CFOs Meet LLMs

John R. Graham[†] Campbell R. Harvey[†] Manish Jha[‡]

[†]Duke University and NBER [‡]Georgia State University

Abstract

Business sentiment is a closely watched economic signal, but measuring it is slow and costly: surveys reach only a few hundred firms, arrive periodically, and take time to compile. We show that large language models hold the potential to address these shortcomings. We prompt an LLM to role-play as the CFO of a specific company at a specific date and focus on the economic-optimism question on the Duke–Federal Reserve CFO Survey over 2002–2025. We find that the LLM reproduces individual human responses: the predicted optimism score significantly forecasts the CFO's actual answer, surviving firm and year-quarter fixed effects and a control for the most recent prior response. Predictive accuracy increases with the amount of information supplied, as both respondent history and firm characteristics improve fit, and the relationship persists under quarterly aggregation. With appropriate conditioning, LLMs may be able to serve as credible digital twins of executives, offering scalable, high-frequency expectations data for financial research and policy.

Keywords: Large language models; CFO survey; Managerial expectations; Business sentiment; Synthetic surveys; Digital twins; Belief formation

JEL Classification: C55, D84, E27, G30, G41

Version: June 11, 2026. The authors have no conflicts of interest to disclose. We thank Manuel Adelino, Alon Brav, Nuno Clara, and Duke's AI Reading Group for helpful comments. Correspondence: Campbell R. Harvey, Duke University, Fuqua School of Business, 100 Fuqua Drive, Durham, NC 27708. Email: cam.harvey@duke.edu.
Latest version: https://ssrn.com/abstract=6920881

Chief Financial Officers occupy a unique vantage point in the economy. Located at the intersection of corporate strategy and the macroeconomic environment, they aggregate information about customers, suppliers, competitors, and capital markets into forward-looking assessments of business activity. Research spanning a quarter century has shown that measured CFO expectations are economically meaningful: Gennaioli et al. (2016) document that CFO expectations (from the Duke survey) predict corporate investment and, in some specifications, outperform Tobin's q; and the Duke–Federal Reserve (Duke-Fed) CFO Optimism Index has historically led both hiring and GDP growth.[1] Graham (2022) further documents the central role that executive expectations play in corporate investment and planning decisions. However, despite their informational value, these expectations are expensive to collect, available only at a fixed periodic frequency, subject to well-documented behavioral biases (Ben-David et al., 2013; Malmendier and Tate, 2015; Boutros et al., 2025), and limited in cross-sectional coverage to the several hundred companies whose executives choose to participate in any given survey wave. These constraints raise a natural question: is it possible to generate credible CFO-style expectations at scale, without the cost, selection bias, modest sample size, or delay of a human survey?

This paper introduces and validates a method for generating synthetic CFO expectations using large language models (LLMs). We prompt the LLM to role-play as the CFO of a specific public company at a specific point in time, supplying the model with the firm's industry, revenue, headcount, and geographic footprint, and imposing an information cutoff matching the date of the actual survey wave. Mimicking an actual question on the Duke-Fed survey, we ask the LLM to state its optimism about the US economy. When a human CFO for this same company has provided this information on the actual Duke-Fed survey, we also supply the model with that person's own prior optimism scores and with the model's own prior forecasts for the same person, always restricted to dates strictly before the survey cutoff. Specifically, the model is asked the canonical Duke CFO Survey question: rate your optimism about the overall U.S. economy on a scale from 0 to 100. For each of 6,075 survey responses from public companies over 2002–2025, we generate three independent LLM forecasts in a few seconds and average them. We match every synthetic score to the response of the actual CFO at the

[1] Fried et al. (2026) demonstrate that the Duke-Fed CFO survey accurately anticipates future inflation; and Baslandze et al. (2026) show that it reasonably anticipates future revenue and wages. See the quarterly survey reports through 2020Q1 archived at `https://cfosurvey.fuqua.duke.edu` and reports since at `https://www.cfosurvey.org`

same firm in the same quarter. Because our panel links responses to specific individuals over time, we can validate at the *individual, within-person* level, complementing distributional validations in prior work (e.g., Hansen et al., 2025).

Our central finding is that the LLM does a good job of mimicking individual CFOs. The matched LLM optimism score is a significant predictor of the actual CFO's response: in a regression of individual CFO optimism on the matched LLM score, the LLM coefficient is more than six standard errors from zero, and the predictive relationship survives the inclusion of firm fixed effects and year-quarter fixed effects, separately and jointly (Table 2). Because the survey responses were never part of the model's training corpus, these predictions are out-of-sample with respect to the specific human answers. This individual-level predictive content is notable in light of prior work showing that managerial expectations are not merely noisy reflections of public information: Gennaioli et al. (2016) show that they contain extrapolative components that predict investment. Our framework approximates how an executive is likely to respond given their information set and prior answers, providing an individual-level benchmark for the predictable component of managerial expectations.

A key methodological contribution relative to earlier persona-based work is the conditioning on each respondent's own history. Supplying the model with a person's prior answers, and with its own prior forecasts for that person, anchors the persona to the individual rather than to a generic firm type and sharply reduces the tendency of instruction-tuned models to collapse toward modal opinions (Santurkar et al., 2023). The effect is visible in the data: the within-quarter dispersion of LLM scores is close to that of the human respondents (a disagreement ratio near one), and the model captures within-firm, over-time variation in optimism. Importantly, the additional information about prior same-human answers is always drawn from dates before the survey cutoff, so it does not introduce look-ahead bias; this CFO-specific information has never been released publicly. To confirm that the LLM is not simply echoing the respondent's most recent answer, we show that the LLM score remains a significant predictor even after controlling directly for the respondent's prior-quarter response (Table 2, columns (5)–(8)).

In addition, we document that the predictive accuracy scales with the information provided to the model. Splitting the sample by how much of a respondent's own history is available, the cross-sectional fit rises monotonically: from an R^2 near 0.10 when no prior responses are available, to roughly 0.33 with a few prior data points, to nearly 0.50 when more than a

handful of prior responses are supplied (Table 3). A parallel pattern holds for the breadth of firm characteristics: prompts that include the full set of profile fields yield better predictions than sparse prompts. These dose–response relationships, where richer conditioning inputs (dose) produce systematically stronger or more accurate model outputs, indicate that the model is using the conditioning information rather than guessing, and they show the conditions under which synthetic executives most faithfully reproduce human responses.

Horton et al. (2023) introduced the concept of *Homo silicus*, the idea that LLMs can be treated as computational models of humans, capable of being assigned endowments, information, and roles and then interviewed at scale (see also Manning et al., 2024). Argyle et al. (2023) show that LLM-generated synthetic survey populations closely match the distributional properties of real human samples, and Suh et al. (2025) demonstrate that fine-tuning language models on scaled survey data substantially improves the prediction of public-opinion response distributions across diverse subpopulations. Manning and Horton (2025) further demonstrate that grounding LLM personas in economic theory and calibrating them against human data substantially improves predictive fidelity in novel settings, providing a methodological foundation for the persona-based approach we adopt here. The work closest to ours is Hansen et al. (2025), who construct synthetic SPF personas and show that LLM-generated forecasts replicate key distributional properties of the human survey. We go further by testing whether the LLM persona predicts the response of the *specific* human at the *same* firm and quarter, and by showing how that fidelity depends on the information the model is given.

Two further strands of the literature sharpen the contrast with our individual-level test. The first simulates a *population* of forecasters to recover the dispersion of beliefs rather than any one person's answer: Bali et al. (2025) represent investors as an ensemble of distinct machine-learning specifications drawing on a common information set, and measure disagreement as the cross-sectional spread of their forecasts. As in the synthetic-SPF design of Hansen et al. (2025), their object of interest is the distribution of simulated beliefs; our object of interest is instead the matched response of an identified human at a specific firm and quarter. The second strand is cautionary: Henning et al. (2025) find that in endogenous experimental asset markets, out-of-the-box LLM traders price close to fundamental value and largely fail to reproduce the bubbles that humans reliably generate. Because LLM agents need not replicate human behavior, our finding of individual-level fidelity is nontrivial; consistent with Henning et al. (2025)'s account, that fidelity emerges in our setting only after conditioning each persona on

the respondent's own history, without which instruction-tuned models collapse toward modal opinions (Santurkar et al., 2023).

Our framework introduces an LLM-based measure of CFO-level economic sentiment. Within the growing literature applying machine-learning methods to finance and economics (see Korinek, 2023, for a survey), which includes sentiment classification (Loughran and McDonald, 2011; Hansen and McMahon, 2016; Tetlock, 2007; Tetlock et al., 2008), financial analysis (Cao et al., 2024), inflation expectations (Wu et al., 2025; Faria-e Castro and Leibovici, 2024), market research (Brand et al., 2023), and asset pricing and portfolio construction (Heaton et al., 2017; Gu et al., 2020; Uddin et al., 2024; Lopez-Lira and Tang, 2023; Wang et al., 2026), we provide the first individual-level, within-firm validation of LLM-simulated survey responses, showing that the model can approximate private responses using publicly available observables together with a respondent's own past answers.

We also address the memorization concerns raised by Lopez-Lira et al. (2025), Sarkar and Vafa (2025), and Gao et al. (2025). Memorization refers broadly to the model's retention of facts from its training corpus; it is broader than look-ahead bias, which is the narrower problem that arises when the model uses memorized information to "predict" outcomes that had already occurred during the training-data window. A model can exhibit memorization without look-ahead bias (e.g., by recalling a firm's historical revenue), but cannot exhibit look-ahead bias without memorization. Our design provides two layers of protection. First, we impose information cutoffs that instruct the model not to access post-survey data, and any respondent history we supply is itself dated strictly before the cutoff. Second, and more fundamentally, the target variable itself (the individual CFO optimism score at the survey date) was never part of the model's training corpus because the survey data were submitted through Duke's institutional agreement with OpenAI and were contractually excluded from training. The model cannot recall what it has never seen. Bybee (2025) similarly uses LLM-generated responses to construct aggregate economic expectations from text, complementing our firm-level approach.

The remainder of the paper proceeds as follows. Section I describes the construction of our LLM-generated forecasts, including the respondent-history conditioning. Section II introduces the data and benchmarks. Section III presents the empirical results. Section IV discusses implications. The final section offers some conclusions.

I. Construction of the LLM Optimism Scores

We construct LLM-generated responses to simulate the answers of individual CFOs participating in The CFO Survey, conducted quarterly by Duke University and the Federal Reserve Banks of Atlanta and Richmond. Our approach employs a large language model to role-play as financial executives at specific companies at a specific point in time, incorporating rich contextual information available at each forecast date. For each survey response, we compile a detailed firm-executive profile and standardize company names to track firms across survey waves (see Appendix A for details on name cleaning, missing-data imputation, and variable definitions).

A. LLM Response Generation

We employ GPT-5.4 (OpenAI, 2026), accessible through Duke University's IT infrastructure, to generate simulated CFO responses. Accessing OpenAI via Duke's institutional agreement ensures that submitted survey data are not used for model training; accordingly, the model does not retain or learn the contents of these surveys across runs. We access the model through OpenAI's Responses API.

For each actual survey response in our sample, we create a corresponding LLM-generated response using two prompts: a *system prompt* that establishes the context, persona, and analytical instructions, and a *user prompt* that provides the firm-specific profile, the respondent's history (when available), and that poses the survey question. The system prompt corresponds to component (i) below, while the user prompt comprises components (ii)–(iv). Among the analytical instructions in the system prompt, we explicitly direct the model to collect and synthesize analyst coverage, including recent analyst reports, price targets, and consensus revenue estimates, when forming its forecast. We also enable the model's web-search tool, which allows it to retrieve and incorporate publicly available information, such as company earnings reports, CFO interviews, and analyst commentary, from the internet prior to the specified cutoff date.[2] The temporal cutoff restriction imposed in the system prompt ensures that even with web search enabled, the model can only access information available before the designated date.

[2]Without web search enabled, the model relies solely on its pre-training data, which may yield less informed forecasts. Enabling web search allows the model to ground its responses in the same types of public information a CFO would have had access to at the time of the survey.

(i) System Instructions. We provide the model with detailed instructions that describe the context of the CFO Survey and the forecasting task (see Appendix B). In particular, we impose an information cutoff date that corresponds to the date the actual survey was completed. The model is instructed to use only information available as of this cutoff, excluding any data or developments occurring afterward. This temporal constraint attempts to level the playing field so that the LLM operates under the same information set as the human CFO at the time of their response, and to address the look-ahead concerns raised by Lopez-Lira et al. (2025) and Fritsch and Mazlish (2026), who show that without such constraints, LLM forecasts may partially recall historical outcomes from training data (see also Wu et al., 2025).

The value of this temporal constraint becomes especially clear in one example from our sample. For the 2020Q1 survey, conducted just weeks before the COVID-19 pandemic triggered a global economic collapse, we restrict the model's information cutoff to February 14, 2020. Even with web search enabled, the LLM's forecasts for this quarter exhibit no reduction in optimism relative to prior quarters, mirroring the sentiment of human forecasters who similarly had little forewarning of the impending crisis. In this particular case, our prompting strategy successfully prevents the model from incorporating knowledge of subsequent events: a model with access to post-cutoff information would almost certainly produce markedly more pessimistic forecasts for this period.

The system prompt directs the model to adopt a comprehensive analytical framework that mirrors the decision-making process of a financial executive. Specifically, we instruct the model to consider historical financial performance (3–5 years of revenue data, growth trends, profit margins), forward-looking company indicators (management guidance, product pipelines, order backlogs, expansion plans), analyst coverage (consensus estimates, price targets, key assumptions), comparable-company analysis where direct company information is limited, market and industry context (growth projections, competitive dynamics, regulatory changes), and macroeconomic factors (GDP forecasts, interest rates, inflation, and sector-specific indicators).

(ii) Company-Specific Profile. For each response, we provide a profile containing all available information about the responding executive and their firm as of the survey cutoff date (see Appendix C for a sample query). This profile includes the executive's name, job title, company name, ticker, industry classification, prior-year sales revenue, total employment, percentage of foreign revenue, headquarters location, state, ownership status, and credit rating. When

certain identifying fields, specifically job title, company name, industry, and headquarters location, are missing for a given wave, we impute them using the respondent's nearest past survey response in time, as determined by the survey completion date. All remaining missing fields are simply omitted from the prompt, allowing the model to draw on its broader knowledge base for context (see Appendix A).

(iii) Respondent History. A central element of our design is that, whenever a respondent has answered the survey in earlier waves, we supply the model with that person's own history, strictly limited to dates before the survey cutoff. Specifically, for each respondent we inject (a) up to the twelve most recent prior optimism scores the person reported (within the preceding twelve calendar quarters), (b) the model's own prior forecasts for the same person (the mean of three runs) over the same window, and (c) the person's lifetime-to-date average optimism, computed only from quarters strictly before the current one. When a respondent has no usable history, this section of the prompt is left blank rather than fabricated. Because the histories are constructed sequentially, processing the earliest quarters first so that later prompts can see the model's previously generated forecasts, each forecast is conditioned only on information that would have been available at the time. The system prompt instructs the model to treat the person's own past survey responses as a strong prior on that individual's style, optimism level, and bias (unless clearly contradicted by new information as of the cutoff) and to treat its own prior forecasts as a secondary prior, with continuity valued but the actual human responses taking precedence.

(iv) Survey Question. The specific survey question we ask is: "Rate your optimism about the overall U.S. economy on a scale from 0–100, with 0 being the least optimistic and 100 being the most optimistic."

We focus on the U.S.-economy optimism question because it is the one item asked in essentially unchanged form in every wave across our 2002–2025 sample. That long and consistent history is precisely what the within-person conditioning requires, namely a single question posed repeatedly to the same respondents over many years. Other survey items rotate across waves or change wording, so they do not support a comparably long matched panel. The method itself is not specific to optimism.

B. Implementation and Stability

For each of the 6,075 survey responses in our sample, we ask the LLM to make a forecast three separate times, using the exact same input each time. Because LLM outputs are stochastic, identical prompts can yield different responses across independent calls. This gives us 18,225 total forecasts (6,075 × 3).

By running three independent forecasts per CFO response, we can measure how much the model's predictions vary from run to run for the same prompt. This within-prompt variation is informative: if the LLM consistently gives very different forecasts for identical inputs, we should be cautious about treating any single forecast as reliable; if the forecasts cluster tightly, we gain confidence that the model is responding to the substance of the input rather than to noise.

Table A2 in Appendix D reports stability statistics for the three independent replications. We find that roughly 96% of total variance is attributable to differences across survey observations rather than stochastic variation within repeated calls, implying that the three forecasts are very highly correlated (pairwise correlations of about 0.96). The mean within-prompt standard deviation is just 2.22 points, compared to the average LLM optimism of 56.6. Because noise partially cancels when independent draws are averaged, the Spearman–Brown formula (Spearman, 1910; Brown, 1910) implies that the reliability of a three-call average rises to 0.985, meaning only about 1.5% of variance in the aggregated scores is attributable to random fluctuation across calls. This reinforces the broader finding that LLM-generated survey data can achieve high internal consistency (Argyle et al., 2023; Li et al., 2024), and represents a marked improvement in stability relative to prompts that omit respondent history.

II. Expectations Data

To contextualize LLM-generated forecasts, we compile two benchmark datasets that capture contemporaneous expectations from professional forecasters and consumers. These benchmarks allow us to compare LLM outputs against two of the most widely used sources of macroeconomic expectations in the literature (Manski, 2004; Coibion and Gorodnichenko, 2012), and to assess whether the LLM tracks CFO optimism over and above what these established gauges capture (Section III).

A. *Macro Forecasts and Consumer Sentiment*

We supplement the CFO survey with two external benchmarks. First, we obtain consensus forecasts from the Survey of Professional Forecasters (SPF), administered quarterly by the Federal Reserve Bank of Philadelphia; we extract mean forecasts for real GDP growth at the year-ahead horizon (`drgdp5`). The SPF is the oldest quarterly survey of macroeconomic forecasts in the United States, drawing on projections from approximately 30–40 professional forecasters at major financial institutions, consulting firms, and academic organizations. Second, we incorporate quarterly sentiment data from the University of Michigan's Surveys of Consumers (`ics_all`), which measures consumer confidence and economic expectations based on interviews with approximately 500 households. The Michigan Survey is benchmarked to 100 (1966 base) and dates to 1946.[3] Consumer sentiment has been shown to contain forward-looking information about consumption and aggregate demand (Ludvigson, 2004), making it a natural complement to the firm-level CFO survey.

B. *Summary Statistics*

Table 1 presents summary statistics for the main variables. Panel A reports individual-level statistics over the full sample of 6,075 matched responses. Mean LLM optimism (56.6) is somewhat below mean CFO Survey optimism (60.1), and the two distributions have comparable dispersion at the individual level (standard deviations of 16.2 and 17.6, respectively).

Panel B reports cross-sectional disagreement at the quarterly level. In each quarter, we compute the standard deviation of optimism scores across all respondents; the table reports the time-series distribution of these within-quarter dispersions. The average within-quarter standard deviation of CFO Survey responses is 16.5, only modestly larger than the corresponding LLM figure of 12.3[4]; the disagreement ratio, the within-quarter CFO standard deviation divided by the within-quarter LLM standard deviation, averages 1.42 with a median of 1.35. This near-parity is a substantial change relative to persona prompts that omit respondent history, which tend to compress the LLM distribution far below the human one. Instruction-tuned models are known to collapse toward modal opinions (Santurkar et al., 2023), a problem

[3] We chose the Michigan survey because it is the oldest and highest profile consumer survey. Alternative sentiment indices include the Conference Board Consumer Confidence Index, first published in 1967 (available monthly from 1978), and the New York Federal Reserve Survey of Consumer Expectations (available from 2013).

[4] Our use of within-quarter cross-sectional dispersion as a disagreement measure connects to a broader literature that analyzes the relation between standard disagreement proxies; see Goulding et al. (2026).

that persists even after explicitly steering models toward specific demographic profiles (also see Vafa et al., 2024); conditioning each persona on its own history substantially restores the cross-sectional heterogeneity of the human panel.

Panel C reports time-series variation by aggregating observations to the quarterly level. Not all quarterly survey waves include the economic-optimism question; restricting to waves that do yields 93 quarters. Quarterly mean LLM optimism (58.6) and CFO Survey optimism (60.8) are similar in level, although the LLM aggregate is somewhat more volatile across quarters (standard deviation of 9.5 versus 6.7). The Michigan Consumer Sentiment Index has a higher mean and is also volatile, reflecting the well-known sensitivity of consumer sentiment to economic conditions (Fuhrer, 1993). The SPF real GDP growth forecast averages 2.70% over the sample.

III. Main Analysis and Results

A. Do LLMs Track Individual CFO Expectations?

We begin by asking whether the LLM digital twins are close to the responses of their human CFOs at the same firm in the same quarter. This is a within-person, matched test that goes well beyond the distributional validation currently in the literature: rather than asking whether the synthetic population looks like the real distribution in the cross-section, we ask whether the digital twin, given the firm's observable characteristics, a temporal information cutoff, and the respondent's own pre-cutoff history, can approximate what the human executive will say.

Table 2 reports the analysis. The dependent variable is the individual CFO's optimism score, and the key regressor is the matched LLM score. In every specification the LLM score is positively and highly significantly associated with the human optimism score. In the baseline (column 1), a one-point increase in the LLM score is associated with a 0.57-point increase in the CFO's reported optimism ($t = 17.98$), and the score alone accounts for 27% of the cross-sectional variation in individual responses. We consider this analysis out-of-sample with respect to the human responses, since the model has never seen the human answer it is predicting.[5]

The next several columns probe the source of this predictive content. Adding firm fixed

[5] Gao et al. (2025) test for look-ahead bias by estimating whether a prompt appeared in an LLM's training corpus. Their concern applies when the predicted quantity was itself in the training corpus. Our setting meets this requirement because the individual CFO responses were never used for training.

Table 1: Summary Statistics

The sample consists of 6,075 matched observations from the Duke–Federal Reserve CFO Survey (Graham and Harvey, 2001; Graham et al., 2020) of public-company CFOs collected between 2002 and 2025. *Panel A* reports individual-level statistics over the full sample. *Panel B* reports cross-sectional disagreement: for each of the 93 quarters in the sample, we compute the standard deviation of optimism scores across all respondents and report the time-series distribution of these within-quarter dispersions; the disagreement ratio is the within-quarter CFO standard deviation divided by the within-quarter LLM standard deviation, computed quarter by quarter. *Panel C* reports time-series variation by aggregating observations to the quarterly level, enabling comparison with macroeconomic benchmarks available at that frequency. LLM Optimism and CFO Survey Optimism are measured on a 0–100 scale; Michigan Consumer Sentiment is an index level; SPF Real GDP Growth is the Survey of Professional Forecasters consensus forecast for one-year-ahead real GDP growth, in percent.

	Obs	Mean	Median	SD
Panel A: Full Sample (Individual Level)				
LLM Optimism	6,075	56.57	59.33	16.19
CFO Survey Optimism	6,075	60.13	60.00	17.62
Panel B: Cross-Sectional Disagreement (Quarterly)				
LLM Optimism (within-quarter SD)	93	12.31	12.21	3.04
CFO Optimism (within-quarter SD)	93	16.53	16.09	2.82
Disagreement Ratio (CFO/LLM)	93	1.42	1.35	0.46
Panel C: Time-Series Variation (Quarterly Aggregates)				
LLM Optimism	93	58.61	60.33	9.47
CFO Survey Optimism	93	60.78	61.09	6.70
Michigan Consumer Sentiment	93	80.53	81.60	12.78
SPF Real GDP Growth	93	2.70	2.72	0.75

effects (column 2) absorbs each firm's time-invariant average optimism. Because the survey is typically answered by a single executive per firm, the firm fixed effect effectively differences out that respondent's own long-run mean, so the coefficient is identified from *within-firm, over-time* variation, equivalently, movements in a given CFO's optimism around their personal average. Estimating off this variation is demanding: it asks not whether the model can sort persistently optimistic executives from persistently pessimistic ones, but whether it tracks when a particular CFO turns more or less optimistic than usual. The LLM score survives this test, remaining large and significant ($t = 12.86$), so its predictive content does not stem merely from persistent differences across firms or executives. Adding year-quarter fixed effects instead (column 3), which absorb all variation common across CFOs within a given quarter, also leaves the score highly significant ($t = 16.72$), implying that the model discriminates across CFOs *within* the same quarter and not merely across quarters. The most demanding specification (column 4) includes firm and year-quarter fixed effects simultaneously, isolating, for each CFO, deviations from their own average that are also not shared by other executives that quarter; even here the LLM score remains strongly significant ($t = 6.67$), explaining nearly half the variation in individual optimism.

Columns (5)–(8) confront the most natural concern about our history-conditioning design: that the LLM may simply echo the respondent's own most recent answer, which is one of the inputs we supply. To address this, we restrict the sample to observations for which the respondent answered in the immediately preceding survey wave and include that prior answer (*Last-Quarter CFO Optimism*) directly as a control. The LLM score remains a strong, significant predictor across all four specifications, with coefficients between 0.29 and 0.53. In other words, the model contributes predictive information about the current response over and above the persistence captured by the respondent's last answer; the LLM is doing more than copying history forward. The prior answer itself is significant in the simplest specification but is largely subsumed once fixed effects are added, whereas the LLM score is not.

Table 2: Predicting Individual CFO Survey Optimism

This table reports OLS coefficients from regressions of individual CFO Survey optimism scores (0–100 scale) on the matched LLM-generated optimism score. The sample covers public-company CFOs from the Duke–Federal Reserve CFO Survey, 2002–2025. Columns (1)–(4) use all matched observations and progressively add firm and year-quarter fixed effects. Columns (5)–(8) repeat these specifications on the subsample of respondents who also answered in the immediately preceding survey wave, additionally controlling for that prior response (*Last-Quarter CFO Optimism*). Standard errors are two-way clustered at the firm and year-quarter levels (Cameron et al., 2011); t-statistics are in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Dep. variable = CFO Survey Optimism							
	No History Control				Control for Last-Quarter Resp.			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
LLM Optimism	0.565***	0.429***	0.578***	0.276***	0.521***	0.527***	0.489***	0.286**
	(17.98)	(12.86)	(16.72)	(6.67)	(7.50)	(7.04)	(3.94)	(2.33)
Last-Quarter CFO Optimism					0.217***	0.0366	0.251**	0.171*
					(3.04)	(0.51)	(2.21)	(1.84)
Firm FE		Y		Y		Y		Y
Year-Quarter FE			Y	Y			Y	Y
N	6,075	5,631	6,075	5,631	1,980	1,818	1,980	1,818
R^2	0.269	0.446	0.305	0.494	0.502	0.612	0.547	0.662

This individual-level predictive content connects to a growing body of evidence that managerial expectations are not simply noisy reflections of publicly available macro data. Using data from the Duke CFO survey, Gennaioli et al. (2016) show that managerial expectations contain extrapolative components that predict investment; Gormsen and Huber (2024) demonstrate that CFO beliefs about the cost of capital vary systematically across firms in ways that affect real decisions; and Jha et al. (2023) find that AI-based measures of executive communication carry incremental information about real activity. Our LLM score, by design, approximates what an AI version of the executive would report given the firm's public footprint and the executive's own track record. That it predicts the actual human response, even after firm and time fixed effects and a control for the prior answer, suggests that the model successfully

reconstructs at least part of the information that makes managerial expectations valuable in the first place.

B. More Information, Better Predictions

If the LLM is genuinely using the information we supply rather than guessing, predictive accuracy should improve as we give the model more to work with. Table 3 confirms this with two complementary information splits. The dependent variable is again individual CFO Survey optimism, and each column reports the baseline (no fixed effects) regression of that variable on the matched LLM score within a subsample defined by the amount of information available.

Columns (1)–(3) split the sample by how much of the respondent's own history the model received. When no prior responses are available, typically a respondent's first appearance in the survey, the LLM score still predicts optimism, but explains only about 10% of the variation ($R^2 = 0.100$). When one to three prior data points are available, the fit more than triples ($R^2 = 0.334$), and when more than three prior responses are supplied, it increases to nearly half ($R^2 = 0.494$). The model appears to be using the respondent's track record to sharpen its individual-level forecasts.

Columns (4)–(5) split the sample by the breadth of firm characteristics included in the prompt. Not every survey wave records all 12 profile fields, and the median prompt contains 10 characteristics, which is the threshold we use to divide the sample. The full list of profile fields are provided in Appendix C. When fewer than 10 are available, the LLM score explains about 21% of the variation, whereas prompts containing ten to twelve characteristics achieve a better fit ($R^2 = 0.317$). Both information dimensions point in the same direction: richer conditioning yields a more faithful synthetic executive. These dose–response patterns are useful both diagnostically, since they indicate that the model responds to substantive inputs, and practically, since they identify the conditions under which synthetic survey responses are most reliable.

Table 3: Predictive Accuracy by Amount of Information

This table reports OLS coefficients from baseline (no fixed effects) regressions of individual CFO Survey optimism on the matched LLM-generated optimism score, estimated separately within subsamples defined by the information available to the model. Columns (1)–(3) partition the sample by the number of the respondent's own prior survey responses supplied in the prompt: none, one to three, and more than three. Columns (4)–(5) partition the sample by the number of firm characteristics included in the prompt: fewer than ten, and ten to twelve. The sample covers public-company CFOs, 2002–2025. Standard errors are two-way clustered at the firm and year-quarter levels; t-statistics are in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Dep. variable = CFO Survey Optimism				
	Respondent History			Firm Characteristics	
	None	1–3	>3	<10	10–12
	(1)	(2)	(3)	(4)	(5)
LLM Optimism	0.367***	0.604***	0.743***	0.535***	0.585***
	(8.65)	(23.29)	(19.62)	(17.08)	(14.56)
N	2,237	2,236	1,602	3,102	2,973
R^2	0.100	0.334	0.494	0.214	0.317

C. Heterogeneity in Fit Across Quarters

The quality of the LLM's individual-level predictions is not uniform across time. To illustrate, we estimate the cross-sectional R^2 from regressing CFO optimism on the matched LLM score *separately within each of the 93 quarters*, and then rank these quarterly R^2 values. The distribution is wide: the median quarter has an R^2 near 0.21, but the figure ranges from close to zero in a few early quarters to above 0.90 in the best-fitting quarter. Consistent with the information-quantity results in Table 3, fit tends to be better in later quarters, when more respondents have accumulated the survey history that the model conditions on.

Figure 1 makes the within-quarter relationship concrete by plotting individual CFO optimism against the matched LLM score for four representative quarters, chosen at the 25th, 50th, 75th, and 100th percentiles of the quarterly R^2 distribution. Even at the 25th percentile

(2020Q1), the relationship is positive; at the median (2015Q1) and 75th percentile (2014Q1) it is visibly tighter; and in the best-fitting quarter (2022Q2) the LLM score lines up almost one-for-one with the human responses. The panels also distinguish respondents by information set: cyan-filled markers identify CFOs for whom the model conditions on prior survey history, while unfilled markers identify first-time respondents. The “first-timer" points in the 50th and 75th percentiles cluster in a narrow horizontal band (in the absence of respondent-specific history, the model collapses toward a common prior and generates little cross-sectional spread), whereas the filled points fan out across the optimism range and trace the upward-sloping fit. This visual pattern is the within-quarter analogue of the dose-response relationship in Table 3: the model's ability to differentiate one CFO from another comes largely from the history it is given.

D. Robustness: Aggregating to the Quarterly Level

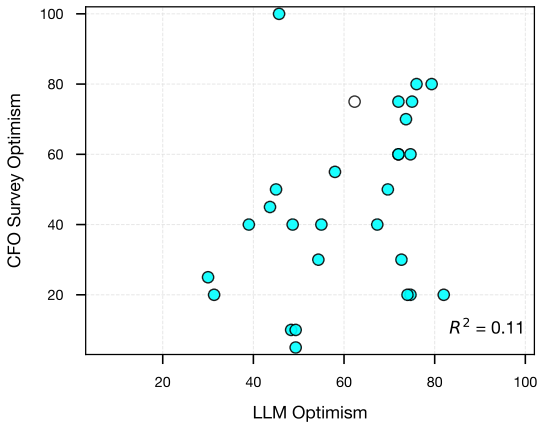
Our individual-level results are estimated off the panel structure of the matched data. As a robustness check, we ask whether the relationship survives once that structure is removed, collapsing the data to the quarterly frequency by averaging both the LLM score and the CFO Survey optimism score within each year-quarter. This forces the LLM to track movements in average CFO optimism over time rather than differences across respondents within a quarter, and lets us benchmark it against the two established macroeconomic gauges. Table 4 reports these regressions of quarterly mean CFO optimism on quarterly mean LLM optimism, the Michigan Consumer Sentiment Index, and the SPF year-ahead real GDP growth forecast.

The LLM score is a strong predictor of aggregate CFO optimism. On its own alongside Michigan sentiment (column 1) it enters with a coefficient of 0.579 ($t = 7.11$); paired with the SPF forecast (column 2) the coefficient is 0.586 ($t = 8.82$). Column (3) shows the fit obtainable from the two human benchmarks alone (Michigan and SPF, $R^2 = 0.316$), and column (4) places all three measures in a horse race. There the LLM score remains highly significant ($t = 6.86$), while neither Michigan sentiment nor the SPF forecast enters significantly once the LLM score is included. The fact that LLM optimism retains independent explanatory power for quarterly CFO sentiment even after controlling for consumer sentiment and professional forecasts indicates that the model captures variation in executive optimism that the standard gauges miss, and confirms that the individual-level results are not an artifact of the panel.

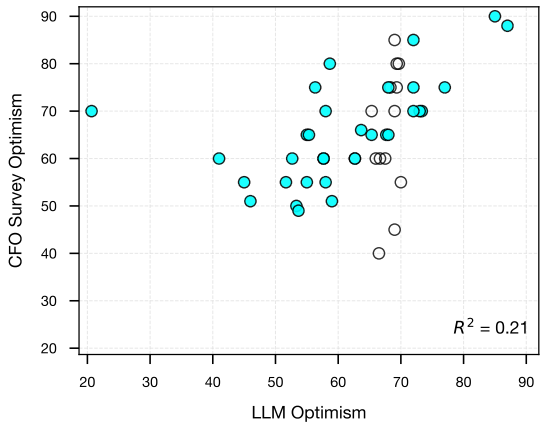
Figure 1: LLM vs. CFO Optimism Across the Distribution of Quarterly Fit

Each panel is a scatter plot of individual CFO Survey optimism (vertical axis) against the matched mean LLM optimism score (horizontal axis) for a single survey quarter, with the OLS fit line shown. *Cyan-filled* markers denote respondents for whom the model conditions on prior survey history (i.e., the respondent has appeared in an earlier wave), while *unfilled* markers denote respondents appearing in the sample for the first time, for whom no own-history is available. Unfilled observations exhibit markedly lower dispersion in the LLM score within a given quarter, as the model defaults toward a narrower prior in the absence of respondent-specific information. The four quarters are chosen at the 25th, 50th, 75th, and 100th percentiles of the distribution of within-quarter R^2 values across the 93 sample quarters: *(A)* 2020Q1 ($R^2 = 0.11$), *(B)* 2015Q1 ($R^2 = 0.21$), *(C)* 2014Q1 ($R^2 = 0.40$), and *(D)* 2022Q2 ($R^2 = 0.92$).

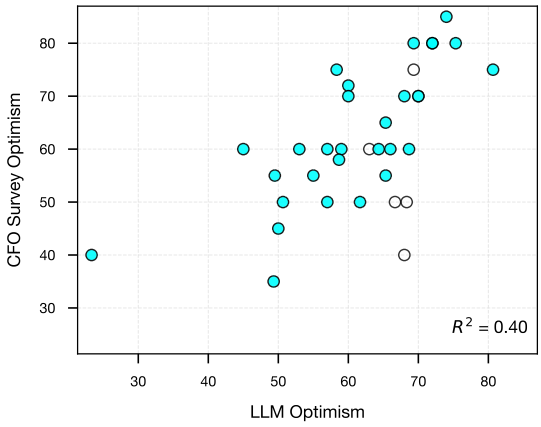
(A) 25th Percentile: 2020Q1



(B) 50th Percentile: 2015Q1



(C) 75th Percentile: 2014Q1



(D) 100th Percentile: 2022Q2

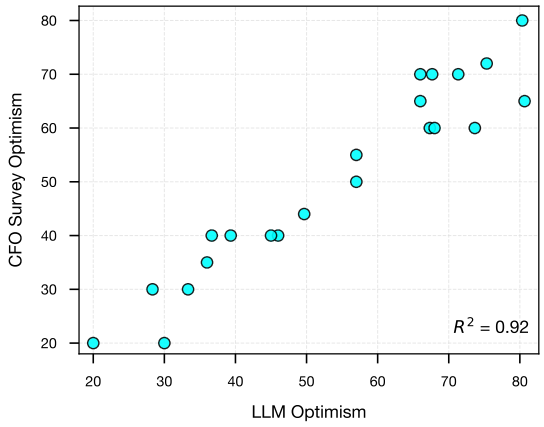


Table 4: LLM Optimism and Quarterly CFO Sentiment

This table reports quarterly-level OLS regressions of mean CFO Survey optimism on mean LLM optimism and macroeconomic benchmarks. The data are collapsed to one observation per year-quarter by averaging within quarter. *LLM Optimism* is the quarterly mean LLM score; *Michigan Consumer Sentiment* is the index level; *SPF Real GDP Growth* is the Survey of Professional Forecasters consensus forecast for one-year-ahead real GDP growth. The sample spans 2002–2025 (93 quarters). Standard errors are Newey–West (Newey and West, 1987) with 4 lags; t-statistics are in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Dep. variable = CFO Survey Optimism (Mean)			
	(1)	(2)	(3)	(4)
LLM Optimism (Mean)	0.579***	0.586***		0.574***
	(7.11)	(8.82)		(6.86)
Michigan Consumer	0.0228		0.264***	0.0163
Sentiment	(0.47)		(3.24)	(0.33)
SPF Real		0.479	1.161	0.432
GDP Growth		(1.18)	(1.10)	(0.98)
N	93	93	93	93
R^2	0.715	0.716	0.316	0.717

IV. Discussion

The results have implications for the academic study of expectations, for policy, and for practice.

A new expectations measure. Macroeconomic expectations research has long been constrained by a tension between depth and breadth. Surveys such as the Duke–Federal Reserve CFO Survey and the SPF offer economically meaningful readings of insider sentiment, but their coverage is narrow, and their frequency is low. Dictionary-based text measures, constructed from earnings calls, 10-K filings, or news coverage (Tetlock, 2007; Loughran and McDonald, 2011; Hansen and McMahon, 2016), provide broader coverage but are largely backward-looking, quantifying the tone of language that has already been written or spoken. More recent work shows that LLMs applied to earnings press releases and financial disclosures can extract forward-looking signals

(Bai et al., 2023; Jha et al., 2023; Siano, 2025). Our approach pushes this logic further: rather than processing an archival corporate disclosure, the LLM itself generates the expectation from its learned representation of the firm, the executive's own track record, and the prevailing macroeconomic environment. Because no source document is required, the measure can be produced for any firm with a public footprint, at any date, and at any frequency – a combination of scalability, timeliness, and individual-level fidelity that is, to our knowledge, unavailable from any existing data source.

The scalability opens several research designs that are currently infeasible. Researchers studying how expectations respond to policy shocks (monetary announcements, trade actions, regulatory changes) could generate daily or weekly LLM sentiment scores for the full cross-section of public firms, circumventing the quarterly- or monthly-frequency constraint that limits event-study designs with traditional survey data. Because the model can be prompted with counterfactual company characteristics (e.g., varying foreign-revenue exposure while holding everything else fixed), it also enables a form of "synthetic survey experiment" that has no natural analog in human data.

The role of conditioning information. A central lesson of our analysis is that the fidelity of a synthetic executive depends heavily on what the model is told. Conditioning each persona on the respondent's own history sharply improves individual-level prediction, restores the cross-sectional heterogeneity that instruction-tuned models otherwise suppress (Santurkar et al., 2023), and stabilizes the model's output across replications. The monotone relationship between information and accuracy (Table 3) suggests that the upper bound on synthetic-survey fidelity is governed less by the model's raw capability than by the quality and quantity of the conditioning data, which is encouraging for applications in which rich firm- and respondent-level context is available.

Policy applications. The Federal Reserve and other central banks regularly monitor the Duke-Fed CFO Optimism Index as a leading indicator of business conditions.[6] An LLM-based analog available at higher frequency and broader coverage could improve the timeliness of assessments in the inter-meeting period between FOMC decisions. AI-generated expectations could also complement the Federal Reserve's Beige Book process, in which regional contacts provide qualitative assessments of economic conditions (Balke and Petersen, 2002). An LLM-generated

[6]The CFO Survey has been jointly conducted with the Federal Reserve Banks of Richmond and Atlanta since 2020; see `https://cfosurvey.fuqua.duke.edu`.

"synthetic Beige Book" covering the universe of public firms, conditioned on information available at the date of each report, would provide a systematic benchmark against which to evaluate the incremental content of human contact reports.

Limitations and caveats. Several limitations merit discussion. First, while the LLM score captures substantial firm- and individual-level variation, the residual variation is still considerable: actual CFOs possess private information (pending orders, internal cost pressures, board-level strategy shifts) that no model conditioned on public data and survey history can fully recover. The LLM is a complement to, not a replacement for, the human survey (Wang et al., 2025). Second, a history-conditioning design requires that respondents appear more than once; for first-time respondents the model has less to work with, and predictive accuracy is correspondingly lower (Table 3, column 1). Third, like any forecasting exercise, the model is a product of its training data, and its "beliefs" may embed historical regularities that break down in structurally novel regimes. Although our information-cutoff strategy and the fact that the target was never in the training corpus address the memorization concern of Lopez-Lira et al. (2025), we cannot fully rule out indirect contamination. Two avenues would further alleviate this concern: training a time-stamped model at each point in time (Sarkar, 2024), which is computationally demanding (He et al., 2025); and out-of-sample validation conducted live at the time of future quarterly surveys, which we plan to undertake in future research.

V. Conclusions

This paper presents a method for generating synthetic CFO expectations using large language models. Prompting the LLM to role-play as the CFO of a specific public company at a specific point in time, and supplying, where available, the respondent's own pre-cutoff history, we produce synthetic optimism scores matched to survey responses spanning 2002–2025. Three main findings emerge.

First, the LLM score is a significant predictor of the actual CFO's response at the individual level. The predictive relationship survives firm fixed effects, year-quarter fixed effects, and a direct control for the respondent's most recent prior answer, indicating that the model contributes information beyond simple persistence.

Second, predictive accuracy scales with the information provided. Both the amount of a respondent's own history and the breadth of firm characteristics in the prompt significantly

improve the fit, with the cross-sectional R^2 increasing from roughly 0.10 to nearly 0.50 as more respondent history is provided. The model demonstrably uses the conditioning information.

Third, the relationship between human CFO sentiment and digital twin sentiment is robust to aggregation. At the quarterly level, mean LLM optimism remains a significant predictor of mean CFO optimism even after controlling for the Michigan Consumer Sentiment Index and the SPF professional forecast, indicating that the model captures executive sentiment that established gauges miss.

Several directions remain for future work. Because own-history conditioning drives individual-level fit (Table 3), firms without survey history produce the weakest synthetic forecasts; one way to relax this is to borrow a comparable peer's human CFO history, for example forecasting Ford while supplying GM's survey responses, extending to human history the comparable-company logic the model already applies to fundamentals. For policy use, where an aggregate indicator is the object of interest, an equal-weighted mean of the twins may not be optimal: weighting each twin by its past accuracy may track aggregate sentiment more closely than a simple average. Live validation against future survey waves, free of any residual contamination concern, is a third natural step.

These results suggest that LLMs might be able to serve as credible synthetic executives, particularly when conditioned on a rich firm- and respondent-level context. For researchers, the method provides a new tool for studying belief formation and information aggregation at the firm level. For practitioners and policymakers, it offers the prospect of a real-time, scalable, individual-level sentiment measure.

References

Argyle, L. P., E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis 31*(3), 337–351.

Bai, J. J., N. M. Boyson, Y. Cao, M. Liu, and C. Wan (2023). Executives vs. chatbots: Unmasking insights through human-AI differences in earnings conference Q&A. SSRN Working Paper 4480056.

Bali, T. G., B. T. Kelly, M. Mörke, and J. Rahman (2025). Machine forecast disagreement. *Review of Financial Studies*. Forthcoming.

Balke, N. S. and D. Petersen (2002). How well does the Beige Book reflect economic activity? Evaluating qualitative information quantitatively. *Journal of Money, Credit and Banking 34*(1), 114–136.

Baslandze, S., Z. Edwards, J. R. Graham, T. McClure, B. H. Meyer, M. Sparks, S. R. Waddell, and D. Weitz (2026). Artificial intelligence, productivity, and the workforce: Evidence from corporate executives. Technical report. Forthcoming.

Ben-David, I., J. R. Graham, and C. R. Harvey (2013). Managerial miscalibration. *Quarterly Journal of Economics 128*, 1547–1584.

Boutros, M., I. Ben-David, J. R. Graham, C. R. Harvey, and J. W. Payne (2025). The persistence of miscalibration. *The Review of Financial Studies*. Advance access, article hhaf070.

Brand, J., A. Israeli, and D. Ngwe (2023). Using GPT for market research. Working Paper, Harvard Business School.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology 3*(3), 296–322.

Bybee, J. L. (2025). The ghost in the machine: Generating beliefs with large language models. Working Paper, University of Chicago Booth School of Business.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics 29*(2), 238–249.

Cao, S., W. Jiang, J. Wang, and B. Yang (2024). From man vs. machine to man + machine: The art and AI of stock analyses. *Journal of Financial Economics 160*, 103910.

Coibion, O. and Y. Gorodnichenko (2012). What can survey forecasts tell us about information rigidities? *Journal of Political Economy 120*, 116–159.

Faria-e Castro, M. and F. Leibovici (2024). Artificial intelligence and inflation forecasts. *Federal Reserve Bank of St. Louis Review 106*(12), 1–14.

Fried, S., T. Graf, S. Malhotra, and S. R. Singh (2026, May). What do financial officers predict for price growth? *FRBSF Economic Letter 2026*(11). Federal Reserve Bank of San Francisco.

Fritsch, G. P. and J. Z. Mazlish (2026, March). High-frequency fiscal shocks. Preliminary working paper.

Fuhrer, J. C. (1993, January). What role does consumer sentiment play in the U.S. macroeconomy? *New England Economic Review*, 32–44.

Gao, Z., W. Jiang, and Y. Yan (2025). A test of lookahead bias in LLM forecasts.

Gennaioli, N., Y. Ma, and A. Shleifer (2016). Expectations and investment. *NBER Macroeconomics Annual 30*, 379–431.

Gormsen, N. J. and K. Huber (2024, Jun). Firms' perceived cost of capital. NBER Working Papers 32611, National Bureau of Economic Research, Inc.

Goulding, C. L., C. R. Harvey, and H. Kurtović (2026, April). Disagreement of disagreement. *SSRN 4647471*.

Graham, J., B. Meyer, N. Parker, and S. R. Waddell (2020, May). Introducing The CFO Survey: A collaboration between Duke University's Fuqua School of Business and the Federal Reserve Banks of Richmond and Atlanta. https://www.richmondfed.org/research/national_economy/cfo_survey/research_and_commentary/2020/20200515_introducing_cfo_survey. Accessed: April 3, 2026.

Graham, J. R. (2022). Presidential address: Corporate finance and reality. *Journal of Finance 77*, 1975–2049.

Graham, J. R. and C. R. Harvey (2001). The theory and practice of corporate finance: Evidence from the field. *Journal of Financial Economics 60*, 187–243.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of Financial Studies 33*, 2223–2273.

Hansen, A. L., J. J. Horton, S. Kazinnik, D. Puzzello, and A. Zarifhonarvar (2025). Simulating the survey of professional forecasters. Working Paper, SSRN 5066286.

Hansen, S. and M. McMahon (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics 99*, S114–S133.

He, S., L. Lv, A. Manela, and J. Wu (2025). Chronologically consistent large language models. Revise and Resubmit at *Journal of Financial Economics*. arXiv:2502.21206.

Heaton, J. B., N. G. Polson, and J. H. Witte (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry 33*(1), 3–12.

Henning, T., S. M. Ojha, R. Spoon, J. Han, and C. F. Camerer (2025). Llm agents do not replicate human market traders: Evidence from experimental finance.

Horton, J. J., A. Filippas, and B. S. Manning (2023, April). Large language models as simulated economic agents: What can we learn from homo silicus?

Jha, M., J. Qian, M. Weber, and B. Yang (2023). Chatgpt and corporate policies. NBER Working Paper 32161.

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

Li, P., N. Castelo, Z. Katona, and M. Sarvary (2024). Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science 43*(2), 254–266.

Lopez-Lira, A. and Y. Tang (2023). Can ChatGPT forecast stock price movements? return predictability and large language models. Working Paper.

Lopez-Lira, A., Y. Tang, and M. Zhu (2025). The memorization problem: Can we trust LLMs' economic forecasts? Working Paper.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-Ks. *Journal of Finance 66*, 35–65.

Ludvigson, S. C. (2004). Consumer confidence and consumer spending. *Journal of Economic Perspectives 18*(2), 29–50.

Malmendier, U. and G. Tate (2015). Behavioral ceos: The role of managerial overconfidence. *Journal of Economic Perspectives 29*, 37–60.

Manning, B. S. and J. J. Horton (2025). General social agents. Working Paper, arXiv:2508.17407.

Manning, B. S., K. Zhu, and J. J. Horton (2024). Automated social science: Language models as scientist and subjects. NBER Working Paper 32381.

Manski, C. F. (2004). Measuring expectations. *Econometrica 72*, 1329–1376.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

OpenAI (2026). Introducing gpt-5.4. Technical report, OpenAI.

Santurkar, S., E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto (2023). Whose opinions do language models reflect?

Sarkar, S. K. (2024). StoriesLM: A family of language models with time-indexed training data. Available at SSRN 4881024.

Sarkar, S. K. and K. Vafa (2025). Lookahead bias in pretrained language models. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Shrout, P. E. and J. L. Fleiss (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin 86*(2), 420–428.

Siano, F. (2025, November). The news in earnings announcement disclosures: Capturing word context using llm methods. *Management Science 71*(11), 9831–9855.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology 3*(3), 271–295.

Suh, J., E. Jahanparast, S. Moon, M. Kang, and S. Chang (2025). Language model fine-tuning on scaled survey data for predicting distributions of public opinions.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance 62*, 1139–1168.

Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy (2008). More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance 63*, 1437–1467.

Uddin, A., X. Tao, and D. Yu (2024). The network factor of equity pricing: A signed graph Laplacian approach. *Journal of Financial Econometrics 22*(5), 1616–1655.

Vafa, K., A. Rambachan, and S. Mullainathan (2024). Do large language models perform the way people expect? measuring the human generalization function.

Wang, M., D. J. Zhang, and H. Zhang (2025). Large language models for market research: A data-augmentation approach.

Wang, Y., H. Gao, C. R. Harvey, Y. Liu, and X. Tao (2026). Machine learning meets Markowitz. Working Paper.

Wu, J. C., J. Xi, and S. Xie (2025, October). Llm survey framework: Coverage, reasoning, dynamics, identification. Working Paper 34308, National Bureau of Economic Research.

Appendix A. Data Construction

This appendix details the procedures used to construct the estimation sample from the raw survey responses from The CFO Survey, conducted by Duke University before June 2020 and jointly by Duke and the Federal Reserve Banks of Atlanta and Richmond since then.

A.1 Sample Selection

We begin with the universe of survey responses collected between 2002 and 2025. We restrict the sample to respondents at publicly traded companies, which allows us to link survey responses to externally verifiable firm characteristics (and allows the LLM to find public information about these firms). We further require a non-missing economic-optimism response. After applying these filters, the matched estimation sample contains 6,075 firm-quarter observations.

A.2 Company Name Standardization

Because the survey is administered repeatedly over time, the same company may appear under slightly different name variations across waves (e.g., "Acme Corp.", "ACME CORPORATION", "Acme, Inc."). To track companies across multiple survey waves, we apply the following standardization procedure to all company names:

1. Convert all characters to uppercase.
2. Remove common legal suffixes, including Inc., Corp., Corporation, LLC, LLP, Ltd., L.P., Co., and their variations (with or without trailing periods).
3. Eliminate parenthetical text (e.g., "(US)" or "(formerly XYZ)").
4. Strip leading and trailing whitespace, collapse consecutive internal spaces to a single space, and remove extraneous punctuation such as trailing commas or periods.

After standardization, we assign a unique firm identifier (PERMCO, a permanent company identifier from the Center for Research in Security Prices) to each distinct cleaned name. This identifier is used to construct the firm panel and to apply firm fixed effects in Table 2.

A.3 Respondent Linkage and History

To construct each respondent's history, we assign a stable person key based on the strongest available identifier, preferring a respondent identifier, then a cleaned contact email, then an

IP address, and propagate identity and profile fields across rows sharing any of these keys. For each respondent and survey date, we then assemble (i) the respondent's own optimism scores from prior waves, (ii) the LLM's prior mean forecasts for that respondent, and (iii) the respondent's lifetime-to-date average optimism. All three are restricted to dates strictly before the survey cutoff and to a window of the preceding twelve calendar quarters; up to twelve prior data points are injected into the prompt. The runner processes quarters in chronological order so that the LLM's prior forecasts are available as history for later quarters of the same respondent.

A.4 Missing Data Imputation

The survey instrument collects several firm-level characteristics at each wave: job title, industry sector, sales revenue, total employment, percentage of foreign sales, and headquarters location. However, not all respondents complete every field in every wave.

To maximize the information available in each LLM prompt, we impute missing values using a within-firm temporal nearest-neighbor procedure. Specifically, for each firm-quarter observation with a missing demographic field, we search for the same firm's responses in adjacent survey waves and fill the gap with the value from the closest available date (preferring the most recent prior response when equidistant). This approach assumes that firm characteristics such as industry classification, headquarters location, and broad revenue and employment categories change slowly relative to the quarterly survey frequency. When a field remains missing after this imputation step, we include only the available information in the LLM prompt, allowing the model to draw on its broader knowledge base for context.

A.5 Variable Definitions

Table A1 summarizes the key variables used in the analysis.

Table A1: Variable Definitions

Variable	Definition
CFO Survey Optimism	Self-reported optimism about the overall U.S. economy on a 0–100 scale, from the Duke–Federal Reserve CFO Survey.
LLM Optimism	GPT-5.4-generated optimism score (0–100) obtained by prompting the model to role-play as the CFO of a specific firm at the survey date, conditioned on the firm's profile and the respondent's pre-cutoff history. Reported as the average of three independent replications.
Last-Quarter CFO Optimism	The same respondent's CFO Survey optimism score from the immediately preceding survey wave (used in Table 2, columns (5)–(8)). The any-prior-quarter analog is used in Table A3.
Michigan Consumer Sentiment	University of Michigan Index of Consumer Sentiment (`ics_all`), quarterly.
SPF Real GDP Growth	Mean forecast of year-ahead real GDP growth (`drgdp5`) from the Survey of Professional Forecasters, Federal Reserve Bank of Philadelphia.

A.6 Quarterly Aggregation

For the aggregate-level analysis in Table 4, we collapse the individual-level panel to one observation per quarter by computing the mean LLM optimism score and the mean CFO Survey optimism score within each year-quarter. Each quarterly observation is then matched to the contemporaneous Michigan Consumer Sentiment Index and the SPF consensus forecast for year-ahead real GDP growth. The resulting dataset yields 93 quarterly observations.

Appendix B. System Prompt

Financial executives provide forecasts about their expectations of the future, from the perspective of their own companies, in The CFO Survey, which is jointly conducted by Duke University and the Federal Reserve Banks of Richmond and Atlanta.

Your task is to prepare forecasts as if you are a financial executive working at a specific

company at a specific point in time, and then repeat the task as if you are a different financial executive working at a different company, and keep repeating this task the specified number of times.

When preparing forecasts: (i) use only the information available as of the `CUTOFF_DATE`; (ii) search for information about the specific company you are role-playing—if detailed information is not available, use information on comparable companies, drawing from financial statements, analyst reports, and industry data; (iii) if the profile includes this person's own past survey responses, treat them as a strong prior on this individual's style, optimism bias, and level—your answer should be consistent with their personal pattern unless clearly contradicted by new information as of the `CUTOFF_DATE`; (iv) if the profile includes the prior LLM forecasts you have previously produced for this same person, treat those as a secondary prior—continuity matters, but the actual human responses (when present) take precedence; (v) explicitly incorporate the following dimensions into your thought process.

Historical Financial Performance: review past 3–5 years of revenue data (quarterly and annual); identify growth trends, seasonality patterns, and cyclical behaviors; analyze revenue composition by business segments, product lines, or geographic regions; examine profit margins and their relationship to revenue growth.

Forward-Looking Company Indicators: latest management guidance and earnings call transcripts; product pipeline, launch schedules, and innovation roadmap; customer acquisition trends, retention rates, and churn analysis; order backlog, booking trends, and contract renewals; expansion plans; sales force changes and marketing investment levels.

Analyst Coverage: collect and synthesize recent analyst reports and price targets; review consensus revenue estimates and any recent revisions; identify key assumptions driving analyst projections; note any divergent views or concerns raised by analysts.

Comparable Company Analysis (if primary data unavailable): focus on industry peers with similar business models, revenue streams, market capitalization, geographic footprint, and customer base; analyze revenue growth patterns, market share trends, and competitive positioning; evaluate responses to industry headwinds and tailwinds; review valuation multiples, growth assumptions, and recent strategic initiatives.

Market and Industry Context: overall industry growth projections and market size; key industry trends, disruptions, and regulatory changes; competitive landscape and market share dynamics; customer demand patterns and buyer behavior shifts; technology adoption rates

and innovation cycles.

Macroeconomic Factors: GDP growth forecasts for relevant geographic markets; interest rate environment and credit conditions; currency exchange rate impacts (for international operations); inflation effects on pricing power and costs; sector-specific economic indicators.

For each simulation, assume the role of one survey participant and provide forecasts. Incorporate all relevant details about overall economic conditions up to the `CUTOFF_DATE` and exclude any data or developments occurring after the `CUTOFF_DATE`. Internally consider historical performance, company guidance, analyst views, industry conditions, macro factors, and the respondent's own history—think step-by-step before answering. Do *not* output or reveal internal chain-of-thought or reasoning. For each simulated CFO produce exactly one line of output in this exact format: `<number>`, where `<number>` is the forecast number only (no percent sign, no decimals unless necessary).

Survey Question: Rate your optimism about the overall U.S. economy on a scale from 0–100, with 0 being the least optimistic and 100 being the most optimistic.

Appendix C. Sample Query

`CUTOFF_DATE` for this exercise: 2024-02-15

Profile for this exercise:

Name: [confidential]
Current role: CFO
Company: [confidential]
Ticker: [confidential]
Industry: [confidential]
Sales revenue: $1–4.9 billion
Number of employees: 2,500–4,999
Percent of sales in foreign countries: More than 50%
Headquarters location: Midwest U.S.
State: [confidential]
Ownership: Public
Credit rating: [confidential]

Respondent history:

This person's own past U.S.-economy optimism responses (0–100), up to 12 most recent prior responses, within the last 12 calendar quarters:

– 2023Q1: 55.0
– 2023Q2: 58.0
– 2023Q4: 60.0

Your (the LLM's) prior forecasts for this same person (mean of 3 runs), up to 12 most recent prior responses, within the last 12 calendar quarters:

– 2023Q1: 54.3
– 2023Q2: 56.7
– 2023Q4: 59.0

This person's average U.S.-optimism response across the last 12 calendar quarters (N=3): 57.7

Appendix D. Forecast Stability

Table A2: Stability of LLM-Generated Forecasts Across Replications

This table reports within-prompt stability statistics for the LLM-generated optimism scores. For each survey observation, three independent forecasts are generated via separate API calls with identical prompts. Pairwise correlations are computed across all replication pairs. The within-prompt standard deviation and coefficient of variation (CV) are computed row-wise. The intraclass correlation coefficient (ICC) is estimated via one-way ANOVA (Shrout and Fleiss, 1979).

Statistic	**Value**
Pairwise Correlations Across Replications	
Replication 1 vs. Replication 2	0.958
Replication 1 vs. Replication 3	0.959
Replication 2 vs. Replication 3	0.957
Within-Prompt Variation	
Mean standard deviation	2.22
Mean coefficient of variation	4.9%
Intraclass Correlation (One-Way ANOVA)	
ICC	0.958
95% confidence interval	[0.956, 0.959]
Reliability of group mean ($n \approx 3$)	0.985

Appendix E. Controlling for Any Prior Response

Table 2 (columns (5)–(8)) controls for the respondent's answer in the *immediately preceding* survey wave. Table A3 reports the analogous specifications controlling instead for the respondent's optimism in *any* prior quarter for which it is available (the respondent's prior optimism on record), which preserves a larger sample. The LLM score remains a strong, significant predictor throughout, confirming that its predictive content is not merely a restatement of the respondent's recent history.

Table A3: Predicting CFO Optimism, Controlling for Any Prior Response

This table repeats the specifications of Table 2, columns (5)–(8), on the subsample of respondents with at least one prior survey response, controlling for the respondent's prior optimism on record (*Prior CFO Optimism*) rather than specifically the immediately preceding wave. The dependent variable is individual CFO Survey optimism. Columns progressively add firm and year-quarter fixed effects. Standard errors are two-way clustered at the firm and year-quarter levels; t-statistics are in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Dep. variable = CFO Survey Optimism			
	(1)	(2)	(3)	(4)
LLM Optimism	0.583***	0.564***	0.507***	0.312***
	(12.92)	(12.56)	(7.60)	(4.32)
Prior CFO	0.0880**	−0.0828**	0.140***	0.0483
Optimism	(2.11)	(−2.07)	(2.64)	(1.09)
Firm FE		Y		Y
Year-Quarter FE			Y	Y
N	3,832	3,624	3,832	3,624
R^2	0.397	0.519	0.434	0.566