Can AI Do Financial Research?

LLM-Guided Hypothesis Discovery in Asset Pricing

Huan Liu Miao Liu Zhizhe Liu Danqing Mei

Google *Boston College* *Columbia* *CKGSB*

2026

The Question

Can an AI research agent autonomously execute the

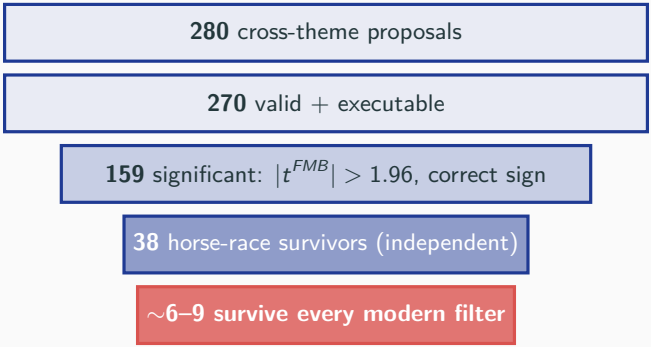
hypothesis-discovery loop *in empirical asset pricing?*

Three nearby strands we are *not* doing:

- ML for return prediction (Gu, Kelly, Xiu 2020; Chen, Pelger, Zhu 2024)
- AI as *paper-writing* tool (Novy-Marx & Velikov 2026)
- ML for hypothesis generation (Ludwig & Mullainathan 2024)

Our object of study
The discovery loop itself: *propose* an interpretable accounting signal → *test* on real CRSP/Compustat data → *reflect* and revise.
Observed end-to-end, including the agent's reasoning trace.

Preview: an honest attrition story



Two signals survive everything:

- `LeveragedPPEGrowth` — debt *level* × PP&E growth.
  Aggressive capacity expansion on an already-leveraged balance sheet.
- `UltimateAlphaComposite` — `SUMRANK` of two orthogonal channels: leverage-level × investment growth, and debt-issuance *flow* × investment growth.

The Discovery Laboratory

Design: human laboratory + AI agent

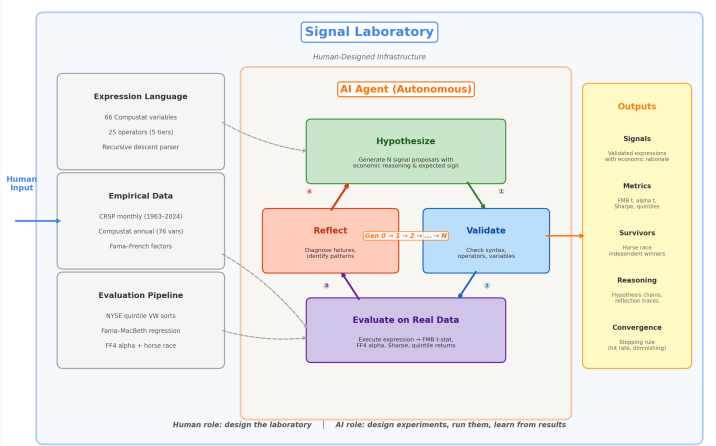
Humans design the laboratory:

- Hypothesis *language*
- Admissible *data*
- *Validation* rules
- *Evaluation* pipeline

AI executes the loop:

- Propose signals
- Observe outcomes
- Reflect → revise

Evaluation standards stay independent of search.



Constrained creativity: the expression language

66 Compustat variables $\times$ 24 operators, max nesting depth 6.

Operator tier	Examples
Algebra / scaling	`ADD, SUB, MUL, RATIO, SCALE`
Composite formation	`SUMRANK` (cross-sectional rank composites)
Temporal transforms	`GROWTH, DELTA, LAG, MA, VOL, TREND, ACCEL`
Industry-relative	`IND_ADJ, ZSCORE, IND_ZSCORE, PERCENTILE`
Nonlinear / robust	`LOG, ABS, NEG, INV, SIGN, WINSOR, INDICATOR`

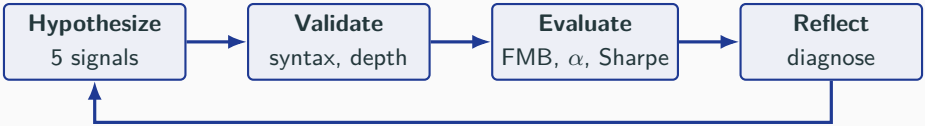
Example proposal (a "cash cow" signal):

```
SUMRANK( IND_ADJ(RATIO(SUB(revt,cogs),at)), NEG(GROWTH(at,1)) )
```

Why constrained?
The narrow language guarantees every proposal is *parseable, executable, and interpretable*. Same logic as AlphaFold: disciplined constraints make creative search tractable.

The propose–test–reflect loop

One *generation*: 5 proposals → 5 evaluations → structured feedback.



The agent: Claude Opus 4.6, extended thinking, 10k thinking tokens/call.
Encrypted reasoning chain is preserved across all 7 generations—the agent truly builds on its prior thought, not just summary stats.

Feedback categories: successful | wrong-sign | weak | error. Each new generation must include at least one revision of a prior failure and one new direction.

Search structure: 8 cross-theme pairs

Theme pair	Economic mechanism
Profitability × Investment	"cash cow" (high profit, low growth)
Valuation × Quality	quality value (cheap *and* fundamentally strong)
Investment × Financing	empire building (growth + external finance)
Profitability × Intangibles	R&D productivity (profitable innovators)
Profitability × Valuation	value-quality wedge
Accruals × Quality	earnings quality premium
Valuation × Financing	disciplined value (cheap, low issuance)
Investment × Accruals	real vs. working-capital growth

8 pairs × 7 generations × 5 proposals = **280 candidate signals**

CRSP/Compustat **July 1963 – Dec 2024**; microcap-excluded; NYSE breakpoints; VW portfolios; Fama–MacBeth w/ Newey–West (6 lags). Single-theme programmatic baseline (99 signals, no LLM) provides the control.

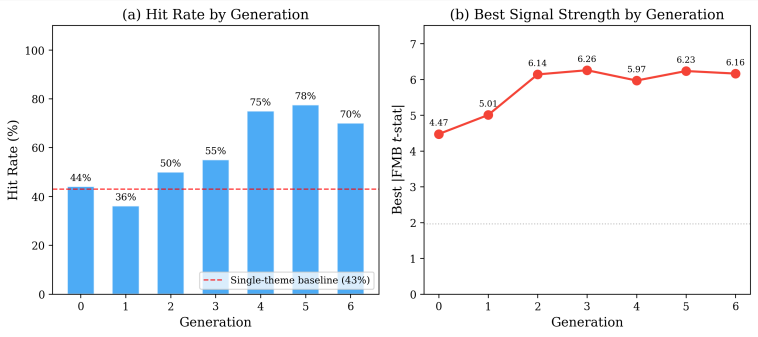
Discovery Results

Aggregate: a 59% hit rate, broad-based

Theme pair	Tested	Sig.	Wrong	Weak	Hit rate	Best $\lvert t \rvert$
Prof × Val	34	25	0	9	**74%**	6.26
Val × Qual	34	24	1	9	71%	5.60
Inv × Fin	34	24	1	9	71%	5.39
Val × Fin	34	22	0	12	65%	6.23
Prof × Inv	34	20	1	13	59%	5.49
Prof × Int	33	17	2	14	52%	6.16
Inv × Acc	34	14	2	18	41%	3.72
Acc × Qual	33	13	1	19	39%	3.07
Total	**270**	**159**	**8**	**103**	**59%**	

- Hit rate **39%–74%** across pairs — broad-based, not concentrated.
- **8 of 270** (3%) wrong-sign — the agent uses these to revise the next round's proposals.
- **10 of 280** (4%) fail validation — the LLM occasionally hallucinates operators outside the 24-op catalog; caught before evaluation.

Generations improve quality



Hit rate: 44% (Gen 0) → 78% (Gen 5). Best $|t|$: 4.47 → 6.26.

The Gen-1 dip is informative: reflection over-corrects before stabilizing. By Gen 4–5 the hit rate roughly doubles and best $|t|$ improves by ∼40%. Stronger *and* more frequent signals.

Does It Survive Modern Validation?

Why post-discovery testing, not train/test?

- Discovery loop = in-sample evolutionary search.
  The agent revises proposals using outcomes computed on the same panel.
- A train/test split would *weaken* the object of study (how the agent reasons and revises).
- Instead: hand the 38 horse-race survivors to the **full modern battery**:
  1. Multiple testing (Harvey, Liu, Zhu 2016; Harvey 2017)
  2. Multi-model factor alphas (FF5, FF6, q-factor)
  3. Spanning vs. **209** published anomalies (Chen & Zimmermann 2022)
  4. Subsample robustness (7 windows)
  5. Permutation tests (1,000 iterations)
- **Goal:** an honest, transparent attrition.

Multiple testing is *not* where the attrition is

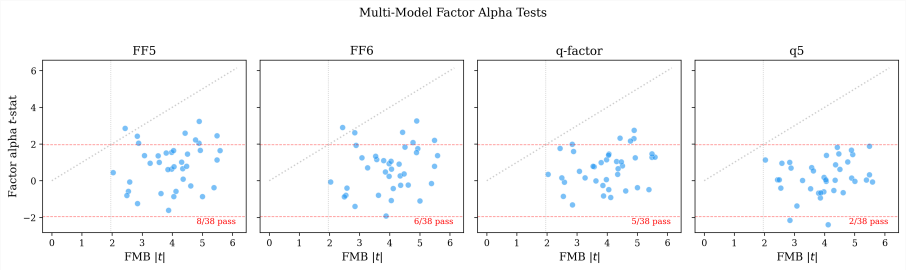
$M = 365$ **unique signals tested.** Counts of horse-race survivors retained:

	$\alpha = 5\%$		$\alpha = 1\%$	
Method	$\lvert t \rvert$ hurdle	Survivors	$\lvert t \rvert$ hurdle	Survivors
Bonferroni (FWER)	3.81	24	4.19	16
Holm (FWER)	3.81	25	4.19	16
BHY (FDR)	2.66	**30**	3.17	**27**

Bayesian (Harvey 2017), $\Pr(H_0 \mid \text{data}) < 5\%$:
skeptical prior **28/38** moderate **30/38** agnostic **37/38**

- **Most survivors clear** multiple-testing hurdles.
- The horse-race survivors are *not* merely an artifact of testing many formulas.
- Attrition lives elsewhere.

Factor absorption *is* the main attrition



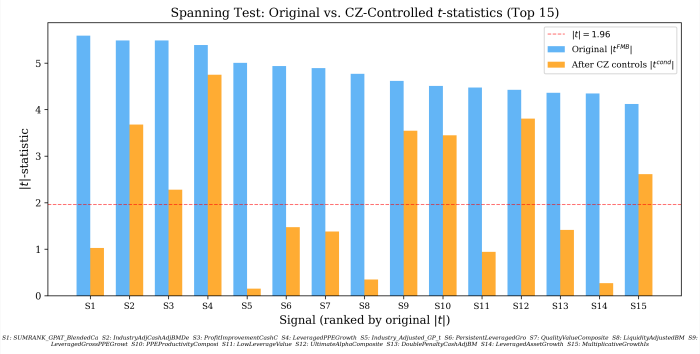
Of 38 survivors retain $|t^{\alpha}| > 1.96$:

- ME-weighted FMB: **13**
- FF5: **8** FF6: **6**
- q-factor: **5** q^5: **2**

Most signals load on profitability- and investment-related premia that factor models already price.

A *characterization* of the landscape, not a failure of the agent: the space of interpretable accounting signals is heavily populated by known effects.

Novelty vs. the 209-anomaly zoo



Conditional $|t|$ after controlling for the 10 most-correlated CZ anomalies: 9/34 retain $|t| > 1.96$; 6/34 retain $|t| > 3.0$.

`LeveragedPPEGrowth`: $t^{cond} = 4.75$ (highest in paper).

`UltimateAlphaComposite`: $t^{cond} = 3.81$, ρ with closest CZ anomaly $= 0.34$.

Survives time and label shuffles

Subsample robustness

7 windows: full, pre/post-1991, pre/post-2000, post-2010, ex-recession.

- Robust in all 7: **12/38**
- Robust in ≥ 6: **22/38**
- Post-2010 (hardest): **16/38**
- Stable / decay / strengthen: **15 / 17 / 6**

Permutation tests

1,000 iterations; shuffle labels, recompute $|t|$.

- Time-series shuffle: **33/38** $p < 0.05$
- Cross-section shuffle: **38/38** $p < 0.05$

Predictability is *not* an artifact of the panel's temporal or cross-sectional structure.

Temporal robustness narrows the credible set; permutation rules out spurious structure.

How the AI Reasons

Case A: a paradigm shift in the agent's reasoning

Investment × Financing pair, Generations 0–3.

- **Gen 0–1**: every successful signal multiplies an investment *flow* (asset growth, capex) by a financing *flow* (debt issuance).
- **Gen 2**: the agent articulates the shift explicitly:

  *"`RATIO(ADD(dltt,dlc),at)` is the leverage ratio—a **STOCK** variable from the financing theme capturing cumulative financing decisions—while `GROWTH(at,1)` captures current asset expansion. Unlike all prior signals that used debt CHANGES (flows), this pairs the debt **LEVEL** with investment growth: highly leveraged firms face higher costs of financial distress, so aggressive investment on an already-strained balance sheet signals managerial recklessness."*

- **Gen 3**: substitutes PP&E growth for asset growth → `LeveragedPPEGrowth`, $t^{FMB} = 5.39$, $t^{cond} = 4.75$, robust in **7/7 subsamples**.

The agent didn't just propose — it *diagnosed* a structural distinction.

Case B: composing orthogonal alpha channels

Same pair, Generation 5–6.

By Gen 5 the agent has two distinct families of successful signals.
Its trace identifies the orthogonality:

> *"The alpha leaderboard reveals two distinct information channels: leverage-**level** conditioning produces the highest FMB t-stats because the **stock** of accumulated leverage is a stable conditioning variable; debt-issuance **flows** paired with investment growth produce the highest alphas because financing flows capture timing decisions orthogonal to standard factors."*

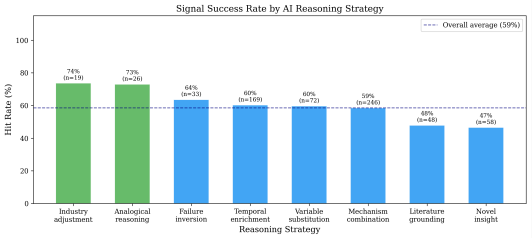
Gen 6 composition: `SUMRANK` of the two channels → `UltimateAlphaComposite`.

	Factor alpha t^{α}			CZ spanning	7/7	permutation
	FF5	FF6	q	t^{cond}	subsamples	$p < 0.05$
`UltimateAlphaComposite`	**2.60**	**2.64**	**2.32**	**3.81**	✓	✓

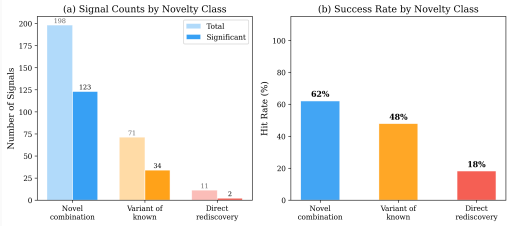
Survives every dimension simultaneously — the only Gen-6 signal to do so.

Where the agent's comparative advantage lies

Reasoning strategy × hit rate



Novelty × hit rate



- Strongest: **industry adjustment** (74%), **analogical reasoning** (73%).
- Weakest: **novel insight from first principles** (47%).
- Novel combinations of known building blocks (71% of all proposals) hit at 62% — vs. 18% for direct rediscoveries.

Bottom line: the agent's edge is *structured recombination*, not novel economic theory.

What LLMs can — and cannot — do (yet)

✔ Can do:

- Search the interpretable-signal space efficiently (44% → 78% hit rate).
- Diagnose wrong-sign results and revise (Case A).
- Identify orthogonal mechanisms and compose them (Case B).
- Produce auditable reasoning across generations.

✖ Cannot do (yet):

- Generate many signals *novel* relative to FF5 / 209 anomalies.
- Escape factor subsumption (**30/38** absorbed by some factor model).
- Reliably theorize from first principles (47% hit rate).
- Replace economic judgment in laboratory design.

AI is the executor of the discovery loop;
humans still design the laboratory.

Thank you.

Code & data: forthcoming on release.