Predicting Earnings-Day Share Prices with Language Models

Can a language model reading an earnings announcement predict how the shares will react?

Alistair Smallwood, Head of Applied AI at Primer

ABSTRACT

Earnings announcements are among the most important scheduled events for a listed company. Investors need a first view quickly: whether the shares will rise or fall against their sector on the day, and whether the move will be large or small for that stock. On a busy morning dozens of companies report before the market opens, and few investors can read them all in time. This paper asks whether a language model can provide that first view, and whether its stated confidence shows when the view is reliable.

We tested nine language models on earnings announcements from US and European companies published between July and October 2026, and settled on Claude Opus 5.5. On the same announcements it was as accurate on direction as the best alternative, and its confidence and size judgements separated reliable calls from unreliable ones better than any other model tested. Opus 5.5 then read all 2,788 announcements, every one published after the end of its training data. For each announcement the model receives the announcement and a short packet of information describing how the shares had traded beforehand and what was expected. It states the probability that the shares outperform their sector on the reaction day, and whether the move will be larger or smaller than the stock's usual results-day move. It calls the direction correctly 66% of the time, against 50% for chance and about 60% for the rule that shares rise when earnings beat the consensus forecast.

The main finding is that the model's confidence is informative. On 52% of results it assigns 60% or more to one direction, and these calls are correct 74% of the time, against 56% for the remaining calls. The size call is stronger still. Of the stocks the model expected to move much more than usual, 90% did so, against 44% of all stocks, and it called the direction correctly for 85% of them. The same pattern appears in four test periods held back from development, with the exception of a September in which simple rules also failed.

We test 33 changes intended to improve the calls. Only a stronger model and information about expectations clearly helped. Other additional information did not, and the search-and-learning methods that perform well in games and coding found no lasting improvement, because each earnings outcome is observed once, a day later, and with substantial noise. The calls do not support a trading strategy at scale, because most of the price move occurs before a large order can be filled. Their value is as a fast first view, with a measured track record, for fundamental investors.

1 Introduction

Results days are usually the most important company-specific news events in a stock's year. Shares often move several times their normal daily range, and an investor's first view, formed before the market opens, shapes the decisions that follow: whether to act, to wait, or to study the results in more detail. That first view has three parts: the direction of the move against the sector, its size relative to what is normal for the stock, and the degree of confidence in both.

A large language model can now read an earnings announcement in about a minute. Earlier research shows that language models can extract information relevant to share prices from news headlines and financial statements (Lopez-Lira and Tang; Kim, Muhn and Nikolaev). Most of this work tests on periods the models may have seen during training, which creates a risk of look-ahead bias (Glasserman and Lin). We test only on results published after the end of the model's training data, and each test was fixed before any outcome was read.

This paper follows our working papers on capital markets days and regime shifts. In the capital markets day study, the model anticipated what management would announce, but that view was largely reflected in prices already. Here the target is the market's reaction on the day itself, and the question is whether the model's reading of the announcement contains information that the share price does not yet reflect.

2 Research questions and scope

We address six questions. First, how often does the model call the direction of the reaction correctly? Second, does its stated confidence identify the reliable calls? Third, can it say whether the move will be larger or smaller than usual, and can that judgement be turned into a figure an investor can use? Fourth, which changes to the model, its information or its instructions improve the calls? Fifth, does the choice of frontier model matter? Sixth, can the calls be traded?

The study covers US, UK and European companies whose shares trade at least USD 2m a day and which reported between July and October 2026. The main results use the 2,788 announcements for which the model answered the full size question: 2,056 in the US and 732 in the UK and Europe. Other tests use the subsets stated alongside each result. Apart from the trading test in section 11, the paper makes no claim about investment returns.

3 Out-of-sample design

Target. We measure the share-price move relative to the company's sector: the return on the stock minus the return on its sector fund, from the last close before the announcement to the close of the first full trading session after it. Measuring against the sector removes market-wide moves on the day.

Choice of model. We compared nine language models on the same earnings announcements: Claude Opus 5.5; GPT-5.6 Luna, a cheaper model; two specialist forecasting models, Jev and OpenAI Decisions; Terra, used to test averaging; and four other frontier models, Grok 4.7, Kimi K3.1 Pro, Gemini 4 Argon and Qwen 3.8 Max. Opus 5.5 was as accurate on direction as the best of them, and its confidence and size judgements separated reliable calls from unreliable ones better than any other (sections 8 and 10). Sections 4 to 7 therefore report Opus 5.5.

Information cut-off. The model receives only information that was public before trading resumed. Its training data end in June 2026, and every result in the study was published in July or later. A recall check on several hundred results found no evidence that the model knew any outcome.

Packet of information. Each announcement comes with a short packet of information: the share-price run-up, the position in the 52-week range, the stock's usual results-day move, the consensus forecast for earnings per share and, where available, Primer's material on the company, such as its results briefing, with tables comparing what analysts expected with what the company reported. Some tests added the call transcript, social media posts, short interest, estimate revisions or options data.

Output. The model states the probability that the shares outperform their sector, the likely size of the move if the shares rise and if they fall, and whether the move will be much smaller, smaller, about usual, bigger or much bigger than usual. The *usual move* is the median absolute sector-relative move on the stock's previous eight results days, known before the announcement.

Pre-specified tests. For each test, the rules, sample and pass criteria were recorded and hashed before any outcome was read, and each held-back sample was used once. Failed tests are reported. Most comparisons are paired: the same results are read with and without a single change. Confidence intervals come from resampling whole trading days 2,000 times, because stocks reporting on the same day share that day's market conditions. Two reads of an identical packet of information differ on about 6% of calls, which changes accuracy by about 1.5 points; a change had to exceed that to count.

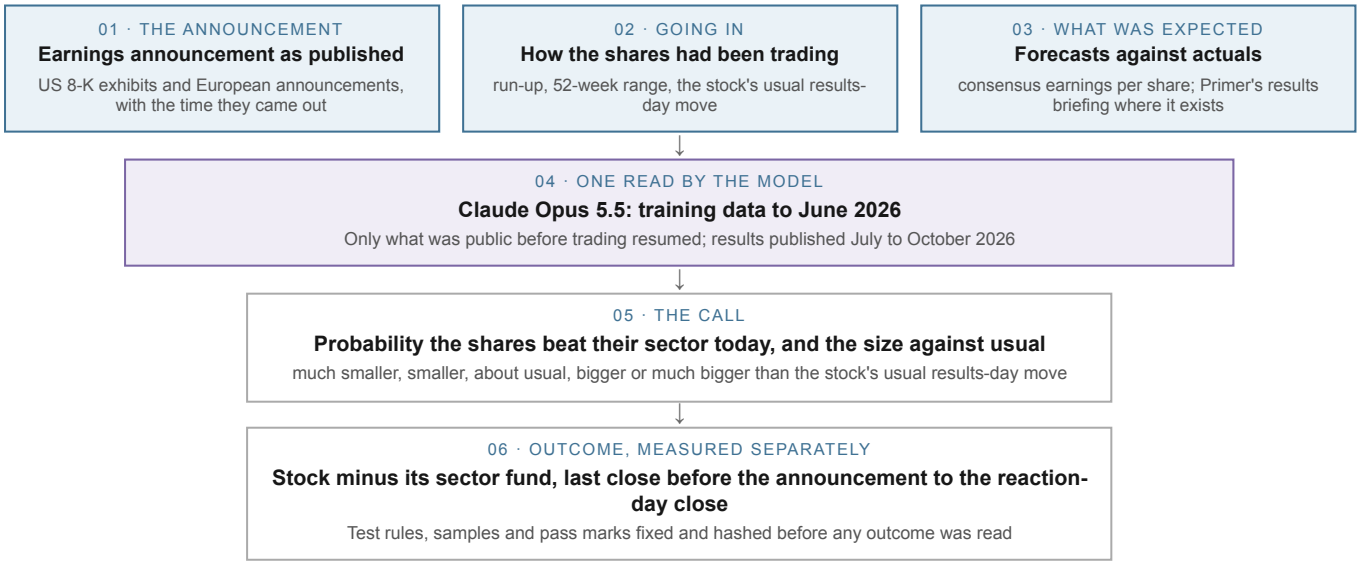


Figure 1. *How each call is produced and scored. The packet of information is cut at the moment trading resumes, and every result came out after the model's training data ends.*

4 Direction accuracy

Headline result. Across all 2,788 results, the shares moved in the direction the model indicated 65.5% of the time: 65.1% in the US and 66.7% in the UK and Europe. Chance would give 50%. The rule that shares rise when earnings beat the consensus forecast gives about 60% in a normal season.

Held-back tests. The set-up was developed on early-summer results and then tested on periods held back until the end (Figure 2). Accuracy was 64% in late August, 66% in October and 67% in Europe. September, with only 138 results, gave 57%. The month was difficult for every method: the earnings-beat rule was correct 52% of the time, against 61% in each September from 2015 to 2025, and price moves were unusually small.

Interpretation. A call that is correct two times in three is not sufficient on its own. Its usefulness depends on the two results that follow: the model can identify which of its calls are reliable, and its judgement of the size of the move is dependable.

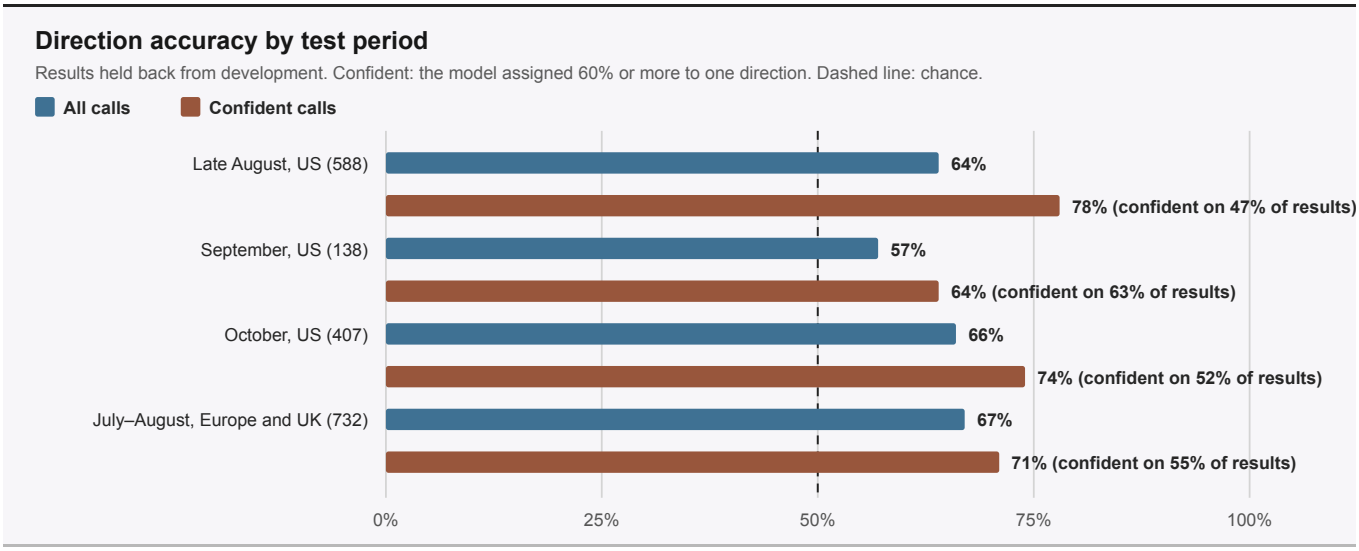


Figure 2. *Every test period beat chance, and in every period the confident calls were more accurate than the rest. September was difficult for every method: the rule that shares rise when earnings beat the forecast was correct only 52% of the time, against 61% in every September from 2015 to 2025.*

5 Confidence identifies the reliable calls

Confident calls. We define a call as confident when the model assigns 60% or more to one direction. This occurred on 52% of results, and these calls were correct 74.2% of the time. The remaining calls were correct 56.1% of the time, only slightly better than chance. The pattern holds in both regions: 75% against 55% in the US, and 71% against 61% in Europe.

> INVESTOR INTERPRETATION
>
> Investors can act on the confident calls and set the rest aside. On about half of results the call was correct three times in four; on the other half it was close to chance, and identifying those cases is part of the value.

Under-confidence. The model is cautious rather than over-confident: when it stated 65%, the shares moved its way about 71% of the time. A cheaper model showed the opposite bias, being confident on 86% of results but correct on only 66% of them. Stretching the model's probabilities to correct for its caution, with the adjustment estimated on summer results and tested once on late August, raised the share of confident calls from 47% to 62% of results, while their accuracy fell only from 78% to 74%.

Instability of the adjustment. In September, when the model itself was less accurate, the same adjustment made it over-confident, and its confident calls were correct only 60% of the time. In October the adjustment helped again. Re-estimating it on the most recent 300 results is safer, giving confident calls on 56% of results at 74% accuracy, and an adjustment estimated for one model cannot be transferred to another. The underlying trade-off has been stable: about half of results receive a confident call that is correct 74 to 78% of the time, and raising that share to 70 to 75% costs a few points of accuracy.

6 Bigger or smaller than usual

The size label. In addition to direction, the model states whether the move will be larger or smaller than the stock's usual results-day move. Without any forecast, 44% of stocks move more than usual. The model's label separates the two groups clearly (Figure 3, Table 1).

Size and direction. The calls for larger moves are also the most accurate on direction. For 'bigger' and 'much bigger' calls the shares moved in the indicated direction 73% of the time, and 79% of the time when the model was also confident (376 calls). For the 57 confident 'much bigger' calls the figure was 93%. When the model expects a smaller move than usual, the shares usually move little and its direction call is weaker.

Comparison with options prices. Options prices indicate how large a move traders expect. On the same US results, the model's probability of a large move ranked the stocks that did move sharply better than the options-implied move did (0.70 against 0.64, on a scale where 0.5 is chance and 1 is a perfect ranking).

Comparison with the opening price. By the time the market opens, the size of the overnight price gap alone identifies a larger-than-usual move about as well as the model (0.71 against 0.70). Combined, the two reach 0.75, so the model adds information beyond the gap, and within each gap size its label still separates the outcomes. However, an investor who waits for the open has already observed much of what the model anticipated.

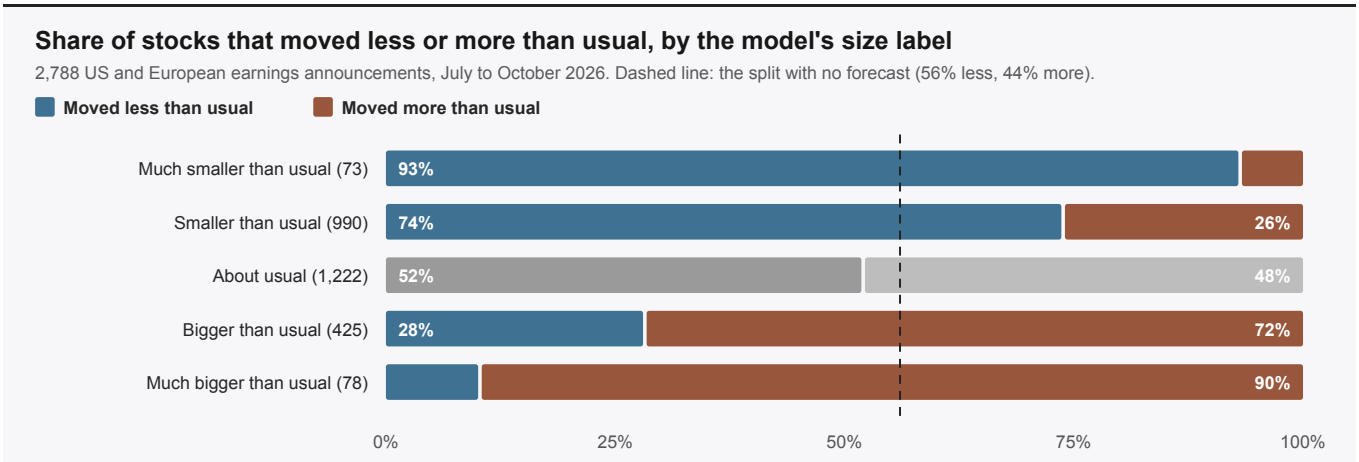


Figure 3. *The size label is ordered, and both extremes were reliable. When the model expected a 'much bigger' move, nine stocks in ten moved more than their usual results-day move; when it expected a 'much smaller' move, 93% moved less than usual.*

Model said the move would be	Results	Moved more than usual	Typical move against usual	Direction right	Direction right when confident (n)
Much smaller	73	7%	a fifth of usual	49%	too few
Smaller	990	26%	about 0.6 times	60%	73% (347)
About usual	1,222	48%	about the same	68%	72% (722)
Bigger	425	72%	about 1.7 times	71%	77% (319)
Much bigger	78	90%	about 4 times	85%	93% (57)

Table 1. *The size label, 2,788 US and European results. 'Typical move against usual' is the median actual move divided by the stock's usual results-day move: 'about 4 times' means a stock that usually moves 3% typically moved about 12%.*

7 Converting the size label into a figure

Label and yardstick. Each label corresponds to a stable multiple of the stock's usual move: about one fifth for 'much smaller', 0.6 times for 'smaller', about one for 'about usual', 1.7 to 1.9 times for 'bigger' and about four times for 'much bigger'. These multiples were estimated on July and August US results and held on September and October and in Europe, although the overall level varies by about a fifth from season to season. The label multiplied by the usual move ranks the size of moves as well as the model's own percentage estimate, but not better.

Thresholds rather than ranges. The spread within each label is wide. For 'bigger' calls the middle half of outcomes lies between 0.9 and 3.2 times the usual move, and for a single stock the size estimate is typically wrong by about half the predicted move. Narrow ranges therefore fail: of confident calls with a predicted move of 7 to 13%, only 19% ended inside that range. One-sided statements are more reliable: of confident 'much bigger' calls, 74% moved in the stated direction by more than twice the usual move.

Presentation to investors. The call is therefore best stated as a threshold, with its yardstick and track record, and only when the model is confident:

ILLUSTRATION ONLY	YARDSTICK	TRACK RECORD
ACME: likely up more than 10% today Up · 72% probability · much bigger than normal	**Normal move 5.1%, the middle of its last 8 results days. 'Much bigger' means more than twice normal, so more than 10%**	**74% of past confident calls like this moved the stated way by more than twice their normal move**

8 Which changes improved the calls

We tested 33 changes intended to improve the calls, covering the choice of model, the information it reads, the wording of the question, automated search and learning from past results, and corrections to its confidence. Figure 4 shows every change for which a paired accuracy comparison is available.

The model. On the same 407 October results, Opus 5.5 was correct 66% of the time, against 63% for GPT-5.6 Luna, a cheaper model. Two specialist forecasting models scored 62 to 63%. When given Opus's written analysis they matched it, but did not exceed it.

Information about expectations. An announcement reports what happened; the share price responds to how that compares with what investors expected. In Europe, where standard forecast data are sparse, removing Primer's material reduced accuracy by 5 points and reduced the share of confident calls from 61% to 44% of results. Primer's results briefing was the first input to pass a pre-specified test, and its tables of forecasts against actual figures performed as well as the full briefing, which indicates that the value lay in the figures rather than the text. In the US, where the model already sees the consensus forecast for earnings per share, a more detailed table of analysts' forecasts added nothing.

Other information. Social media posts, options data, short interest, estimate revisions, previews, key debates and notes on earlier reactions added nothing. With all of them included in October, accuracy fell by half a point. The effect of the call transcript is unresolved: where the call took place the previous evening, accuracy was 10 points higher in late August and 7 points higher in September, but no higher in October.

Question design. One change to the question helped. Asking for the size in parts (if the shares rise, if they fall, the most likely outcome, and the comparison with the usual move) corrected the model's size estimates, which had been far too small: 1.2% when stocks moved 3.6%, and 4.5% afterwards. Written rules of thumb, of the kind a senior analyst might give a junior, made the model more cautious rather than more accurate, and worse at identifying large moves. When asked to grade results on a fixed scale from 'quiet' to 'shock', it placed 125 of 150 in 'normal'.

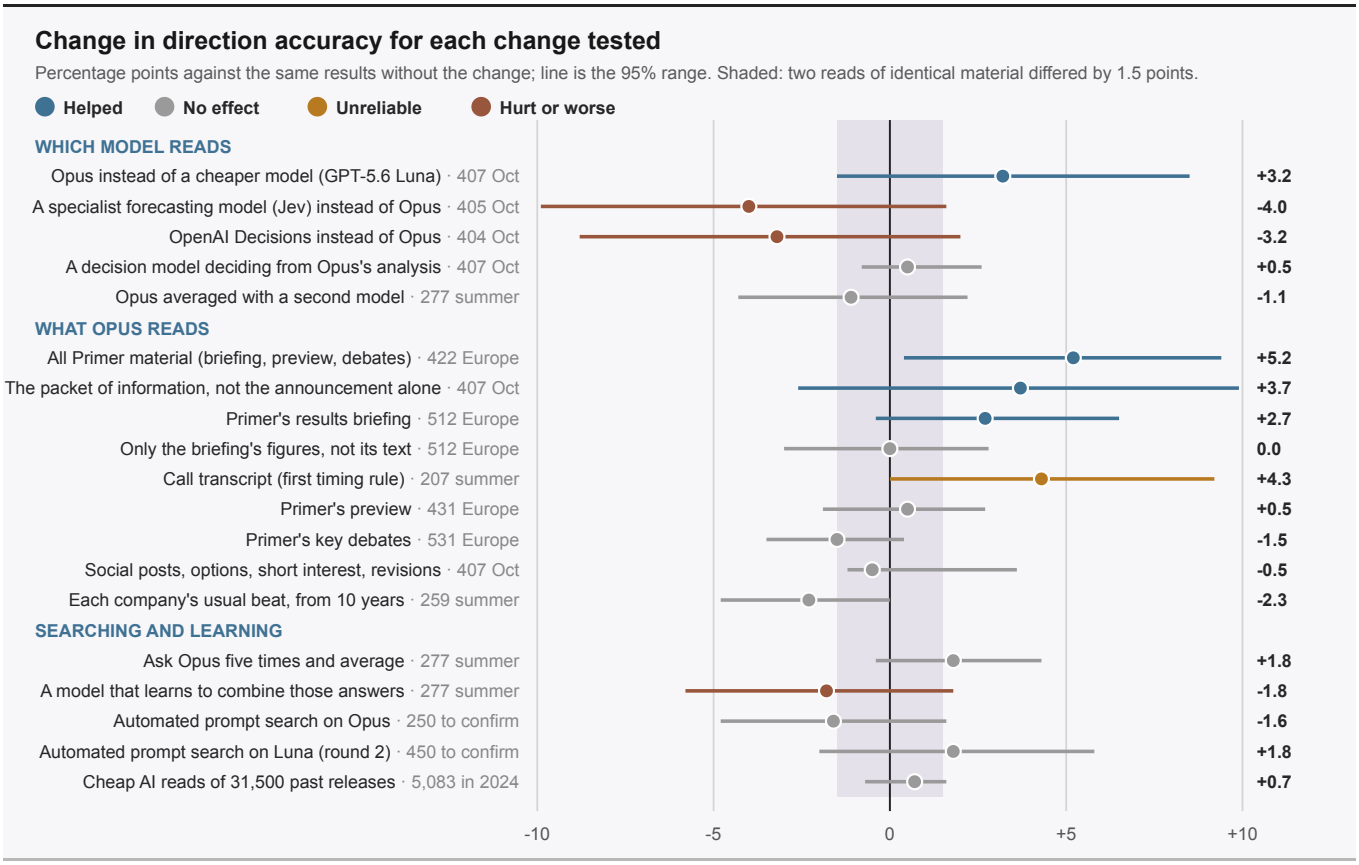


Figure 4. *Only the choice of model and information about expectations changed accuracy clearly. Most changes fell within the band that two reads of identical material produce by chance. Some changes marked 'helped' were assessed on probability quality or on the confident calls; on plain accuracy their intervals still include zero.*

9 Why search and learning did not transfer

Sutton's 'bitter lesson' (2019) argues that general methods which scale with computing power, namely search and learning, eventually outperform knowledge built in by hand. The argument accounts for AlphaGo and for AI systems that write code and prove theorems. Before building this stage we reviewed more than 100 papers and system reports across five families of method, checked each against its original source, and applied the most promising.

Methods tested. Asking the model five times and averaging the answers added about 2 points, within the range expected by chance, at five times the cost; the five answers agreed on direction 82% of the time. A model trained to combine the answers reduced the accuracy of confident calls by 7 points. An automated loop that examined the largest errors, rewrote the instructions and kept only changes that also improved a second sample found apparent improvements on 120 results that disappeared on the next 250. A statistical model trained on 94,000 past results from 2015 to 2024 was correct 64 to 66% of the time on 2024 but 60% on 2026, and added nothing to the language model.

Reasons for the failure. Each success of search and learning depends on a loop that can run millions of times because every attempt is checked instantly and exactly (Table 3). Gains from self-consistency level off after five to ten answers (Wang et al.). Without an exact checker, sampling more answers stops helping (Brown et al.; Stroebl et al.). Prompts found by search score better on the examples used to find them than on new ones (Yang et al.). DeepSeek-R1 replaced learned reward models and tree search with simple rule-based rewards. Learning from forecasting outcomes does work, but the gains are modest (Turtel et al.), and random rewards can produce apparent gains, so any improvement needs a placebo test (Shao et al.).

What does transfer. Hand-built knowledge reaches a limit: our rulebooks made a strong model worse. The scarce resource is resolved outcomes, not computing power, so outcomes should be spent on tests specified in advance. New information is worth more than additional effort on the same information: in one forecasting system, removing news retrieval cost about seven times as much as ensembling added (Halawi et al.).

What search and learning need	Go, code, maths	Earnings-day reactions
An exact, instant check of each answer	Win or lose; the code runs or it does not; the proof checks	None. The outcome arrives once, the next day, and one outcome cannot prove a probability right or wrong
More practice from more computing	AlphaGo Zero played 4.9 million games against itself in three days	Impossible. Results arrive at calendar speed, a few thousand a year, whatever is spent
The same input gives the same answer	Yes: the rules decide	No. The same news can move shares either way. Even knowing the opening price, the close goes the same way only 73% of the time
Rules that stay fixed	Yes	No. In September 2026 'beat forecasts, shares rise' worked 52% of the time, against 61% normally
Skill easy to separate from luck	Yes: winning by accident is rare	No. A direction call is right half the time by luck, so a learning loop mostly learns noise

Table 3. *Why the recipe behind AlphaGo and AI coding systems does not transfer. Each earnings day gives a single, noisy answer, and computing power cannot make more of them.*

10 Comparison of frontier models

Four other frontier models read the same 250 October results, all released on or after 13 October so that none could have appeared in their training data, with identical packets of information and questions (Table 2).

Direction. Grok 4.7 (64.8%) and Kimi K3.1 Pro (64.4%) matched Opus (64.4%); Gemini 4 Argon and Qwen 3.8 Max were about 4 points lower. None of these differences exceeds what chance would produce on 250 results.

Confidence. The other models classed far more results as confident (85% for Kimi and 94% for Qwen, against 52% for Opus) and were correct less often on those calls. When each model's most confident quarter and half of calls are compared at equal coverage, Opus scores highest on both, although the differences are within noise. In practice, Opus's 60% threshold can be used as it stands; the other models would need a threshold set by the user.

Size and combination. Opus's 'bigger' calls were followed by larger-than-usual moves more often (59% against 45 to 54%), on small samples. Averaging Opus with any other model added nothing: where two models disagreed, each was correct about half the time. The other models cost about half as much per read. Kimi K3.1 Pro was released on 23 October, after most of these results; its training data are very unlikely to include them, but this is less certain than for the other models.

Model	Direction right	Called confident	Right when confident	Most confident 25% right	Most confident 50% right	Brier (lower is better)	'Bigger' calls that moved more than usual (n)	Cost per read
Claude Opus 5.5	64.4%	52%	70.8%	75%	71%	0.224	59% (39)	11.5¢
Grok 4.7	64.8%	65%	65.6%	73%	68%	0.226	54% (52)	6.4¢
Kimi K3.1 Pro	64.4%	85%	66.2%	68%	69%	0.229	48% (115)	4.2¢
Gemini 4 Argon	60.0%	65%	67.9%	70%	69%	0.235	49% (74)	4.8¢
Qwen 3.8 Max	60.8%	94%	62.1%	74%	68%	0.233	45% (33)	6.0¢

Table 2. *Five frontier models, one read each of the same 250 October 2026 US results, released on or after 13 October, with identical material and question. The 'most confident' columns compare each model's own best calls at the same coverage; ties at the cut are shared out evenly.*

11 Tradability after the open

We tested whether the model's large-move calls can be traded. We simulated buying or selling 25% of a stock's average daily trading value behind each confident 'bigger' or 'much bigger' call in the US, starting at the open, taking no more than 15% of the volume in any five-minute interval, hedged with the sector fund and closed at the end of the day.

Results. If filled at the opening price, the 303 trades would have returned 1.5% each before costs. Filled at a realistic pace, which took a median of 215 minutes, they returned 0.45% each (95% interval 0.02 to 0.89%) and were profitable only 52% of the time. The total dollar result depended on two trades. Much smaller positions, filled within the first ten minutes, retained about 1.1% per trade.

Interpretation. Most of the reaction occurs overnight and in the first minutes of trading, before a meaningful position can be built. The calls therefore do not support a trading strategy at scale. Their value lies in a fast first view for investors deciding what to do with positions they already hold or follow.

12 Implications for investors

A pre-market call with a track record. Before the open, each result would receive a single call, stated as a threshold with its yardstick and the frequency with which such calls have held in the past. Only confident calls would be shown. On a morning when 40 companies report, this indicates which results to read first and how the market is likely to respond.

Continuous scoring. Quick reactions to results are published every morning but are rarely scored. A system of this kind is scored on every call against later outcomes, and its confidence adjustment is re-checked against recent months rather than fixed once. It does not replace reading the results. It provides a first view that indicates when it is likely to be wrong.

13 Limitations

One season and one model generation. The study covers four months of results, and the difficult September shows how far market behaviour can change.

Retrospective design. Every outcome postdates the model's training data, but the set-up was developed on early-summer results. The held-back periods address this. Table 1 is weighted towards the summer, although its multiples held in later months.

Speed of the market. Automated traders react to headline figures within seconds. We have not tested whether the call adds information to the after-hours price a few minutes after the announcement, only to the next opening price.

Coverage and sample size. The study includes only shares trading at least USD 2m a day; European announcement times come from Primer's timestamps and were not independently checked; 'much bigger' calls are rare (78 in total, 57 confident); and the five-model comparison rests on one read of each of 250 results.

Investment returns. Apart from the trading test, which is gross of costs, the paper measures the accuracy of calls rather than investment performance. Licensed price and consensus values are not reproduced.

14 Conclusion

A language model can provide a useful first reading of an earnings announcement. It calls the direction of the reaction correctly about two times in three and, more importantly, its confidence is informative: its confident calls were correct 74% of the time. Its judgement of whether the move will be larger or smaller than usual is reliable enough to be stated as a figure relative to the stock's own history.

The improvements came from a stronger model and from information about expectations. Search and learning, which require an exact and immediate measure of success, found no lasting gain. The next steps are a live season with the system fixed in advance, and a test against the after-hours price a few minutes after each announcement.

R Evidence and references

1. Smallwood, A. (2026). *Forecasting Capital Markets Day Targets with Language Models*. Primer working paper. https://www.primerapp.com/blog/forecasting-capital-markets-day-targets-with-language-models/
2. Smallwood, A. (2026). *Identifying Regime Shifts with Language Models*. Primer working paper. https://www.primerapp.com/blog/identifying-regime-shifts-with-language-models/
3. Lopez-Lira, A. and Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. arXiv:2304.07619.
4. Kim, A., Muhn, M. and Nikolaev, V. (2024). Financial statement analysis with large language models. arXiv:2407.17866.
5. Glasserman, P. and Lin, C. (2023). Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. arXiv:2309.17322.
6. Sutton, R. (2019). The bitter lesson. Essay, incompleteideas.net.
7. Silver, D. et al. (2017). Mastering the game of Go without human knowledge. *Nature* 550.
8. Wang, X. et al. (2023). Self-consistency improves chain of thought reasoning in language models. ICLR 2023.
9. Brown, B. et al. (2024). Large language monkeys: scaling inference compute with repeated sampling. arXiv:2407.21787.
10. Stroebl, B., Kapoor, S. and Narayanan, A. (2024). Inference scaling flaws: the limits of LLM resampling with imperfect verifiers. arXiv:2411.17501.
11. Yang, C. et al. (2024). Large language models as optimizers (OPRO). ICLR 2024.
12. DeepSeek-AI (2025). DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv:2501.12948.
13. Turtel, B. et al. (2025). Outcome-based reinforcement learning to predict the future. arXiv:2505.17989.
14. Shao, R., Li, S. S. et al. (2025). Spurious rewards: rethinking training signals in RLVR. arXiv:2506.10947.
15. Halawi, D., Zhang, F., Yueh-Han, C. and Steinhardt, J. (2024). Approaching human-level forecasting with language models. NeurIPS 2024.
16. Research records: projects/earnings-day-forecast (test plans, reports, ledgers, scripts and derived results).

Research status. Working paper, prepared 1 November 2026. All results are retrospective; every outcome postdates the forecasting model's training data. Held-back tests follow plans recorded before results were seen; analyses described as exploratory were added later. The paper makes no claim of investment returns. Licensed price and consensus values are not reproduced.