Scaling Core Earnings Measurement with Large Language Models

Matthew Shaffer*

University of Southern California

Charles C.Y. Wang

Harvard Business School

March 21, 2025

Abstract

We study the application of large language models (LLMs) to the estimation of core earnings, i.e., a firm's persistent profitability from its central business activities. This construct is central to investors' assessments but increasingly challenging to quantify as financial disclosures become more "bloated" and accounting standards increase non-recurring impacts on GAAP net income. LLMs, with their ability to interpret unstructured text and mimic human reasoning, may be well-suited for this task. Using 10-K filings from U.S. public companies between 2000 and 2023, we employ LLMs with two prompting strategies: (i) a baseline "one-shot" approach providing only a definition of core earnings and the full 10-K, and (ii) a sequential approach using the model to summarize unusual losses, then gains, and then aggregate them. Under the baseline approach, the LLM conflates core earnings with other financial concepts (e.g., EBITDA). However, the sequential approach yields a measure that outperforms GAAP Net Income and Compustat's alternatives in key tests of the properties a valid core earnings measure should have. Our findings demonstrate how LLMs can reduce information processing costs for tasks requiring judgment over complex disclosures, while highlighting the importance of careful prompt engineering.

Keywords: Large Language Models, Core Earnings, Financial Statement Analysis, Generative AI, Accounting, Textual Analysis, Machine Learning, Qualitative Reasoning

JEL: G12, G14, M41, C45, C63, D83

*Shaffer is an Assistant Professor, University of Southern California (`mdshaffe@marshall.usc.edu`). Wang is the Tandon Family Professor of Business Administration at Harvard Business School (`cwang@hbs.edu`).

We thank conference participants at The 1st Workshop on LLMs and Generative AI for Finance at the ACM International Conference on AI in Finance (ICAIF), workshop participants at the University of Virginia Darden School of Business, and National Taiwan University, and Richard Sloan, Ethan Mollick, Ethan Rouen, Wilbur Chen, Alex Kim, Brian Baik, and Jesse Gardner for insightful comments. We thank Freyaa Chawla for research assistance. All errors are our own.

1 Introduction

Core earnings, the persistent profitability from a firm's main business activities, is a concept that is fundamental to investors' performance assessment and valuation of firms. However, it is not a standardized or reported GAAP metric, and quantifying it has become increasingly challenging for investors. This difficulty stems from two main factors: the proliferation of complex, verbose financial disclosures ("disclosure bloat") and accounting standards that have increased non-recurring items in GAAP Net Income. The chasm between GAAP earnings and what investors consider "core" earnings has widened, and bridging it has become more challenging (Rouen, So, and Wang, 2021).

Estimating core earnings is a judgment-intensive process that goes beyond numerical calculations or factual retrieval. There is no "one-size-fits-all" algorithmic approach applicable to all companies. It requires integrating unstructured information scattered across financial statements, footnotes, and the Management Discussion and Analysis (MD&A) section. A single income statement line item might combine recurring operational costs and one-time write-offs, requiring cross-referencing with footnotes to quantify each. The task demands not only an understanding of accounting principles but also general business knowledge and the ability to contextualize data points within a firm's operations.

Non-GAAP earnings measures from company management and sell-side analysts are not "neutral" (Gu and Chen, 2004) and often include adjustments for distinct accounting concepts, such as stock-based compensation, or measures like EBITDA. Standardized data providers like Compustat also offer inconsistent measures, sometimes conflating non-recurring and non-"operational" items (Gardner, Sloan, and Yoon, 2024). Thus, valid core earnings measures have historically required human expertise and effort on a firm-by-firm basis, rendering them partially subjective and prone to error — whether of attention (e.g., missing data points in long filings) or reasoning. Despite its fundamental importance, producing valid, neutral measures of core earnings remains a significant and costly endeavor.

Recent advances in large language models (LLMs) could change this. Unlike traditional rule-based ("symbolic") artificial intelligence, LLMs are based on neural networks trained through deep learning on vast amounts of human-generated text. They can "reason" flexibly in ways that mirror human judgment, possess general background knowledge, and have a strong facility with text. These capabilities seem well-suited to core earnings analysis, which requires integrating dispersed textual information, contextualizing it with general background knowledge, and applying judgment.

On the other hand, the nature of language models suggests potential pitfalls for such tasks. Fundamentally, their outputs are sequences of words predicted by a neural network estimated from associations in its training data. This could lead to blending of distinct concepts that are frequently referenced in its training corpus (e.g., core earnings vs. EBITDA vs. cash analysis), causing model outputs to overlook fine distinctions, even while their fluent and articulate outputs make them appear confident and authoritative, persuading human decision makers. Or, crowding the model's input with such long context (i.e., entire 10-K filings) could cause models to fail on the other components of the task, such as making appropriate distinctions among accounting concepts and following instructions – even if the models could be capable of the latter components of the task with a smaller context.

These considerations raise fundamental questions about the applicability of LLMs to estimating core earnings and other tasks of a similar nature. Whether they can effectively perform such analyses could significantly inform the potential of LLMs to reduce investors' information processing costs – an important issue for capital market participants and researchers, especially in light of the increasing complexity and length of financial reports (Blankespoor, deHaan, and Marinovic, 2020). It is also relevant to broader questions about the implications of using such models across professions: as workers offload cognitive labor onto these models, could this come with a hard-to-detect catch?

We develop a process to use LLMs to estimate core earnings from the annual 10-K filings of a large sample of U.S. public companies: scraping filings from EDGAR, extracting the HTMLs to clean text, and making calls to the GPT-4o API provided by OpenAI. Our sample includes roughly 2,000 companies over the 25-year period from 2000 to 2024. At the outset, we expected the LLM's performance on our task to be sensitive to our prompting strategy. However, since our task requires inputting entire 10-K filings, with a median length of over 45,000 words, running analyses for our sample is expensive. Therefore, to balance cost efficiency with our goal of assessing the impact of prompting on model performance, we selected two strategies at opposite ends of the spectrum: one representing "out of the box" performance and the other designed to maximize potential utility.

In the first, the baseline approach, we provide the LLM with a one-shot prompt including a definition of core earnings, the entire 10-K text, and instructions to estimate core earnings and provide its reasoning, without further guidance. This strategy captures how the model performs end-to-end without procedural oversight. The performance of the model under this approach was of interest to us for two reasons: (1) It hinges on the model's "native" reasoning capability, rather than expert guidance; (2) It models a human analyst delegating the entire process to current AIs with minimal effort. Therefore, assessing the model's performance under this strategy can inform us of the consequences of workers using LLMs in this way. We refer to this strategy as the "lazy analyst approach" as a mnemonic.

The second is a structured approach, where we provide the LLM with a full 10-K and instruct it to follow three sequential steps: (1) identifying unusual expenses/losses, then (2) identifying unusual income/gains, and then (3) tabulating them all and quantifying core earnings. We refined this approach in early small-sample experiments intended to yield maximum potential performance. We refer to this strategy as the "sequential prompt." The full prompts are reproduced in Section 3.2, and representative examples of the LLM's output under each strategy are in Appendices A and B, respectively.

Qualitatively, as seen in the example in the appendix, we observed that the LLM's output from the lazy analyst approach frequently made conceptual errors, appearing to conflate core earnings with other distinct financial analysis concepts (e.g., EBITDA and cash flow analysis, etc.). For instance, under this strategy, the model often adjusted for interest expense (a recurring cost of debt financing), stock-based compensation (a recurring cost for employee retention), and depreciation and amortization (recurring allocations of the costs of required investments).[1] In contrast, the outputs from the sequential prompting strategy

[1]This occurred despite the prompt's definition of core earnings being intentionally crafted to make precise distinctions and anticipate these conflations. The prompt specifies that core earnings represent the recurring "owner's earnings" arising from "ongoing activities, exclusive of ancillary items and one-time shocks." An

appeared aligned with our notion of core earnings in the samples we read. However, since such qualitative evaluations are neither scalable nor entirely neutral, we proceed to quantitative evaluations in large-sample tests.

A good core earnings measure should exhibit three empirical properties. First, it should have high persistence – notably higher than that of net income – if it has successfully excluded net income's transitory components. Second, its adjustments – the difference between the measure and net income – should have a low degree of persistence, if those exclusions are indeed for transitory items.[2] Third, it should help predict future net income, if it indeed tracks its persistent components.

Following prior literature, we perform such tests with the earnings measures scaled in per-share terms, at the firm level and in the pooled panel, and over both short- and medium-term horizons (Barth and Clinch, 2009). In each test, we benchmark the measures against GAAP Net Income – the most readily available summary of period income for investors – as well as Compustat's *OPEPS* and *OIADP*. These Compustat measures, which rely on human analysts parsing financials, are widely used as standardized metrics and, thus, a natural benchmark. Prior research suggests some imperfections in these measures: they may not always capture relevant non-core gains and losses (Rouen et al., 2021) and can be inconsistent in the accounting items they include or exclude (Gardner et al., 2024). Nevertheless, our chosen benchmarks represent the most standard basic ("free") and sophisticated ("costly") alternatives available to investors and researchers.

Our empirical analyses yield four main results. First, we show that the LLM-generated core earnings measures are "smoother" than net income. In pooled autoregressive (AR_1) regressions, they exhibit significantly higher persistence parameters than net income. Sec-

accounting expert would interpret these keywords to mean that recurring costs of credit financing should not be excluded ("owner's earnings"), and the absence of terms like "operational" avoids ambiguities related to depreciation and amortization. However, the model's output appears highly fluent and could be judged authoritative by users without the expertise or confidence to discriminate.

[2] Rouen et al. (2021) motivates these properties. It explains why neither of the first two properties alone is sufficient as a test. For instance, earnings figures higher up the income statement (e.g., revenues) may exhibit higher persistence, but by excluding significant, persistent core expenses integral to the firm's operations, they become poor measures of core earnings despite their high persistence. Conversely, if a measure excludes only a small portion of transitory earnings from net income, the adjustments may show low persistence, but the resulting core earnings measure would still be inadequate.

ond, we show that the core earnings measure generated from the sequential LLM approach effectively removes only transitory items from net income, in contrast to most of the other earnings measures. Its adjustments relative to net income have a statistically insignificant persistence coefficient in autoregressive models, indicating that they are non-recurring.

Third, we show that the LLM-generated core earnings measures perform relatively well in predicting future net income. We employ several approaches to test this property, starting with the most straightforward: comparing the absolute 'error' of each earnings measure relative to the next period's Net Income. The sequential LLM prompt's core earnings measure has a mean absolute error of $1.62 per share, outperforming contemporaneous Net Income but slightly trailing Compustat's *OPEPS*. We then conduct regression tests to provide a more flexible and general assessment of the measures' predictive ability. In regressions comparing the earnings measures' ability to predict net income one period ahead, with year fixed effects included to adjust for common time-series factors, the LLM-based core earnings measures yield the largest coefficients. Further, when we extend the horizon to predict net income in two years, the relative performance of the core earnings estimate from the sequential prompt improves. This suggests that this approach may be more effective at identifying and removing non-core gains and losses from company operations.[3] Finally, we implement within-firm predictive regressions, i.e., estimating separate intercept and slope parameters for each firm. These models arguably best reflect what a fundamental investor might consider the "predictive validity" of an earnings measure – that is, the predictive value of earnings innovations within a particular firm rather than across the entire sample. In these tests, we see a similar pattern: the core earnings estimate derived from the LLM sequential prompt has similar performance to the Compustat measures at the horizon of one period ahead, but outperforms at a horizon of t+2.

Finally, we examine how the core earnings measures relate to market valuations. In efficient markets, stock prices reflect investors' expectations of future earnings. Therefore,

[3] As Rouen et al. (2021) explain, non-core earnings comprise both purely transitory items with very low short-run persistence and non-core operational items that may exhibit some short-run persistence but not over longer horizons. If *OPEPS* excludes fewer of these latter items from net income than our LLM-based measure, it may perform better at the one-year horizon but not over a two-year average.

an effective core earnings measure – which better predicts future earnings – should have a stronger association with stock prices. However, "warranted" valuation multiples vary significantly across industries and time, with systematic differences in growth and discount rates. Thus, we estimate a regression model of future prices on current earnings that accounts for industry-year variation, essentially estimating the average valuation multiple for each industry in each year for each earnings measure. We then compare how well the different earnings measures explain variation in stock prices at several different horizons. In the first test, we compare how well the earnings measures explain market prices at the end of the fiscal year. Here, the Lazy Analyst Core Earnings per Share and OIADP — i.e., the higher income-statement level constructs — slightly outperform, though all measures produced similar adjusted R^2 values. However, when we extend the horizon to two years ahead, the Sequential Prompt Core Earnings measure once again performs relatively better, achieving the highest adjusted R^2 and outperforming Compustat's OPEPS and net income by economically meaningful margins. These results are consistent with the LLM-based core earnings measures better capturing the persistent components of earnings that are reflected in market valuations over longer horizons.

Overall, our results suggest that the earnings measure from the sequential prompting strategy is a valid measure of core earnings, and does just as well as or often outperforms Compustat's *OPEPS*, its closest competitor, in key validation tests. The performance of this LLM-based measure – based on an API call costing less than one dollar and one minute of compute time on average per firm – is striking, particularly given the time- and cost-intensive processes associated with the alternatives. S&P Global makes significant investments in collecting and standardizing financial data. Our results suggest that large language models have enormous potential for lowering the costs associated with processing and analyzing the increasingly bloated financial disclosures of publicly traded companies. GPT-4 was the first accessible model to be even plausibly considered for this task. Given the current level of interest, investment, and research progress in AI systems based on LLMs, it seems possible that, in the years to come, using procedures like ours with future models could become standard practice for quantifying constructs like these.

Our study contributes to the literatures on core earnings and AI applications in accounting and finance in several ways. First, we provide empirical evidence on the utility of LLMs in core earnings estimation. While Rouen et al. (2021) demonstrated the value of core earnings adjustments in predicting future firm performance using a proprietary dataset generated by human-assisted "robo" analysts, we offer a computational, reproducible, and scalable approach. Second, our paper represents a significant step in the academic study of these models' application in finance and accounting. Several prior working papers in this nascent and fast-growing area (see Section 2.2 for a review) have scoped out LLMs' performance on a range of separate task categories, e.g., numerical, qualitative, and factual. We believe the most distinctive and relevant application of future LLMs will be in tasks that, like ours, blend background knowledge, reasoning, and integration across unstructured text, tasks mirroring white-collar professionals' work.

Our findings should be of direct interest to investors, as we provide a practical guide for employing these models in core earnings estimation. Our results offer empirical support for anecdotal claims that these models can fail when used "out of the box," on complex tasks without sufficient guidance; but can perform remarkably well when properly guided. While we used the API to scale up for our empirical study, investors and analysts could one day adopt our strategy using user-friendly interfaces like the ChatGPT app (if and when OpenAI expands the allowable textual input length on the app to match the 128k context window accommodated by the API). Many such professionals are eager to offload some of their cognitive labor to AIs – but the consequences for the quality of their work are yet to be seen. Thus, our study is germane to broader conversations about AI's employment in white-collar knowledge work.

The remainder of this paper is structured as follows: Section 2 provides a review of relevant literature, covering both prior approaches to core earnings analysis and recent applications of LLMs in accounting and finance. Section 3 describes our sample and variable construction, including our approach to LLM prompt design and analysis. Section 4 presents our findings. Section 5 discusses some important caveats. In particular, we discuss the question of look-ahead bias in our study and others of its type. As we note, because such large neural networks are not mechanistically interpretable, it is impossible to answer this question with certainty either way, but there are several reasons to think it should not cause significant bias in our study. Even so, for the purpose of retrospective analysis of financial performance, it would not undermine the usefulness of our measure. Indeed, Compustat's *OPEPS*, which is widely used by researchers and which we use as the most competitive benchmark in our tests, explicitly incorporates retrospective adjustments by design (Rouen et al., 2021). Finally, Section 6 concludes.

2 Related Literature

2.1 Core earnings

Core earnings represent a firm's sustainable and recurring profitability derived from its primary business activities, excluding the effects of transitory shocks, non-recurring events, and ancillary activities. This concept is central to investors' assessments of performance and valuation judgments. However, quantifying core earnings has become increasingly challenging. Over time, components of GAAP earnings stemming from ancillary business activities or transitory shocks have grown significantly in both frequency and magnitude, and financial reports have become more bloated, making it difficult for investors and analysts to discern a company's core economic performance from them (Rouen et al., 2021).

Different practitioners have developed different approaches to address this challenge. Sell-side analysts often report and forecast firms' earnings on a non-GAAP basis, commonly referred to as "street earnings." These measures typically exclude items deemed transitory or not reflective of central business activities. Gu and Chen (2004) studied analysts' specific choices of inclusions and exclusions in their street earnings measures and found that these choices largely reflect legitimate economic rationales. However, they also noted inconsistencies across different firms, which could indicate either bias or appropriate firm-specific judgments. Similarly, managers frequently report non-GAAP "pro forma" earnings, excluding items they consider unimportant for understanding firm performance. Curtis, Mcvay, and Whipple (2014) examined managers' choices regarding their preferred non-GAAP metrics and demonstrated that these measures are often more informative than GAAP earnings. Nonetheless, they also identified that an economically significant proportion of firms appear opportunistic, disclosing non-GAAP earnings information primarily when it enhances investors' perceptions. These findings underscore that there is a demand for alternative earnings measures that better track a firm's ongoing operations, but the measures from analysts and managers lack consistency and neutrality.

Academic researchers have aimed to systematically estimate core earnings to address these limitations. Gardner et al. (2024) provide a thorough discussion of the demand for core earnings measures that are adjusted for nonrecurring items, the various proxies and measures that have been used by practitioners and academics, and their limitations. These proxies include I/B/E/S "actuals," which refer to realized earnings based on non-GAAP metrics commonly forecasted by analysts, and several measures from Compustat – all of which have apparent inconsistencies. The authors employ unobserved components modeling (UCM) with a Kalman smoother, incorporating future earnings data to distinguish firms' recurring and nonrecurring earnings components. This approach quantifies a latent core earnings estimate on an *ex post* basis for a large sample. They find that market prices behave as if investors react to the modeled recurring earnings measure, i.e., they price stocks as if on these measures. Additionally, they flag cases with significant deviations between their modeled recurring earnings and standard measures to identify common categories of nonrecurring impacts on net income not captured by standardized sources.

As they emphasize, their measure is inherently based on *ex post* information and is not one that investors or academics could construct and use *ex ante*. However, their study underscores the demand for core earnings measures, investors' implicit use of this construct, the deficiencies of currently used surrogates, as well as the fact that the errors in Compustat measures are identifiable based on disclosures in public 10-K filings — findings that help motivate our approach.

2.2 Textual Analysis and Large Language Models

Accounting researchers have long sought to incorporate the extensive textual information, beyond summary financial statement measures, in firm disclosures. However, it has been challenging to do this empirically. Studies employing textual analysis prior to the advent of LLMs often relied on predefined dictionaries or simple word-counting methods, which may not fully capture the contextual meaning of language as a human reader would interpret it (Davis, Piger, and Sedor, 2012).

Advances in Large Language Models (LLMs) have the potential to dramatically change textual analysis of corporate disclosures. Unlike traditional rule-based (i.e., 'symbolic') artificial intelligence (AI), LLMs' fundamental architecture is a neural network estimated by deep learning on vast amounts of human-generated text. As such, they can "reason" flexibly in ways that mirror human judgment, have general background knowledge, and are especially capable working with text. At the same time, the adoption of LLMs comes with challenges and tradeoffs. One concern is interpretability. Unlike "symbolic" AI, the decision-making processes of LLMs – through vast arrays of billions of parameters with dense interactions – are almost impossible to interpret mechanistically (Dong, Stratopoulos, and Wang, 2024). Despite these challenges, these models' potential practical utility is evident, which has motivated researchers to begin studying their application to financial and accounting analysis tasks.

One practical application for empirical researchers is to employ LLMs to automate and improve "rote" data collection and classification from textual documents, such as IRS audits (Choi and Kim, 2024) and non-answers on conference calls (de Kok, 2024). However, the models have also shown promise in more analytical and judgment-intensive tasks. Kim, Muhn, and Nikolaev (2023a) investigated the potential of generative AI tools in assessing and uncovering corporate risks by analyzing earnings call transcripts. They developed new measures of firm-level risk exposure related to political, climate, and AI-related risks using the GPT-3.5 model, finding that these AI-generated risk measures were more informative and predictive of firm-level volatility and other economic outcomes compared to traditional methods. Kim, Muhn, and Nikolaev (2023b) explore the potential of generative AI tools, such as ChatGPT, in assisting investors with processing complex corporate disclosures. They found that AI-generated summaries of Management Discussion and Analysis (MD&A) sections and earnings call transcripts were more effective at explaining stock market reactions than the original documents, suggesting that LLMs can distill and interpret the information in a manner consistent with investor judgments. Kim, Muhn, and Nikolaev (2024) explored the potential of GPT-4 in performing financial statement analysis traditionally carried out by professional human analysts. Their study found that the LLM, particularly when using a chain-of-thought prompt, outperformed both human analysts and traditional logistic regression models in predicting the direction of future earnings changes.

In a concurrent study, Kim and Nikolaev (2023) develop a measure of 'contextualized profitability' by conditioning quantitative profitability with embeddings of MD&A text using BERT (Bidirectional Encoder Representations from Transformers). Unlike conversational AI models that have dominated headlines in recent years (such as ChatGPT and Claude), BERT is an 'encoder-only' language model, converting text into high-dimensional vector representations (embeddings) of semantic meaning. They show that conditioning profitability measures on BERT embeddings improves traditional asset pricing models. While we share the general goal of leveraging textual information to refine the use of quantitative financial-statement measures, our study differs significantly. We employ the entire 10-K filing and use GPT-4o, a widely-used general purpose model, to directly analyze the text. This directly quantifies an interpretable core earnings measure with an explicit adjustment process for each firm, aligning with how fundamental analysts might use AI for financial analysis.

In sum, recent findings show that LLMs can perform well on a range of task categories individually: numerical (earnings prediction), qualitative (sentiment classification), factual (professional certification exams), or rote (data collection). However, we believe the most distinctive and relevant application of future LLMs may be in tasks that, like ours, blend background knowledge, reasoning, integration of text, and judgment – tasks that mirror those of human knowledge workers.

3 Empirical Design

3.1 Sample Construction

Our research question requires us to construct a sample of 10-K filings for processing by LLMs and link those analyses to fundamental and market variables for validation. We therefore begin our sample construction with a Compustat-CRSP Merged (CCM) annual panel spanning FYs 2000 to 2023. We first exclude observations lacking a valid Central Index Key (CIK), since that is necessary to link the panel to SEC filings. From there, given the high cost of LLM calls with large contexts, we apply several additional filters to define a constrained, cost-effective sample suitable for our tests. Most notably, we exclude firms with fewer than 15 consecutive years in the CCM sample – a restriction that reduces sample size and costs and seemed appropriate in that many of our tests relate to earnings persistence. Additionally, we exclude observations from financial services and REITs (SIC codes 6000–6999) and utilities (SIC codes 4900–4999) due to their distinct earnings properties. Finally, we remove observations missing values for total assets, net income, or PERMCO. Together, these leave a filtered sample of 45,593 CCM observations.

Next, we scrape the corresponding 10-K filings from EDGAR for these firms to extract textual data necessary for our analysis. We successfully retrieve filings for 35,060 unique CIK-year pairs. We then extract plain text from the 10-K filings using `Beautiful Soup`, a Python library for parsing HTML and XML documents and clean them up for easier processing by language models. For the LLM analyses, we structured our API calls to ensure that the entire input would fit within the token limit of the model used. We estimated that only a small fraction ($<$5%) of the cleaned text version of the 10-K filings in our sample would exceed gpt-4o's token limit. Upon manual inspection of those cases, it appeared that it was predominantly Part 3 of the 10-Ks, which covers governance issues, such as the company's directors, executive officers, and related-party transactions, that would exceed the token limit, while data relevant to adjusted earnings analysis were earlier. Therefore, we judged that truncating this subset of files to fit within the model's token limit was unlikely to materially affect our analysis.

Finally, we employ two different calls or prompting strategies to large language models, described below, which return identifiable adjusted earnings measures for 33,765 and 32,183 of those observations across those two separate approaches. Table 1 quantifies the sample attrition at each stage.

3.2 LLM Calls: Two Prompting Strategies

Our study uses a state-of-the-art large language model (LLM) to estimate firms' core earnings from their 10-K filings. Our motivation is twofold: (1) to understand how these models would perform on this task with their native reasoning capabilities and (2) to scope out how they could best be used to aid practitioners and inform academics' understanding of core earnings. At the outset, we realized that the quality of model outputs would vary with the prompting strategy. However, due to the substantial costs of processing our 10-K sample with state of the art LLMs (exceeding $25k per run on average), we could not implement a large number of different approaches.

We therefore settled on two prompting strategies at opposite ends of the spectrum. The first is what we refer to as the "Lazy Analyst" approach, as a mnemonic shorthand. This was our starting benchmark to see how the model performed when prompted with a high-level definition of core earnings and provided with an entire 10-K, without further procedural or other guidance on *how* to perform the task. At the opposite end of the spectrum, in order to scope the potential utility of these models for estimating core earnings, we refined an approach through experiments on a small holdout sample, where we iteratively read and evaluated the LLM's results. We refer to this as the "Sequential Prompt" approach as a shorthand and describe its procedure in full below.

3.2.1 Lazy Analyst Approach

The Lazy Analyst approach simulates an analyst who relies on the LLM to estimate core earnings without providing detailed guidance or breaking down the task into components. The prompt provides a high-level definition of core earnings and instructs the model to analyze the 10-K text to estimate it. The prompt we used is reproduced in its entirety below:

> You are a financial analyst tasked with determining a company's core earnings based on its 10-K filing. Core earnings represent the persistent profitability of the company's central and ongoing activities, exclusive of ancillary items and one-time shocks. This concept aims to capture the owner's earnings – the sustainable, recurring profitability that accrues to equity holders.
>
> Please analyze the provided 10-K text and estimate the company's core earnings. Start with the reported GAAP net income and make adjustments you deem necessary based on the information in the 10-K. Provide a clear explanation of your reasoning for each adjustment.
>
> Additionally, to make it possible to extract your answer later, please include the following tag at the end of your response, after you finish your reasoning and calculation: "*Core Earnings Calculation (final) = $[your determination]" where [your determination] is the final core earnings amount you calculate.
>
> Here is the full 10-K text: [Cleaned 10-K Text Inserted Here]

API Call Parameters:

- **Model Used**: `gpt-4o`
- **Temperature**: 0.0
- **Max Output Tokens**: 4,000
- **System Message**: "You are a financial expert."

This approach allows us to evaluate the LLM's ability to estimate core earnings without explicit guidance on the procedure. However, we chose the words in the prompt carefully to convey a precise and discriminatory definition of core earnings, and not lead the model down the path of other concepts. For example, the double emphasis in the phrase "owner's earnings – the sustainable, recurring profitability that accrues to equity holders" was intentionally chosen to distinguish it from firm-level earnings concepts such as EBIT and

EBITDA. Similarly, we intentionally excluded the terms "operations" and "operational" in order to avoid ambiguities about depreciation and amortization.

We provide a representative example of the model's output for this approach in Appendix A. Notably, in this case, the model's stated analysis included adjustments for Interest Expense, Depreciation, Amortization, and Stock-Based Compensation. We consider these conceptual errors, given the precise definition of core earnings offered above. This was typical: We observed such adjustments in a large fraction of the analyses from this approach.

3.2.2 Sequential Prompt Approach

The Sequential Prompt approach is the method we refined in tests on a small holdout sample to yield maximum validity: Instead of requesting the core earnings calculation directly, we decompose the task into structured steps. Further, instead of defining core earnings conceptually, we simply ask the model to identify and summarize "unusual items" and then aggregate them into a new adjusted earnings measure. More specifically, we use three threaded API calls, each with its own prompt and response. Note that previous messages are included as context for the next messages in the thread.

1. **Call 1: Identification of Non-Recurring Losses/Expenses**:

   > You are an expert financial analyst with extensive experience.
   >
   > Here's a 10-K. Are there any nonrecurring/unusual expenses in the income statement, cash flow statement, footnotes, or MD&A? Be comprehensive and check your work twice.
   >
   > [Cleaned 10-K Text Inserted Here]

2. **Call 2: Identification of Non-Recurring Gains/Income**:

   > Are there any nonrecurring/unusual income in the income statement, cash flow statement, footnotes, or MD&A? Be comprehensive and check your work twice.

3. **Call 3: Computation of Adjusted Earnings**:

Based on the above, compute a new earnings measure. Start with net income, add back nonrecurring/unusual expenses, subtract nonrecurring/unusual income. No hypothetical values. Express in $ Millions.

Provide a summary table like this example:

CIK	Company Name	Fiscal Year-End	Item	Amount ($m)	Location of Disclosure	Description	Net of Tax
0001234567	XYZ Corp	2024-12-31	Net Income	$113.5	Income Statement	Reported net income.	Yes
0001234567	XYZ Corp	2024-12-31	Tax Cuts and Jobs Act of 2017	+$129.2	MD&A	One-time tax expense due to changes in tax legislation.	Yes
0001234567	XYZ Corp	2024-12-31	(Loss) income from discontinued operations	+$1.5	Income Statement	Loss from operations that have been disposed of or are held for sale.	Yes
0001234567	XYZ Corp	2024-12-31	Asbestos-related benefit, net	-$19.9	Income Statement	Nonrecurring benefit related to asbestos liabilities.	Yes

API Call Parameters:

- **Model Used**: `gpt-4o`
- **Temperature**: 0.0
- **Max Output Tokens**: 4,000
- **System Message**: “You are an expert financial analyst with extensive experience.”

We refined this strategy through iterative experimentation on a small holdout sample at an early stage of the project, intended to maximize performance in estimating core earnings. Our initial intuition in considering this approach is that it leveraged the core capabilities of language models as such in “encoding” the meaning of text, as in summarization (Kim et al., 2023b). In other words, this prompting strategy can be interpreted as an iterated summarization task: First summarize the unusual losses from the text, then summarize the unusual gains, then use each to generate a summary set of adjustments to net income.

We provide a representative example of the model’s output for this approach in Appendix B. Compared to the analysis from the lazy analyst approach, it is less verbose, and does not have the same conceptual errors. This is consistent with our observations from the sample that we read. Qualitatively, based on reading the analyses in full in our initial iterations on our holdout sample, we believed that the sequential prompt approach yielded analyses that better tracked the construct of core owners' earnings, and had fewer conceptual errors. Our interpretation was that in the unstructured baseline approach, the model struggled to simultaneously integrate over long context, extract relevant amounts, and do the requisite reasoning end-to-end. Breaking the approach down into these component "summarization" tasks seemed to leverage the language model's core capabilities while constraining other kinds of errors.

However, these are our interpretations from small-sample experiments and should be considered cautiously. We cannot truly trace the "reasoning process" through the model; we can only observe and evaluate its outputs. With this in mind, we now turn to empirical tests that do so quantitatively.

4 Findings

Below, we report descriptive statistics and validation tests of the properties of the core earnings measures produced by the LLM under the two prompting strategies, benchmarked to GAAP Net Income (NI) itself, as well as two related widely-used measures available from Compustat: OPEPS and OIADP. All of our tests are implemented in per-share terms. We refer to all five of these measures – Net Income per Share, Lazy Analyst Core Earnings per Share, Sequential Prompt Core Earnings per Share, Compustat OIADP per Share, and Compustat OPEPS – collectively as "the earnings measures."

4.1 Descriptive Statistics

Figure 1 plots the measures over time. The core earnings measure from the sequential prompting strategy appears smoother over time than GAAP Net Income and *OPEPS*, while staying closer to the average level of Net Income than *OIADP*. Panels A and B of Table 2 report summary statistics for the raw earnings measures in absolute terms (millions of dollars) and on a per-share basis, respectively. The Lazy Analyst and Sequential Prompt Core Earnings measures have higher mean values ($555.7 million and $487.8 million, respectively) compared to Net Income ($425.0 million). This is unsurprising given that one-time losses are more common than one-time gains under the conservatism and timely loss recognition principles of accounting, and given that the Lazy Analyst approach tends to make other adjustments, i.e., not limiting itself to non-recurring items. Analogously, Compustat's OPEPS is higher than the GAAP Net Income benchmark; and Operating Income After Depreciation (OIADP) is even higher, with the highest mean ($706.9 million), which is unsurprising given that it excludes some recurring expenses by design.

The next panels report statistics in terms of the "adjustments" of the alternate earnings measures, i.e., their deviation from GAAP Net Income. Panel C shows the percentage of observations where adjustments are income-increasing, income-decreasing, or result in no change. The sequential prompt core earnings measure is higher than GAAP Net Income (i.e., income-increasing) only 59.8% of the time, compared to 87.5% for Compustat's OIADP and 63.7% for OPEPS. Finally, Panel D reports the magnitude of the adjustments to Net Income per Share. The sequential prompt measure exceeds GAAP Net Income by less than 10 cents per share at the median (specifically, 9.3 cents), versus 88.5 cents per share for OIADP. These are desirable features in that they indicate that this measure tracks the construct of bottom-line profitability on average while smoothing out over time.

4.2 Earnings Persistence and Predictive Ability

In this section, we present our empirical tests which compare the core earnings measures we derived from our LLM calls vs. GAAP Net Income and the two standard Compustat alternatives. None of these tests alone provides a definitive or conclusive measure of the "best" earnings metric (Dechow, Ge, and Schrand, 2010). For example, firms' total assets exhibit high persistence but represent a fundamentally different construct than profitability. Therefore, AR(1) coefficients alone do not offer a normative ranking of candidate earnings measures. Similarly, a firm's revenue will predict its earnings in regression models; but revenue is an inherently different construct from owner's profits, so its predictive ability doesn't make it a good measure of the latter. No single test statistic can definitively answer the question. Instead, we draw our inferences from the findings as a whole, after comparing the relative performance of the measures on each of the tests of properties that a valid core earnings measure should have.

4.2.1 Persistence of Each Measure

We first estimate autoregressive (AR(1)) models for each measure on a per-share basis. Specifically, we estimate pooled regressions of each earnings measure on its own first lag. The model is specified as:

$$\text{Earnings Measure}_{i,t} = \alpha_i + \beta \times \text{Earnings Measure}_{i,t-1} + \epsilon_{i,t},$$

where Earnings $\text{Measure}_{i,t}$ represents the current value of the earnings measure for firm i, and $\epsilon_{i,t}$ is the error term. The earnings measures are Net Income per Share, Lazy Analyst Core Earnings per Share, Sequential Prompt Core Earnings per Share, Compustat OIADP per Share, and Compustat OPEPS.

Table 3 presents the results of these pooled autoregressive regressions. The persistence coefficients (β) and explanatory power (R^2) vary significantly among the measures. GAAP Net Income per Share exhibits the lowest persistence in terms of both the slope estimate (0.50) and the overall explanatory power (R^2 = 0.20). Both LLM-generated measures – Lazy Analyst and Sequential Prompt Core Earnings per Share – have higher persistence (0.83 and 0.71, respectively) and improved explanatory power (R^2 = 0.59 and 0.43, respectively). Further, both LLM-generated measures have higher persistence and explanatory power than Compustat's OPEPS (0.63, R^2 = 0.31). However, Compustat OIADP has the highest overall persistence (0.94, R^2 = 0.70), followed by the Lazy Analyst core earnings measure—i.e., the "higher income statement levels" appear to have the most pronounced autocorrelation in this pooled test.

In short, both LLM-generated core earnings measures are smoother and more persistent than GAAP Net Income; the sequential prompt core earnings measure is smoother and more persistent than Compustat's OPEPS; and OIADP and the Lazy Analyst core earnings, respectively, have the highest persistence in terms of the slope coefficient and R^2. We present this analysis for completeness but note that persistence metrics alone should not be used to rank candidate earnings measures, for the reasons discussed above. However, the results so far do indicate that the measure yielded from the sequential prompt LLM strategy has the desired properties of tracking the construct of bottom-line owners' profitability on average while smoothing out transitory shocks.

4.2.2 Persistence of Adjustments

One property of a high-quality core earnings measure is that the adjustments made relative to net income should capture non-recurring or transitory items, which by definition should not persist into future periods. We now empirically assess whether the alternative earnings measures have this property. For each such measure, we calculate the adjustment as the difference between the measure and GAAP Net Income (all on a per-share basis). We then estimate autoregressive (AR(1)) models of this measure.

Table 4 presents the results of these regressions estimated on the pooled sample. The adjustments for the Sequential Prompt Core Earnings per Share measure have a persistence coefficient that is statistically insignificant, indicating that these adjustments do not predictably persist into future periods. The adjustments for the Lazy Analyst Core Earnings per Share have a persistence coefficient of 0.076 (significant at the 5% level) indicating that its adjustments are persistent, i.e., not for transitory items. The adjustments for Compustat's OIADP exhibit the highest persistence coefficient of 0.31 (significant at the 1% level), suggesting that these adjustments include more persistent components that recur over time. The adjustments for Compustat OPEPS per Share have an insignificant and near-zero persistence coefficient of -0.0093.

In sum, the Sequential Prompt Core Earnings per Share measure has the desired property that its adjustments are not significantly persistent. Compustat's OPEPS does as well, while both the Lazy Analyst measure of core earnings and Compustat's OIADP lack this property and have recurring adjustments.

4.2.3 Predictive Accuracy of Earnings Measures

Another test of candidate core earnings measures is their predictive ability for future Net Income. GAAP has some “anti-smoothing” components (such as asset impairments) that lead to volatility in any given period. However, GAAP accounting mechanics ensure that accruals, including one-time events, eventually reconcile with related cash flows, and Net Income averages out to equity owners’ residual over time. Therefore, testing how well various earnings measures predict expected Net Income can provide a better assessment of their performance as core owners’ earnings measures. To the extent that a measure includes persistent, but not transitory, components of Net Income, it should predict its future values.

Therefore, in our next set of tests we begin analyzing the relation of these earnings measures not to their own future values, but to future bottom-line Net Income. There are several potential ways to operationalize this in a test. We begin with the most direct and simple such test: the absolute “error” of each earnings measure relative to the next period’s Net Income. To provide a very direct measure of different earnings measures’ utility in predicting future bottom-line profitability, we quantify the absolute difference between the next period’s Net Income per Share and the current period’s earnings measure:

$$\text{Absolute Prediction Error}i,t = \left|\text{NI}i,t+1 - \text{Earnings Measure}_{i,t}\right|$$

where $\text{NI}i,t+1$ is the next period’s Net Income per Share. Table 5 presents the mean and median absolute prediction errors for each measure.

Compustat OPEPS has the lowest mean and median absolute errors ($1.55 and $0.52 per share, respectively). Closely following is the Sequential Prompt Core Earnings per Share, with mean and median errors of $1.62 and $0.63 per share. This indicates that our structured LLM approach results in a measure that tracks future profitability more closely than GAAP Net Income itself (mean $1.78, median $0.60). The Lazy Analyst measure (mean $1.73, median $0.65) does not outperform raw Net Income on median prediction error. This is unsurprising given our observations that it often made exclusions for recurring expenses. Relatedly, Compustat OIADP per Share has the highest prediction errors (mean

$2.13, median $1.00), as it does not track bottom-line Net Income.

4.2.4 Predictive Ability for Future Earnings

The absolute prediction error analysis has obvious limitations: Investors generally expect and work to forecast growth in earnings power over time. Growing firms will inherently have differences between their current and future earnings even if and when their earnings trajectories are highly predictable. A standard method for validating candidate earnings that remedies that limitation is to estimate regression models of how the measures in one period predict Net Income in later periods. If a candidate core earnings measure includes the persistent components of Net Income, as desired, that should be reflected in predictive value in regressions, even when earnings are growing.

We first estimate a pooled panel regression of forward Net Income on each earnings measure, including year fixed effects to isolate common time-series factors:

$$\text{NI}_{i,t+1} = \beta_0 + \beta_1 \times \text{Earnings Measure}_{i,t} + \gamma_t + \epsilon_{i,t},$$

Here, $\text{NI}_{i,t+1}$ is the Net Income per Share for firm i in period $t+1$, $\text{Earnings Measure}_{i,t}$ represents the current period earnings measure per share for firm i, β_0 is the intercept term, β_1 estimates the change in future Net Income predicted by a unit difference in each earnings measure, γ_t represents year fixed effects that control for time-specific factors common to all firms, and $\epsilon_{i,t}$ is the error term. We cluster standard errors at the firm level to account for serial correlation.

Table 6 presents the regression results. For Net Income per Share, the coefficient is 0.44, with an R^2 of 0.16. This implies that a $1 difference in Net Income per Share predicts on average a $0.44 difference in next period's Net Income per Share, and that the model explains about 16% of the variance in future Net Income. The Sequential Prompt Core Earnings per Share has a stronger predictive relationship, with a coefficient of 0.64 and an R^2 of 0.23, also outperforming Compustat's OPEPS (coefficient 0.54, R^2 of 0.19) on this measure. The Lazy Analyst Core Earnings per Share has a coefficient of 0.65 and an R^2 of

0.23, while Compustat's OIADP has a coefficient of 0.60 and an R^2 of 0.29.

4.2.5 Longer-Term Predictive Ability

We next examine how well the earnings measures predict Net Income per Share two fiscal periods ahead, again including year fixed effects. Specifically, we estimate:

$$\mathrm{NI}_{i,t+2} = \beta_0 + \beta_1 \times \text{Earnings Measure}_{i,t} + \gamma_t + \epsilon_{i,t},$$

where $\mathrm{NI}_{i,t+2}$ is the Net Income per Share for firm i two periods after the current reporting period, and γ_t represents year fixed effects that adjust for common time series factors. Table 7 presents the results, with standard errors clustered by firm.

Compared to the one-year-ahead regressions, we find that the Sequential Prompt Core Earnings per Share measure improves its relative performance at this longer horizon, with the highest coefficient, of 0.66, and an R^2 of 0.22. Compustat's OIADP has a coefficient of 0.58 and an R^2 of 0.22, i.e., tied on overall explanatory power but with a lower predictive relationship coefficient. The Lazy Analyst Core Earnings measure has a coefficient of 0.60 with an R^2 of 0.17, while Compustat's OPEPS has a coefficient of 0.54 and an R^2 of 0.15. Finally, Net Income per Share itself shows the lowest predictive capacity in these two-year-ahead forecasts, with a coefficient of 0.38 and an R^2 of 0.10. Notably, the relative performance of the "higher income statement level" measures, OIADP and the Lazy Analyst core earnings estimate, declines at these horizons, consistent with our theory, as they remove some persistent components.

4.2.6 Firm-Level Earnings Persistence

Next, we estimate firm-specific predictive regressions of one-period-ahead Net Income per Share on the earnings measures, i.e., estimating separate intercept and slope parameters for each firm. Doing so isolates firm-specific earnings evolutions from the influence of cross-sectional variation on the estimated parameters of interest. This modeling approach maps more closely to what investors might consider the "predictive validity" of an earnings measure, i.e., the predictive value of earnings innovations within the firm itself rather than as estimated across the entire sample.

We estimate the following regression model for each firm i with a minimum of four observations in the main panel:

$$\text{NI}_{i,t+1} = \alpha_i + \beta_i \times \text{Earnings Measure}_{i,t} + \epsilon_{i,t},$$

where $\text{NI}_{i,t+1}$ is the Net Income per Share for firm i in period $t+1$, α_i is the firm-specific intercept, $\text{Earnings Measure}_{i,t}$ is the specified earnings measure per share for firm i in period t, β_i is the firm-specific coefficient representing the predictive relationship between the earnings measure and future Net Income, and $\epsilon_{i,t}$ is the error term. We collect the estimated coefficients (β_i) and R^2 values from the individual firm-level regressions. We then compute the average coefficient, average R^2, and the number of firms to summarize the predictive performance by earnings measure.

Table 8 presents the results. The Sequential Prompt Core Earnings per Share measure has an average R^2 at 0.27, indicating that firm-specific deviations in this measure explain approximately 27% of the variation in future Net Income per Share. It also has a coefficient of 0.42. Compustat's OIADP (average R^2 of 0.28) and OPEPS (average R^2 of 0.28) have slightly higher explanatory power, with OPEPS having the highest coefficient (0.54). Net Income per Share (average R^2 of 0.24, coefficient of 0.40) and the Lazy Analyst measure (average R^2 of 0.25, coefficient of 0.42) have lower predictive ability.

4.2.7 Extending the Horizon to $t+2$

We next extend our firm-level regression analysis to a two-period-ahead horizon. Specifically, we estimate the following model for each firm i with a minimum of four observations:

$$\text{NI}_{i,t+2} = \alpha_i + \beta_i \times \text{Earnings Measure}_{i,t} + \epsilon_{i,t},$$

where $\text{NI}_{i,t+2}$ is the Net Income per Share for firm i in period $t+2$, α_i is the firm-specific intercept, $\text{Earnings Measure}_{i,t}$ is the specified earnings measure at period t, β_i is the firm-specific slope coefficient relating the current earnings measure to net income two periods later, and $\epsilon_{i,t}$ is the error term. As before, we compute each firm's coefficient (β_i) and R^2, then report the average values across firms.

Table 9 reports the results. The Sequential Prompt Core Earnings per Share measure has a relatively better predictive advantage at this longer horizon, with the highest average R^2 at 0.22 and a coefficient of 0.31. Compustat's OIADP performs second best in terms of explanatory power (average R^2 of 0.19) but has a lower coefficient (0.19), while Compustat's OPEPS has a similar coefficient to the Sequential Prompt measure (0.31) but lower explanatory power (average R^2 of 0.19). The Lazy Analyst measure (average R^2 of 0.18, coefficient of 0.22) and Net Income per Share (average R^2 of 0.16, coefficient of 0.25) show the weakest predictive ability at this extended horizon.

4.3 Association with Market Valuations

Next, we examine the earnings measures' associations with firms' valuations, all scaled on a per-share basis. This approach seeks to capture which earnings measures are most closely aligned with the constructs investors implicitly use in pricing equities. The conceit is that investors value shares based on their notion of recurring and future profitability multiplied by a common price-to-earnings (P/E) multiple. Regressing market prices on the earnings measure implicitly estimates a common multiple for them. Therefore, comparing the R^2 from models using different earnings measures can be interpreted as a measure of which construct they are pricing stocks "as if" on.

However, regressing price on earnings measures alone has an obvious limitation: That regression model implicitly applies a uniform earnings multiple across all firms, but warranted multiples vary with growth expectations and discount rates. Relatedly, in practice, investors typically benchmark their valuations to industry peers when using price-to-earnings multiples.

4.3.1 Goodness of Fit with Industry-Specific Valuation Multiples

To address the limitations of imposing a single common earnings multiple across all firms and to better capture the variation in "warranted" multiples due to differences in growth and discount rates, we estimate models in which each earnings measure is interacted with industry-year dummy variables. This approach allows the valuation multiple to vary by industry and year, and better aligns with how investors benchmark multiples.

We estimate the following regression model for each earnings measure:

$$\text{Price}_{i,t+1} = \sum_j \beta_j \left(D_{j,i,t} \times \text{Earnings Measure}_{i,t} \right) + \epsilon_{i,t},$$

where $\text{Price}_{i,t+1}$ is the stock price per share for firm i at the close of the next fiscal year (i.e., relative to the fiscal year of the annual 10-K used to derive the earnings measures), $\text{Earnings Measure}_{i,t}$ is the earnings measure per share, $D_{j,i,t}$ is a dummy variable for industry-year group j (based on two-digit SIC codes and fiscal year) to which firm i belongs at time t, β_j is the estimated coefficient representing the industry-year-specific valuation multiple on the earnings measure, and $\epsilon_{i,t}$ is the error term.

Table 10 presents the adjusted R^2 values from these regressions for each earnings measure. (The coefficients are not reported, since there is a separate point estimate for each of the 1,342 industry-year groups.) The adjusted R^2 values are high in general, indicating that industry-year level earnings multiples explain a large portion of the variation in valuations. In comparing the candidate earnings measures, Lazy Analyst Core Earnings per Share has the highest adjusted R^2, Compustat OPEPS follows, and the Sequential Prompt Core Earnings per Share measure is third, though all three are very close. Net Income per Share has the lowest adjusted R^2 at 0.73. Although the Lazy Analyst measure performs best in this test, the differences among the top four measures are very small.

4.3.2 Extending the Horizon: Assessing Value Relevance at $t+2$

Next, we extend the horizon for the price variable. This approach recognizes that markets may not be perfectly efficient in the short term due to information processing costs, limited investor attention, or other frictions. In particular, as we argue, getting valid core earnings estimates is “costly” for investors, both in terms of information processing and/or procuring data access to core earnings measures produced by others. If investors gradually incorporate information about firms’ persistent earnings components into stock prices over time – as they receive additional signals and updates about fundamentals – it is informative to assess whether these earnings measures are associated with market valuations at longer horizons. Extending the horizon of our value relevance tests allows us to evaluate how well the earnings measures capture information that gets reflected in market valuations as the market assimilates more information over time.

As in the previous test at the shorter horizon, we interact the earnings measures with industry-year dummies, to allow the valuation multiples to vary across industries and years, accommodating differences in “warranted” multiples from variation in industry level expected growth and discount rates at any point in time. Observations are included only if the earnings measure is positive and non-missing (since negative valuation multiples are not interpretable), and the forward price at $t + 2$ is available and strictly positive.

Table 11 presents the adjusted R^2 values from these regressions. (The coefficients are not reported, since there is a separate point estimate for each industry-year group.) In this test, the Sequential Prompt Core Earnings per Share measure has the highest adjusted R^2 of 0.76, suggesting that this measure is most closely associated with stock prices two years ahead. The Lazy Analyst Core Earnings per Share measure follows with an adjusted R^2 of 0.74. Both LLM-derived measures outperform the Compustat measures (OIADP per Share and OPEPS), which have adjusted R^2 values of 0.73 and 0.72, respectively. Net Income per Share has the lowest adjusted R^2 of 0.68.

At this longer horizon, performance differences are economically more meaningful. For example, Sequential Prompt Core Earnings per Share produced an adjusted R^2 that is higher than those produced by the Compustat measures by 5-6% and by net income per share by 12%. The results of Table 11 are consistent with the LLM-based measures better capturing the persistent components of earnings that are reflected in market valuations over longer horizons, mirroring our earlier net income forecast regression results.

5 Caveats

Our conclusions are subject to two important caveats: First, our quantitative tests identify several desirable properties for a core earnings measure in large-sample tests. But we cannot be confident that the analyses produced by current LLMs will yield valid core earnings measures for any particular case, or meet the threshold of reliability that would be required for high-stakes decisions such as making concentrated investments. Our results support the strict claims we make – and no more.

Second, a potential inferential threat we face, like all other studies applying recent vintages of LLMs to financial analysis tasks on archival data, is that they were trained on data that spans the sample period, including annual reports, news articles, etc. from subsequent periods relative to the index dates of the 10-K filings we have the LLM analyze. Therefore, it is natural to question whether our validation tests are confounded by "look-ahead bias," i.e., whether the predictive validity of the LLM's core earnings measures reflects its knowledge of the future rather than general reasoning capabilities. Answering this question with scientific certainty is nearly impossible. LLMs are based on neural networks consisting of billions of parameters with complex, dense interactions. Mechanistic interpretation is challenging even for smaller models; with OpenAI's state-of-the-art offerings, it is impossible, as the model weights are not accessible to external researchers.

We offer two arguments for why look-ahead bias is unlikely to detract from the significance of our findings. First, we do not think it is probable that look-ahead bias explains our results. The Sequential Prompt strategy guided the model through a step-by-step process (collecting unusual gains, then unusual losses, then tabulating and aggregating), and we are able to verify that the model's outputs adhere to those steps, as we collect its response to each. Importantly, nothing in the prompts indicated that the model's output would be evaluated based on its relationship with future profitability. Indeed, in the Sequential Prompt strategy, the first two prompts did not even reveal that the final task would be aggregation into an adjusted earnings measure. The initial prompts simply instructed it to collect unusual losses/expenses and then gains/income, and the third prompt instructed it to aggregate the results of the previous chats. Thus, it seems less plausible that look-ahead bias would influence even the initial data collection components prior to the aggregation into an adjusted earnings measure.

However, to offer some empirical assurance on this point, we performed an additional robustness test. Specifically, we anonymized a random sample of 1,000 10-Ks from our original analysis sample (due to cost constraints), using regular expressions to replace all iterations of the filing company's name with "Anonymous Company," and re-ran the Sequential Prompt strategy. We then tested whether the *difference* between the original core earnings measure and the result when using this anonymization was associated with future changes in Net Income. We report the results of this test in Table 12.

We note at the outset that this anonymization is not an entirely satisfactory approach in general. Even with the company name redacted, this does not eliminate the possibility that the LLM could infer the company's true identity from all the other context in the document. On the other hand, it could be argued that anonymization is an "unfair" benchmark because a human analyst—or an LLM used in practice—going forward would have access to the company's name, identity, history, and contemporaneous news, for use in contextualizing its financial statements. Finally, substituting the company name with "Anonymous Company" could activate parts of the language model's neural network that lead to other variances. In short, this anonymization cannot be considered a perfect proxy for how the model would perform without any look-ahead bias or knowledge of the future relative to the index date.

Therefore, while we did not apply this anonymization procedure to the full sample, it is valuable for testing the specific concern about look-ahead bias. This hypothesis posits that the model uses foreknowledge of the firm's future profitability relative to the index year to inform its core earnings calculation, thereby improving performance on the earnings prediction tests. If this were true, we would expect the difference between the original Sequential Prompt core earnings measure and the anonymized one to be positively associated with future changes in Net Income. In other words, the LLM measure derived from the non-anonymized 10-K – where the model knows the firm's identity and potentially its future outcomes – would be biased toward future earnings changes.

However, as indicated in Table 12, the difference between the original Sequential Prompt core earnings measure and the anonymized one is *not* statistically significantly associated with future changes in Net Income. Indeed, the point estimate is negative, which is the opposite of what we would expect under the look-ahead bias hypothesis. Thus, based on the reasoning laid out above, the nature of the sequential prompt, and the results of this robustness test, we do not see strong reasons to believe that look-ahead bias explains our results. However, ultimately, given the near impossibility of mechanistic interpretation of these models, the definitive test will be out-of-sample empirical validation (Dong et al., 2024; Kim et al., 2024) once sufficient post-periods after the models' knowledge cutoff dates occur. Pending that, the anonymization procedure above provides the empirical evidence we can offer for now.

A second argument is that these LLM-derived core earnings measures would be of significant value *even in the presence of visibility of future information*. One use of these measures is to inform the retrospective analysis of historical financial performance. For such purposes, having knowledge of future events would not undermine the utility of our measure. Indeed, Compustat's benchmark operating earnings measures incorporate retrospective adjustments (Rouen et al., 2021): notably, it applies a "three-year rule" for retroactively determining whether an item is non-recurring. Despite the presence of such look-ahead bias, financial researchers use these measures to aid their historical analyses of firm performance. At a minimum, our results suggest that large language models have enormous potential in facilitating this type of work, at scale, on a more timely basis, and at a relatively low cost.

6 Conclusion

Quantifying core earnings – the persistent profitability from a firm's primary activities – is a fundamental task of financial analysis. However, it has become increasingly challenging to deal with "bloated" financial disclosures and the prevalence of non-recurring items affecting GAAP net income. Large Language Models, whose advancements and user interfaces have dominated news headlines for the past two years, seem potentially well-suited for this kind of task, given their facility working with text and their ability to mirror human reasoning and context-specific judgments. We show that a current leading LLM, when provided with carefully designed prompts, can produce estimates of core earnings that appear conceptually sound, track bottom-line owners' earnings, and have several desirable properties, outperforming widely-used proxies on several standard tests.

Our work has significant implications for both research and practice. Extensions of our approach with future models could eventually provide a new standard for core earnings estimates that are consistent, neutral, and scalable, resolving a longstanding challenge in financial analysis. For practitioners, it highlights the potential and pitfalls of using current AI tools for white-collar knowledge work. As many professionals begin integrating AI into their workflows, and offloading some of their cognitive effort, our findings provide timely lessons. While LLMs possess remarkable capabilities, the effectiveness of current models depends on how they are piloted.

References

Barth, M. E. and G. Clinch (2009). Scale effects in capital markets-based accounting research. *Journal of Business Finance and Accounting 36*(3-4), 253–288. ISBN: 0306686X.

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

Choi, G.-Y. and A. G. Kim (2024). Firm-Level Tax Audits: A Generative AI-Based Measurement.

Curtis, A. B., S. E. Mcvay, and B. C. Whipple (2014). The disclosure of non-gaap earnings information in the presence of transitory gains. *Accounting Review 89*(3), 933–958.

Davis, A. K., J. M. Piger, and L. M. Sedor (2012, September). Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language*. *Contemporary Accounting Research 29*(3), 845–868.

de Kok, T. (2024). ChatGPT for Textual Analysis? How to use Generative LLMs in Accounting Research.

Dechow, P., W. Ge, and C. Schrand (2010). Understanding earnings quality: A review of the proxies, their determinants and their consequences. *Journal of Accounting and Economics 50*(2-3), 344–401.

Dong, M. M., T. C. Stratopoulos, and V. X. Wang (2024). A Scoping Review of ChatGPT Research in Accounting and Finance. *SSRN Electronic Journal*.

Gardner, J., R. G. Sloan, and J. S. Yoon (2024, August). Distinguishing between recurring and nonrecurring components of earnings using unobserved components modeling. *Journal of Accounting and Economics 78*(1), 101687.

Gu, Z. and T. Chen (2004). Analysts' treatment of nonrecurring items in street earnings. *Journal of Accounting and Economics 38*, 129–170. ISBN: 01654101.

Kim, A., M. Muhn, and V. Nikolaev (2023a, October). From Transcripts to Insights: Uncovering Corporate Risks Using Generative AI. arXiv:2310.17721 [cs, econ, q-fin].

Kim, A. and V. V. Nikolaev (2023). Context-based interpretation of financial information. *Chicago Booth Research Paper* (23-08).

Kim, A. G., M. Muhn, and V. V. Nikolaev (2023b). Bloated Disclosures: Can ChatGPT Help Investors Process Information? *SSRN Electronic Journal*.

Kim, A. G., M. Muhn, and V. V. Nikolaev (2024). Financial Statement Analysis with Large Language Models.

Rouen, E., E. C. So, and C. C. Wang (2021, December). Core earnings: New data and evidence. *Journal of Financial Economics 142*(3), 1068–1091.

Figures

Fig. 1.
Average Earnings Measures Over Time (Per Share)

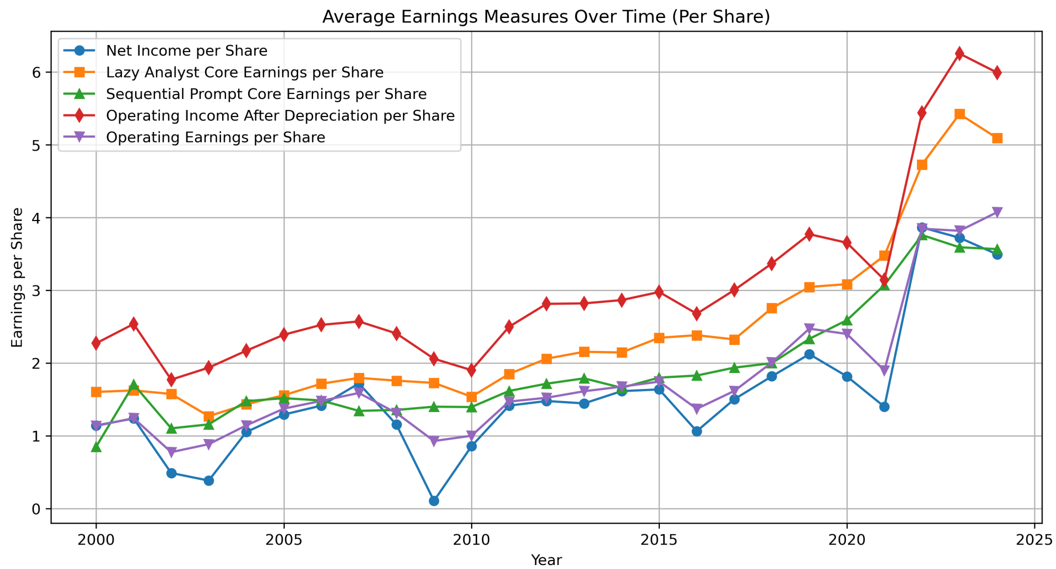

Figure 1 graphs the average earnings measures on a per-share basis from 2000 through 2023. The earnings measures plotted are:

- **Net Income per Share**: Net income divided by the number of shares outstanding
- **Lazy Analyst Core Earnings per Share**: Core earnings estimated using the "Lazy Analyst" LLM prompt, divided by the number of shares outstanding
- **Sequential Prompt Core Earnings per Share**: Core earnings estimated using the sequential LLM prompt approach, divided by the number of shares outstanding
- **Operating Income After Depreciation per Share**: Compustat's OIADP divided by the number of shares outstanding
- **Operating Earnings per Share**: Compustat's OPEPS

Fig. 2.
Distribution of LLM Core Earnings and GAAP Earnings (Per Share)

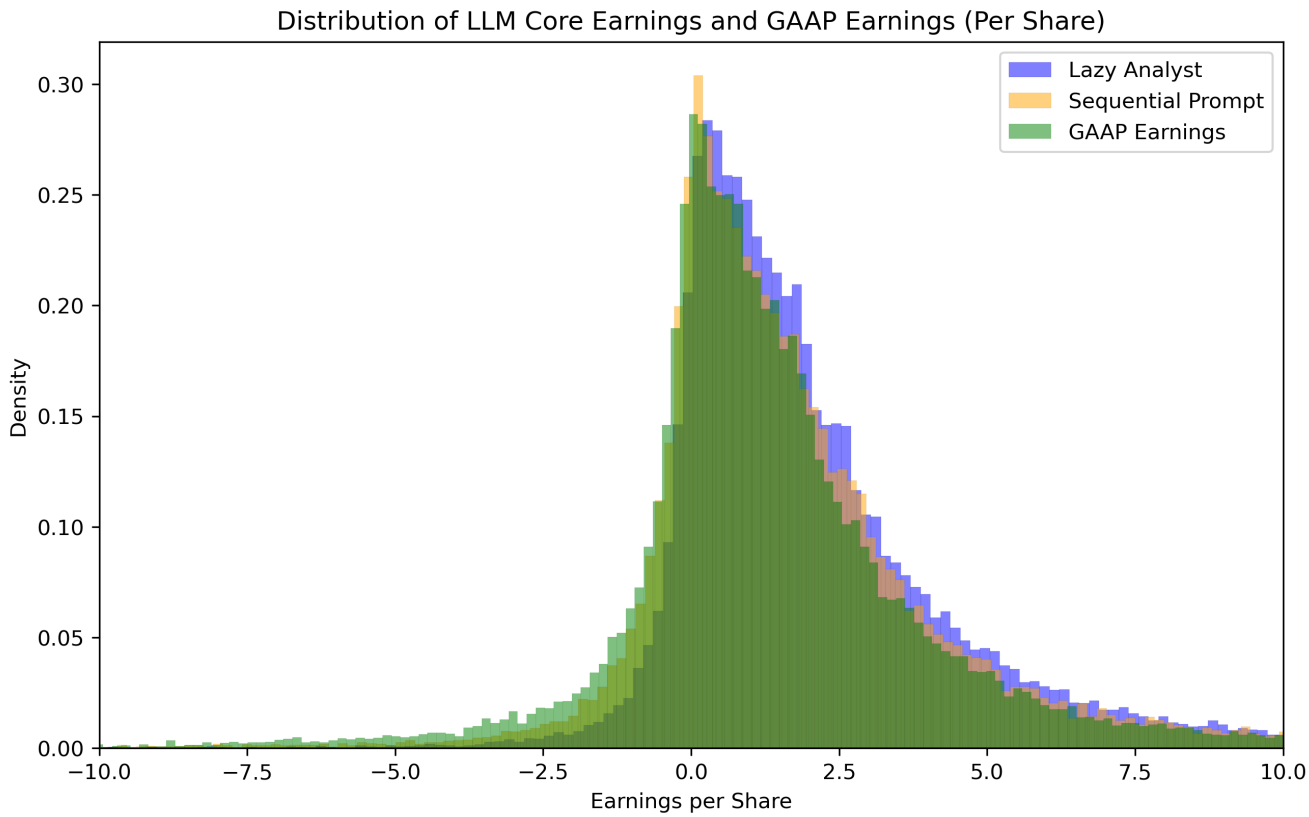

Figure 2 displays histograms of the distribution of GAAP Net Income and the two LLM-based alternative earnings measures, all in per-share terms.

- **Lazy Analyst Core Earnings per Share**: Core earnings estimated using the "Lazy Analyst" LLM prompt, divided by the number of shares outstanding
- **Sequential Prompt Core Earnings per Share**: Core earnings estimated using the sequential LLM prompt approach, divided by the number of shares outstanding
- **GAAP Earnings (Net Income per Share)**: Net Income divided by the number of shares outstanding

Table 1.
Sample Construction

Table 1 quantifies steps of attrition in our sample construction.

Description	Observations
Initial Compustat-CRSP sample, 2000-2023	162,740
Less Observations with missing CIK	(13,053)
Less Firms with fewer than 15 consecutive years of data	(72,983)
Less Financial firms and REITs (SIC 6000-6999)	(16,732)
Less Utilities (SIC 4900-4999)	(5,548)
Less Observations with missing total assets, net income, or PERMCO	(8,831)
Filtered Compustat-CRSP sample (base CIK-year pairs)	45,593
CIK-year pairs with posted and accessible 10-Ks via EDGAR	35,060
Non-missing core earnings measures returned from LLM call with Sequential Prompt approach	33,765
Non-missing core earnings measures returned from LLM call with Lazy Analyst approach	32,183
Observations with valid measures from both approaches	**31,375**

Table 2. Summary Statistics and Adjustments to Earnings Measures

Panel A: Summary Statistics: Raw Earnings Measures ($ Millions)
This panel reports summary statistics for the earnings measures, in millions of dollars. The earnings measures included are:

- **Net Income**: The net income of firms as reported in their financial statements
- **Lazy Analyst Core Earnings**: Core earnings estimated using the Lazy Analyst approach.
- **Sequential Prompt Core Earnings**: Core earnings estimated using the Sequential Prompt approach
- **Compustat OIADP**: Operating Income After Depreciation as reported by Compustat

	count	mean	std	min	25%	50%	75%	max
Net Income	34608	424.951	2521.505	-98696.000	0.524	33.798	198.000	99803.000
Lazy Analyst Core Earnings	32183	555.685	2590.206	-8417.000	10.000	62.427	269.550	107097.000
Sequential Prompt Core Earnings	33765	487.796	2535.262	-74659.000	3.300	45.900	227.700	99803.000
Compustat OIADP	34608	706.898	3148.408	-28387.000	8.415	79.000	358.124	119437.000

Panel B: Summary Statistics: Earnings Measures in Per-Share Terms
This panel reports summary statistics for the earnings measures scaled to a per-share basis. The earnings measures included are:

- **Net Income per Share**: The net income of firms as reported in their financial statements, scaled by shares outstanding
- **Lazy Analyst Core Earnings per Share**: Core earnings estimated using the Lazy Analyst approach, scaled by shares outstanding
- **Sequential Prompt Core Earnings per Share**: Core earnings estimated using the Sequential Prompt approach, scaled by shares outstanding
- **Compustat OIADP per Share**: Operating Income After Depreciation as reported by Compustat, scaled by shares outstanding
- **Compustat OPEPS**: Earnings Per Share from Operations as reported by Compustat

	count	mean	std	min	25%	50%	75%	max
Net Income per Share	34608	1.599	5.806	-270.379	0.033	1.021	2.514	223.115
Lazy Analyst Core Earnings per Share	32183	2.457	4.827	-128.736	0.475	1.513	3.139	157.251
Sequential Prompt Core Earnings per Share	33765	1.974	4.962	-148.604	0.184	1.228	2.809	140.220
Compustat OIADP per Share	34608	3.070	5.942	-296.616	0.428	1.946	4.240	197.482
Compustat OPEPS	34607	1.813	5.315	-269.830	0.170	1.150	2.610	189.530

Table 2. Summary Statistics and Adjustments to Earnings Measures (cont.)

Panel C: Direction of Adjustments

This panel reports the percentages of observations where net adjustments in terms of the deviation between the alternative earnings measure and Net Income were income-increasing, income-decreasing, or neither.

	Income Increasing	Income Decreasing	No Change
Total Adjustments per Share, Lazy Analyst Prompt	68.681	22.951	1.361
Total Adjustments per Share, Sequential Prompt	59.769	37.448	0.347
Total Adjustments per Share, Compustat OIADP	87.462	12.512	0.026
Total Adjustments per Share, Compustat OPEPS	63.696	36.298	0.003

Panel D: Summary Statistics of Amounts of Adjustments

This panel provides summary statistics for the amounts of adjustments, in terms of the deviation between the alternative earnings measure and Net Income, for each alternative measure.

	count	mean	std	min	25%	50%	75%	max
Total Adjustments per Share, Lazy Analyst Prompt	32183	0.737	4.758	-185.106	0.000	0.254	0.928	282.177
Total Adjustments per Share, Sequential Prompt	33765	0.395	5.300	-217.441	-0.172	0.093	0.661	280.994
Total Adjustments per Share, Compustat OIADP	34608	1.471	4.199	-200.078	0.224	0.885	2.024	139.430
Total Adjustments per Share, Compustat OPEPS	34607	0.214	2.844	-176.085	-0.029	0.025	0.260	101.047

Table 3.
Autoregressive Regressions

Table 3 presents the results of pooled autoregressive (AR(1)) regressions for various earnings measures on a per-share basis. Specifically, each earnings measure is regressed on its own first lag. The earnings measures analyzed here and throughout are:

- **Net Income per Share**: Net income divided by the number of shares outstanding.
- **Lazy Analyst Core Earnings per Share**: Core earnings per share estimated using the Lazy Analyst approach.
- **Sequential Prompt Core Earnings per Share**: Core earnings per share estimated using the Sequential Prompt approach.
- **Compustat OIADP per Share**: Operating Income After Depreciation per share as reported by Compustat, divided by the number of shares outstanding.
- **Compustat OPEPS**: Operating Earnings Per Share as reported by Compustat.

The model specification for each regression is:

$$\text{Earnings Measure}_{i,t} = \beta_0 + \beta \times \text{Earnings Measure}_{i,t-1} + \epsilon_{i,t}$$

where Earnings $\text{Measure}_{i,t}$ is the current value of the earnings measure for firm i at time t. The model is estimated using OLS on the full pooled sample. Standard errors are clustered at the firm level to account for serial correlation and are reported in parentheses below the coefficient estimates. Significance levels are indicated by *, **, and ***, corresponding to the 10%, 5%, and 1% significance levels, respectively, for two-tailed tests of the null hypothesis that the coefficient is equal to zero.

	(1)	(2)	(3)	(4)	(5)
Net Income per Share (Lagged)	0.4963***				
	(0.0625)				
Lazy Analyst Core Earnings per Share (Lagged)		0.8306***			
		(0.0475)			
Sequential Prompt Core Earnings per Share (Lagged)			0.7136***		
			(0.0483)		
Compustat OIADP per Share (Lagged)				0.9421***	
				(0.0277)	
Compustat OPEPS (Lagged)					0.6253***
					(0.1048)
Observations	32540	28794	31378	32540	32539
$Adj.R^2$	0.2013	0.5853	0.4292	0.7003	0.3052

Table 4.
Autoregressive Regressions of Adjustments

Table 4 presents the results of pooled autoregressive (AR(1)) regressions of the adjustments for each alternative earnings measure relative to Net Income. Specifically, for each core earnings measure, the adjustment is calculated as:

$$\text{Adjustment}_{i,t} = \text{Earnings Measure}_{i,t} - \text{NI}_{i,t}$$

For each adjustment, we estimate the following AR(1) regression model:

$$\text{Adjustment}_{i,t} = \beta_0 + \beta \times \text{Adjustment}_{i,t-1} + \epsilon_{i,t}$$

where:

- $\text{Adjustment}_{i,t}$ is the adjustment for firm i at time t.
- $\text{Adjustment}_{i,t-1}$ is the adjustment for firm i at time $t-1$.
- β_0 is the intercept term.
- β is the coefficient capturing the persistence of the adjustment.
- $\epsilon_{i,t}$ is the error term.

The regressions are estimated on the pooled sample using OLS. Standard errors are clustered at the firm level to account for serial correlation within firms and are reported in parentheses below the coefficient estimates. Significance levels are indicated by *, **, and ***, corresponding to the 10%, 5%, and 1% significance levels, respectively, for two-tailed tests of the null hypothesis that the coefficient is equal to zero.

	(1)	(2)	(3)	(4)
Adjustments for Lazy Analyst CE (Lagged)	0.0758**			
	(0.0349)			
Adjustments for Sequential Prompt CE (Lagged)		0.0063		
		(0.0307)		
Adjustments for Compustat OIADP (Lagged)			0.3131***	
			(0.0587)	
Adjustments for Compustat OPEPS (Lagged)				-0.0093
				(0.0371)
Observations	28807	31372	32535	32534
$Adj.R^2$	0.0047	0.0000	0.0906	0.0001

Table 5.
Prediction Error Metrics: Mean, Median, and MSE

Table 5 reports how close each earnings measure is to next-period Net Income. For firm i in period t, we measure this deviation in two ways, absolute error and squared error:

$$\mathrm{AE}_{i,t} = \left|\mathrm{NI}_{i,t+1} - \text{Earnings Measure}_{i,t}\right|, \qquad \mathrm{SE}_{i,t} = \left(\mathrm{NI}_{i,t+1} - \text{Earnings Measure}_{i,t}\right)^2,$$

where $\mathrm{NI}_{i,t+1}$ is the Net Income per Share for firm i in period $t+1$ and Earnings Measure$_{i,t}$ is the specified earnings measure at time t.

We then summarize each measure's overall error in three ways:

- **Mean Absolute Error (MAE)**: the sample mean of $\mathrm{AE}_{i,t}$,
- **Median Absolute Error**: the sample median of $\mathrm{AE}_{i,t}$,
- **Mean Squared Error (MSE)**: the sample mean of $\mathrm{SE}_{i,t}$.

	MAE	Median Absolute Error	MSE
Net Income per Share	1.7847	0.6027	37.8516
Lazy Analyst Core Earnings per Share	1.7335	0.6493	29.7298
Sequential Prompt Core Earnings per Share	1.6175	0.6343	27.7778
Compustat OIADP per Share	2.1279	0.9964	30.7332
Compustat OPEPS	1.5527	0.5246	32.7222

Table 6.
Future Net Income Prediction

Table 6 presents the results of pooled regressions of next period Net Income per Share on current period earnings measures. ($\mathrm{NI}_{i,t+1}$). The model specification for each regression is:

$$\mathrm{NI}_{i,t+1} = \beta_0 + \beta_1 \times \text{Earnings Measure}_{i,t} + \gamma_t + \epsilon_{i,t}$$

where:

- $\mathrm{NI}_{i,t+1}$ is the Net Income per Share for firm i at time $t+1$
- $\text{Earnings Measure}_{i,t}$ is the current period earnings measure for firm i
- β_0 is the intercept term
- β_1 is the coefficient measuring the predictive relationship
- γ_t represents year fixed effects
- $\epsilon_{i,t}$ is the error term

The regressions are estimated on the full pooled sample using OLS with year fixed effects. Standard errors are clustered at the firm level to account for serial correlation within firms and are reported in parentheses below the coefficient estimates. Significance levels are indicated by *, **, and ***, corresponding to the 10%, 5%, and 1% significance levels, respectively, for two-tailed tests of the null hypothesis that the coefficient is equal to zero.

	Dependent variable: Future Net Income per Share				
	(1)	(2)	(3)	(4)	(5)
Net Income per Share	0.4411***				
	(0.0657)				
Lazy Analyst Core Earnings per Share		0.6474***			
		(0.0656)			
Sequential Prompt Core Earnings per Share			0.6381***		
			(0.0639)		
Compustat OIADP per Share				0.6041***	
				(0.0338)	
Compustat OPEPS					0.5403***
					(0.1001)
Observations	32535	30423	31731	32535	32534
$Adj.R^2$	0.1576	0.2329	0.2499	0.2895	0.1896
Year FE	Yes	Yes	Yes	Yes	Yes

Table 7.
Future Net Income Prediction (Two Years Ahead)

Table 7 presents the results of pooled regressions of two-periods-ahead Net Income per Share on the current period earnings measures. The dependent variable in each regression is the Net Income per Share for firm i at time $t+2$ ($\mathrm{NI}_{i,t+2}$). The model specification for each regression is:

$$\mathrm{NI}_{i,t+2} = \beta_0 + \beta_1 \times \text{Earnings Measure}_{i,t} + \gamma_t + \epsilon_{i,t},$$

where:

- $\mathrm{NI}_{i,t+2}$ is the Net Income per Share for firm i two periods ahead
- $\text{Earnings Measure}_{i,t}$ is the specified current-period earnings measure
- β_0 is the intercept term
- β_1 is the coefficient measuring the predictive relationship
- γ_t represents year fixed effects
- $\epsilon_{i,t}$ is the error term

The regressions are estimated on the full pooled sample using OLS with year fixed effects. Standard errors are clustered at the firm level to account for serial correlation within firms and are reported in parentheses below the coefficient estimates. Significance levels are indicated by *, **, and ***, corresponding to the 10%, 5%, and 1% significance levels, respectively, for two-tailed tests of the null hypothesis that the coefficient is zero.

	Dependent variable: Future Net Income per Share (t+2)				
	(1)	(2)	(3)	(4)	(5)
Net Income per Share	0.3830***				
	(0.0708)				
Lazy Analyst Core Earnings per Share		0.6018***			
		(0.0719)			
Sequential Prompt Core Earnings per Share			0.6573***		
			(0.0828)		
Compustat OIADP per Share				0.5766***	
				(0.0439)	
Compustat OPEPS					0.5355***
					(0.0878)
Observations	30547	28600	29784	30547	30546
$Adj.R^2$	0.1024	0.1762	0.2154	0.2179	0.1537
Year FE	Yes	Yes	Yes	Yes	Yes

Table 8.
Firm-Level Regressions Predicting Net Income (One Period Ahead)

Table 8 presents the results of firm-level regressions assessing each earnings measure's ability to predict one-period-ahead Net Income per Share. Within each firm with sufficient data (minimum of four observations), we estimate the following model:

$$\mathrm{NI}_{i,t+1} = \alpha_i + \beta_i \times \text{Earnings Measure}_{i,t} + \epsilon_{i,t},$$

where:

- α_i is the firm-specific intercept term.
- β_i is the firm-specific slope coefficient relating the current period's earnings measure to next period's net income.
- $\epsilon_{i,t}$ is the error term.

We collect the results from each firm-level regression. The table below reports the following summary statistics:

- **Average Coefficient**: The mean of the estimated slope coefficients (β_i) across firm-level regressions.
- **Average R-squared**: The mean of the R^2 values across firm-level regressions.
- **Number of Firms**: The number of firms included in the analysis.

Earnings Measure	Coef	Ave R^2	N. Firms
Net Income per Share	0.4012	0.2420	1930
Lazy Analyst Core Earnings per Share	0.4169	0.2463	1908
Sequential Prompt Core Earnings per Share	0.4229	0.2740	1907
Compustat OIADP per Share	0.4071	0.2798	1930
Compustat OPEPS	0.5406	0.2796	1930

Table 9.
Firm-Level Regressions Predicting Net Income (Two Periods Ahead)

Table 9 presents the results of firm-level regressions assessing each earnings measure's ability to predict net income per share two periods ahead. Within each firm with sufficient data (minimum of four observations), we estimate the following model:

$$\text{NI}_{i,t+2} = \alpha_i + \beta_i \times \text{Earnings Measure}_{i,t} + \epsilon_{i,t},$$

where:

- α_i is the firm-specific intercept term.
- β_i is the firm-specific slope coefficient relating the current period's earnings measure to net income two periods later.
- $\epsilon_{i,t}$ is the error term.

We collect the results from each firm-level regression. The table below reports:

- **Average Coefficient**: The mean of the estimated slope coefficients (β_i) across firm-level regressions.
- **Average R-squared**: The mean of the R^2 values across firm-level regressions.
- **Number of Firms**: The number of firms included in the analysis.

Earnings Measure	Coef	Ave R^2	N. Firms
Net Income per Share	0.2538	0.1616	1910
Lazy Analyst Core Earnings per Share	0.2146	0.1844	1882
Sequential Prompt Core Earnings per Share	0.3078	0.2227	1885
Compustat OIADP per Share	0.1992	0.1938	1910
Compustat OPEPS	0.3058	0.1875	1910

Table 10.
Goodness of Fit with Market Valuations: Adjusted R^2 from Regressions of Stock Price per Share on Interactions of Candidate Earnings Measures with Industry-Year Dummies

Table 10 reports the adjusted R^2 values from regressions of firms' stock prices on interactions between candidate earnings measures and industry-year dummy variables. The dependent variable is the firm's stock price per share at the end of the next fiscal year (relative to the fiscal year of the index 10-K analyzed by the LLMs), denoted as $\text{Price}_{i,t+1}$.

For each earnings measure, the regression model estimated is:

$$\text{Price}_{i,t+1} = \sum_j \beta_j \left(D_{j,i,t} \times \text{Earnings Measure}_{i,t} \right) + \epsilon_{i,t}$$

where:

- $\text{Price}_{i,t+1}$ is the stock price per share of firm i at the end of the next fiscal period.
- $\text{Earnings Measure}_{i,t}$ is the specified earnings measure per share for firm i at time t.
- $D_{j,i,t}$ is a dummy variable for industry-year group j to which firm i belongs at time t, based on two-digit SIC codes and fiscal year.
- β_j represents the mean valuation multiple specific to industry-year group j.
- $\epsilon_{i,t}$ is the error term.

The regressions are estimated without an intercept term since the industry-year dummies span the entire sample. The adjusted R^2 values reported reflect the proportion of variance in the stock prices explained by each model, adjusted for the number of predictors. The number of observations used in each regression is also reported in the table. (Coefficients are not reported because there are 1,348 separate coefficients for each of the same number of separate industry-year groupings.)

Earnings Measure	Adjusted R-squared	Observations
Net Income per Share	0.7330	24916
Lazy Analyst Core Earnings per Share	0.7841	26767
Sequential Prompt Core Earnings per Share	0.7751	25780
Compustat OIADP per Share	0.7585	27585
Compustat OPEPS	0.7756	26267

Table 11.

Goodness of Fit with Market Valuations at $T + 2$: Adjusted R^2 from Regressions of Forward Market Price per Share on Interactions of Candidate Earnings Measures with Industry-Year Dummies

Table 11 reports the adjusted R^2 values from regressions of firms' stock prices at the end of the subsequent fiscal period ($t + 2$) on the earnings measures interacted with industry-year dummy variables. The earnings measures analyzed are the same as in previous tables.
For each earnings measure, the regression model estimated is:

$$\text{Price}_{i,t+2} = \sum_j \beta_j \left(D_{j,i,t} \times \text{Earnings Measure}_{i,t} \right) + \epsilon_{i,t}$$

where:

- $\text{Price}_{i,t+2}$ is the firm's stock price per share at the end of the fiscal period subsequent to the fiscal year in which the index 10-K was posted
- $\text{Earnings Measure}_{i,t}$ is the specified earnings measure per share for firm i at time t.
- $D_{j,i,t}$ is a dummy variable for industry-year group j to which firm i belongs at time t, based on two-digit SIC codes and fiscal year.
- β_j represents the average multiple for industry-year group j.
- $\epsilon_{i,t}$ is the error term.

The regressions are estimated without an intercept term since the industry-year dummies span the entire sample. Observations are included only if the earnings measure is positive and non-missing, and the forward price at $t + 2$ is available and strictly positive. The number of observations used in each regression is also reported in the table. (Coefficients are not reported because there are separate coefficients for each of the separate industry-year groupings.)

Earnings Measure	Adjusted R-squared	Observations
Net Income per Share	0.6786	23528
Lazy Analyst Core Earnings per Share	0.7400	25238
Sequential Prompt Core Earnings per Share	0.7605	24389
Compustat OIADP per Share	0.7255	26007
Compustat OPEPS	0.7199	24803

Table 12.

Robustness Test: Difference in Anonymized vs. Non-Anonymized Results and Future Earnings Changes

Table 12 presents the robustness test regression where the dependent variable is the difference between the core earnings measure outputted via the Sequential Prompt approach for the full 10-K vs. the anonymized version (in per share terms). This variable is constructed as:

$$\text{Difference}_{i,t} = \text{Original Measure}_{i,t} - \text{Anonymized Measure}_{i,t}$$

where Original Measure$_{i,t}$ is the core earnings measure outputted via the Sequential Prompt approach for the full 10-K and Anonymized Measure$_{i,t}$ is the measure outputted via the Sequential Prompt approach for the anonymized 10-K.
Standard errors are clustered at the firm level to account for serial correlation and are reported in parentheses below the coefficient estimates. Significance levels are indicated by *, **, and ***, corresponding to the 10%, 5%, and 1% significance levels, respectively, based on two-tailed t-tests of the null hypothesis that the coefficient is equal to zero.

	Dependent variable: Difference in Measure Between Original 10-K vs. Anonymized
	(1)
Change in Future Net Income per Share	-0.0025
	(0.0151)
Intercept	0.2724***
	(0.0851)
Observations	873
R^2	0.0001

Appendix A Example of LLM Analysis Using "Lazy Analyst" approach

Lawson Products, Inc. (2018)

To determine the company's core earnings, we start with the reported GAAP net income and make adjustments based on the provided 10-K text. Here's the step-by-step calculation and reasoning:

1. **Starting Point: GAAP Net Income**
   - Reported GAAP Net Income for 2017: $29,688,000
2. **Adjustment for One-Time Gains:**
   - Gain on sale of property: $5,422,000
   - Reasoning: This is a one-time gain and not related to the company's core operations.
   - Adjusted Net Income: $29,688,000 - $5,422,000 = $24,266,000
3. **Adjustment for Non-Core Items:**
   - Environmental remediation expense: $0 (no new expense reported for 2017)
   - Legal and settlement costs: Not explicitly mentioned, assume $0 for 2017
   - Acquisition-related expenses: $700,000
   - Reasoning: These are one-time costs associated with acquisitions.
   - Adjusted Net Income: $24,266,000 - $700,000 = $23,566,000
4. **Adjustment for Tax Benefits:**
   - Tax benefit from valuation allowance release: $21,228,000
   - Reasoning: This is a deferred tax asset valuation allowance release, which is a one-time event.
   - Adjusted Net Income: $23,566,000 - $21,228,000 = $2,338,000
5. **Adjustment for Interest Expense:**
   - Interest expense: $622,000
   - Reasoning: Interest expense is a recurring cost and should be considered in core earnings.
   - Adjusted Net Income: $2,338,000 + $622,000 = $2,960,000
6. **Adjustment for Depreciation and Amortization:**
   - Depreciation and amortization: $6,770,000
   - Reasoning: These are non-cash charges and part of ongoing operations.
   - Adjusted Net Income: $2,960,000 + $6,770,000 = $9,730,000
7. **Adjustment for Stock-Based Compensation:**
   - Stock-based compensation: $3,106,000
   - Reasoning: This is a non-cash charge but part of ongoing employee compensation.
   - Adjusted Net Income: $9,730,000 + $3,106,000 = $12,836,000
8. **Adjustment for Foreign Exchange Gains:**
   - Foreign exchange gain: $780,000
   - Reasoning: This is a non-core item as it relates to currency fluctuations.
   - Adjusted Net Income: $12,836,000 - $780,000 = $12,056,000
9. **Adjustment for Non-Operating Income:**
   - Other non-operating income: Assume $0 (no specific mention)

Final Calculation:

- Core Earnings Calculation (final) = $12,056,000

Appendix B Example of LLM Analysis Using "Sequential Structured Prompt" approach

ACCO Brands Corporation, FY 2009 10-K. Note: minor formatting changes made to plain .txt output from LLM call for formatting in paper .pdf environment. Compressed description of items in table and duplicative information (e.g., original LLM output included company name, CIK and Year in each row each time).

Here is the summary table for ACCO Brands Corporation, including the computation of a new earnings measure:

Item	Amount (m)	Description
Net Income	-$126.1	Reported net income
Goodwill and Asset Impairment	+$1.8	Nonrecurring impairment
Restructuring Charges	+$17.4	Nonrecurring restructuring costs
Loss on Discontinued Ops	+$0.8	Nonrecurring loss on sale
Debt Refinancing Costs	+$9.1	Nonrecurring loss on early debt extinguishment
Valuation Allowance on DTAs	+$108.1	Nonrecurring charge for deferred tax assets
Other Charges	+$4.6	Nonrecurring restructuring-related charges
Gain on Debt Extinguishment	-$4.9	Nonrecurring gain on note repurchase
Pension Curtailment Gain	-$1.0	Nonrecurring gain from pension freeze

Computation of New Earnings Measure

Net Income: -$126.1 million

Add Back Nonrecurring/Unusual Expenses:

- Goodwill and Asset Impairment: +$1.8 million
- Restructuring Charges: +$17.4 million
- Loss on Discontinued Operations: +$0.8 million
- Debt Refinancing Costs: +$9.1 million
- Valuation Allowance on Deferred Tax Assets: +$108.1 million
- Other Charges: +$4.6 million

Subtract Nonrecurring/Unusual Income:

- Gain on Debt Extinguishment: -$4.9 million
- Pension Curtailment Gain: -$1.0 million

New Earnings Measure Calculation:

- Net Income: -$126.1 million
- Add Back Nonrecurring/Unusual Expenses: +$141.8 million
- Subtract Nonrecurring/Unusual Income: -$5.9 million

New Earnings Measure: -$126.1 million + $141.8 million - $5.9 million = $9.8 million

Appendix C Description of Variables

This table defines variables used in our analysis.

Description of Variables

Variable	Description	Computation
Net Income	Net Income (in $ millions)	Compustat variable NI
Net Income per Share	Net Income scaled by shares outstanding	Compustat NI and CRSP CSHO
Net Income per Share (t+i)	Net Income per Share one fiscal period ahead	Compustat NI for period t+i divided by CSHO, with leads indexed by gvkey
Lazy Analyst Core Earnings	Core Earnings estimated using the Lazy Analyst approach	Extracted from returned result from LLM call described in main text
Sequential Prompt Core Earnings	Core Earnings estimated using the Sequential Prompt approach	Extracted from returned result from LLM call described in main text
Lazy Analyst Core Earnings per Share	Lazy Analyst Core Earnings scaled by shares outstanding	Lazy Analyst Core Earnings divided by Common Shares Outstanding from CRSP
Sequential Prompt Core Earnings per Share	Sequential Prompt Core Earnings scaled by shares outstanding	Sequential Prompt Core Earnings divided by Common Shares Outstanding from CRSP
Compustat OIADP per Share	Operating Income After Depreciation per share	OIADP from Compustat divided by CSHO from CRSP
Compustat OPEPS	Operating Earnings Per Share as reported by Compustat	Directly from Compustat data item OPEPS
Total Adjustments per Share, Lazy Analyst	Net difference between the core earnings estimated using this LLM approach and net income	Lazy Analyst Core Earnings per Share minus Net Income per Share
Total Adjustments per Share, Sequential Prompt	Net difference between the core earnings estimated using this LLM prompting approach and Net Income per Share	Sequential Prompt Core Earnings per Share minus Net Income per Share
Total Adjustments per Share, Compustat OIADP	Net difference between Compustat OIADP and Net Income, on a per share basis	Compustat OIADP per Share minus Net Income per Share
Total Adjustments per Share, Compustat OIADP	Net difference between Compustat OIADP and Net Income, on a per share basis	Compustat OIADP per Share minus Net Income per Share

Continued on next page

Description of Variables (continued)

Variable	Description	Computation
Total Adjustments per Share, Compustat OPEPS	Net difference between Compustat OPEPS and Net Income, on a per share basis	Compustat OPEPS minus Net Income per Share
$Price_{i,t+1}$	Stock price at fiscal year-end	PRCC_F from Compustat-CRSP