Information Systems and Economics (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2026.070108
ISSN 2523-6407 Vol. 7 Num. 1

Research on the Integration Mechanism of Large Language Models and Analyst Consensus: An Empirical Study Based on Financial Forecasting

Jiqi Li

School of Economics and Management, Tongji University, Shanghai, China
jessiel@tongji.edu.cn

Keywords: Large Language Models, Human-AI Collaboration, Financial Forecasting, Prompt Engineering, Analyst Consensus

Abstract: When both a human analyst signal and large language model (LLM) judgment are available for a financial forecasting task, what governs the quality of the result is the way the two are combined rather than the mere fact of combining them. This paper compares three prompt-defined integration mechanisms implemented through a single commercial LLM over 6,363 U.S. firm-year observations. The mechanisms differ only in how each handles the analyst signal. One withholds it entirely and asks the model to read the statements on its own. Structured integration supplies the consensus alongside simple reliability metadata. Critical integration goes further and requires the model to challenge that signal through an explicit multi-step deliberative protocol. The engaged-use-of-AI principle from the human-AI collaboration literature predicts that the most deliberative mechanism should perform best, yet the evidence points the other way. Structured integration attains the highest accuracy at about 59%, while critical integration is the lowest at about 54%, falling below even the signal-free baseline at about 55%, and the gaps are statistically significant. The structured advantage concentrates in financially healthier and non-loss firms, where analyst signals are most reliable. Viewed alongside the emerging literature on LLM overthinking and the long-standing verbal-overshadowing effect, this reversal marks a boundary condition of the engaged-use principle. Critical engagement adds value when a human expert interrogates an AI, because the expert brings independent domain knowledge to the critique; transposed onto an LLM asked to scrutinize a human signal, the same protocol supplies no new information and introduces deliberative interference instead.

1. Introduction

The diffusion of large language models such as GPT-4, Claude, and DeepSeek into investment research has sharpened a question that predates them: when human judgment and machine judgment are both at hand, how should they be joined? Cao et al. [1] show that machine models beat many human analysts on average yet lose their edge in firms with heavy intangibles, financial distress, and opaque information, and that human-machine combinations outperform either input

used alone. Kim, et al. [2] show that an LLM guided by a chain-of-thought prompt can read anonymized financial statements and predict the direction of next-year earnings at a level approaching purpose-built neural networks. The design problem that sits between these two lines remains open. Once an analyst signal and an LLM are both available, the integration between them can be engineered in many ways, and those ways need not be equally good.

This paper takes that design problem as its subject and compares three integration mechanisms, all implemented through the same model so that any performance gap is attributable to the prompt rather than to the model. Pure model judgment receives only anonymized statements, reproducing the setup of Kim et al. [2]. Structured integration adds the analyst consensus direction along with two pieces of reliability metadata, the number of covering analysts and their level of agreement. Critical integration gives the model the same inputs but instructs it to form an independent view, locate conflicts with the consensus, decide how much to trust each side, and only then commit to a forecast. This last design follows the engaged-use principle of Lebovitz et al. [3], whose radiologists obtained better diagnoses precisely when they interrogated the AI rather than deferring to it.

The headline result runs against that motivation. Structured integration is the strongest of the three, and critical integration is the weakest, trailing structured integration by a wide and statistically reliable margin and falling below even the signal-free baseline. Two complementary literatures make sense of the pattern. The overthinking phenomenon documented for LLMs in computer science [4], [5] shows that longer deliberation can lower accuracy, while the verbal-overshadowing effect in cognitive psychology [6], [7] shows that putting a pattern-based judgment into words degrades it. Forcing an LLM to narrate an independent judgment, adjudicate a conflict, and justify a weighting appears to displace the direct pattern matching that produces its best forecasts.

The study makes three contributions. It locates the LLM-side boundary of the engaged-use principle, a principle examined until now only from the human side, and by doing so it carries the overthinking literature out of general reasoning benchmarks and into a structured professional task in finance. The practical payoff is a deflationary design lesson for investment-research tools, since a plain prompt that hands the model an analyst signal with a note on its reliability outperforms an elaborate protocol that asks the model to think harder.

2. Related Literature and Hypotheses

LLMs have moved quickly from headline sentiment [8] and central-bank language [9] to firm-level signals drawn from filings [10] and to earnings-direction prediction from statements alone [2], with text-based features that capture far more return-relevant information than dictionary methods [11]. The same literature is candid about the limits. Machine earnings predictions inherit the overreaction once thought uniquely human [12], hallucination and look-ahead bias persist [13][14], and the gains from generative AI accrue to analysts who already have judgment rather than replacing them[15].

That combination need not help is itself a finding. Dellermann et al. [16] frame hybrid intelligence around complementarity, and Fuegener et al. [17] show that the gains depend on the human's metacognitive accuracy. Lebovitz et al. [3] supply the engaged-use principle that motivates the critical design. Against this optimism, Vaccaro et al. [18] pool 370 studies and find that human-AI combinations on average underperform the better of the two on decision tasks, which means collaborative value is not automatic but has to be engineered.

A recent strand qualifies the assumption that more reasoning is better. Shorter LLM reasoning chains can dominate longer ones [4] , and explicit chain-of-thought can lower accuracy on the very

tasks where deliberation also hurts humans [5]. The pattern mirrors verbal overshadowing, where verbalizing a face or one's reasons for a preference [7] degrades the underlying judgment. For statement-based earnings-direction prediction, a task amenable to direct pattern matching, the implication is that heavy deliberative scaffolding may interfere rather than help.

Analyst signals are informative and biased at once. The classic record documents optimism [19], herding [20], and conflicts of interest [21], while recent work formalizes predictable over- and under-reaction within a diagnostic-expectations framework [22], documents the sticky expectations behind the profitability anomaly [23], and benchmarks analyst expectations against a real-time machine standard [24]. The consensus therefore carries both genuine private-information signal and predictable, bias-driven noise. Structured integration lets the model use this dual-natured signal as a light prior, whereas critical integration forces it to narrate the filtering. Whether the narration helps is the empirical question this paper resolves.

These strands converge on two testable predictions. The reasoning about deliberative interference, set against the engaged-use rationale, generates a clear ordering among the mechanisms once we recognize that an LLM interrogating a human signal contributes no independent information source to the adjudication.

Hypothesis 1 (Mechanism Ordering). Structured integration attains the highest forecasting accuracy, whereas critical integration does not improve on, and may fall below, pure model judgment.

The reasoning about signal quality generates a complementary prediction about where the structured advantage should be largest. Because the value of structured integration derives from the external analyst signal it injects, that value should rise with the signal's reliability and with the firm's information opacity to the model itself.

Hypothesis 2 (Heterogeneity). The advantage of structured integration over pure model judgment is larger in firms where the analyst signal is more reliable, namely financially healthier and non-loss firms, and in firms whose information is sparser to the model, namely smaller firms.

3. Research Design

The sample is built from Compustat, I/B/E/S, and CRSP through Wharton Research Data Services, restricted to U.S. firms with December fiscal year-ends, total assets above one million dollars, year-end price above one dollar, and coverage by at least three analysts within thirty days of the earnings announcement, following Kim et al. [2]. After stratified random sampling on year, market-capitalization tercile, and intangible-ratio tercile, the analysis sample comprises 6,363 firm-year observations over 2015 to 2022, which yields 19,089 long-format rows across the three mechanisms. The target is the direction of next-year EPS, coded one when next-year EPS exceeds current EPS and zero otherwise. Statements are anonymized to relative years with company names, industry codes, and absolute years removed.

The three mechanisms share inputs and output schema and differ only in the prompt. Pure model judgment receives the anonymized statements alone. Structured integration adds the consensus direction together with the number of covering analysts and a high, moderate, or low agreement label, and asks the model to run trend analysis, ratio analysis, a consensus evaluation, and an integrated prediction. Critical integration provides the same information but instructs the model not to anchor on the consensus and to proceed through four explicit steps, an independent judgment formed from the statements alone, a conflict-detection step comparing that judgment with the consensus, a trust-attribution step that names which signal is more reliable, and an integrated final forecast. Every forecast comes from GPT-4o at temperature zero, so the comparison isolates the integration design rather than the model or the output format.

The empirical strategy follows the two hypotheses. For Hypothesis 1, the analysis reports accuracy, balanced accuracy, and F1 for each mechanism, applies pairwise McNemar tests to the matched-sample disagreements, and estimates a panel fixed-effects regression on the long-format data with pure model judgment as the baseline, firm characteristics as controls, industry and year fixed effects, and standard errors clustered at the firm level. For Hypothesis 2, the analysis regresses each mechanism's accuracy improvement over pure model judgment on standardized firm characteristics.

4. Empirical Results

Structured integration records the highest accuracy and balanced accuracy of the three mechanisms, as shown in Table 1. Pure model judgment lands close to the human analyst consensus, while critical integration trails the field and sits below the signal-free baseline. The Naive rule that always predicts an increase attains the highest F1 but a balanced accuracy of exactly 0.5, which confirms that F1 alone is misleading under class imbalance and that balanced accuracy is the more honest yardstick here.

Table 1: Predictive accuracy across models (N = 6,363).

Model	Accuracy	Balanced Acc.	F1
Naive (always increase)	0.5603	0.5000	0.7182
Human analyst consensus	0.5694	0.5521	0.6441
Pure model judgment	0.5513	0.5517	0.5780
Structured integration	**0.5887**	**0.5918**	**0.6068**
Critical integration	0.5403	0.5464	0.5474

Pairwise McNemar tests confirm that these gaps are not noise. Structured integration beats pure model judgment with $p = 0.0016$, critical integration trails structured integration with $p < 0.0001$, and critical integration is significantly worse than pure model judgment with $p = 0.0310$. The panel fixed-effects regression delivers the same verdict in a form that controls for firm characteristics and for industry and year effects. Structured integration carries a coefficient of +0.0369 with a t-statistic of 4.25, an improvement of roughly 3.7 accuracy points over the baseline, whereas the critical coefficient of -0.0111 with a t-statistic of -1.27 is statistically indistinguishable from zero and points, if anything, slightly downward. Among the controls, financial health enters strongly positive, with the standardized Z-score coefficient at +0.0401 and a t-statistic of 10.89, and the loss indicator enters strongly negative at -0.0667 with a t-statistic of -7.22, which reproduces the familiar difficulty of calling earnings direction for loss-making firms. These results support Hypothesis 1.

The heterogeneity regressions support Hypothesis 2 and sharpen its economic reading. The improvement that structured integration delivers over pure model judgment grows with financial health, with a standardized Z-score coefficient of +0.1039 and a t-statistic of 10.53, and collapses for loss firms, with a loss coefficient of -0.2109 and a t-statistic of -9.52. The same improvement is larger for smaller firms, with a standardized size coefficient of -0.0682 and a t-statistic of -6.87, which fits the intuition that the analyst signal carries the most incremental information precisely where the model's own knowledge of the firm is thinnest. The pattern locates the value of structured integration where it should be, in firms whose analyst signal is cleanest and whose public footprint is smallest, and it cautions against expecting the same payoff in distressed or loss-making firms whose signals are themselves unreliable.

5. Discussion and Conclusions

Three mechanisms for joining analyst consensus to LLM judgment do not perform alike, and the

simplest of the integrative designs wins. Structured integration, an analyst signal handed over with a note on its reliability, outperforms both the signal-free model and the elaborate critical protocol, and it does so most clearly where the signal is reliable.

The reading I favor draws the overthinking and verbal-overshadowing literatures together. Asking the model to state an independent view, name its conflict with the consensus, weigh the two, and reconcile them substitutes a long deliberative narrative for the direct pattern match that yields its sharpest forecasts. This locates a boundary of the engaged-use principle. Engagement helps a human expert because the expert brings independent knowledge to bear on the AI's output, whereas an LLM asked to interrogate a human signal brings no comparable independent source, since its critique is generated from the same parameters as its first read, so the protocol adds deliberation without adding information.

For practice the lesson is deflationary and useful. A research tool that passes the model an analyst signal with a reliability tag will, on this evidence, beat one that scripts the model through a critical-thinking routine. Elaboration is not free, and in a structured prediction task it can cost accuracy.

The study is bounded in ways that mark out the next steps. It rests on one model and one task, the direction of annual EPS, and the prompts are particular instances of broad mechanism families rather than the full design space, so reasoning-first models and other forecasting targets may shift the balance among mechanisms.

References

[1] Cao, S., Jiang, W., Wang, J.L. and Yang, B. (2024) From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses. Journal of Financial Economics, 160, 103910.

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

[3] Lebovitz, S., Lifshitz-Assaf, H. and Levina, N. (2022) To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. Organization Science, 33, 126-148.

[4] Hassid, M., Synnaeve, G., Adi, Y. and Schwartz, R. (2026) Don't Overthink It: Preferring Shorter Thinking Chains for Improved LLM Reasoning. arXiv preprint, arXiv:2505.17813.

[5] Liu, R., Geng, J., Wu, A.J., Sucholutsky, I., Lombrozo, T. and Griffiths, T.L. (2025) Mind Your Step (by Step): Chain-of-Thought Can Reduce Performance on Tasks Where Thinking Makes Humans Worse. arXiv preprint, arXiv:2410.21333.

[6] Schooler, J.W. and Engstler-Schooler, T.Y. (1990) Verbal Overshadowing of Visual Memories: Some Things Are Better Left Unsaid. Cognitive Psychology, 22, 36-71.

[7] Wilson, T.D. and Schooler, J.W. (1991) Thinking Too Much: Introspection Can Reduce the Quality of Preferences and Decisions. Journal of Personality and Social Psychology, 60, 181-192.

[8] Lopez-Lira, A. (2024) Can ChatGPT Forecast Stock Price Movements? The Predictive Edge: Outsmart the Market Using Generative AI and ChatGPT in Financial Forecasting, 121-133.

[9] Hansen, A.L. and Kazinnik, S. (2023) Can ChatGPT Decipher Fedspeak? SSRN Journal.

[10] Jha, M., Qian, J., Weber, M. and Yang, B. (2024) ChatGPT and Corporate Policies. NBER Working Paper, 32161.

[11] Siano, F. (2025) The News in Earnings Announcement Disclosures: Capturing Word Context Using LLM Methods. Management Science.

[12] Frank, M.Z., Gao, J. and Yang, K. (2025) Behavioral Machine Learning? Regularization and Forecast Bias. arXiv preprint, arXiv:2303.16158.

[13] Kang, H. and Liu, X.Y. (2023) Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination. arXiv preprint, arXiv:2311.15548.

[14] Sarkar, S.K. and Vafa, K. (2024) Lookahead Bias in Pretrained Language Models. SSRN Working Paper, 4754678.

[15] Bertomeu, J., Lin, Y., Liu, Y. and Ni, Z. (2023) Capital Market Consequences of Generative AI: Early Evidence from the Ban of ChatGPT in Italy. SSRN Journal.

[16] Dellermann, D., Ebel, P., Soellner, M. and Leimeister, J.M. (2019) Hybrid Intelligence. Business & Information Systems Engineering, 61, 637-643.

[17] Fügener, A., Grahl, J., Gupta, A. and Ketter, W. (2021) Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI. MIS Quarterly, 45, 1527-1556.

[18] Vaccaro, M., Almaatouq, A. and Malone, T. (2024) When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis. Nature Human Behaviour, 8, 2293-2303.
[19] Lim, T. (2001) Rationality and Analysts' Forecast Bias. The Journal of Finance, 56, 369-385.
[20] Welch, I. (2000) Herding Among Security Analysts. Journal of Financial Economics, 58, 369-396.
[21] Michaely, R. and Womack, K.L. (1999) Conflict of Interest and the Credibility of Underwriter Analyst Recommendations. The Review of Financial Studies, 12.
[22] Bordalo, P., Gennaioli, N., Porta, R.L. and Shleifer, A. (2019) Diagnostic Expectations and Stock Returns. The Journal of Finance, 74, 2839-2874.
[23] Bouchaud, J., Krüger, P., Landier, A. and Thesmar, D. (2019) Sticky Expectations and the Profitability Anomaly. The Journal of Finance, 74, 639-674.
[24] Van Binsbergen, J.H., Han, X. and Lopez-Lira, A. (2023) Man versus Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases. The Review of Financial Studies, 36, 2361-2396.