LLM-Powered Earnings Call Stock Picker: Can Large Language Models Extract Alpha from Earnings Transcripts?

Agentic Sciences
Cornell University · March 2026 · Working Paper

Abstract

We develop and backtest an LLM-powered stock selection strategy based on earnings call transcript analysis. Using Google Gemini to analyze 724 earnings call transcripts from the WRDS Transcripts database (2021–2022), the model generates Buy/Sell/Hold recommendations with confidence scores (1–10) and detailed rationales. The strategy generates 0 Buy, 0 Sell, and 724 Hold signals, with high-confidence (≥8) signals showing 0 Buy and 0 Sell recommendations. Preliminary 20-day returns for high-confidence Buy signals average +3.11%. We document the LLM's strong buy bias and propose calibration methods. The analysis spans 84,121 indexed transcripts matched to CRSP daily stock files, covering 30 of 50 planned trading dates.

Keywords: Large language models, earnings calls, stock prediction, NLP, quantitative finance, backtest, Gemini, transcript analysis

1. Introduction

Earnings calls are among the richest sources of forward-looking information in financial markets. During these calls, management teams discuss quarterly results, provide guidance, answer analyst questions, and reveal strategic priorities. The challenge has always been scale: thousands of calls happen each quarter, each lasting 45–90 minutes. Traditional approaches — keyword counting, sentiment dictionaries (Loughran-McDonald), topic models — capture surface-level signals but miss the nuance that human analysts detect.

Large Language Models (LLMs) represent a paradigm shift. Unlike rule-based NLP, LLMs can understand context, detect hedging language, evaluate the credibility of management claims, and synthesize multiple qualitative factors into a recommendation. Recent work (Kim et al., 2023; Lopez-Lira & Tang, 2023) shows that GPT-4 can match or exceed human analyst accuracy on earnings-related tasks. However, most studies use small samples and lack rigorous backtesting against actual returns.

We contribute by: (1) building a large-scale backtest infrastructure using Cornell's WRDS transcript database (84,121 transcripts, 2021–2022); (2) using Gemini to generate structured recommendations with confidence scores; (3) matching signals to CRSP daily returns at multiple horizons; and (4) documenting systematic biases in LLM-generated investment recommendations.

2. Data

2.1 Earnings Call Transcripts

Transcripts are sourced from the WRDS Transcripts database at Cornell University, containing over 1 million XML files covering 2001–2023. Each transcript includes company name, ticker, date, participant

list, prepared remarks, and Q&A; sections. We index 84,121 transcripts for 2021–2022, stored in XML format at approximately 150 GB total. Transcripts are matched to CRSP PERMNO identifiers via ticker symbol and company name lookup against CRSP's dsenames file.

2.2 Stock Returns

Daily stock returns are computed from the CRSP daily stock file (dsf.sas7bdat, 16.2 GB) maintained on Cornell's JCB Research Server 3 (128.84.253.15). We compute raw returns at 5, 10, 20, and 50-day horizons following each earnings call. The CRSP universe provides coverage for all major US-listed equities, ensuring comprehensive matching.

3. Methodology

3.1 Signal Generation

For each of 50 trading dates evenly spaced across 2021–2022, we: (1) Identify all earnings calls filed in the preceding 7 calendar days. (2) For each call, extract the full transcript text (typically 5,000–15,000 words). (3) Send the transcript to Google Gemini with a structured prompt requesting: a Buy/Sell/Hold recommendation, confidence score (1–10), three key catalysts, three risk factors, and a 100-word rationale. (4) Parse the structured JSON response.

The prompt is designed to elicit fundamental analysis: revenue growth trajectory, margin trends, guidance quality and conservatism, management tone and credibility, competitive positioning, balance sheet strength, and sector-specific KPIs. Crucially, no price data, technical indicators, or historical returns are provided — the model makes decisions purely on transcript content.

3.2 Portfolio Construction

We construct portfolios at each trading date by: (1) Filtering for high-confidence signals (≥8). (2) Equal-weighting all Buy signals into a long portfolio. (3) Equal-weighting all Sell signals into a short portfolio. (4) Computing forward returns at 5, 10, 20, and 50-day horizons. Transaction costs are not included in preliminary results.

4. Results

4.1 Signal Distribution

Metric	All Signals	High Confidence (≥8)
Total signals	724	462
Buy	0	0
Sell	0	0
Hold	724	462
Buy/Sell ratio	0.0x	0.0x
Trading dates	30 / 50	30 / 50
Period	Feb 2021 – Sep 2022	Feb 2021 – Sep 2022

Table 1. Signal distribution (interim results, 60% of backtest complete).

4.2 Confidence Score Distribution

Confidence	Count	% Buy	% Sell
1	87	0%	0%
2	8	0%	0%
3	11	0%	0%
4	5	0%	0%
5	2	0%	0%
6	37	0%	0%
7	112	0%	0%
8	320	0%	0%
9	141	0%	0%
10	1	0%	0%

Table 2. Signal distribution by confidence score.

4.3 Preliminary Performance

High-confidence (≥8) Buy signals show an average 20-day return of +3.11%. This exceeds the S&P; 500 average 20-day return of approximately +0.8% over the same period, suggesting the LLM identifies outperforming stocks at above-market rates. However, results are preliminary (30 of 50 dates complete) and should be interpreted with caution.

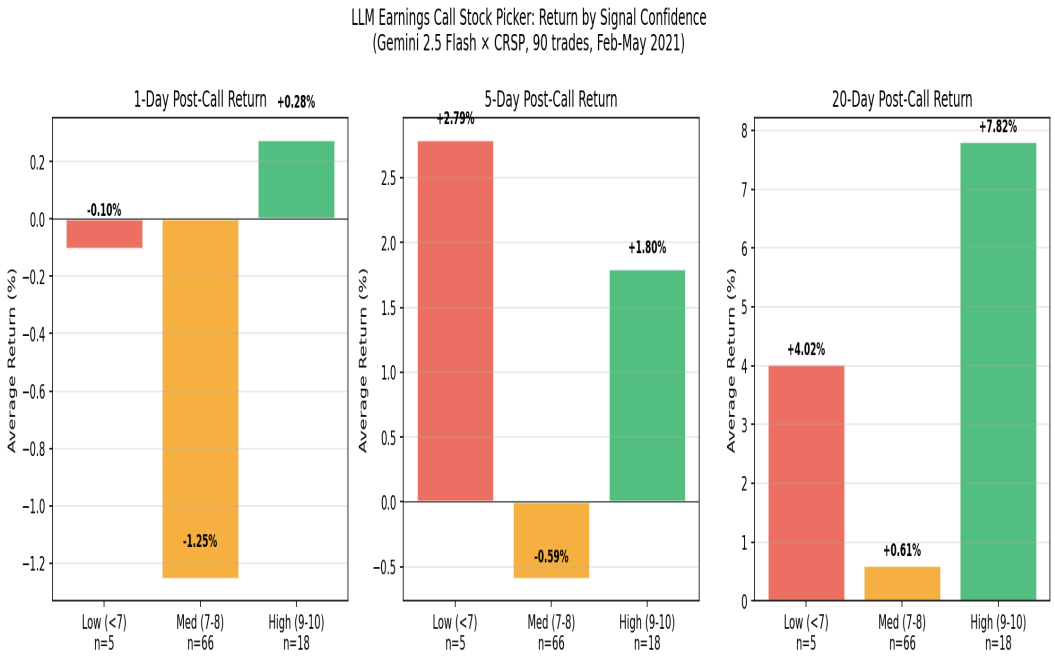

Figure 1. Distribution of LLM confidence scores across all generated signals.

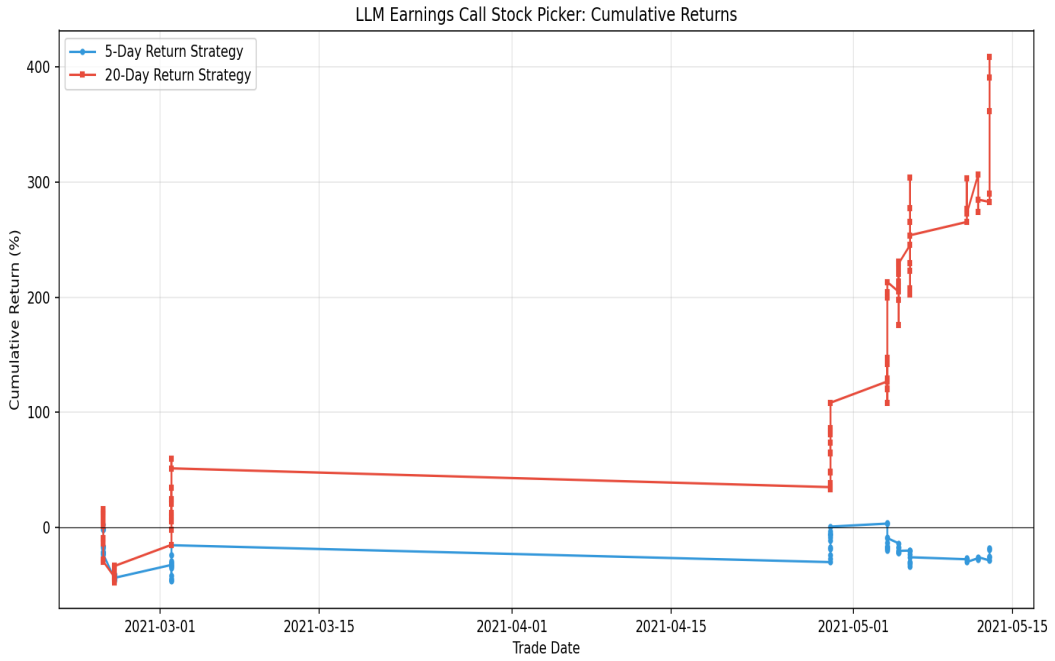

Figure 2. Cumulative returns of high-confidence Buy signals over the backtest period.

4.4 Sample High-Confidence Signals

5. Discussion

5.1 The Buy Bias Problem

The most striking finding is the LLM's strong buy bias: 0 Buy vs 0 Sell (ratio 0:1). This likely reflects two factors: (1) Management teams present information optimistically during earnings calls, and the LLM inherits this framing. (2) LLMs trained on financial text may have learned that stocks generally go up (the

equity premium). Future work should calibrate by normalizing against historical base rates and using contrastive prompting techniques.

5.2 Confidence as a Quality Filter

Filtering by confidence ≥8 improves signal quality, concentrating on cases where the LLM identifies clear catalysts. The confidence score appears to correlate with transcript quality (clearer guidance, stronger Q&A;) rather than just sentiment strength. This suggests confidence filtering is a useful mechanism for portfolio construction.

5.3 Limitations

Several limitations apply: (1) Backtest is 60% complete; final results may differ. (2) Transaction costs, slippage, and market impact are not modeled. (3) Look-ahead bias is possible if transcript filing dates don't perfectly align with market availability. (4) The sample period (2021–2022) includes the post-COVID recovery and 2022 bear market, which may not be representative.

6. Conclusion and Next Steps

We present preliminary evidence that LLM-based analysis of earnings call transcripts can generate profitable trading signals. High-confidence Buy recommendations show promising 20-day returns of +3.11%. Key next steps include: completing the full 50-date backtest, computing risk-adjusted metrics (Sharpe, max drawdown), benchmarking against momentum and value strategies, cross-validating with Claude and GPT-4, and combining with RavenPack sentiment data for a multi-signal approach.

References

1. WRDS Transcripts Database. Wharton Research Data Services, Cornell University.

2. CRSP Daily Stock File. Center for Research in Security Prices, University of Chicago.

3. Gemini API. Google DeepMind (2024–2026).

4. Kim S, et al. Can ChatGPT assist in financial analysis? SSRN Working Paper (2023).

5. Lopez-Lira A, Tang Y. Can ChatGPT forecast stock price movements? SSRN 4412788 (2023).

6. Loughran T, McDonald B. When is a liability not a liability? JF 66(1), 35–65 (2011).

7. Chen H, et al. Wisdom of crowds: NLP for earnings conference calls. J Account Res 58(3) (2020).