LLM-generated predictions more reliable when historical context is included. The book-to-market ratio is positively associated with the match rate, indicating that value firms benefit more from the inclusion of historical context than growth firms. This result is consistent with prior research showing that growth firms have more uncertain and narrative-driven valuations (Damodaran, 2009), which may limit the effectiveness of additional context.

The coefficients for beta, volatility, and momentum are not statistically significant, suggesting that risk-related factors do not materially influence the relative benefits of adding context. Earnings variability, however, is positively associated with match rate in some specifications, implying that firms with more volatile earnings benefit from contextual information when forming return predictions.

Article length is negatively associated with match rate across models, suggesting that longer news articles may introduce noise or irrelevant details that reduce predictive accuracy. The inclusion of year-by-industry fixed effects modestly increases the explanatory power of the model, but the overall explanatory power remains low. These results reinforce the conclusion that the economic value of context depends largely on firm size, analyst coverage, and valuation characteristics rather than traditional risk factors.

6. Conclusion

This paper identifies a fundamental constraint in how large language models process financial information. While additional context initially enhances predictive performance, excessive information degrades it, producing an inverted U-shaped relationship between context size and accuracy. This phenomenon is robust across both corporate finance and asset pricing tasks, suggesting it reflects an inherent limitation of LLMs rather than a domain-specific artifact.

We argue that this limitation is likely to be systematic across all LLMs, as it can be traced to two fundamental sources common to their design.

First, the architectural constraints of LLMs create informational bottlenecks that limit their ability to retain and prioritize relevant details as context expands. The layered structure of the neural networks within LLMs enforces these constraints, as each layer inside the LLM has a fixed dimension, effectively functioning as an informational bottleneck. At each stage, preceding layers collectively serve as implicit encoders, processing incoming information and transferring it in a compressed form to optimize the efficiency of subsequent encoders. An informational bottleneck alone is not a sufficient condition for the information overload documented in this paper. In principle, a perfect encoder would filter out redundant or

Table 8: Determinants of News-Level Match Rate

This table presents the results of a pooled regression where the dependent variable is the individual news-level match rate, defined as the difference between C_1 and C_0. Match rate is determined based on whether the model's prediction aligns with the sign of the concurrent stock return. The regression is estimated using financial alerts published between 9:00 AM and 4:00 PM. Independent variables are mean-variance standardized. Standard errors, clustered at the firm level, are shown in parentheses. Variable definitions are provided in the appendix. Statistical significance is denoted by *, **, and *** at the 10%, 5%, and 1% levels.

	Llama Small		Llama Medium		Llama Large	
	(1)	(2)	(3)	(4)	(5)	(6)
Coverage	0.007***	0.007***	0.007***	0.008***	0.006***	0.007***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Size	0.009***	0.007***	0.008***	0.006**	0.007***	0.004
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Book-to-market	0.001	0.001	0.004***	0.002	0.004***	0.004***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Beta	−0.001	−0.000	−0.001	0.000	0.000	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Volatility	−0.000	−0.001	−0.001	−0.001	−0.002	−0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Momentum	0.001	0.001	0.001	0.001	−0.000	−0.001
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
Earnings variability	0.002	0.001	0.003	0.001	0.003**	0.001
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
Aricle length	−0.001	−0.002	−0.003	−0.003*	−0.003	−0.004**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Constant	−0.016***	−0.016***	−0.016***	−0.016***	−0.011***	−0.011***
	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)
Year × Industry FE	No	Yes	No	Yes	No	Yes
Observations	416,715	416,684	416,715	416,684	416,715	416,684
R-squared	0.001	0.010	0.001	0.010	0.001	0.011

irrelevant information, ensuring that the fixed layer size prevents the full utilization of context without degrading performance. With perfect encoders, the relationship between information volume and model performance would follow a logarithmic pattern, where additional context either enhances or has no effect on performance. However, despite their high parametrization, these implicit encoders are unlikely to be perfect. When LLMs are trained, theory offers no guarantee of convergence to a global optimum, and it is statistically likely that most LLMs are trained until they converge to a local minimum. As a result, the encoders are best thought of as imperfect approximations. As context expands, those imperfections compound with the limited layer size, preventing LLMs from fully filtering out less relevant elements. Consequently, space allocated to unimportant information comes at the expense of valuable content, ultimately diminishing output quality.

Second, LLM training procedures prioritize expanding context windows but do not optimize for handling long prompts effectively. While training ensures that models can retrieve deeply embedded information, it does not guarantee that they filter redundant or misleading input, leading to performance deterioration as more context is introduced.

Our asset pricing results further indicate that while larger and more advanced LLMs mitigate the effects of information overload, they do not eliminate them. This suggests that technological improvements can expand the usable context window but do not resolve the underlying problem. More powerful architectures help preserve information efficiency over longer sequences, but the fundamental constraints of token processing and positional embeddings persist. Contrary to the assumption that more data invariably improves decision-making, our findings highlight the need for empirical validation when determining how much context should be incorporated. The optimal amount of information depends on the specific task, model size, and how information is encoded.

These results have direct implications for financial research and AI-driven decision-making. The non-monotonic relationship between context size and predictive accuracy challenges the prevailing assumption that maximizing an LLM's context window necessarily improves performance. Instead of assuming that broader input leads to better analysis, researchers and practitioners must treat context inclusion as an optimization problem, balancing informativeness against computational frictions.

This research builds on Chen et al. (2025), emphasizing the importance of studying AI in finance and advancing beyond black-box approaches when identifying biases or limitations that may affect economic outcomes. Specifically, we argue that future research should move beyond psychological analogies and develop systematic methods to determine the right amount of context for LLMs. Just as econometrics warns against indiscriminately adding variables due to risks of overfitting and multicollinearity, AI-driven financial analysis must

recognize that context selection is a crucial empirical decision. The growing integration of AI into financial markets raises important questions about how LLMs process economic data, whether firms and investors can strategically manage information flow, and how AI-induced inefficiencies shape market outcomes. Identifying the precise mechanisms that govern context-dependent performance is essential to ensuring that LLMs enhance, rather than distort, financial decision-making.

References

Abis, S., Veldkamp, L., 2024. The Changing Economics of Knowledge Production. The Review of Financial Studies 37, 89–118, https://doi.org/10.1093/rfs/hhad059.

Aldasoro, I., Gambacorta, L., Korinek, A., Shreeti, V., Stein, M., 2024. Intelligent financial system: how AI is transforming finance. Working paper, Bank for International Settlements, https://www.bis.org/publ/work1194.htm.

Ashokkumar, A., Hewitt, L., Ghezae, I., Willer, R., 2024. Predicting Results of Social Science Experiments Using Large Language Models. Working paper, New York University, https://docsend.com/view/ity6yf2dansesucf.

Babina, T., Fedyk, A., He, A., Hodson, J., 2024. Artificial intelligence, firm growth, and product innovation. Journal of Financial Economics 151, 103745, https://doi.org/10.1016/j.jfineco.2023.103745.

Bybee, L., 2023. Surveying generative AI's economic expectations. Working Paper 2305.02823, arXiv, https://doi.org/10.48550/arXiv.2305.02823.

Cao, S., Jiang, W., Wang, J., Yang, B., 2024. From Man vs. Machine to Man + Machine: The art and AI of stock analyses. Journal of Financial Economics 160, 103910, https://doi.org/10.1016/j.jfineco.2024.103910.

Chen, H., Didisheim, A., Somoza, L., 2025. Out of the (Black) Box: AI as Conditional Probability. Working Paper 5012852, Social Science Research Network, https://dx.doi.org/10.2139/ssrn.5012852.

Chen, S., Green, T. C., Gulen, H., Zhou, D., 2024a. What Does ChatGPT Make of Historical Stock Returns? Extrapolation and Miscalibration in LLM Stock Return Forecasts. Working Paper 2409.11540, arXiv, https://doi.org/10.48550/arXiv.2409.11540.

Chen, Y., Kelly, B. T., Xiu, D., 2024b. Expected returns and large language models. Working Paper 4416687, Social Science Research Network, https://ssrn.com/abstract=4416687.

Cheng, Q., Lin, P., Zhao, Y., 2024. Does Generative AI Facilitate Investor Trading? Evidence from ChatGPT Outages. Singapore Management University School of Accountancy Research Paper No. 2025-186, Singapore Management University, https://dx.doi.org/10.2139/ssrn.4872189.

Chernev, A., Böckenholt, U., Goodman, J., 2015. Choice overload: A conceptual review and meta-analysis. Journal of Consumer Psychology 25, 333–358, https://doi.org/10.1016/j.jcps.2014.08.002.

Damodaran, A., 2009. Valuing young, start-up and growth companies: Estimation issues and valuation challenges. Working Paper 1418687, Social Science Research Network, https://dx.doi.org/10.2139/ssrn.1418687.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al., 2025. DeepSeek-V3 Technical Report. Working Paper 2412.19437, arXiv, https://doi.org/10.48550/arXiv.2412.19437.

Didisheim, A., Ke, S. B., Kelly, B. T., Malamud, S., 2024. APT or "AIPT"'? The Surprising Dominance of Large Factor Models. Working Paper 33012, National Bureau of Economic Research, http://www.nber.org/papers/w33012.

Dimoka, A., 2012. How to Conduct a Functional Magnetic Resonance (fMRI) Study in Social Science Research. MIS quarterly 36, 811–840, https://doi.org/10.2307/41703482.

Eisfeldt, A. L., Schubert, G., 2024. AI and Finance. Working Paper 33076, National Bureau of Economic Research, http://www.nber.org/papers/w33076.

Eppler, M. J., Mengis, J., 2004. The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. The Information Society 20, 325–344, https://doi.org/10.1080/01972240490507974.

Fama, E. F., MacBeth, J. D., 1973. Risk, Return, and Equilibrium: Empirical Tests. Journal of Political Economy 81, 607–636, https://www.jstor.org/stable/1831028.

Fedyk, A., Kakhbod, A., Li, P., Malmendier, U., 2024. ChatGPT and Perception Biases in Investments: An Experimental Study. Working Paper 4787249, Social Science Research Network, https://dx.doi.org/10.2139/ssrn.4787249.

Francis, J., LaFond, R., Olsson, P. M., Schipper, K., 2004. Costs of Equity and Earnings Attributes. The Accounting Review 79, 967–1010, https://doi.org/10.2308/accr.2004.79.4.967.

Giglio, S., Kelly, B., Xiu, D., 2022. Factor Models, Machine Learning, and Asset Pricing. Annual Review of Financial Economics 14, 337–368, https://doi.org/10.1146/annurev-financial-101521-104735.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al., 2024. The Llama 3 Herd of Models. Working Paper 2407.21783, arXiv, https://doi.org/10.48550/arXiv.2407.21783.

Hong, H., Lim, T., Stein, J. C., 2000. Bad News Travels Slowly: Size, Analyst Coverage, and the Profitability of Momentum Strategies. The Journal of Finance 55, 265–295, https://doi.org/10.1111/0022-1082.00206.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366, https://doi.org/10.1016/0893-6080(89)90020-8.

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J., Wang, H., 2024. Large Language Models for Software Engineering: A Systematic Literature Review. ACM Transactions on Software Engineering and Methodology 33, 1–79, https://doi.org/10.1145/3695988.

Inbar, Y., Pizarro, D. A., Bloom, P., 2009. Conservatives are more easily disgusted than liberals. Cognition and Emotion 23, 714–725, https://doi.org/10.1080/02699930802110007.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. The Journal of Finance 48, 65–91, https://doi.org/10.1111/j.1540-6261.1993.tb04702.x.

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica 47, 363–391, https://doi.org/10.2307/1914185.

Kelly, B., Malamud, S., Zhou, K., 2024. The virtue of complexity in return prediction. The Journal of Finance 79, 459–503, https://doi.org/10.1111/jofi.13298.

Kim, A., Muhn, M., Nikolaev, V., 2025. Financial Statement Analysis with Large Language Models. Working Paper (This paper has been withdrawn by Valeri Nikolaev) 2407.17866, arXiv, https://doi.org/10.48550/arXiv.2407.17866.

Korinek, A., 2023. Generative AI for Economic Research: Use Cases and Implications for Economists. Journal of Economic Literature 61, 1281–1317, https://doi.org/10.1257/jel.20231736.

Korinek, A., 2024. Generative AI for Economic Research: LLMs Learn to Collaborate and Reason. Working Paper 33076, National Bureau of Economic Research, https://www.nber.org/papers/w33198.

Lopez-Lira, A., Tang, Y., 2024. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. Working Paper 4412788, Social Science Research Network, https://dx.doi.org/10.2139/ssrn.4412788.

Ludwig, J., Mullainathan, S., 2024. Machine learning as a tool for hypothesis generation. The Quarterly Journal of Economics 139, 751–827, https://doi.org/10.1093/qje/qjad055.

Muth, J. F., 1961. Rational Expectations and the Theory of Price Movements. Econometrica pp. 315–335, https://doi.org/10.2307/1909635.

Radner, R., 1972. Existence of Equilibrium of Plans, Prices, and Price Expectations in a Sequence of Markets. Econometrica 40, 289–303, https://doi.org/10.2307/1909407.

Schroder, H. M., Driver, M. J., Streufert, S., 1967. Human information processing: Individuals and groups functioning in complex social situations. Holt, Rinehart and Winston New York.

Simon, H. A., 1955. A Behavioral Model of Rational Choice. The Quarterly Journal of Economics 69, 99–118, https://doi.org/10.2307/1884852.

Simon, H. A., 1990. Bounded rationality. Utility and probability pp. 15–18, https://link.springer.com/chapter/10.1007/978-1-349-20568-4_5.

Stigler, G. J., 1961. The Economics of Information. Journal of political economy 69, 213–225, https://www.jstor.org/stable/1829263.

Thaler, R. H., 2018. From Cashews to Nudges: The Evolution of Behavioral Economics. American Economic Review 108, 1265–1287, https://doi.org/10.1257/aer.108.6.1265.

Appendix A. Appendix

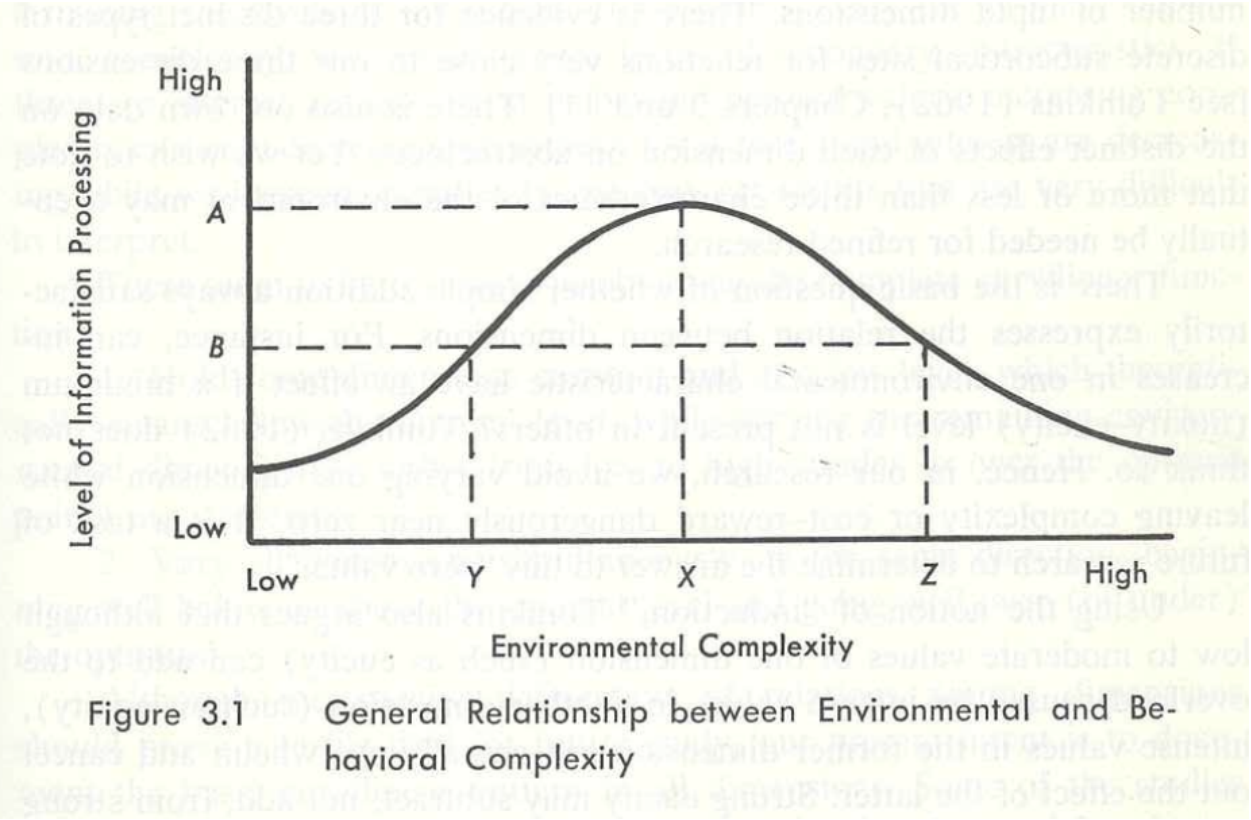


Figure 3.1 General Relationship between Environmental and Behavioral Complexity

Fig. A1. Human Information Overload

The figure shows the inverted U-shape documented on page 37 the book Schroder et al. (1967) describing information overload in humans.

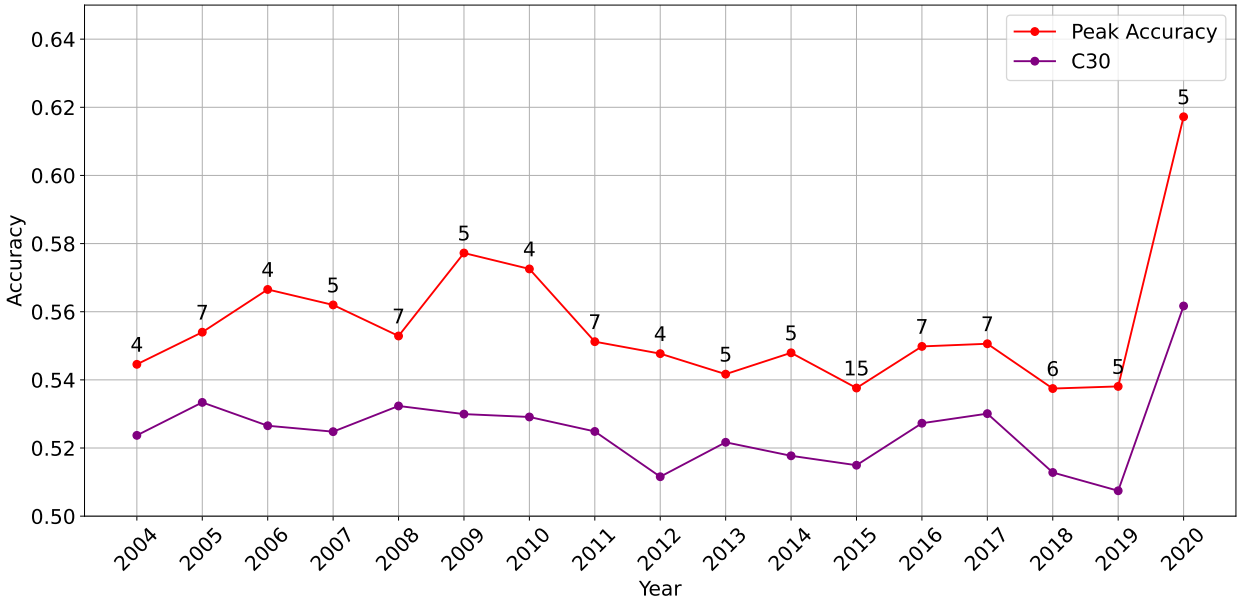


Fig. A2. Temporal Variation in Peak Performance

This figure presents the match rate of the LLM's predictions over time. The red line represents the peak match rate in each year, with the corresponding number indicating the number of earnings call sections included to achieve optimal performance. For instance, a value of 5 means that 50 percent of the current quarter's earnings call transcript was used. The purple line represents the match rate when the model is provided with the full transcript of the current and two prior quarters (C30). The x-axis denotes the year of prediction, while the y-axis represents the match rate.

Table A1: Annualized Average Returns for Long-Short Portfolios

This table reports annualized average returns for the Long, Short, and Long-Short portfolios based on daily returns, comparing the no-context (C_0) and context (C_1) strategies. Results are presented for the three Llama models: Llama Small, Llama Medium, and Llama Large.

		Equally Weighted		Value Weighted	
		C_0	C_1	C_0	C_1
Long	Llama Small	0.46	0.45	0.25	0.20
	Llama Medium	0.50	0.47	0.26	0.23
	Llama Large	0.49	0.49	0.26	0.24
Short	Llama Small	-0.43	-0.40	-0.03	0.02
	Llama Medium	-0.45	-0.44	-0.02	0.03
	Llama Large	-0.49	-0.48	-0.02	0.03
Long-Short	Llama Small	0.89	0.86	0.28	0.18
	Llama Medium	0.95	0.91	0.27	0.20
	Llama Large	0.98	0.97	0.28	0.22

Table A2: Annualized Standard Deviations for Long-Short Portfolios

This table reports annualized standard deviations for the Long, Short, and Long-Short portfolios based on daily returns, comparing the no-context (C_0) and context (C_1) strategies. Results are presented for the three Llama models: Llama Small, Llama Medium, and Llama Large.

		Equally Weighted		Value Weighted	
		C_0	C_1	C_0	C_1
Long	Llama Small	0.23	0.23	0.20	0.21
	Llama Medium	0.23	0.23	0.21	0.20
	Llama Large	0.23	0.23	0.21	0.21
Short	Llama Small	0.29	0.29	0.25	0.25
	Llama Medium	0.29	0.29	0.24	0.25
	Llama Large	0.28	0.29	0.24	0.24
Long-Short	Llama Small	0.17	0.17	0.15	0.14
	Llama Medium	0.17	0.17	0.15	0.15
	Llama Large	0.17	0.17	0.15	0.14

Table A3: Equal-weighted Portfolio Returns: Model Comparison

This table presents the regression results of $r_t(C_1, M)$ on the Fama-French five factors and $r_t(C_0, M)$ across different LLM sizes for equally weighted portfolios. The table displays the estimated intercept (Alpha) and coefficients for the market excess return (Mkt-RF), size (SMB), value (HML), profitability (RMW), and investment (CMA) factors. The coefficient on $r_t(C_0, M)$ captures the relationship between the contextualized and non-contextualized return signals. Standard errors are shown in parentheses. Statistical significance is indicated by *, **, and *** at the 10%, 5%, and 1% levels.

	Llama Small		Llama Medium		Llama Large	
	(1)	(2)	(3)	(4)	(5)	(6)
Alpha	0.003***	0.000***	0.004***	0.000***	0.004***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mkt-RF	−0.064***	−0.004	−0.071***	−0.010*	−0.054***	−0.011*
	(0.015)	(0.007)	(0.015)	(0.006)	(0.015)	(0.006)
SMB	−0.011	0.006	−0.015	−0.011	−0.015	−0.006
	(0.030)	(0.013)	(0.029)	(0.012)	(0.029)	(0.012)
HML	−0.046*	0.019*	−0.061**	0.002	−0.078***	−0.016
	(0.026)	(0.011)	(0.025)	(0.010)	(0.025)	(0.011)
RMW	0.152***	0.021	0.137***	0.005	0.117***	−0.002
	(0.040)	(0.017)	(0.039)	(0.016)	(0.039)	(0.016)
CMA	−0.010	−0.046**	−0.015	−0.009	0.007	−0.009
	(0.054)	(0.023)	(0.053)	(0.022)	(0.053)	(0.022)
$r_t(C_0, M)$		0.902***		0.899***		0.912***
		(0.007)		(0.006)		(0.007)
Observations	4,015	4,015	4,015	4,015	4,015	4,015
R-squared	0.014	0.821	0.017	0.830	0.013	0.827

Table A4: Value-weighted Portfolio Returns: Model Comparison

This table presents the regression results of $r_t(C_1, M)$ on the Fama-French five factors and $r_t(C_0, M)$ across different LLM sizes for value-weighted portfolios. The table displays the estimated intercept (Alpha) and coefficients for the market excess return (Mkt-RF), size (SMB), value (HML), profitability (RMW), and investment (CMA) factors. The coefficient on $r_t(C_0, M)$ captures the relationship between the contextualized and non-contextualized return signals. Standard errors are shown in parentheses. Statistical significance is indicated by *, **, and *** at the 10%, 5%, and 1% levels.

	Llama Small		Llama Medium		Llama Large	
	(1)	(2)	(3)	(4)	(5)	(6)
Alpha	0.001***	0.000	0.001***	0.000	0.001***	0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mkt-RF	−0.034***	−0.015	−0.031**	−0.000	−0.039***	−0.022**
	(0.013)	(0.010)	(0.013)	(0.011)	(0.013)	(0.010)
SMB	−0.001	0.010	−0.012	−0.001	−0.032	−0.003
	(0.025)	(0.020)	(0.025)	(0.021)	(0.025)	(0.020)
HML	−0.094***	−0.032*	−0.120***	−0.082***	−0.069***	−0.029*
	(0.022)	(0.018)	(0.022)	(0.018)	(0.022)	(0.017)
RMW	0.062*	0.028	0.060*	0.047*	0.019	−0.001
	(0.034)	(0.028)	(0.034)	(0.028)	(0.034)	(0.027)
CMA	0.085*	0.017	0.111**	0.098***	0.018	0.031
	(0.045)	(0.037)	(0.046)	(0.037)	(0.045)	(0.036)
$r_t(C_0, M)$		0.561***		0.573***		0.593***
		(0.013)		(0.012)		(0.013)
Observations	4,015	4,015	4,015	4,015	4,015	4,015
R-squared	0.011	0.340	0.014	0.354	0.009	0.362

A.1. Context prompt

User Prompt

Here is the header of a piece of news: `headline`

Here are the news for the same company from the previous `N` days: `context`

Appendix B. Variable definitions

Unless otherwise notes, capitalized text in brackets refer to Compustat variable names.

Variable	Definition and description
Earnings Per Share	Earnings Per Share (Diluted) - Excluding Extraordinary Items (EPSFX)
Earnings variability	Follows Francis, LaFond, Olsson, and Schipper (2004)
Assets	Total Assets (AT)
Beta	60-month beta following Fama and MacBeth (1973)
Book Equity	Stockholders equity plus deferred taxes (TXDB) plus investment tax credit (ITCB) minus post-retirement benefit asset (PRBA) minus preferred stock. Stockholders equity is either total stockholders equity (SEQ), or if missing then total common equity (CEQ) plus preferred stock par value (PSTK), or if missing then total assets (AT) minus total liabilities (LT) plus minority interest (MIB). Preferred stock is either preferred stock redemption value (PSTKRV), or if missing then preferred stock liquidating value (PSTKL), or if missing the preferred stock carrying value (PSTK).
Coverage	monthly percentile rank of the number of headlines and articles for the same firm over the past 30 days
Market capitalization	Number of shares outstanding (SHROUT) multiplied by the absolute value of price (PRC) in CRSP at the end of the month closest to Compustat DATADATE
Momentum	Follows Jegadeesh and Titman (1993)
Risk	idiosyncratic volatility, estimated as the standard deviation of daily market-adjusted returns, multiplied by the square root of the number of observations in the 365 calendar days leading up to the firm's fiscal year-end
Volatility	Standard deviation of stock returns over the most recent ten available trading days ending one day prior to the event.