Introducing Forward Model MK1: a purpose-built LLM for the financial markets

Ehsan Yousefzadeh
Forward Model, Inc.
ehsan@forwardmodel.ai

10 November 2026

Abstract

In this paper we introduce Forward Model MK1, a purpose-built LLM trained on the financial markets: a proprietary model, trained by Forward Model, Inc. We set out to create, research and develop the best model for stock picking, as the primary goal, while keeping it a general-purpose model for the wider investment workflow. We ask whether a model built for that purpose reads the same evidence better than the best general frontier models. On seven trading days from 2 to 11 September 2026, all after the stated knowledge cutoff of every model tested, Forward Model MK1 and four frontier models (GPT-6 Astra, Gemini 4 Argon, Claude Opus 5.5 and Claude Fable 5.1) scored the same 500 stocks a day from the same context, and the scores were graded 21 trading days later by their rank correlation with market- and factor-adjusted returns. The design and the pass bars were written before any rival's answer was read. Forward Model MK1 reads +0.105, against +0.059 for GPT-6 Astra on the pooled 1,000-name sample the rule called for, +0.047 for Gemini 4 Argon, +0.040 for Claude Opus 5.5 and +0.038 for Claude Fable 5.1. It led every frontier model over the week, three of them on every one of the seven days and GPT-6 Astra on five of the seven, and against the two Anthropic models the margin clears the pre-registered bar. The report also sets out the evaluation itself, every stock, every day, graded by the market on one adjusted yardstick, which we believe is a faster and fairer test of a model in the markets than the alpha of deployed capital measured over years. The evaluation covers one week, about one independent window at 21 days.

1 Introduction

Forward Model is an AI research lab focused on the financial markets. In this paper we introduce Forward Model MK1, a proprietary large language model purpose-built for the financial markets, trained and owned by Forward Model, Inc. We set out to create, research and develop the best model for stock picking: given the evidence on a company, to judge better than any other model how its stock will do over the weeks ahead. Stock picking was the primary goal, and it is what this paper evaluates. We also kept the model general-purpose, able to carry out the wider tasks of an investment workflow, from reading a filing to comparing a company with its peers, so the stock-picking judgment lives inside a model an investor can work with.

Markets have always kept score: every call, were you right? We hold our model to that. Every score it publishes sees only what was knowable that morning, is stamped with its date and version, and is graded by the market on days none of the models was trained on, and the graded record is public. We believe this is a faster and fairer way to evaluate a model in the markets than measuring the alpha of deployed capital over years: it grades thousands of calls a week on one adjusted yardstick, where a portfolio's returns offer a few hundred data points a year and carry the manager's sizing and timing with them. The evaluation is the second subject of this report.

General frontier models are very capable readers of text, so the natural question is whether a model built for the financial markets reads the same evidence better than they do when the question is which stocks will do well over the next month. This report gives Forward Model MK1 and four frontier models the same context on the same stocks and compares the grades. The evaluation is forward in a precise sense: every test day falls after the stated knowledge cutoff of every model tested, so nothing about these days can have been seen by any model beforehand, and the comparison is a test of reading rather than of memory. This matters because a model whose training data runs past a test day may have seen how that month turned out, and a good score could then be recall rather than judgment; testing only on days after every model's stated cutoff closes that door for all of them at once.

What the evaluation validates is how well each model ranks stocks by their next 21 trading days when given the same evidence, on one week. It does not measure a model's general ability, its performance with its own tools or retrieval, or how it would do on other weeks; those limits are set out in section 7.

This report does not describe how Forward Model MK1 was built. Its training method, architecture and size are ours, and no claim here depends on any of them: the model is evaluated on its outputs alone, with the same context and the same grading as the rivals.

2 Evaluation

The test days are the seven trading days from Wednesday 2 to Friday 11 September 2026, with Monday 7 September closed for Labor Day. On each day 500 stocks were drawn at random, by a rule fixed before the first read, from the roughly 2,800 US stocks Forward Model MK1 scores every trading day, the same 500 for every model, and each pair of models is read on the names both answered. The horizon is 21 trading days.

The design was pre-registered.[1] The test, the sample rule, the bars and the top-up rule were written before any rival's answer was read, and the sample size was chosen by the rule before the first read. A second disjoint draw of 500 names on the same days is made only when a first read lands inside the noise band, and the pooled 1,000 names are then final. The rule was triggered once, for GPT-6 Astra (section 6).

[1] The pre-registration, with the dates its rules were approved and the hashes of the commits that carry them, is published at forwardmodel.ai/research/pre-registration.

3 Models

Table 1 lists the models, the exact version tested and the stated knowledge cutoff or release date. Every test day falls after every published cutoff and after every open-weight release. The frontier models were read on 9 November 2026 through their vendors' APIs, and the open-weight models between 28 October and 1 November 2026 through OpenRouter.

Table 1: The models, the exact version tested and the stated knowledge cutoff or release date.

Model	Version	Stated knowledge cutoff
Forward Model MK1 (ours)	forward-model-mk1	31 August 2026
GPT-6 Astra (OpenAI)	gpt-6-astra	30 April 2026
Claude Opus 5.5 (Anthropic)	claude-opus-5-5	June 2026
Gemini 4 Argon (Google)	gemini-4-argon	March 2026
Claude Fable 5.1 (Anthropic)	claude-fable-5-1	June 2026
Qwen3.5-122B-A10B (Alibaba), open weights	qwen3.5-122b-a10b	weights public 24 February 2026
GLM 5.3 (Z.ai), open weights	glm-5.3	API 14 August 2026, weights 28 August 2026
DeepSeek V4 Flash 0731 (DeepSeek), open weights	deepseek-v4-flash-0731	version of 31 July 2026

4 Method

The centerpiece of the method is that every model tested, Forward Model MK1 included, is given exactly the same context and asked exactly the same question. For each stock on each day the context is the evidence recorded that morning: the stock's price and market value as of the prior close, company facts, a financial summary, filings, news, peers, sector context, valuation metrics and analyst estimates. It is rendered by the same neutral prompt (version 3.1), which asks for a structured assessment and a score, with no prior score and no memory of earlier days. No model receives extra context, retrieval, tools or examples, none is tuned for the test, and Forward Model MK1 receives nothing the rivals do not. We call this fixed context and prompt our call harness, and it is proprietary; the word is used here only in that sense.

The frontier models, on which the claim rests, ran at their default settings; the two Anthropic models had output allowances of 24,000 tokens (Opus) and 40,000 tokens (Fable), which no answer reached. The open-weight models are reported for completeness and are not part of the claim; they were read at the settings that returned answers in the schema form, thinking off for Qwen3.5 and low reasoning effort for GLM and DeepSeek. The number of answers from each model that carry a grade is in Table 2.

5 Grading

A score is graded on what the stock did next, after taking out what the market, the stock's sector and its style would have done to any stock like it. For each test day d and each stock i scored that day, let $r_{i,d}$ be the stock's return over 21 trading days from the last close in its evidence packet, and let m_d be the median of those returns across all the stocks scored on day d. The market-relative return $\tilde{r}_{i,d} = r_{i,d} - m_d$ is then regressed, across the stocks of that day, on an intercept, a set of sector indicators and six style scores measured on day d (momentum, value, size, quality, growth and low volatility):

$$\tilde{r}_{i,d} = \alpha_d + \sum_s \beta_{s,d}\, 1[i \in s] + \sum_{k=1}^{6} \gamma_{k,d}\, f_{k,i,d} + \varepsilon_{i,d}.$$

The residual $\varepsilon_{i,d}$ is the stock's adjusted return: the part of its move that neither the market, nor its sector, nor its style explains. A stock missing a style score on the day is fitted on its sector alone. The six are the style factors of the asset-pricing literature, each built from public prices and fundamentals as a z-score across the universe on the day: momentum from the trailing return, value from earnings, book and sales yields, size from market value, quality from margins, returns on equity and capital, and leverage, growth from revenue growth, and low volatility from realized volatility. The construction is fixed, public in its inputs and the same for every model, so it cannot favor one; it is also conservative for Forward Model MK1, since anything a model gets right that these factors already explain earns it no credit. A model's skill on day d is the Spearman rank correlation between its scores $q_{i,d}$ and the adjusted returns across the stocks it scored,

$$\text{skill}_d = \rho_S\big(q_{\cdot,d},\, \varepsilon_{\cdot,d}\big),$$

and the week's figure is the mean over the seven days. Because the correlation is taken on ranks, the size of an extreme move does not matter, only its place in the ordering. The adjusted returns are the same for every model, graded once as a property of the stock and the day, so two models differ only in their scores. This construction isolates the stock-picking value of a model's judgment in its purest form: what the model knew about each stock that the market, its sector and its style did not already say.

A case from the test week. On 8 September 2026 Forward Model MK1 scored AMZN 8.2 out of 10, the highest score among that day's 500. Over the next 21 trading days, to 7 October, AMZN rose 0.5 percent, while the median stock scored that day fell about 5.1 percent over the same 21 days, so relative to the market AMZN was up 5.6 percent; its sector and its style scores account for a further 1.5 points, leaving an adjusted return of +4.1 percent. That is the figure the grade sees, not the 0.5. Among the 466 names graded that day AMZN ranked first by score and 141st by adjusted return, and the day's skill is the rank correlation between those two orderings over all 466 names, +0.055 on that day. No score is ever graded alone.

Each rival is compared with Forward Model MK1 on the names both answered, and the difference in skill is reported in Table 2 as Δ, Forward Model MK1 minus the model, with its t statistic. Seven consecutive 21-day windows overlap almost entirely and amount to about 1.1 independent windows, so every t quoted is divided by the overlap correction. The pre-registered bar for saying that Forward Model MK1 beat a model is a positive Δ with a corrected t of 2 or better; the count of days on which Forward Model MK1's skill was higher is a description, not a verdict.

6 Results

Table 2 gives the 21-day results. Forward Model MK1's own figure is read on the full sample, 466 to 474 of the 500 names graded each day. Of the 201 names across the week without a grade, 121 lacked a price in our record on one of the two days, or sat outside the day's graded cohort, and 80, or 2.3 percent, returned no usable answer from Forward Model MK1's read. An answer comes back as a structured record in a fixed form; one that arrives malformed, or a call the provider does not complete, is dropped for that name on that day and never retried into a different answer. The rivals' rates were between zero and 0.3 percent. Names without a grade are dropped, not filled in, and a dropped name is a property of the call, not of the stock. Each rival's figure is read on the names both it and

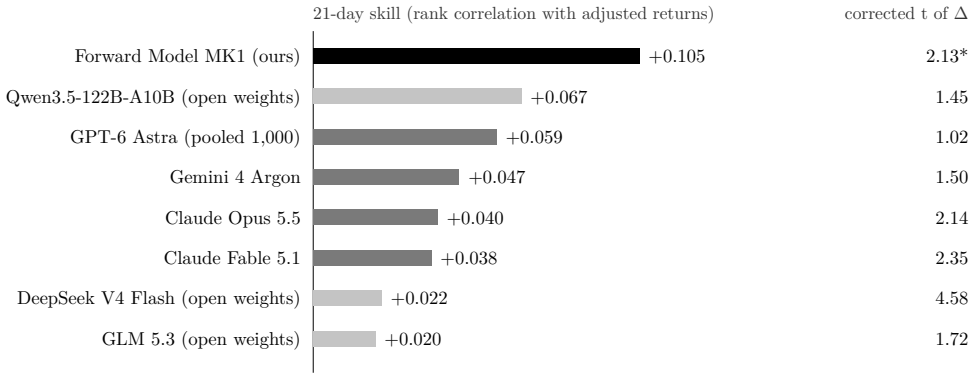
Forward Model MK1 answered; for GPT-6 Astra that is the pooled 1,000 names a day, the first draw and the second that the rule called for.

Table 2: 21-day skill on the test week, every model on the same days and the same names as Forward Model MK1. **Δ** is Forward Model MK1's skill minus the model's on the names both answered; t is the overlap-corrected t statistic of **Δ**, and of the skill itself on Forward Model MK1's row.

Model	Names graded	Skill	**Δ**	t	Days ours ahead
Forward Model MK1 (ours)	3,299 of 3,500	+0.105		2.13	
Claude Opus 5.5	3,289	+0.040	+0.066	2.14	7 of 7
GPT-6 Astra (pooled 1,000 a day)	6,592	+0.059	+0.033	1.02	5 of 7
Gemini 4 Argon	3,287	+0.047	+0.057	1.50	7 of 7
Claude Fable 5.1	3,292	+0.038	+0.065	2.35	7 of 7
Qwen3.5-122B-A10B, open weights	3,270	+0.067	+0.041	1.45	6 of 7
GLM 5.3, open weights	3,291	+0.020	+0.084	1.72	6 of 7
DeepSeek V4 Flash, open weights	3,257	+0.022	+0.080	4.58	7 of 7

Forward Model MK1 led every frontier model over the week. Against Claude Opus 5.5 and Claude Fable 5.1 the pre-registered bar is met, and it was ahead of both, and of Gemini 4 Argon, on every one of the seven days. Against GPT-6 Astra the first 500-name draw left the gap inside the pre-registered noise band, so the rule's second disjoint draw of 500 was made; on the pooled 1,000 names, which the rule makes final, Forward Model MK1 leads by +0.033 with a corrected t of 1.02 and was ahead on five of the seven days, below the bar on this window. Against Gemini 4 Argon the gap lies outside the band and the first read stands, below the bar on this one window. The three open-weight rows are the latest model from each family, read with the same context in late October; the others read then are in the record.

Where a model came close: on the pooled sample GPT-6 Astra read above Forward Model MK1 on 2 September (+0.058 against +0.048) and on 3 September (+0.107 against +0.094); on its first 500 names alone it trailed on every day.



* the corrected t of Forward Model MK1's own skill; every other figure is the corrected t of Δ, Forward Model MK1 minus the model.

Figure 1: 21-day skill on the test week by model. Each bar is the model's own skill on the names it shares with Forward Model MK1; the figure beside each rival is the overlap-corrected t of its Δ against Forward Model MK1, and Forward Model MK1 carries the corrected t of its own skill. Open-weight models are marked.

7 Limitations

The evaluation covers one week. Seven trading days are about 1.1 independent windows at 21 days after the overlap correction, so while the direction on this week is clear, the size of the margin is not settled to the bar for every rival. The sample is 500 names a day rather than the universe, which makes the pairing fair, since every model is read on the same names on the same days, but noisier than the universe would be.

Every rival read our context through our prompt. A different context or prompt could change its figures, and the evaluation measures each model as a reader of the same information, not each model at its own best. The open-weight rows were read at the settings that let each answer in the schema form and are not part of the claim. A model that publishes no knowledge cutoff cannot be shown to be out of sample and is excluded.

8 Conclusion

On seven trading days after every model's stated knowledge cutoff, given the same context and graded the same way, Forward Model MK1 read the stock market better than the frontier models we tested. Its 21-day skill of +0.105 stood above GPT-6 Astra, Gemini 4 Argon, Claude Opus 5.5 and Claude Fable 5.1 over the week, above three of them on every one of the seven days and above GPT-6 Astra on five, and against the two Anthropic models the margin clears the pre-registered bar. One week is one independent window, so the size of the margin against GPT-6 Astra and Gemini 4 Argon is not yet settled; the direction is.

The result is also a case for the evaluation itself. Scoring every stock every day and grading each score against what the market, the sector and the style did not explain produced a verdict on five models in seven trading days, on a yardstick that favors none of them. Forward Model MK1's scores continue to be published daily and graded on the same terms, so the record that this report opens keeps growing after it is read, and each new version of the model earns its own.

Appendix A. The grade under simpler designs

A reader may ask whether the adjustment in section 5 was chosen because it favors Forward Model MK1. The simpler grades answer that. Table 3 regrades the same scores, the same names and the same days in three ways: on the raw 21-day return; on the return after the sector's mean is taken out, a fit on an intercept and the sector indicators only; and with the full market, sector and style adjustment of section 5. Subtracting the day's market median moves every stock by the same amount, so a grade on returns alone and a grade on market-relative returns rank the stocks identically and give the same figure. The names are those with a recorded price at both ends of the window, 3,261 to 3,266 of the pairs in Table 2 and 6,529 of the 6,592 pooled for GPT-6 Astra, so the full-adjustment row differs from Table 2 by a few thousandths; Table 2 is the record.

Table 3: 21-day skill under three grades, the mean over the seven test days. The first row is Forward Model MK1's own skill on the 500-name sample; each other row is Forward Model MK1 minus the model on the names both answered, with the corrected t and the days on which Forward Model MK1 was ahead.

	Returns only	Sector only	Full (section 5)
Forward Model MK1, skill on the sample	+0.149	+0.156	+0.110
Difference vs GPT-6 Astra (pooled 1,000 a day)	+0.081 (t 1.99, 7/7)	+0.078 (t 1.82, 7/7)	+0.036 (t 1.11, 5/7)
Difference vs Gemini 4 Argon	+0.105 (t 2.11, 7/7)	+0.110 (t 2.21, 7/7)	+0.060 (t 1.51, 7/7)
Difference vs Claude Opus 5.5	+0.099 (t 2.19, 7/7)	+0.105 (t 2.44, 7/7)	+0.070 (t 2.18, 7/7)
Difference vs Claude Fable 5.1	+0.097 (t 2.37, 7/7)	+0.103 (t 2.52, 7/7)	+0.067 (t 2.35, 7/7)

The simpler grades credit Forward Model MK1 with more, not less. About a quarter of what a grade on returns alone would award it is market, sector and style exposure that the full adjustment removes, while the rivals' own figures barely move, so every margin in Table 2 is smaller than the simpler grades would show. Under all three grades the ordering of the five models and the direction of every difference are the same. The published design is the conservative one.

References

[1] OpenAI. GPT-6 Astra model page (knowledge cutoff April 30, 2026). https://platform.openai.com/docs/models

[2] Anthropic. Claude Opus 5.5 and Claude Fable 5.1 model overviews (knowledge cutoff June 2026). https://docs.anthropic.com/en/docs/about-claude/models

[3] Google DeepMind. Gemini 4 Argon model card (knowledge cutoff March 2026). https://deepmind.google/models/