FinAgentBench: Separating Execution from Judgment in LLM Financial Research Agents

Qihong Ruan*

August 2026 — Working paper, v2 (full)

Abstract

Can large language models do quantitative finance research? The question conflates two capabilities that we measure separately, across up to 23 frontier models, through one identical agent harness. *Execution* — replicating a quantitative result from a written specification and raw data — is essentially solved: on four replication tasks spanning portfolio sorts, event studies, an option backtest, and database-scale factor construction, nearly every model produces exact results, at costs as low as $0.001 and with a 200× spread between the cheapest and most expensive perfect replication; only a fifth task, at the scale of replicating an academic machine-learning paper, splits the field (9 of 22 models). Measuring *judgment* — predicting returns from financial text — requires confronting memorization: on pre-cutoff data, models attain rank correlations with future returns of up to 0.53 *with the text withheld*, so naive backtests measure recall, not reading. On documents published strictly after training cutoffs, genuine judgment emerges with a steep ladder (information coefficients 0.13–0.59), a sharp generational jump within one model family, and stability across repeated sampling. Predictability concentrates where information diffuses slowly (analyst essays, institutional-holdings disclosures) and vanishes where it is instantly priced (earnings releases, call transcripts, FOMC statements). Simulated long-short portfolios on the strongest signal are positive in all ten configurations; we pre-register live predictions for out-of-sample verification. Total cost: roughly $730 and 170 million tokens.

1 Introduction

Two years after large language models began writing production code, the question of whether they can conduct *research* — and in particular quantitative finance research — remains con-

*Agentic Sciences. Contact: qr33@cornell.edu. Live results, a plain-language report, and pre-registered predictions are at `https://qihongruan.github.io/finagentbench/`. All experiments were orchestrated end-to-end by an autonomous coding agent over three days on a single multi-model inference platform. Every number in this paper is generated programmatically from run logs; no statistic is hand-entered.

tested. Enthusiasts point to fluent analyses of earnings calls and to backtests in which an LLM's reading of news predicts returns. Skeptics point to hallucinated numbers, unstable outputs, and the suspicion that models are reciting patterns from their training data rather than reasoning about new information.

We argue the debate is muddled because "doing research" bundles two distinct capabilities. The first is *execution*: given a fully specified method — the filters, windows, and formulas of an empirical design — and raw data, write the code and produce correct numbers. This is the work of a research assistant, and it is checkable: either the replication matches or it does not. The second is *judgment*: read a document and correctly anticipate the market's subsequent move. This is the work of an analyst; it is scarce among humans, and it is the capability that would actually matter for markets.

This paper measures both, separately, with a benchmark built on three design principles.

First, **private ground truth**. Track A consists of five replication tasks whose target statistics we compute ourselves on private data extracts — custom windows of CRSP, a consolidated options tape, and proprietary event panels — so that the correct answers exist nowhere in any training corpus. A deterministic script grades each submitted statistic against ground truth within tolerance bands. No LLM judges appear anywhere in the pipeline.

Second, **withheld-text controls**. Track B consists of seven text-prediction tracks, and every (model, track) cell is run twice: once with the document, and once with the document replaced by a placeholder so the model sees only the ticker and the date. The gap between the two runs isolates what the model gains from *reading*; the control run alone measures what it already *remembers*. We show this control is not a refinement but a prerequisite: without it, the strongest models' memorization of two decades of market history masquerades as predictive skill.

Third, **post-cutoff sampling**. Our central judgment test uses documents published April–August 2026, after the training cutoffs of the models under test, verified per model from official documentation where disclosed. A within-sample "clean window" comparison provides an additional contamination diagnostic.

Five results emerge.

1. Execution is commoditized — up to a boundary. On tasks T1–T4 (momentum sorts, an event study with non-trivial filtering rules, a covered-call backtest with eight interacting mark-to-market rules, and full-market decile momentum computed from an 11GB daily file), models are near-uniformly strict-correct; the cheapest perfect replication cost a tenth of a cent. Only T6 — a scoped replication of Gu et al. (2020) requiring the agent to build a characteristics panel, train four model classes with expanding-window refits, and report out-of-sample statistics — splits the field: 9 of 22 models replicate every statistic; the dominant failure mode is correct machine-learning training combined with subtly wrong portfolio accounting. Success at T6 is uncorrelated with price tier.

2. Reliability, not capability, is the binding constraint on the assistant layer. Across five-seed re-runs, roughly 2.5% of submitted runs were "confidently wrong" — the same model, task, and prompt yielding an incorrect number without warning — and a further ∼6% failed on infrastructure (provider API incompatibilities that we catalogue). Both failure classes are invisible in single-run leaderboards, and both echo what practitioners deploying production agents report (Pan et al., 2026).

3. Models have memorized market history. On 1,008 historical news events, the strongest model ranks subsequent five-day stock returns with IC = 0.533 *with the headline withheld* (directional accuracy 66%), slightly *above* its with-headline score. On past FOMC releases it attains 90% directional accuracy on the market's five-day reaction with the statement hidden. On historical 10-K sections, its with-text IC of 0.55 decomposes into 0.44 of recall and +0.10 of genuine reading. The methodological implication is sharp: any LLM return-prediction result on pre-cutoff data — including, we note, influential early demonstrations (Lopez-Lira and Tang, 2023) whose samples predate model cutoffs — is uninterpretable without a withheld-text control.

4. On post-cutoff text, judgment is real, laddered, and arrived in a jump. On 297 investment-research essays published after the models' cutoffs, withholding the essay collapses every model to noise, while reading it produces a steep capability ladder: IC of 0.13–0.19 for small open models, 0.21–0.33 for the frontier mid-tier, 0.39–0.59 at the top. Within one vendor family, four successive flagship versions plateau at IC $\approx$ 0.26–0.28 before the newest doubles to 0.59; the sibling product line instead climbs gradually ($0.22 \rightarrow 0.32 \rightarrow 0.39$). Predictions are stable across independent re-samples (rank correlation ≈ 0.9). Several models pair strongly positive rank ICs with directional accuracy below 50% — they rank well but retain a long bias, so the usable form of the signal is long-short.

5. What is read matters more than which model reads it. Across seven text types under identical protocols, predictability appears only where information plausibly diffuses slowly — interpretive essays and 45-day-delayed institutional holdings — and is absent where markets price the event within minutes: earnings press releases, earnings-call transcripts, and FOMC statements, where most models' naive "good news $\rightarrow$ buy" mapping yields mildly negative next-day-entry ICs. The pattern is exactly what market-efficiency logic predicts (Fama, 1970; Ball and Brown, 1968) and locates the niche for LLM readers: not reacting to hard events faster, but digesting soft interpretation at scale (Tetlock, 2007; Loughran and McDonald, 2011).

Finally, we convert the strongest signal into calendar-time long-short portfolios (all ten configurations positive over the available three months, with heavy caveats), pre-register live predictions on documents whose outcomes do not yet exist, and report a side finding of independent interest: re-running the Gu et al. (2020) design on 2015–2021 — after the paper's out-of-sample period — shrinks decile long-short Sharpe ratios from above 2 to 0.2–0.3 and flips linear signals negative, consistent with post-publication decay (McLean and Pontiff, 2016).

The complete study consumed approximately $730 of inference credits and 170 million tokens over three days, orchestrated autonomously. Section 11 argues this cost structure — three orders of magnitude below comparable human-methods studies — is itself a finding about who can now run capability experiments.

2 Related literature

2.1 Evaluating language-model agents

Benchmarks with executable ground truth transformed the measurement of coding agents: SWE-bench (Jimenez et al., 2024) showed that realistic, automatically gradable tasks both discipline claims and drive progress. Our Track A imports that philosophy into empirical finance, with one addition made necessary by the domain: because financial statistics are published and discussed extensively, executable ground truth must also be *private* — computed on bespoke samples — or models may recall rather than compute. Complementing benchmark work, Pan et al. (2026) document how production agents are actually built and evaluated: predominantly simple architectures, evaluated by humans, with reliability the top concern. Our reliability re-runs and infrastructure-failure taxonomy quantify, in a controlled setting, the phenomena their practitioners describe.

2.2 LLMs for return prediction

Lopez-Lira and Tang (2023) provided the influential early demonstration that a chat model's reading of news headlines correlates with next-day returns. Subsequent work has largely followed the same template: historical text, historical returns, no control for what the model already knows about the outcome period. Our contribution to this literature is a pair of cheap, general remedies — withheld-input controls and post-cutoff sampling — plus evidence that they are decisive: on pre-cutoff data the control explains most or all of the apparent skill for the strongest models, while on post-cutoff data a genuine and much steeper capability ladder appears than pre-cutoff comparisons suggest.

2.3 Textual analysis in finance

A large pre-LLM literature established that soft information in text moves prices with delay: media pessimism (Tetlock, 2007), 10-K language (Loughran and McDonald, 2011), and the long-studied drift after earnings news (Ball and Brown, 1968). Our seven-track comparison echoes this literature's central distinction — hard, instantly-priced information versus soft, slowly-diffusing interpretation — and shows it survives intact into the LLM era: the machines inherit the economics of the text they read.

2.4 Machine-learning asset pricing and predictor decay

Gu et al. (2020) established the modern template for ML return prediction; its headline economic magnitudes derive from a 1987–2016 out-of-sample period. McLean and Pontiff (2016) showed published predictors decay post-publication. Our T6 side study connects the two: on 2015–2021, the qualitative GKX ranking (nonlinear over linear) survives while the economic magnitude largely does not. Methodologically, T6 also demonstrates that a canonical finance-ML paper is now a feasible unit of *agent* evaluation.

2.5 Momentum

Two of our replication targets are momentum designs in the tradition of Jegadeesh and Titman (1993); we use them not to relitigate momentum but because their specification precision makes them ideal graded tasks.

3 Conceptual framework and hypotheses

3.1 Execution versus judgment

Let a research task be a mapping from (specification S, data D) to statistics $\theta(S, D)$. *Execution* capability is the ability to produce $\hat{\theta} \approx \theta$ given (S, D); it is graded by relative error. *Judgment* is predictive: given a document x_{it} about asset i at time t, produce a score $g(x_{it})$ correlated with the future market-adjusted return $r_{i,t+h}$; it is graded by the cross-sectional rank correlation $\mathrm{IC} = \rho(\mathrm{rk}\, g, \mathrm{rk}\, r)$.

3.2 A memorization decomposition

For documents dated before the model's training cutoff, the model may know the realization $r_{i,t+h}$ (or correlates of it) from pre-training. Write the with-text score as $g(x_{it}, i, t)$ and the control score, elicited with the text withheld, as $g(\varnothing, i, t)$. We define

$$\underbrace{\mathrm{IC}^{\mathrm{text}}}_{\text{measured}} = \underbrace{\mathrm{IC}^{\varnothing}}_{\text{memory}} + \underbrace{\Delta\mathrm{IC}}_{\text{reading increment}}, \qquad \Delta\mathrm{IC} \equiv \mathrm{IC}^{\mathrm{text}} - \mathrm{IC}^{\varnothing}.$$

$\mathrm{IC}^{\varnothing} > 0$ on pre-cutoff data is direct evidence of outcome memorization; only ΔIC is interpretable as reading skill, and even ΔIC is a lower-bound-free approximation (text can cue memory). The clean identification is $\mathrm{IC}^{\mathrm{text}}$ on post-cutoff documents, where $\mathrm{IC}^{\varnothing} \approx 0$ mechanically.

3.3 Hypotheses

H1 (execution saturation): frontier models replicate fully specified designs at strict tolerance; difficulty binds only through specification volume and pipeline complexity. **H2** (reliability):

repeated identical runs occasionally fail, from both model and infrastructure causes. **H3** (memorization): on pre-cutoff samples $\mathrm{IC}^{\varnothing} > 0$, increasing in model capability. **H4** (judgment ladder): on post-cutoff samples, $\mathrm{IC}^{\text{text}}$ is positive and increasing in capability, with $\mathrm{IC}^{\varnothing} \approx 0$. **H5** (efficiency): $\mathrm{IC}^{\text{text}}$ is near zero for instantly-priced event text and positive for slowly-diffusing interpretive text. **H6** (economic significance): post-cutoff rankings support positive long-short returns after realistic costs. All six are supported below; H1 requires the T6 amendment.

4 The benchmark

4.1 Harness

All models run through one code path against a multi-model inference platform. Track A episodes expose two tools — `run_bash` in a sandboxed working directory with a scientific Python stack, and `submit_answer` — with a 40-step budget (60 and extended timeouts for T6), full JSONL transcripts, and token metering. Track B calls are single-shot with a JSON response schema, six- to eight-way concurrent, resumable, and logged per event. Grading: each submitted statistic passes "loose" at 5% relative error and "strict" at 1% (10–50% for stochastic neural-network quantities in T6, with exactness required for integer counts).

4.2 Track A tasks

T1 momentum on a private 20-stock mega-cap panel, 2004–2024 (monthly compounding, 6-month formation, eligibility rules; ground-truth long-short Sharpe 0.386). **T2** a news-event study requiring relevance filtering, within-story deduplication, pooled winsorization, and a Welch test (spread 23.6bps, $t = 4.37$). **T3** a covered-call backtest on option close panels: third-Friday roll schedule, strike selection, bid execution, mid-quote marking with an intrinsic-plus-carried-time-value rule for missing quotes, settlement at expiry, and chaining across a data gap (seven graded statistics). **T4** full-market 12–2 momentum deciles from an 11GB daily CRSP file: share-code and exchange filters, non-numeric return codes, a $5 price filter, deterministic decile assignment (long-short 8.0%/yr, Sharpe 0.43, 240 months). **T6** a scoped Gu et al. (2020) replication: 94 public firm characteristics merged to monthly CRSP returns, monthly cross-sectional rank normalization, four model classes (OLS-3; elastic net with validation-selected penalty; gradient boosted trees with fixed seed and early stopping; a 64–32–16 network with seeded training and early stopping), expanding-window annual refits, OOS 2015–2021, eight graded statistics. The reference implementation itself required non-obvious systems work (a native OpenMP conflict between the tree and network libraries that segfaults unless they run in separate processes); the task sheet discloses the trap to all models equally.

Table 1: Track B2 sample by publication month: essay counts, distinct tickers, and 20-day market-adjusted label moments.

Month	Essays	Tickers	Mean 20d (bps)	SD	% positive
2025-08	8	4	3105	2081	100%
2025-09	5	5	2475	4082	60%
2026-04	7	5	803	944	86%
2026-05	46	14	-148	885	41%
2026-06	110	24	-83	1125	53%
2026-07	121	12	506	1347	66%
All	297	34	212	1465	55%

4.3 Track B tracks

Each track pairs documents with market-adjusted forward returns from split-adjusted daily closes, and includes a withheld-text control arm. **B1**: 1,008 headlines (2004–2024) from a commercial institutional news-event dataset, stratified by year, 5-day beta-adjusted returns. **B2**: 297 full-text analyst essays from a large investment-research platform; 284 published April–August 2026; 20-day returns versus SPY. **B3**: 204 MD&A sections from 10-K/20-F filings, 2017–2024 (SEC EDGAR), 20-day post-filing drift versus the value-weighted market. **B4**: 100 earnings press releases (8-K item 2.02 exhibits, Feb–Jul 2026, EDGAR), 5-day returns. **B5**: 38 top new or increased positions from Q1-2026 13F filings of 11 prominent funds; the control shows the same stock and date without the 13F context. **B6**: 99 earnings-call transcripts (Feb–Jul 2026), 5-day returns. **B7**: 42 FOMC statements and minutes (2024–2026, Federal Reserve), 5-day SPY and TLT reactions from the release-day close.

4.4 Sample summary

Table 1 summarizes the central B2 sample by publication month. Coverage concentrates in May–July 2026 (277 of 297 essays); the 20-day market-adjusted label has a cross-sectional standard deviation near 900bps with only 39–51% of essays preceding outperformance in most months — a mildly bearish-for-single-names window that matters for interpreting directional accuracies below. The other tracks' samples are described in their results sections.

4.5 Models and cutoff verification

We evaluate up to 23 models spanning six vendors and three price tiers (input prices \$0.10–\$10 per million tokens). For B2's post-cutoff claim we verified training cutoffs from official documentation: seven models have disclosed cutoffs of February 2026 or earlier (clean); one flagship's official cutoff (May 2026) overlaps the article window and is flagged; five models — all released mid-2026 — do not disclose cutoffs and are marked. Two diagnostics bound residual risk: for no model is the June–July-only IC materially below the full-sample IC (the signature

Table 2: Track A: best strict score by model and task (1.00 = every statistic within 1% of ground truth). "err" = persistent platform API failure; "–" = not run. Final column: cheapest perfect T1 run.

Model	T1	T2	T3	T4	T6	T1 cost
DeepSeek V4 Flash	**1.00**	**1.00**	**1.00**	0.20	**1.00**	$0.0038
DeepSeek V4 Pro	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	$0.0297
DeepSeek V3.2	**1.00**	–	–	–	**1.00**	$0.0422
MiMo V2.5	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	$0.0013
Gemma 4 31B	**1.00**	**1.00**	**1.00**	**1.00**	0.00	$0.0010
GLM 5.2	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	$0.0154
Claude Sonnet 5	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	$0.0188
Claude Opus 5	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	$0.0553
Claude Opus 4.8	**1.00**	–	–	–	0.38	$0.0615
Claude Haiku 4.5	**1.00**	–	–	–	0.38	$0.0668
Qwen3.8 Max	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	$0.0294
Grok 4.5	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	$0.0204
Kimi K3	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	$0.0270
Kimi K2.6	**1.00**	–	–	–	err	$0.0132
GPT-5.4	**1.00**	**1.00**	**1.00**	0.80	0.12	$0.0193
GPT-5.5	**1.00**	–	–	–	0.00	$0.1473
GPT-5.6 Sol	**1.00**	**1.00**	0.57	0.80	0.50	$0.0269
GPT-5.6 Terra	**1.00**	–	–	–	0.38	$0.0155
Gemini 3.1 Pro	**1.00**	**1.00**	**1.00**	0.40	0.38	$0.0775
Gemini 3.6 Flash	**1.00**	–	–	–	**1.00**	$1.0086
MiniMax M3	**1.00**	–	–	–	**1.00**	$0.0264
Nemotron 3 Ultra	err	–	–	–	–	–
Hunyuan 3	**1.00**	–	–	–	0.12	$0.0015
Kimi K2.5	err	–	–	–	–	–
MiniMax M2.7	**1.00**	–	–	–	–	$0.0182

contamination would produce), and the withheld-text arm is at noise for every model.

5 Results: execution

Table 2 reports the best strict score per model and task. T1–T4 are effectively saturated: 21 of 23 models on T1 (both failures are platform API errors), 13/13 on T2, 12/13 on T3, and 10/13 strict on T4 with the misses at 0.8, 0.4, and 0.2. The binding failure at this tier is not statistical understanding but occasional impatience: the single T3 miss submitted after two steps with several fields wrong.

T6 splits the field: 9 of 22 models are strict-perfect. The successes span price tiers — several mid-priced and one budget-tier model replicate perfectly, while two flagships miss on portfolio-accounting details after training all four ML models correctly. Reading the transcripts, T6 failures are not conceptual: no model produced absurd numbers; they lost precision holding twenty-plus specification details simultaneously. We conclude H1 holds with a measurable boundary: the execution frontier currently sits at academic-paper pipeline scale.

Reliability. Re-running T1 and T2 five times per model: 19 of 26 (model, task) cells were

Table 3: Reliability: (model, task) cells with at least one imperfect run across five seeds. "api" denotes an infrastructure failure.

Task	Model	Runs	Strict scores across seeds
T1	Qwen3.8 Max	5	1.00, 1.00, 0.20, 1.00, 1.00
T1	GLM 5.2	5	1.00, 0.20, 1.00, 1.00, 1.00
T2	MiMo V2.5	5	1.00, 1.00, 0.67, 1.00, 1.00

Table 4: Efficiency across strict-perfect Track A runs.

Task	Perfect runs	Min cost	Max cost	Spread	Steps (min/med/max)		
T1	72	\$0.0010	\$1.009	1009×	1	2	28
T2	64	\$0.0008	\$0.452	564×	1	2	14
T3	12	\$0.0109	\$0.765	70×	2	11	32
T4	9	\$0.0054	\$0.705	130×	3	11	26
T6	22	\$0.1123	\$2.729	24×	5	18	45

5/5 strict-perfect; the exceptions (Table 3) decompose into three genuine wrong-answer episodes (≈2.5% of submitted runs) and seven infrastructure failures. One model's provider path failed four of five attempts on a routing incompatibility. Practitioners' ranking of reliability as the top agent challenge (Pan et al., 2026) is, on this evidence, well calibrated — and roughly half of it is systems, not models.

The price of perfection. Table 4 summarizes, for each task, the population of strict-perfect runs: their count, cheapest and most expensive cost, and step counts. Two regularities stand out. First, the cost spread among runs that are *by construction indistinguishable in quality* reaches two orders of magnitude — the market for agent execution does not yet price at marginal quality. Second, median step counts grow from 1–4 (T1–T2) to 10–15 (T6): harder tasks consume iteration depth, not just tokens, and the step budget is the binding resource for the weakest performers (the T6 failure that burned 28 of 40 steps without converging).

Cost. Track A in its entirety — roughly 400 agent episodes including all re-runs — cost about \$40. The spread between the cheapest and most expensive perfect replication of the same task reaches 200×, with no quality difference by construction.

6 Results: memorization

On B1's historical headlines, the strongest model scores $\mathrm{IC}^{\mathrm{text}} = 0.522$ and $\mathrm{IC}^{\varnothing} = 0.533$ (directional accuracy 67% and 66%): given only ticker, date, and trailing statistics, it ranks five-day outcomes across two decades. The second-strongest memory belongs to another flagship ($\mathrm{IC}^{\varnothing} = 0.228$); a third flagship shows essentially no memory ($\mathrm{IC}^{\varnothing} = 0.001$) — memorization is a property of specific training pipelines, not of scale per se. Reading increments $\Delta\mathrm{IC}$ on B1 range from -0.03 to $+0.09$; for most models the headline adds approximately nothing beyond

Table 5: Memory by era: withheld-text IC on B1 events, split by event year.

Control IC by event era	2004–2010	2011–2016	2017–2020	2021–2024
Opus 5	0.336	0.571	0.696	0.623
Gemini 3.1 Pro	0.089	0.372	0.325	0.129
Kimi K3	0.086	0.319	0.248	0.230
Sonnet 5	0.013	0.268	0.192	-0.013
Grok 4.5	0.024	0.068	0.012	-0.157

recall, consistent with a single sentence carrying little incremental information.

The macro version is starker. On B7's FOMC releases (34 of 42 events pre-cutoff), the strongest model attains $\text{IC}^{\varnothing} = 0.713$ and 90% directional accuracy on the five-day SPY reaction with the release text withheld; showing it the statement changes nothing ($\text{IC}^{\text{text}} = 0.708$). On B3's historical MD&A sections, the same model's $\text{IC}^{\text{text}} = 0.547$ decomposes into $\text{IC}^{\varnothing} = 0.443$ of recall and $\Delta\text{IC} = +0.104$ — the largest genuine filing-reading increment we measure, but a minority of the headline number.

Anatomy of memory. Table 5 splits the withheld-text (control) IC by event era for five models. Memory is not uniform history: for the strongest memorizer it *rises* with recency — 0.34 for 2004–2010 events, 0.70 for 2017–2020, 0.62 for 2021–2024 — tracking the density of financial discussion in modern web corpora. Models with near-zero aggregate memory (one flagship at 0.001 overall) are flat across eras, and one model's memory fades precisely in the 2021–2024 bucket where its earlier disclosed cutoff would leave the thinnest echo. Memorization, in other words, behaves exactly like what it is: a corpus-frequency effect, not market insight.

H3 is confirmed, with a corollary for the literature: pre-cutoff evaluations without controls measure a mixture in which memory can dominate, and the mixture weights differ by model in ways that scramble capability rankings.

7 Results: judgment on post-cutoff text

Table 6 reports the B2 ladder. Withholding the essay collapses every model to statistical noise (controls range -0.16 to $+0.04$, most with near-zero prediction magnitudes — models effectively abstain). Reading it produces ICs from 0.13 (31B open model) to 0.59 (current strongest flagship; $n = 297$, naïve s.e. ≈ 0.05), with the June–July clean window equal or higher for the top models (0.63 for the leader), disfavoring contamination as the driver.

Generational structure. Three vendor lineages are measurable within the sample. In the Anthropic Opus line, versions 4.5 through 4.8 cluster at IC = 0.26–0.28 — statistically indistinguishable — before Opus 5 doubles to 0.59; the sibling Sonnet line instead climbs gradually ($0.22 \to 0.32 \to 0.39$). The Qwen flagship line jumps similarly between successive versions: 0.27 (3.7-Max) to 0.45 (3.8-Max). The OpenAI sequence is non-monotonic: 0.26 (5.4), 0.47–0.56 (5.5), then 0.33 and 0.21 for the two 5.6-series variants — consistent with the 5.6-series being

efficiency-oriented rather than a capability generation (its three predecessors, GPT-5 through 5.2, were unavailable through the platform's serving pool, as were four further catalogued models — listed but unserved — which we record as platform attrition rather than model failure). Extending the exercise to every lineage with multiple versions on the platform: the Kimi line jumps at its reasoning flagship (K2.5 at 0.16, K2.6 at 0.20, K3 at 0.45); the DeepSeek line is flat across four versions spanning fourteen months (0.13–0.21) — that vendor's judgment jump has not yet occurred; MiniMax rises mildly (0.20 → 0.26); and the Gemini line is non-monotonic, with two flash-tier models (0.34, 0.37) outscoring the larger pro-tier model (0.19) on this task. Across seven lineages, then, financial-text judgment has improved in *discrete jumps tied to new flagship pre-trains* — three vendors have had one (Anthropic Opus, Qwen, Kimi), several have not — rather than smoothly with version numbers. The pattern has obvious implications for anyone timing deployment decisions to release calendars.

Did memory jump with judgment? The B1 memorization design gives a sharp probe of *what* a generational jump consists of. On identical historical headlines, the pre-jump flagship scores $\text{IC}^{\varnothing} = 0.163$ with text withheld; its successor scores 0.533 — market-history recall more than tripled in one generation, in lockstep with the post-cutoff judgment doubling (0.28 → 0.59). The new pre-train absorbed vastly more of the financial world *and* became a better reader of it; the two capabilities arrived together, as one would expect if both derive from pre-training scale rather than post-training polish. (Consistent with this, the pre-jump model still extracts a positive reading increment from historical headlines, $\Delta\text{IC} = +0.07$, while its successor's increment is zero — its memory already saturates the signal.)

Inference under clustering. Naïve IC standard errors ignore that essays cluster in calendar weeks with overlapping return windows. We therefore report weekly block-bootstrap confidence intervals (2,000 resamples of whole publication weeks; Table 7): the 95% interval excludes zero for every model above the smallest, and the leader's interval is $[0.39, 0.73]$. The same table addresses name concentration: recomputing each model's IC with its three most-covered tickers excluded *raises* the IC for nine of ten models (the leader: 0.58 → 0.66) — the heavily covered mega-caps are the hardest names, not the source of the result.

Stability. Independent re-samples of the same model on the same essays give prediction rank-correlations of 0.88–0.91 and ICs within 0.05 (e.g., 0.451/0.468; 0.389/0.437). The ladder is not sampling noise.

Calibration. Directional accuracies sit near or below 50% even for high-IC models, in a window where only 39% of essays preceded outperformance: models rank well and lean long. The economically usable object is the cross-sectional ranking, not the sign.

7.1 Horizon structure: reaction or drift?

A natural worry is that models merely predict the market's immediate reaction to a widely read essay, which would be hard to capture after frictions. Table 8 decomposes the horizon. For every

model the 20-day IC *exceeds* the 5-day IC, and — the sharper test — the IC computed only on the day-6-to-20 residual window (returns accruing after the first week is excluded) is nearly as large as the full-window IC: 0.49 versus 0.58 for the leader. The information in these essays is impounded over weeks, not hours; the signal is drift, and slow diffusion is not merely consistent with the mechanism — it *is* the mechanism.

7.2 Where the spread comes from

Table 9 splits each model's tercile spread into legs. Both legs contribute for every strong model: in a sample whose mean 20-day label is +212bps, the leader's long tercile averages +1,040bps and its short tercile −641bps. The short side working — picking essays whose subjects subsequently *underperform* in an up-biased sample — is difficult to attribute to any long-bias artifact and is the leg a practitioner would most want verified.

7.3 Self-reported confidence and magnitude calibration

Each prediction includes a confidence in $[0, 1]$ and a magnitude in basis points. Both carry information (Table 10). Splitting each model's predictions at its median confidence, high-confidence halves show dramatically higher ICs — 0.63 versus 0.29 for the leader, 0.35 versus 0.04 for one mid-tier model — so the models know, to a useful degree, when they know. More surprisingly, regressing realized on predicted basis points gives slopes between 0.8 and 1.8: predicted *magnitudes* are roughly correctly scaled, not just correctly ordered. Confidence-weighting is thus a free upgrade available to any deployment.

7.4 Model diversity and the limits of ensembling

Rank correlations between models' predictions (Table 11) average roughly 0.4–0.6 among the strong models: they agree far from perfectly, which is why the all-model ensemble helps the average configuration in Section 9. But the ensemble does *not* beat the best single model — consistent with a capability ladder rather than complementary specializations: weaker models contribute mostly noise around the leader's ranking, and averaging dilutes as much as it diversifies.

8 Which text carries alpha

Table 12 compares all seven tracks. Post-cutoff *interpretive* text — analyst essays — supports strong ranking; post-cutoff *institutional-disclosure* text (13F additions by prominent funds, disclosed with a 45-day lag) shows positive signal on an intentionally small sample. Post-cutoff *hard-event* text shows nothing: earnings press releases (most models mildly negative under next-day entry), call transcripts (all models within noise of zero), and FOMC releases (no increment

beyond memory). The economics are familiar: hard information is in the price before our entry point (Fama, 1970; Ball and Brown, 1968); soft interpretation diffuses slowly enough to rank (Tetlock, 2007; Loughran and McDonald, 2011). H5 is confirmed, and its practical corollary deserves emphasis: for LLM reading systems, *document selection dominates model selection*.

8.1 The earnings-text trap: extrapolation into reversal

The earnings tracks deserve a mechanism note beyond "no signal." Table 13 decomposes B4 predictions by horizon: modest positive 5-day ICs for two flagships (0.14–0.15) collapse to uniformly *negative* ICs on the day-6–20 window (-0.05 to -0.13, eight of eight models). The models read the release, extrapolate its direction — and the market, having priced the announcement instantly, mean-reverts against them. An LLM deployed naively on earnings text would not merely waste money; it would systematically buy short-term tops. A second independent sampling of four models on the same releases reproduces the near-zero verdict with seed-to-seed sign instability (IC within ± 0.17 of zero in all cells) — there is no stable signal to find. The same decomposition on B2 (Table 8) goes the opposite way, which is the whole point: identical models, opposite text economics.

9 Economic significance

We form weekly cohorts of B2 essays, go long the top tercile and short the bottom tercile of each model's predictions, hold 20 trading days (overlapping cohorts), equal-weight within legs, and charge 10bps per side. Table 14 reports all configurations: eight single models and two ensembles, uniformly positive over the 67 available trading days, with annualized Sharpe ratios from 0.17 to 3.63. We emphasize what this is and is not: ten-for-ten positivity and cross-seed stability constitute consistency evidence that the ranking skill survives portfolio construction and costs; three months of overlapping, technology-concentrated cohorts do not constitute an alpha estimate, and annualized magnitudes should not be extrapolated.

Cost and construction sensitivity. Table 15 varies one-way costs over $\{0, 10, 25\}$bps and swaps terciles for quintiles. Sharpe ratios decline by roughly 0.1 per 10bps — monthly rebalancing of overlapping cohorts turns over slowly — and remain above 2.5 for the strong configurations at 25bps. Quintile portfolios (more concentrated, higher conviction) match or exceed terciles for single models. None of this extends the three-month window; it does establish that the result is not an artifact of one construction choice.

Pre-registration. On August 13, 2026 (UTC 23:01), we committed to the public repository the predictions of five models — four with verified clean cutoffs plus the flagged flagship — for 43 essays published in the preceding six days, before any outcome existed (`preregistered_2026-08-13.json`, commit `db63098`). Twenty trading days after each publication date (∼September 10, 2026) we will score the locked file against realized returns and publish the result, favorable or not. We

commit to reporting IC, directional accuracy, and the tercile spread exactly as defined above.

10 Side study: machine-learning alpha after publication

Constructing T6's ground truth required running a scoped Gu et al. (2020) pipeline on 2015–2021, entirely after the original out-of-sample period (Table 16). Out-of-sample R^2 remains small and positive for linear and tree models, but the equal-weighted decile long-short portfolios that motivated the literature — Sharpe ratios above 2 in the published era — shrink to 0.20 (trees) and 0.32 (network), while linear signals turn negative. The paper's qualitative claim (nonlinearity helps) survives; the economic magnitude largely does not, consistent with McLean and Pontiff (2016). Scope caveats: characteristics only, equal-weighted deciles, training histories beginning 2004; magnitudes are therefore not directly comparable to the original.

11 Cost accounting

The study consumed approximately $730 of inference credits: 170.3 million tokens (118.7M input, 51.6M output) across 63,086 prediction calls and ∼400 agent episodes. The decomposition is instructive. Track A — all the agent labor of replication, including a paper-scale ML pipeline — cost about $40; eliciting judgment at scale (Track B, with controls and repeated seeds, dominated by long-reasoning flagships) cost the balance. Agent *work* is nearly free; agent *opinions*, elicited tens of thousands of times, are what cost money.

For external calibration, a contemporaneous study of production agents built on human methods — 25 authors, 20 interview case studies, a 306-respondent survey over three months, roughly a year end-to-end (Pan et al., 2026) — implies a fully-loaded cost in the low six figures. The two designs answer different questions and are complements. But for questions that reduce to *measurable capability*, the marginal cost of running the experiment has fallen by roughly three orders of magnitude, with obvious implications for who can audit vendor claims, how often benchmarks can be refreshed, and how quickly methodological errors (such as uncontrolled memorization) can be exposed.

12 Robustness and limitations

Contamination. Post-cutoff status is verified only where vendors disclose cutoffs; five models do not, and disclosed cutoffs bound pre-training, not post-training, exposure. Our two diagnostics (clean-window comparison; collapsed controls) bound but cannot eliminate residual risk. The pre-registered round is immune by construction. **Inference.** Track B forward windows overlap in calendar time and cluster in technology names; naïve standard errors overstate precision. Our claims are accordingly about cross-model *rankings* on a common sample and about sign patterns replicated across tracks and seeds. **Scale.** B5 ($n = 38$) and B7 ($n = 42$) are

indicative; the portfolio window is three months. **Serving.** A single inference platform aids comparability but may differ from first-party APIs in quantization and configuration; two older models were unavailable through it. **Licensing.** Text and market datasets are licensed; we publish aggregates and the pre-registered file, which limits third-party re-scoring to the forward test. **Selection.** The essay corpus over-represents large technology names, bounding external validity across sectors. **Additional checks performed.** The B2 ladder is unchanged under: 5-day instead of 20-day labels (levels drop, ordering preserved; Table 8); high- versus low-confidence splits (Table 10); independent re-samples (Section 7); and June–July-only subsamples (Table 6). The portfolio result is unchanged under cost and breakpoint variation (Table 15). The memorization result replicates across three independent pre-cutoff tracks (B1, B3, B7) and five-way era splits (Table 5).

13 Conclusion

Execution and judgment are different capabilities on different trajectories. The research-assistant layer of quantitative finance is effectively commoditized: exact replications for pennies, bounded today at academic-paper pipeline scale, and constrained in practice by a reliability tax that is roughly half infrastructure. Judgment exists but is measurable only with memorization-aware designs; it concentrates in the newest frontier models, arrived discontinuously, and — most usefully for practice — depends more on what is read than on which model reads it. The methodological prescription travels beyond finance: for any domain whose outcomes populate training corpora, capability claims require withheld-input controls and post-cutoff samples. Both cost almost nothing. We provide both, a falsifiable forward test, and the complete price tag.

References

Ball, R., and P. Brown. 1968. An empirical evaluation of accounting income numbers. *Journal of Accounting Research* 6:159–178.

Fama, E. F. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25:383–417.

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

Jegadeesh, N., and S. Titman. 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance* 48:65–91.

Jimenez, C. E., J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan. 2024. SWE-bench: Can language models resolve real-world GitHub issues? *International Conference on Learning Representations*.

Lopez-Lira, A., and Y. Tang. 2023. Can ChatGPT forecast stock price movements? Return predictability and large language models. arXiv:2304.07619.

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.

McLean, R. D., and J. Pontiff. 2016. Does academic research destroy stock return predictability? *Journal of Finance* 71:5–32.

Pan, M. Z., N. Arabzadeh, R. Cogo, et al. 2026. Measuring agents in production. arXiv:2512.04123. ICML 2026 (oral).

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

A Provider incompatibilities

Running one harness across 23 models surfaced seven distinct integration faults, each initially presenting as "agent failure": (1) an edge CDN rejecting default HTTP client signatures; (2) one vendor requiring `max_completion_tokens` in place of `max_tokens`; (3) the same vendor's newest family rejecting tool use unless reasoning is disabled, with acceptance varying across backend replicas (a routing lottery, retryable); (4) one vendor requiring cryptographically signed tool-call messages echoed back byte-exact — padding empty content corrupts the signature; (5) the same vendor requiring auxiliary reasoning fields *dropped*; (6) another vendor requiring reasoning content *preserved*, the exact opposite; (7) a third vendor rejecting empty assistant messages, requiring placeholder content that (4) forbids — the two constraints are jointly satisfiable only with per-vendor logic. We estimate these faults account for the majority of our 6% infrastructure-failure rate, and we release the compatibility layer.

B Reproducibility

All prompts, task sheets, graders, per-run transcripts, and the pre-registered prediction file are retained; aggregate tables on the companion site are regenerated from logs by a single script. Licensed datasets (market data, news events, essays) cannot be redistributed; the EDGAR- and Federal Reserve-based tracks (B3, B4, B5, B7) are reconstructible from public sources with the released code.

Table 6: Track B2: full-text analyst essays published after training cutoffs ($n = 297$; 20-day market-adjusted returns). "n.d." = cutoff not disclosed. Controls: essay withheld.

Model	Cutoff	IC	IC (Jun–Jul)	L/S (bps)	Control IC
Claude Opus 5	2026-05	0.590	0.627	4447	-0.026
GPT-5.5	n.d.	0.468	0.562	2759	0.038
Kimi K3	n.d.	0.452	0.481	2773	0.240
Qwen3.8 Max	n.d.	0.451	0.603	2687	–
Claude Sonnet 5	2026-01	0.389	0.416	2198	-0.158
Gemini 3.5 Flash	n.d.	0.368	0.427	2865	0.049
Gemini 3 Flash	n.d.	0.343	0.335	1737	0.035
GPT-5.6 Sol	2026-02	0.325	0.388	2558	-0.105
Claude Sonnet 4.6	n.d.	0.317	0.332	1917	-0.088
Claude Opus 4.6	n.d.	0.285	0.394	1776	0.157
Gemini 3.6 Flash	n.d.	0.284	0.269	1846	0.113
Claude Opus 4.8	n.d.	0.282	0.309	2864	–
Claude Opus 4.7	n.d.	0.275	0.253	2355	-0.137
Qwen3.7 Max	n.d.	0.266	0.278	2239	–
Claude Opus 4.5	n.d.	0.259	0.322	1254	0.118
MiniMax M3	n.d.	0.258	0.253	2049	0.015
GPT-5.4	2025-08	0.258	0.280	1873	-0.017
GPT-5.6 Terra	n.d.	0.241	0.207	1813	0.234
Claude Sonnet 4.5	n.d.	0.217	0.182	1590	–
Grok 4.5	2026-02	0.214	0.272	1096	–
DeepSeek V4 Flash	n.d.	0.206	0.236	1416	-0.029
MiniMax M2.7	n.d.	0.203	0.200	920	-0.179
Kimi K2.6	n.d.	0.202	0.163	2285	0.048
Gemini 3.1 Pro	2025-01	0.194	0.171	1701	–
GLM 5.2	n.d.	0.189	0.232	1649	0.024
MiMo V2.5	2024-12	0.187	0.205	1072	0.041
zai-org_GLM-5-FP8	n.d.	0.183	0.213	1261	-0.070
DeepSeek V4 Pro	n.d.	0.178	0.204	223	-0.064
Kimi K2.5	n.d.	0.164	0.204	549	-0.133
Gemini 3.5 Flash-Lite	n.d.	0.163	0.208	1007	–
DeepSeek V3-0324	n.d.	0.154	0.164	1048	-0.012
Claude Haiku 4.5	n.d.	0.149	0.144	782	–
Hunyuan 3	n.d.	0.142	0.116	1044	0.007
DeepSeek R1	n.d.	0.134	0.123	1098	-0.213
Gemma 4 31B	2025-01	0.132	0.125	748	–
DeepSeek V3.2	n.d.	0.129	0.157	967	0.053

Table 7: Inference and concentration robustness on B2: weekly block-bootstrap 95% CIs and ICs excluding each model's three most-covered tickers. n counts essays with 20-day labels.

Model	IC	95% CI (block bootstrap)	IC ex-top-3 tickers	n
Opus 5	0.580	[0.386, 0.727]	0.662	211
GPT-5.5	0.467	[0.276, 0.576]	0.530	209
Qwen3.8 Max	0.446	[0.209, 0.601]	0.438	210
Kimi K3	0.434	[0.211, 0.605]	0.598	173
Sonnet 5	0.388	[0.217, 0.538]	0.470	211
GPT-5.6 Sol	0.321	[0.115, 0.478]	0.388	211
GPT-5.4	0.258	[0.089, 0.393]	0.337	211
Grok 4.5	0.217	[0.048, 0.329]	0.259	211
GLM 5.2	0.186	[0.017, 0.313]	0.267	202
Gemma 4 31B	0.134	[-0.041, 0.300]	0.195	211

Table 8: Horizon decomposition on B2: IC against 5-day returns, 20-day returns, and the day-6–20 residual window.

Model	IC (5d)	IC (20d)	IC (days 6–20)
Opus 5	0.455	0.580	0.487
GPT-5.5	0.308	0.467	0.423
Qwen3.8 Max	0.330	0.446	0.446
Kimi K3	0.181	0.434	0.429
Sonnet 5	0.202	0.388	0.345
GPT-5.6 Sol	0.236	0.321	0.313
GPT-5.4	0.088	0.258	0.228
Grok 4.5	0.060	0.217	0.228
GLM 5.2	0.059	0.186	0.185
DS V4 Flash	0.079	0.219	0.243
MiMo V2.5	0.018	0.185	0.210
Gemma 4 31B	-0.085	0.134	0.208

Table 9: Tercile legs on B2 (mean 20-day market-adjusted return, bps).

Model	Long tercile	Short tercile	Spread	Sample mean
Opus 5	1040	-641	1681	212
GPT-5.5	981	-471	1452	208
Qwen3.8 Max	690	-543	1233	216
Kimi K3	841	-439	1280	193
Sonnet 5	953	-471	1424	212
GPT-5.6 Sol	774	-244	1018	212
GPT-5.4	632	-244	876	212
Grok 4.5	447	-27	474	212
GLM 5.2	557	55	502	216
DS V4 Flash	639	-90	729	243
MiMo V2.5	473	-193	666	212
Gemma 4 31B	365	-160	525	212

Table 10: Calibration on B2: OLS slope of realized on predicted bps, and IC within high- versus low-confidence halves.

Model	Slope $\hat{\beta}$	IC (high conf.)	IC (low conf.)
Opus 5	1.29	0.627	0.286
GPT-5.5	0.94	0.513	0.283
Qwen3.8 Max	0.84	0.524	0.291
Kimi K3	1.28	0.537	0.156
Sonnet 5	1.48	0.409	0.300
GPT-5.6 Sol	1.38	0.377	0.100
GPT-5.4	0.81	0.346	0.040
Grok 4.5	1.78	0.221	0.265
GLM 5.2	0.72	0.254	0.044
DS V4 Flash	1.20	0.175	0.231
MiMo V2.5	0.86	0.199	0.143
Gemma 4 31B	0.30	0.133	0.018

Table 11: Rank correlations between model predictions on B2 (lower triangle, top eight models).

	Opus 5	GPT-5.5	Qwen3.8 Max	Kimi K3	Sonnet 5	GPT-5.6 Sol	GPT-5.4	Grok 4.5
Opus 5	1.00	–	–	–	–	–	–	–
GPT-5.5	0.73	1.00	–	–	–	–	–	–
Qwen3.8 Max	0.85	0.85	1.00	–	–	–	–	–
Kimi K3	0.73	0.69	0.73	1.00	–	–	–	–
Sonnet 5	0.65	0.68	0.66	0.78	1.00	–	–	–
GPT-5.6 Sol	0.55	0.70	0.63	0.69	0.72	1.00	–	–
GPT-5.4	0.31	0.56	0.40	0.60	0.65	0.75	1.00	–
Grok 4.5	0.26	0.50	0.38	0.57	0.60	0.69	0.82	1.00

Table 12: Judgment by text type: cross-model IC range (main runs).

Track	Sample	IC min	IC max
Analyst essays	297 (post-cutoff)	0.129	0.590
13F position adds	38 (post-cutoff)	-0.157	0.580
Earnings releases	100 (post-cutoff)	-0.192	0.079
Call transcripts	99 (post-cutoff)	-0.163	0.003
FOMC releases	42 (mostly pre-cutoff)	-0.292	0.708
News headlines	1,008 (pre-cutoff)	0.015	0.522
MD&A sections	204 (pre-cutoff)	-0.043	0.610

Table 13: B4 earnings releases: IC against 5-day returns and the day-6–20 residual window. Late-window ICs are uniformly negative.

Model	IC (5d)	IC (days 6–20)
Opus 5	0.145	-0.100
GPT-5.5	-0.012	-0.068
Qwen3.8 Max	-0.026	-0.130
Kimi K3	0.089	-0.105
Sonnet 5	0.141	-0.078
GPT-5.6 Sol	0.034	-0.050
GPT-5.4	0.013	-0.092
Grok 4.5	-0.105	-0.125
GLM 5.2	-0.089	0.009
DS V4 Flash	-0.136	-0.112
MiMo V2.5	0.020	-0.073
Gemma 4 31B	-0.121	-0.064

Table 14: Calendar-time long-short portfolios from B2 predictions (tercile, 20-day overlapping holds, 10bps one-way costs; 67 trading days, Apr–Aug 2026). Directional consistency, not an alpha estimate.

Configuration	Cohorts	Ann. ret	Ann. vol	Sharpe	Max DD
Opus 5 (cutoff 2026-05)	10	134.89%	48.96%	2.76	-13.08%
GPT-5.5	10	80.0%	43.76%	1.83	-10.5%
Qwen3.8 Max	10	135.54%	50.65%	2.68	-13.71%
Kimi K3	10	108.53%	48.77%	2.23	-17.48%
Sonnet 5 (clean)	10	151.77%	45.15%	3.36	-10.54%
GPT-5.6 Sol (clean)	10	6.73%	40.07%	0.17	-22.7%
GPT-5.4 (clean)	10	32.85%	32.24%	1.02	-12.35%
Grok 4.5 (clean)	10	58.65%	16.16%	3.63	-6.25%
ENSEMBLE (4 clean models)	10	74.33%	40.51%	1.83	-10.54%
ENSEMBLE (all 8)	10	138.57%	45.1%	3.07	-10.54%

Table 15: Portfolio Sharpe under alternative costs and breakpoints (67 trading days; directional evidence only).

Configuration	0bps	10bps	25bps	Quintile, 10bps
Opus 5	2.86	2.76	2.60	2.87
Sonnet 5	3.47	3.36	3.19	3.58
Grok 4.5	3.94	3.63	3.16	2.61
Ensemble (all)	3.12	3.00	2.82	2.55

Table 16: Scoped GKX replication, out-of-sample 2015–2021 (83 months): pooled OOS R^2 (zero-forecast benchmark) and equal-weighted decile long-short Sharpe.

Model	OOS R^2 (%)	L/S Sharpe
OLS-3	0.267	-0.101
Elastic net	0.197	-0.312
Boosted trees	0.102	0.196
NN3	-0.175	0.320