Can AI Do Financial Research?
LLM-Guided Hypothesis Discovery in Asset Pricing*

Huan Liu† Miao Liu‡ Zhizhe Liu§ Danqing Mei¶

April, 2026

Abstract

We study whether an AI research agent can autonomously execute the hypothesis-discovery loop in empirical asset pricing. We place a large language model inside a human-designed research environment comprising a symbolic language of interpretable accounting formulas, an automated validation layer, and a fixed empirical evaluation pipeline. The agent searches over interactions between economic themes such as profitability, investment, valuation, and quality, and updates its proposals over successive generations, where each generation is a new round of proposal, testing, and revision based on standardized empirical feedback. Across eight theme pairs and seven generations, the system proposes and evaluates 280 candidate signals on a microcap-excluded universe; 159 clear a conventional significance screen in the predicted direction, and 38 survive multivariate horse races designed to isolate signals with independent predictive content. We then subject these survivors to a full battery of modern asset pricing tests, including multiple testing corrections, multi model factor spanning, and novelty tests against 209 published anomalies, and identify a small set that carries genuinely incremental information and survives the strictest filters. The paper introduces a transparent architecture for AI-guided hypothesis discovery in finance in which human researchers design the laboratory environment and the AI research agent autonomously carries out the discovery loop.

Keywords: Asset pricing, AI agent, large language models, scientific discovery, hypothesis generation

JEL Classification: G12, G14, C45, C58

*All errors are our own. Code and data will be made publicly available.

†Google: lhuan@google.com

‡Carroll School of Management, Boston College: miao.liu@bc.edu

§Columbia University: zhizhe.liu@columbia.edu

¶Cheung Kong Graduate School of Business: dqmei@ckgsb.edu.cn

1 Introduction

Artificial intelligence is beginning to reshape the research production function. Recent research argues that generative AI can assist with idea generation, model design, coding, and empirical analysis, while newer work studies AI more directly as a mechanism for hypothesis generation and scientific search (Korinek, 2023; Ludwig and Mullainathan, 2024; Manning et al., 2024; Agrawal et al., 2026). In finance, machine learning has long been used to predict returns and estimate asset-pricing objects, and recent studies show that large language models can help process financial text, extract signals, and even generate complete draft finance papers (Gu et al., 2020; Chen et al., 2024; Novy-Marx and Velikov, 2026). Yet one question remains comparatively unexplored: can an *AI research agent*[1] autonomously execute the core discovery loop in empirical asset pricing when hypotheses must be interpretable, economically legible, and testable using standard finance methods?

We study that question in a human-designed scientific environment built around structured accounting data. At the center of the design is a symbolic expression language: a constrained grammar that allows the AI research agent to build candidate signals from 66 Compustat variables and 24 operators, including ratios, differences, growth rates, moving averages, rolling volatilities, trend slopes, and industry adjustments. The agent uses the economic context of a pair of themes, such as profitability and investment or valuation and quality, to generate interpretable accounting formulas as candidate hypotheses that can be read as economic statements about firms. These symbolic hypotheses are then evaluated through a fixed CRSP and Compustat pipeline using portfolio sorts and Fama-MacBeth regressions, and the resulting empirical summaries are returned to the agent for the next round of reflection and revision. The division of labor is deliberate. Human researchers design the laboratory by defining the hypothesis space, selecting the accounting primitives, choosing the theme pairs, fixing the empirical protocols, and specifying the battery of tests against which discovered signals will be evaluated. The AI research agent then executes the inner loop of discovery inside that laboratory by proposing, testing, and refining hypotheses in light of the evidence. This separation ensures that the evaluation standards are not influenced by the discovery process.

Asset pricing provides a demanding and informative laboratory for this exercise. The domain combines a long historical panel (1963–2024), well-established empirical protocols, and a

[1]We use the term *AI research agent* to denote a large language model embedded in a feedback-driven workflow that can propose structured hypotheses, observe empirical outcomes, and revise subsequent proposals.

rich benchmark literature on anomaly discovery, replication, and the statistical challenges of multiple testing (Harvey et al., 2016; Harvey, 2017; Harvey and Liu, 2020; Hou et al., 2020; Chen and Zimmermann, 2022). It is therefore a setting in which proposed hypotheses can be evaluated against well-defined standards rather than through ad hoc case studies. In our design, the AI research agent operates across eight economically motivated cross-theme pairs and seven generations of propose–test–reflect iterations, proposing 280 candidate signals in total. Following the main specification recommended by Hou et al. (2020), all evaluations exclude microcap stocks and use value-weighted portfolios. Of the 280 proposals, 159 clear a conventional significance screen in the predicted direction. After multivariate horse races designed to isolate signals with independent predictive content, 38 survive.

Our baseline evidence suggests that the framework supports systematic and interpretable signal discovery. As the agent progresses through evolutionary generations with iterative reflection, the quality of proposed signals improves meaningfully: the share of proposals that generate significant cross sectional return predictability rises from 44% in Generation 0 to 78% in Generation 5, and the strongest signal becomes materially sharper, with its absolute t statistic increasing from 4.47 to 6.26. Relative to a programmatic benchmark that searches within a single theme, the cross-theme AI-guided approach produces a substantially larger share of significant signals, higher Sharpe ratios, and stronger overall signal quality, consistent with the view that economically motivated interactions across themes are a fruitful source of cross-sectional variation in returns.

The discovery trace also provides evidence on how the search adapts to feedback. Because each proposal is accompanied by an explicit mechanism and later rounds are conditioned on earlier results, we observe systematic revision patterns rather than only final winners. Across generations, the agent increasingly adopts refinements such as industry adjustment, temporal smoothing, and multi-theme recombination, and in several cases it revises mistaken early intuitions into stronger later specifications, suggesting that empirical feedback helps discipline subsequent search.

The main finding, however, is more disciplined than the baseline evidence. Because the agent revises proposals using outcomes computed on the same historical panel, the discovery loop is inherently an in-sample evolutionary search. We therefore place the inferential weight on a post-discovery validation framework rather than on the initial screen itself, and evaluate the discovered signals using multiple testing corrections (Harvey et al., 2016), skeptical Bayesian

updating (Harvey, 2017), false discovery calibration (Harvey and Liu, 2020), factor spanning against leading benchmark models (Hou et al., 2020), conditional tests against 209 published anomalies (Chen et al., 2024), subsample robustness, and permutation tests. The progressive attrition, from 280 proposals to 159 significant to 38 independent to approximately 6–9 that survive the strictest modern statistical filters, is informative. It demonstrates, with transparent methodology, what happens when LLM-discovered signals face the full battery of tests that the asset pricing literature now considers standard. The attrition is large, but it is not total: a small set of signals carry genuinely incremental information.

Next, we shift focus to the AI reasoning process itself. Because the entire discovery trace is observed, including the model's chain-of-thought reasoning preserved across all seven generations, we can study how an AI research agent updates hypotheses in response to empirical feedback. Two case studies illustrate the agent's reasoning. In the first, the agent discovers that conditioning PP&E growth on existing *leverage levels* (a stock variable) produces far stronger predictability than the flow×flow interactions it had been constructing, a conceptual shift from financing flows to balance-sheet fragility that it articulates explicitly in its reasoning trace. In the second, the agent initially hypothesizes that R&D *productivity* (profit per R&D dollar) should predict returns, but after observing a strongly wrong signed result in an early generation, it revises that intuition through reflection, recognizes that R&D *intensity* is the more relevant margin, and restructures the signal accordingly, ultimately producing a composite with strong five factor alpha.

To our knowledge, this is the first study to combine LLM-guided signal discovery with the full battery of modern multiple testing corrections, multi-model factor spanning, cross-anomaly novelty tests, and subsample robustness analysis recommended by the recent methodological literature, while leaving a full record of hypothesis proposals, failures, and revisions. The paper makes four contributions. First, we introduce a finance-native discovery framework that combines a human-designed research architecture with an autonomous AI research agent. Second, we show that this architecture can discover a set of interpretable cross-sectional signals with substantial predictive content, including a smaller set that retains significance under demanding tests. Third, we subject all discoveries to the rigorous testing standards of Harvey et al. (2016), Harvey (2017), Hou et al. (2020), and Chen and Zimmermann (2022), providing an honest accounting of what survives and what does not. Fourth, because the entire discovery and reasoning trace is observed, we provide direct evidence on how an AI research agent forms,

tests, and revises financial hypotheses, including its capacity for diagnostic reasoning, structural insight, and self-correction. The paper therefore speaks not only to the economics of the discovered signals, but also to the organization of research itself.

Our paper relates to several strands of work. It complements the machine-learning asset-pricing literature, which primarily studies black-box prediction or stochastic discount factor estimation (Gu et al., 2020; Freyberger et al., 2020; Chen et al., 2024; Kelly et al., 2025). It is closely related to the emerging literature on automated factor mining and LLM-based financial discovery (Zhang et al., 2020; Cui et al., 2021; Yu et al., 2023; Wang et al., 2025; Shi et al., 2025a; Tang et al., 2025; Shi et al., 2025b; Kou et al., 2025; Weng et al., 2026), but differs in its emphasis on structured accounting variables, theme-conditioned economic mechanisms, auditable reasoning traces, and comprehensive post-discovery empirical validation using state-of-the-art testing methods. It also differs from work that uses LLMs as tools for financial text analysis and investor assistance (Lopez-Lira and Tang, 2023; Dong et al., 2024; Bernard et al., 2026), and from work that automates the packaging of empirical return patterns into full finance papers (Novy-Marx and Velikov, 2026). Our object of study is the discovery loop itself: how a human-designed, AI-executed system generates, evaluates, and revises interpretable financial hypotheses, and what fraction of those hypotheses survives the tests that modern asset pricing demands.

The remainder of the paper proceeds as follows. Section 2 situates the paper in the relevant literatures. Section 3 describes the symbolic language, evaluation pipeline, and discovery loop. Section 4 outlines the experimental design. Section 5 reports the discovery results. Section 6 subjects the discoveries to the rigorous testing framework. Section 7 analyzes the reasoning and revision traces through case studies. Section 8 discusses interpretation, limitations, and implications for AI-assisted research in finance.

2 Related Literature

2.1 AI, hypothesis generation, and scientific discovery

A growing literature studies how artificial intelligence changes the process of scientific inquiry itself. Korinek (2023) surveys how generative AI can assist economists across idea generation, modeling, coding, and exposition. Ludwig and Mullainathan (2024) show how machine learning can be used as a disciplined tool for hypothesis generation. Manning et al. (2024) automate

the generation and testing of social-scientific hypotheses using structural causal models and LLM-based agents. At a broader level, Agrawal et al. (2026) conceptualize AI as expanding search over large combinatorial spaces within a multi-stage model of scientific productivity, and Wang et al. (2023) review the emerging role of AI in scientific discovery across disciplines.

Natural-science examples such as AlphaFold and AI Feynman illustrate how AI can solve or rediscover highly structured scientific objects (Jumper et al., 2021; Udrescu and Tegmark, 2020). Our setting differs in two respects. First, empirical asset pricing is an observational social-science environment in which hypotheses are evaluated with noisy historical data rather than laboratory or simulated data. Second, the object to be discovered is an interpretable accounting-based signal, not a protein structure or a physical law. Our contribution to this literature is to study an AI research agent in a domain with real financial data, standardized empirical tests, a transparent hypothesis language, and a comprehensive post-discovery validation framework that disciplines the inferential claims.

2.2 Machine learning and AI in asset pricing

A substantial literature applies machine learning to return prediction and asset pricing. Gu et al. (2020) show that nonlinear machine-learning methods improve empirical asset-pricing performance relative to traditional linear models. Freyberger et al. (2020) use flexible non-parametric methods to study nonlinearities in the cross section of returns. Chen et al. (2024) develop deep-learning methods for the stochastic discount factor, and Kelly et al. (2025) introduce transformer-based asset-pricing models that embed cross-asset context directly into the pricing kernel.

Our paper is complementary to this literature but differs in both its unit of analysis and its objective. The black-box machine-learning asset-pricing literature estimates high-dimensional mappings from firm characteristics to returns or prices. By contrast, our AI research agent proposes explicit symbolic hypotheses that are interpretable ex ante as accounting statements about profitability, valuation, financing, investment, or intangibles. The goal is not to maximize predictive fit subject to model complexity, but to generate economically meaningful candidate signals, evaluate them with standard asset-pricing tests, and observe how the agent revises its ideas in response to the evidence. This distinction matters because it shifts the contribution from improved prediction to transparent hypothesis production.

2.3 Automated factor discovery and formulaic alpha mining

Within quantitative finance, a separate literature studies automated discovery of formulaic alpha factors. Earlier work uses evolutionary search or reinforcement learning to navigate large spaces of symbolic formulas. Examples include Zhang et al. (2020), Cui et al. (2021), and Yu et al. (2023), which show that algorithmic search can discover profitable and weakly correlated formulaic alphas.

More recent work explicitly places LLMs and agents inside the alpha-mining loop. Wang et al. (2025) introduce human–AI interactive alpha mining. Shi et al. (2025a) propose a two-stage framework that mines and dynamically combines formulaic alphas. Tang et al. (2025) use multiple LLM agents and regularized exploration to search for decay-resistant factors. Shi et al. (2025b) integrate LLMs with Monte Carlo tree search to refine symbolic formulas using backtesting feedback. Kou et al. (2025) broaden the setting to multi-agent strategy construction using multimodal data, and Weng et al. (2026) make market logic itself an explicit object of generation and refinement.

Our paper is closest to this literature, but the differences are central to our contribution. First, our hypothesis language is built from structured accounting primitives (balance sheet and income statement items) rather than generic technical indicators or broad alpha libraries. Second, the search is conditioned on explicit economic theme pairs, which makes the AI research agent generate hypotheses about interactions such as profitability and investment or valuation and quality, rather than search the entire formula space without economic structure. Third, the evaluation uses canonical cross-sectional asset-pricing tools rather than focusing only on rank correlations or trading metrics. Fourth, because every proposal contains a symbolic expression, a directional hypothesis, and a written economic mechanism, we can analyze the trace of reasoning and revision itself. In this sense, our paper studies not only whether automated factor search can work, but also how an AI research agent learns within a finance-native discovery environment.

Most importantly, our paper goes substantially beyond the existing automated alpha-mining literature in post-discovery validation. Where prior work typically reports in-sample performance metrics (information coefficients, Sharpe ratios, turnover), we apply the full suite of multiple testing corrections (Harvey et al., 2016), Bayesian inference (Harvey, 2017), multi-model factor spanning (Fama–French five-factor, six-factor, and Hou–Xue–Zhang q-factor models), cross-anomaly novelty tests against the 209-signal Chen and Zimmermann (2022) open-source

library, subsample robustness analysis across seven time periods, and non-parametric permutation tests. This validation framework transforms the exercise from "can AI find signals that backtest well?" to "can AI find signals that survive the tests the academic literature now considers standard?"

2.4 AI-generated finance scholarship and research production

Novy-Marx and Velikov (2026) show that LLMs can generate complete academic finance papers once empirically promising return predictors have been identified. Their pipeline mines a very large set of signals from accounting data, filters them using the Novy-Marx and Velikov (2024) *Assaying Anomalies* protocol, and then uses LLMs to generate multiple full papers with alternative theoretical narratives and citations. That contribution is important because it demonstrates how AI can scale the production of finance research outputs and, at the same time, clarifies how AI can industrialize post hoc storytelling.

Our paper studies an earlier and different stage of the research pipeline. The object in our setting is not the finished paper but the discovery loop that precedes it. The AI research agent begins with an economically structured prompt, proposes a hypothesis before the empirical result is known for that proposal, observes the resulting test statistics, and revises its next set of ideas. This sequencing makes it possible to study whether an AI system can participate in the generation, testing, and refinement of interpretable hypotheses.

2.5 Multiple testing, anomaly replication, and statistical inference

Our paper speaks directly to the long-running literature on the anomaly zoo and its statistical interpretation. Harvey et al. (2016) argue that the large number of proposed return predictors requires explicit adjustment for multiple testing and propose t-statistic hurdles well above the conventional 1.96. Harvey (2017) introduces Bayesian minimum Bayes factors and Bayesianized p-values to provide a more interpretable assessment of statistical significance under varying prior beliefs. Harvey and Liu (2020) develop a double-bootstrap procedure that jointly calibrates Type I and Type II error rates using the actual correlation structure among tested signals. Hou et al. (2020) replicate 452 published anomalies and document that 65% fail to hold up under standardized procedures (NYSE breakpoints, value-weighted returns, micro-cap exclusion), with the failure rate rising to 82% under multiple testing adjustments. Chen and Zimmermann (2022) build an open-source library of cross-sectional predictors that provides

the infrastructure for large-scale replication and novelty testing. McLean and Pontiff (2016) show that anomaly returns decline after publication, a pattern we confirm in our setting, where 17 of 38 horse-race survivors exhibit statistically significant decay trends in rolling 60-month windows. Harvey et al. (2026) argue that dependence across tests implies materially higher significance thresholds than the conventional rule.

This literature provides both the motivation and the methodology for our testing framework. We adopt the Hou et al. (2020) main specification (microcap exclusion, NYSE breakpoints, value-weighted portfolios) as the baseline for all evaluations. We apply the Harvey et al. (2016) multiple testing corrections (Bonferroni, Holm, BHY) and the Harvey (2017) Bayesian framework to the full set of tested signals. We use the Chen and Zimmermann (2022) anomaly library as the benchmark for spanning tests, and we evaluate subsample robustness across time periods that include the particularly demanding post-2010 window. Rather than treating significance as the final word, we use these tools to produce an honest, multi-dimensional assessment of which LLM-discovered signals survive modern empirical standards, and which do not.

Viewed through this lens, our contribution to the anomaly literature is twofold. First, we provide a case study of what happens when a new source of signal proposals (an AI research agent) is subjected to the full battery of tests that the literature has developed precisely to discipline such proposals. Second, we demonstrate that the honest attrition narrative, 280 proposals narrowing to approximately 6–9 survivors under the strictest filters, is itself informative about the boundaries of AI-guided discovery in finance.

3 Framework: The Discovery Engine

This section describes the discovery architecture. The central design choice is to separate the scientific environment from the search process itself. Human researchers specify the hypothesis language, the admissible data, the validation rules, and the empirical evaluation pipeline. The AI research agent operates inside that environment: it proposes accounting-based hypotheses in symbolic form, observes the empirical outcomes of those hypotheses, and updates subsequent proposals in light of the evidence. This division of labor preserves interpretability while allowing the agent to search a large space of economically structured signals.

Figure 1 summarizes this architecture by showing the human-designed laboratory, the autonomous AI research agent loop, and the outputs generated by the system; we discuss each

component in detail in the subsections that follow. We organize the framework around four components. First, we define a structured hypothesis space for accounting-based signals. Second, we map every valid expression into a fixed empirical evaluation pipeline. Third, we embed the agent in a feedback-driven propose–test–revise loop. Fourth, we direct the search toward economically motivated interactions across themes. The resulting system is not merely a software workflow. It is a scientific architecture: human researchers design the laboratory, and the AI research agent autonomously executes the discovery loop within it.

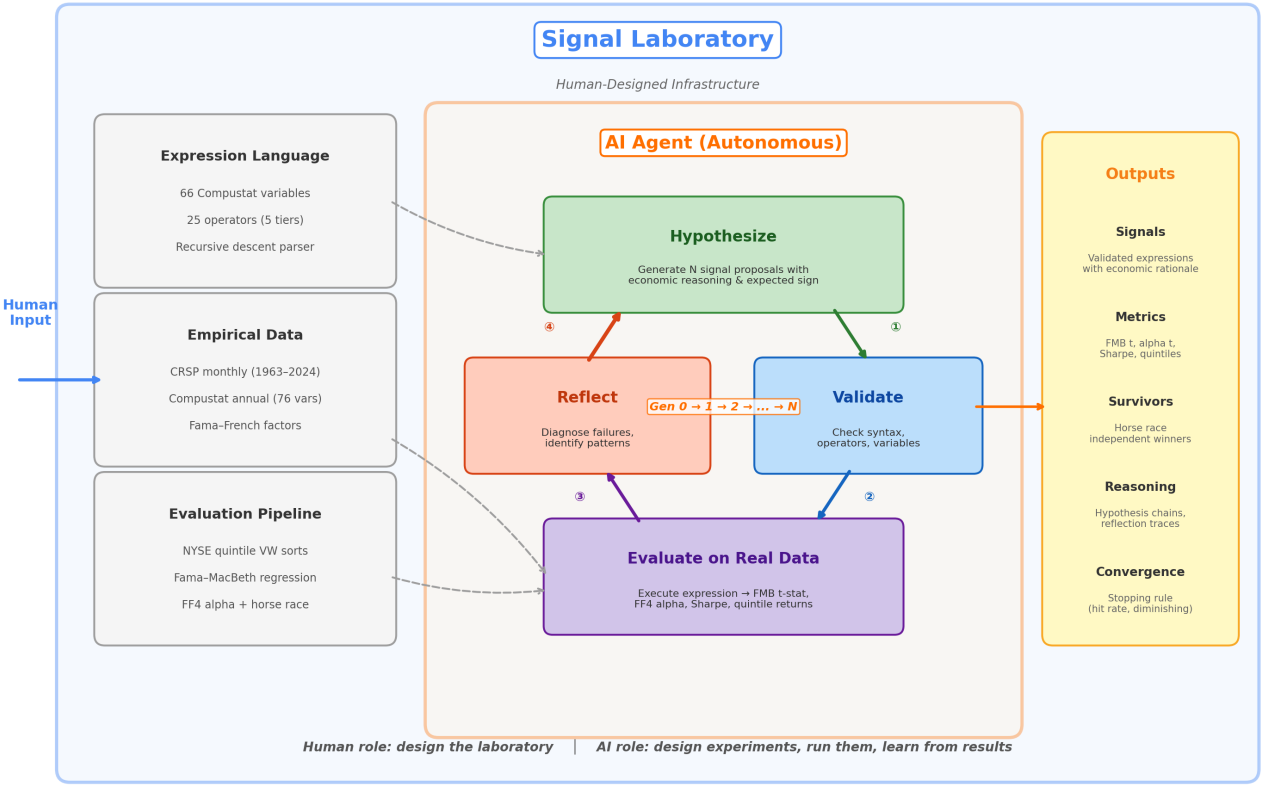

Figure 1: Finance laboratory for AI-guided hypothesis discovery. Human researchers design the laboratory environment—the expression language, empirical data inputs, and evaluation pipeline—while the autonomous AI research agent operates within that environment. In each generation, the agent iterates through four steps: it hypothesizes candidate signals with economic rationale, validates them against the symbolic grammar, evaluates valid expressions on real data using the fixed empirical pipeline, and reflects on the resulting successes and failures before proposing the next round. The framework produces not only validated signals and their performance metrics, but also horse-race survivors, reasoning traces, and convergence diagnostics, making the discovery process itself transparent and auditable.

3.1 Expression Language

The symbolic language is the key interface between accounting structure and AI-guided discovery. It is deliberately narrower than a general-purpose programming language. The language comprises 66 Compustat annual variables, 24 composable operators, and a recursive descent parser that compiles symbolic expressions into executable firm×year panels. It is therefore a structured hypothesis space: rich enough to support genuine discovery, but constrained enough to keep the search economically meaningful and empirically tractable.

Variables. The language includes 66 Compustat annual items spanning the balance sheet (e.g., `at`, `ceq`, `ppent`, `che`, `invt`, `intan`, `gdwl`), income statement (e.g., `revt`, `cogs`, `ib`, `sale`, `xsga`, `xrd`, `dp`), cash flow statement (e.g., `oancf`, `capx`, `dpc`), and financing items (e.g., `dltt`, `dlc`, `sstk`, `dltis`, `prstkc`). Each variable is annotated with its financial statement, a plain-English description, and common scaling denominators. These annotations matter because they give the AI research agent a finance-native vocabulary rather than a raw set of column names.

Operators. The language provides 24 operators with explicit arity and parameters. Operators can be grouped conceptually into five tiers of complexity:

1. *Basic algebra and scaling*: `ADD`, `SUB`, `MUL`, `RATIO`, `SCALE`;
2. *Composite formation*: `SUMRANK`—computes within-year cross-sectional ranks and sums them to form composites;
3. *Temporal transforms*: `GROWTH`, `DELTA`, `LAG`, `MA` (moving average), `VOL` (rolling volatility), `TREND` (linear slope), `ACCEL` (second difference);
4. *Cross-sectional and industry-relative transforms*: `IND_ADJ` (industry-median adjustment), `ZSCORE`, `PERCENTILE`, `IND_ZSCORE`;
5. *Nonlinear and robustness transforms*: `LOG`, `ABS`, `NEG`, `INV`, `INDICATOR`, `SIGN`, `WINSOR`.

Operators compose recursively up to a maximum nesting depth of 6. For instance, the expression

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

computes a composite of industry-adjusted gross profitability and negated asset growth, a "cash cow" signal that combines the profitability and investment themes.

Design rationale. The symbolic language is a research-design choice, not merely a computational convenience. It is deliberately narrower than unrestricted Python, SQL, or free-form statistical modeling. Without such scope restrictions, the AI research agent would face an effectively unbounded search space: it could propose arbitrary code, unstable data manipulations, or mathematically admissible expressions that are economically meaningless. In that setting, a large share of the computational budget would be spent on syntax errors, data-pipeline bugs, and open-ended debugging rather than on hypothesis discovery.

The structured language imposes just enough discipline to eliminate this unproductive search while preserving genuine flexibility. Because operators can be nested recursively up to depth 6, the admissible space still contains millions of candidate signals. At the same time, every proposal is guaranteed to be parseable, executable, and economically interpretable once it passes validation. The guiding principle is that appropriately chosen constraints enable, rather than inhibit, discovery. As in other scientific AI systems, disciplined constraints make creative search tractable by ruling out malformed outputs without dictating the answer (Jumper et al., 2021).[2]

In this sense, the human role is not to handcraft the final signals or steer the search proposal by proposal. It is to define the laboratory: what constitutes a well-formed hypothesis, which accounting primitives are available, and how empirical performance is measured. The AI research agent then searches autonomously within that laboratory.

3.2 Evaluation Pipeline

The evaluation pipeline is fully standardized across all proposals. Once a valid symbolic expression is submitted, the downstream mapping from annual accounting information to monthly portfolios and cross-sectional regressions is fixed. This standardization is important for interpretation: differences in signal performance reflect differences in the proposed accounting expressions rather than changes in portfolio construction or test design. Following the main

[2]This logic parallels AlphaFold. Rather than allowing unrestricted coordinate search, the model operates within a structured space shaped by geometric and physicochemical constraints, so that search is concentrated on biologically plausible protein structures. Our expression language is the financial analogue of those constraints: it restricts search to parseable, executable, and economically interpretable hypotheses without dictating which signal the agent will discover.

specification of Hou et al. (2020), all evaluations exclude microcap stocks, defined as stocks below the 10th percentile of NYSE market equity in a given month, and use NYSE-only breakpoints with value-weighted portfolios.

Given a signal expression, the evaluation pipeline proceeds in five stages.

Stage 1: Signal computation. A recursive descent parser evaluates the expression against Compustat annual data merged with CRSP via the CCM link table. The output is a firm×year panel of signal values. Temporal operators (e.g., `GROWTH`, `LAG`) compute within-firm shifts using pre-computed lagged columns, ensuring that the resulting signal uses only information available at the time of portfolio formation.

Stage 2: Monthly panel construction. The firm-year signal is mapped to monthly frequency using the Fama–French annual sorting convention. Accounting data from fiscal year $t-1$ become available in July of year t and are held constant through June of year $t+1$. Specifically, each CRSP month is assigned a "formation year": months July–December map to the current calendar year, and months January–June map to the previous calendar year. The signal is then merged to monthly CRSP by `(permno, formation year)`, producing a panel in which each firm's signal is constant over the July-to-June holding period. This convention ensures a minimum six-month gap between fiscal-year-end and portfolio formation, preventing look-ahead bias in the use of accounting data.

Stage 3: Portfolio formation. Each month, stocks are sorted into quintile portfolios based on the signal. Breakpoints are computed using NYSE-listed stocks only, following Fama and French (1993). All eligible stocks (NYSE, NASDAQ, and AMEX) are then assigned to portfolios using those breakpoints. Stocks below the 10th NYSE market-equity percentile are excluded before breakpoint computation and portfolio assignment, following the microcap exclusion protocol of Hou et al. (2020). This convention prevents the proliferation of small, illiquid stocks from distorting breakpoints and inflating return spreads.

Stage 4: Return computation. Value-weighted portfolio returns are computed using market capitalizations lagged by one month:

$$R_{p,t} = \sum_{i \in p} \frac{\text{ME}_{i,t-1}}{\sum_{j \in p} \text{ME}_{j,t-1}} \times R_{i,t}, \tag{2}$$

where portfolio assignments are also based on month $t-1$ information. A long–short portfolio is constructed as Q5 minus Q1 (high signal minus low signal).

Stage 5: Statistical evaluation. We assess each signal using two complementary tests:

1. *Time-series factor regression.* We regress the long–short portfolio's excess returns on the Carhart (1997) four factors:

$$R_t^{LS} - R_t^f = \alpha + \beta_{\text{MKT}}\text{MktRF}_t + \beta_{\text{SMB}}\text{SMB}_t + \beta_{\text{HML}}\text{HML}_t + \beta_{\text{UMD}}\text{UMD}_t + \varepsilon_t. \tag{3}$$

   We report the annualized alpha, its t-statistic, and the annualized Sharpe ratio.

2. *Fama–MacBeth cross-sectional regression.* Each month t, we run a cross-sectional regression of individual stock returns on the lagged, cross-sectionally ranked signal:

$$R_{i,t} = a_t + b_t \times \text{Signal}_{i,t-1} + \varepsilon_{i,t}. \tag{4}$$

   The time-series average of the slope coefficients $\bar{b}$ and its Newey–West t-statistic (6 lags) provide a measure of cross-sectional return predictability that is robust to heteroskedasticity and autocorrelation.[3]

For purposes of the discovery loop, a proposal is classified as *successful* if its Fama–MacBeth $|t| > 1.96$ with the predicted sign. We use this rule as a standardized screening criterion that can be fed back to the agent in later rounds. As we discuss in Section 6, this threshold serves as an initial screen; the rigorous testing framework applies substantially more demanding criteria.

3.3 The AI Research Agent Discovery Loop

The discovery loop is feedback-driven but constrained. Once the theme pair, symbolic language, and evaluation protocol are fixed, the AI research agent proposes signals, receives structured summaries of their empirical outcomes, and revises subsequent proposals without further human intervention. Each iteration of this propose–test–reflect cycle is a *generation*.

[3]We use 6 Newey–West lags throughout, following standard practice in monthly cross-sectional studies (Petersen, 2008). This correction is applied consistently across all evaluations, including the main discovery loop, factor model tests, spanning regressions, and subsample analyses.

Generation 0: Initial proposals. The AI research agent receives a structured prompt containing:

- the full variable catalog (66 items with descriptions and common scalers);
- the operator library (24 operators with arity, parameters, and examples);
- an economic context describing the two themes being combined, their interaction mechanism, and example interaction ideas in plain English; and
- baseline results from single-theme testing (the top signals per theme from a prior programmatic search), which establish a benchmark for the cross-theme exercise.

The agent then proposes 5 signals. Each proposal consists of a name, a symbolic expression, a one-sentence directional hypothesis, and a short explanation of why the interaction is expected to predict returns beyond what either theme captures alone.

Validation and evaluation. Each proposed expression undergoes mechanical validation before any empirical testing occurs. The validation screen checks: (i) balanced parentheses, (ii) recognized operators, (iii) variables contained in the catalog, (iv) correct operator arity, and (v) nesting depth no greater than 6. Invalid expressions are discarded. Valid signals are then executed through the fixed evaluation pipeline described in Section 3.2.

Generation $k \geq 1$: Reflection and refinement. In later generations, the AI research agent receives an updated prompt containing all prior outcomes, organized into four categories: successful signals (with Fama–MacBeth t-statistic, alpha t-statistic, Sharpe ratio, and quintile returns), wrong-sign signals, weak signals ($|t| \leq 1.96$), and errors. The agent is asked to perform three tasks: (i) identify which structural features distinguish stronger from weaker proposals, (ii) reflect on the economic mechanisms suggested by the observed results, and (iii) propose 5 new signals informed by that analysis. The prompt explicitly requires at least one proposal that attempts to improve on a prior failure and at least one that explores a new mechanism.

Extended thinking and reasoning continuity. The AI research agent (Claude Opus 4.6) operates with extended thinking enabled, allocating a budget of 10,000 tokens per call for internal chain-of-thought reasoning before producing its visible response. Critically, the agent's encrypted reasoning chain is preserved across generations via the API multi-turn mechanism:

each generation's response, including the thinking blocks with their cryptographic signatures, is passed back as part of the conversation history for the next generation. This means the agent can build on the full depth of its prior reasoning, not merely on the summary statistics visible in the prompt. The thinking traces are also stored for post-hoc analysis (Section 7).

Convergence. The loop continues until one of three stopping criteria is met: (i) the hit rate falls below 20% for two consecutive generations, (ii) the best |FMB t| in the latest generation falls below 50% of Generation 0's best, or (iii) a maximum of 7 generations is reached. In practice, none of our eight theme pairs triggered the first two criteria within 7 generations.

3.4 Cross-Theme Interaction Design

A distinctive feature of our framework is that the AI research agent searches over interactions between economic themes rather than within a single theme. We begin with eight broad themes (profitability, valuation, investment, financing, accruals, quality, intangibles, and distress), which yield 28 possible pairs. We study eight priority pairs, selected *ex ante* based on prior finance research suggesting that the return implications of one dimension often depend on another and that both can be measured cleanly with annual accounting data. This design keeps the search economically structured: rather than mining arbitrary formulas, the agent explores settings in which the literature already suggests a plausible interaction mechanism.

Theme pairs. Table 1 lists the eight cross-theme pairs, their economic mechanisms, and the number of generations. We briefly describe each:

1. **Profitability × Investment.** Prior work identifies profitability and investment as two central dimensions of the cross section of expected returns. Gross profitability predicts returns, while aggressive asset growth and investment are associated with lower subsequent returns; both the Fama–French five-factor model and the Hou–Xue–Zhang framework place these variables at the center of expected-return variation (Novy-Marx, 2013; Fama and French, 2015; Hou et al., 2015; Cooper et al., 2008). This pair therefore targets the idea that profitability is most informative when it is not accompanied by aggressive reinvestment. Firms with strong operating profitability and restrained investment resemble "cash cows," whereas high profitability combined with rapid expansion may reflect empire building or weaker future returns.

2. **Valuation × Quality.** The value literature suggests that cheap stocks earn high average returns, but prior work also shows that value is much more informative when conditioned on fundamentals. Piotroski (2000) shows that fundamental strength helps separate winners from losers among high book-to-market firms, Piotroski and So (2012) emphasize expectation errors within value and glamour strategies, and Asness et al. (2019) formalize the broader interaction between quality and cheapness. This pair therefore targets firms that are inexpensive relative to fundamentals yet not low quality in profitability, safety, or earnings quality, distinguishing genuine value opportunities from distress-driven "value traps."

3. **Investment × Financing.** A large literature links aggressive expansion and external financing to weak subsequent returns. Capital investment is less favorably received when firms appear to overinvest, and both debt and equity issuance have been shown to predict low future returns (Titman et al., 2004; Richardson, 2006; Cooper et al., 2008; Pontiff and Woodgate, 2008). This pair is therefore motivated by the idea that financing sharpens the interpretation of investment: growth financed externally is more likely to reflect weak capital discipline, managerial overreach, or temporarily overpriced securities than growth financed internally.

4. **Profitability × Intangibles.** Intangible investment creates a natural measurement problem in accounting-based asset pricing because R&D and related expenditures are largely expensed rather than capitalized. As a result, conventional profitability measures can understate the economics of innovative firms and mismeasure the capital base against which profitability should be scaled (Lev and Sougiannis, 1996; Eisfeldt and Papanikolaou, 2013; Peters and Taylor, 2017). This pair is therefore designed to search for profitability signals that become more informative once conditioned on intangible intensity, R&D effort, or other proxies for unbooked capital.

5. **Profitability × Valuation.** This pair sits at the intersection of the profitability and value premia. Profitability predicts returns even controlling for traditional value measures, while value is more compelling when low multiples do not simply reflect poor underlying economics (Novy-Marx, 2013; Fama and French, 2015; Piotroski, 2000; Asness et al., 2019). The interaction therefore targets firms that are simultaneously cheap and fundamentally strong. In economic terms, profitability helps distinguish undervalued firms from low-

multiple firms with weak prospects, while valuation helps identify profitable firms whose strength is not yet fully capitalized into prices.

6. **Accruals × Quality.** The accrual anomaly shows that the market tends to overweight the accrual component of earnings relative to the cash-flow component (Sloan, 1996). A related literature on earnings quality argues that accrual reliability, cash-flow backing, and earnings persistence vary systematically across firms, with lower-quality earnings being less informative about future performance (Dechow and Dichev, 2002; Richardson et al., 2005). This pair therefore targets firms where high accruals coincide with weak earnings quality, a setting in which reported performance may be especially prone to mispricing.

7. **Valuation × Financing.** Financing behavior is often especially informative within the value universe. Piotroski (2000) shows that leverage, liquidity, and equity issuance help separate strong from weak value firms, while Pontiff and Woodgate (2008) document that issuance is itself negatively related to future returns. This pair therefore targets the idea that cheap stocks that are simultaneously reducing reliance on external finance may reflect genuine undervaluation and managerial discipline, whereas cheap firms that continue to issue debt or equity may be closer to classic value traps.

8. **Investment × Accruals.** The investment and accruals literatures are closely related because both concern the composition and quality of balance-sheet growth. Asset growth predicts low future returns, and part of that growth may come through working-capital accruals rather than productive capacity expansion (Cooper et al., 2008; Sloan, 1996; Fairfield et al., 2003; Richardson, 2006). This pair therefore asks whether the market distinguishes sufficiently between "real" investment and lower-quality expansion concentrated in receivables, inventories, and other accrual components. In that sense, accruals act as a conditioning variable for the quality of investment growth.

Horse race. After the evolutionary loop, all significant signals within a theme pair enter a multivariate horse race. Cross-theme proposals are often nearby variants of the same underlying mechanism, so our goal is to isolate signals with incremental predictive content rather than count every successful variant as a separate discovery. We first deduplicate signals with pairwise cross-sectional rank correlations above 0.95, retaining the version with the larger univariate $|t|$-statistic. We then run simultaneous Fama–MacBeth regressions and apply backward stepwise

elimination until all remaining signals have conditional $|t| > 1.96$. The microcap exclusion is applied to both signals and returns in the horse race, consistent with the main evaluation specification. The surviving signals are therefore the pair-level mechanisms that remain informative after accounting for redundancy across related discoveries.

4 Experimental Design

This section describes the empirical design used to evaluate the discovery framework. The core elements of the design are fixed ex ante: the data sources, sample filters, portfolio construction rules, regression procedures, theme pairs, generation limits, and the outcome summaries fed back to the AI research agent. We hold this environment constant across all proposed signals so that variation in performance can be attributed to variation in the hypotheses themselves rather than to changes in implementation.

4.1 Data

We use monthly stock returns from the Center for Research in Security Prices (CRSP) and annual accounting data from Compustat, merged through the CRSP–Compustat Merged (CCM) link table. The sample spans July 1963 to December 2024. These sources define the fixed empirical environment in which all candidate signals are evaluated.

CRSP. We include all common stocks (share codes 10 and 11) listed on NYSE, NASDAQ, and AMEX. Monthly returns incorporate delisting returns following Shumway (1997): performance-related delistings (codes 500–599) with missing delisting returns are assigned a return of -30%. The resulting panel contains approximately 3.5 million firm-month observations across roughly 26,000 unique securities.

Compustat. We extract 66 annual accounting items spanning the balance sheet (33 items), income statement (16 items), and cash flow and financing activities (17 items). Key items include total assets (`at`), book equity (`ceq`), revenue (`revt`), cost of goods sold (`cogs`), operating cash flow (`oancf`), capital expenditures (`capx`), R&D spending (`xrd`), long-term debt (`dltt`), and equity issuance (`sstk`), among others. Observations are matched to CRSP through the CCM link table using standard primary-link filters (linktype LC/LU, linkprim P/C), with link

dates respected. Each firm-year observation is assigned to the fiscal year in which the reporting period ends.

Microcap exclusion. Following the main specification of Hou et al. (2020), we exclude microcap stocks from all signal evaluations. Microcap stocks represent approximately 60% of the stock count but only about 3% of aggregate market capitalization. As shown by Hou et al. (2020), they have the highest cross-sectional return dispersion and disproportionately influence equal-weighted and all-exchange-breakpoint results. Excluding them yields conservative estimates that are more representative of the investable universe. This exclusion is applied to the signal panel, the return panel, and the market-equity weights simultaneously, ensuring that breakpoints, portfolio assignments, and regressions all operate on the same universe.

Financial firm exclusion. For five of the eight theme pairs whose signals primarily involve operating cash-flow or investment variables (Profitability×Investment, Investment×Financing, Profitability×Intangibles, Profitability×Valuation, and Investment×Accruals), we exclude financial firms (SIC codes 6000–6999) from signal computation, following standard practice for accounting-ratio analyses. Financial firms are retained for the remaining three pairs (Valuation×Quality, Accruals×Quality, Valuation×Financing), where the economic mechanisms center on valuation multiples and accrual quality for which financial firms are economically valid subjects. In all cases, financial firms remain in the CRSP return and market-equity panels used for factor model estimation.

Factor data. Monthly Fama–French factors (Mkt-RF, SMB, HML, RMW, CMA) and the momentum factor (Mom) are obtained from Kenneth French's data library. The Hou–Xue–Zhang q-factor (Mkt, ME, I/A, ROE) and q^5 (adding expected growth) monthly returns are obtained from `global-q.org`. For the rigorous testing framework (Section 6), we use the Fama–French five-factor and six-factor models alongside the q-factor and q^5 models. The Chen–Zimmermann open-source cross-sectional predictor library, containing 209 firm-level characteristic signals, is obtained from `openassetpricing.com` (Chen and Zimmermann, 2022).

4.2 Evaluation Metrics

We evaluate each signal using both cross-sectional and portfolio-based tests. The Fama–MacBeth slope t-statistic captures whether the signal forecasts cross-sectional return variation,

while factor-adjusted alpha and the Sharpe ratio summarize the economic magnitude of the associated long–short portfolio.

Each signal is summarized using four metrics:

1. **FMB t-statistic**: the Fama–MacBeth cross-sectional regression t-statistic, computed with Newey–West standard errors (6 lags). This is the primary measure of cross-sectional return predictability.

2. **Alpha t-statistic**: the t-statistic for the Carhart four-factor alpha of the long–short (Q5−Q1) portfolio. This measures return after controlling for market, size, value, and momentum exposures.

3. **Sharpe ratio**: the annualized Sharpe ratio of the long–short portfolio. This measures risk-adjusted return per unit of volatility.

4. **Coverage**: the fraction of firm-month observations for which the signal is non-missing. This guards against proposals that appear successful only because they apply to a narrow subset of firms.

A proposal is classified as *successful* if its Fama–MacBeth $|t| > 1.96$ and the estimated sign matches the ex ante directional hypothesis. We use this rule as a screening criterion for the discovery loop: it determines how outcomes are categorized and summarized for later generations. We interpret it as an initial discovery screen rather than as a complete inferential standard. The rigorous testing framework in Section 6 applies substantially more demanding criteria, including multiple testing corrections, multi-model factor alphas, spanning tests against 209 published anomalies, subsample robustness analysis, and permutation tests.

4.3 Single-Theme Programmatic Baseline

To provide a benchmark for the cross-theme AI-guided search, we construct a set of 99 single-theme signals programmatically without any LLM involvement. These signals are generated by enumerating canonical accounting ratios, growth rates, and composites within each of the eight individual themes (profitability, valuation, investment, accruals, quality, financing, intangibles, and distress). Each signal is evaluated through the same pipeline used for the cross-theme proposals: microcap-excluded universe, NYSE quintile breakpoints, value-weighted portfolios, Fama–MacBeth regressions with Newey–West standard errors. The single-theme benchmark

therefore provides a controlled comparison: it uses the same data, the same evaluation protocol, and the same accounting variable catalog, but searches within a single theme rather than across theme interactions and does not benefit from LLM-guided reflection or iterative refinement.

4.4 AI Research Agent Configuration

We use a single frontier language model throughout the main experiments in order to keep the discovery process internally consistent across theme pairs and generations. The model is Anthropic's Claude Opus 4.6 (`claude-opus-4-6`), accessed through the Anthropic API with extended thinking enabled. No fine-tuning, task-specific few-shot examples, or retrieval-augmented generation is employed. The agent operates with general pretrained financial knowledge together with the domain-specific context encoded in each prompt.

Importantly, the AI research agent has no direct access to the raw CRSP/Compustat panel and does not alter the empirical protocol. It proposes symbolic expressions; all validation, data construction, portfolio formation, and statistical evaluation are carried out externally by the fixed pipeline. This separation limits the channel through which in-sample information enters the search process: the agent revises hypotheses based on standardized outcome summaries—including categorized results, t-statistics, Sharpe ratios, and quintile return spreads, rather than direct access to the underlying return panel. This design means that the discovery process constitutes an in-sample evolutionary search: the agent sees results computed on the same dataset it is implicitly optimizing over. We do not claim out-of-sample validation from the discovery loop itself. Instead, we rely on the rigorous post-discovery testing framework (Section 6) to discipline the inferential claims.

5 Discovery Results

This section reports the discovery results in five steps. We first summarize aggregate success across all theme pairs and generations. We then document how proposal quality evolves across generations. Third, we isolate the subset of signals that survives redundancy reduction in multivariate horse races. Fourth, we examine the economically strongest signals. Finally, we compare the cross-theme discoveries with the single-theme baseline. This structure separates the volume of successful proposals from the smaller set of discoveries that appear to contain distinct information. Section 6 then subjects the survivors to a rigorous battery of post-discovery

tests.

5.1 Aggregate Performance

We begin with the broadest question: does the discovery architecture generate a substantial number of successful proposals? Table 2 summarizes the results across all eight theme pairs.

Across 280 proposals evaluated on the microcap-excluded universe, 270 produce valid (non-error) results and 159 clear the conventional significance screen ($|t^{FMB}| > 1.96$ with Newey–West standard errors, 6 lags) in the predicted direction, yielding a 59% hit rate. Eight proposals are statistically significant but in the wrong direction, 103 are weak ($|t| \leq 1.96$), and 10 fail due to expression execution errors (e.g., the LLM used an operator not in the catalog).

Two results stand out from the aggregate evidence. First, the hit rate is broad-based rather than concentrated in a single theme pair: it ranges from 39% (Accruals×Quality) to 74% (Profitability×Valuation). Second, the small set of wrong-sign proposals (8 of 270, or 3%) is also informative. These proposals do not simply represent null results; in several cases they reveal that the underlying mechanism operates in the opposite direction from the agent's initial conjecture. As we document in the case studies of Section 7, the agent treats such reversals as diagnostic evidence and uses them to revise subsequent proposals.

5.2 Generation-by-Generation Improvement

The next question is whether the reflection loop improves proposal quality over time. Appendix table 5 and Figure 2 report hit rates and the best signal strength by generation. Hit rates show an overall upward trend, rising from 44% in Generation 0 to 78% in Generation 5, though the path is not monotonic: Generation 1 dips to 36% before the improvement takes hold in Generations 2–5. Generation 6 settles at 70%, modestly below the peak. Best signal strength climbs from $|t| = 4.47$ in Generation 0 to $|t| = 6.26$ in Generation 3 and remains in the 6.1–6.3 range thereafter.

This pattern is consistent with the intended role of the reflection loop. The agent does not simply generate fresh variants at random; rather, it appears to redirect its search toward variable combinations, operators, and economic mechanisms that performed well in earlier rounds. The Generation 1 dip is informative: it suggests that the agent's initial attempt at reflection sometimes overshoots, discarding viable approaches before learning which revisions are productive. By Generation 4–5, the accumulated feedback stabilizes proposal quality at

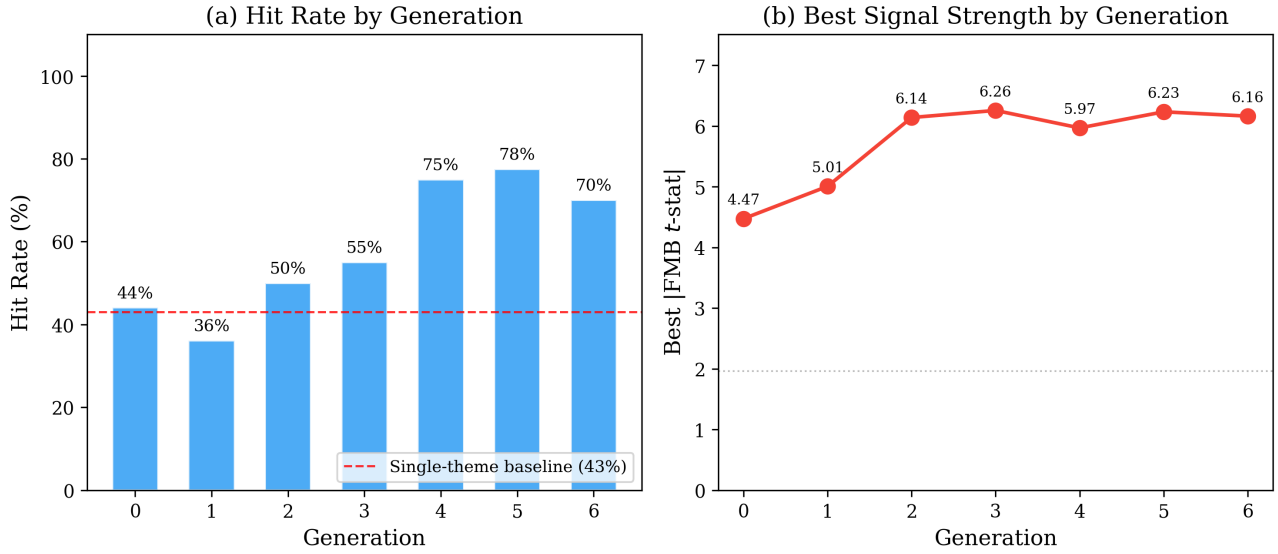

Figure 2: Signal quality by generation. Panel (a): hit rate (fraction of non-error proposals achieving $|t^{FMB}| > 1.96$ with predicted sign) by generation, aggregated across all 8 theme pairs. The dashed line indicates the single-theme programmatic baseline (34%). Panel (b): best $|t^{FMB}|$ achieved in each generation across all pairs.

roughly double the Generation 0 baseline. Importantly, the improvement is not driven by a shift toward weaker, easier-to-find variants. The best $|t|$ per generation increases alongside the hit rate, indicating that the agent produces both more and stronger signals as the search progresses.

5.3 Horse Race: Independent Survivors

Aggregate success does not by itself tell us how many distinct mechanisms the system has uncovered. Many successful proposals are nearby variants of the same underlying idea. For example, multiple versions of "leveraged asset growth" with slightly different operator choices or variable substitutions. We therefore turn to the multivariate horse race described in Section 3, which isolates signals with incremental predictive content.

Table 3 reports the results within each theme pair. Of 159 successful proposals, 38 survive the stepwise backward elimination procedure. The surviving signals are not evenly distributed across theme pairs. Valuation×Quality produces the largest number of independent survivors (13), reflecting a rich set of distinct interactions between cheapness and earnings quality. Investment×Financing produces 11 survivors, consistent with the multiple channels through which investment and external financing interact (leverage conditioning, issuance timing, capex funding). By contrast, Accruals×Quality and Investment×Accruals each contribute

only 1 survivor, suggesting that these theme interactions produce narrower sets of economically distinct mechanisms.

Figure 3 visualizes the hit rate and survivor count across all eight pairs.

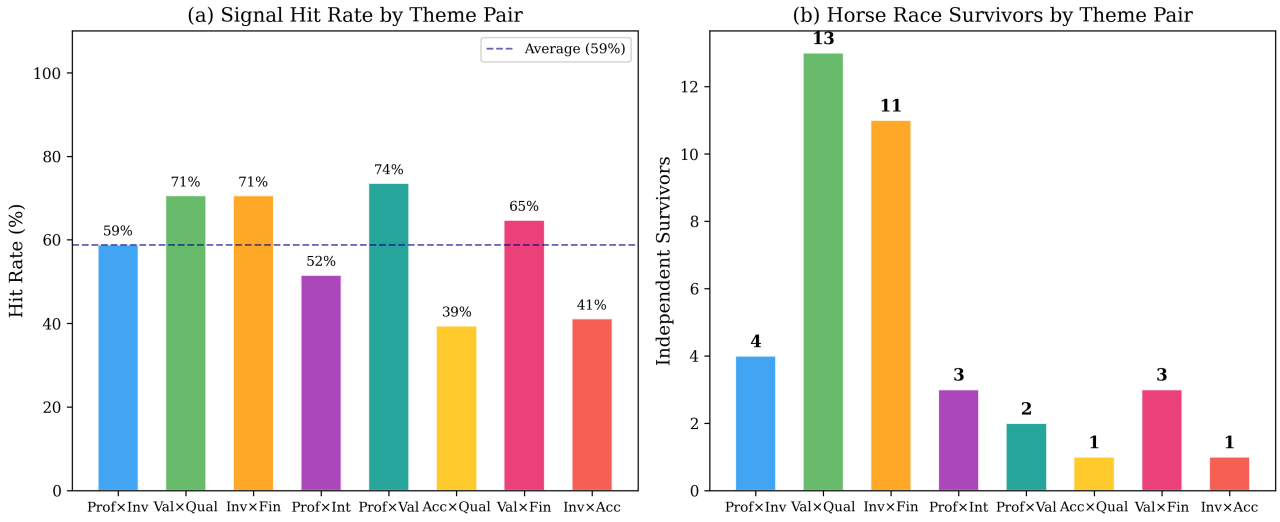

Figure 3: Signal performance by theme pair. Panel (a): hit rate by pair. Panel (b): number of independent horse-race survivors by pair.

5.4 Economically Strongest Signals

We next ask which of the surviving signals are economically strongest. Appendix table 6 reports the ten survivors with the highest long–short Sharpe ratios. Several features of Appendix table 6 are worth noting. First, 7 of the top 10 signals are first proposed in Generations 3–6, indicating that the strongest discoveries often emerge only after repeated feedback and revision. Second, the list includes both composite (`SUMRANK`) and ratio signals. Composite signals exploit the joint ranking of related mechanisms, while ratio signals sometimes survive because they capture a more integrated accounting relation. Third, the magnitudes are economically meaningful: the leading signal, `ConditionalSP_Quality`, achieves an annualized Sharpe ratio of 0.54, and several others exceed 0.45. These values place the strongest survivors in a range that is not merely statistically detectable but also economically large.

5.5 Cross-Theme versus Single-Theme Comparison

The final question in this section is how the cross-theme discovery exercise compares with the single-theme baseline. Table 4 reports summary statistics for the cross-theme AI-guided

search and for the 99-signal programmatic search over individual themes, evaluated on the same data environment with the same microcap exclusion and evaluation protocol.

The cross-theme discovery exercise outperforms the single-theme baseline on all summary measures. The hit rate rises from 34% (single-theme) to 59% (cross-theme), a 25 percentage-point advantage. The best Fama–MacBeth $|t|$ improves from 5.60 to 6.26, and the best Sharpe ratio from 0.40 to 0.54. The cross-theme search also produces 38 horse-race survivors, whereas the single-theme baseline is not subjected to a horse race (its signals are not generated through an iterative process). We interpret this comparison as evidence that the combined architecture, namely theme-conditioned search plus AI-guided refinement, is productive. At the same time, the comparison does not separately identify the contribution of each component: the cross-theme search benefits simultaneously from theme interaction, iterative refinement, extended thinking, and the LLM's pretrained financial knowledge. Table 4 should therefore be read as a benchmark of overall discovery yield rather than as a clean decomposition of why the gain arises.

6 Rigorous Testing

Section 5 shows that the discovery framework produces many statistically significant candidates and that 38 signals survive within-pair horse races. That evidence is useful, but it is not the standard by which an anomaly should be judged. Because the agent revises proposals using performance summaries from the same historical sample, the relevant question is not how many signals clear an initial $|t| > 1.96$ screen, but how many remain once they are evaluated under the validation standards that modern asset pricing now treats as essential. This section applies that framework along five dimensions: multiple-testing adjustment, factor-model absorption, spanning against published anomalies, temporal robustness, and permutation inference. The resulting pattern is one of sharp but informative attrition. The horse-race survivors are not simply a large-M artifact, but many are absorbed by known factor structures or closely related published signals. However, a small subset remains strong across several dimensions simultaneously. Appendix D reports supplementary placebo and related robustness diagnostics.

6.1 Multiple Testing Corrections and Bayesian Inference

Multiple-testing adjustment is a necessary first filter, but it is not the endpoint of the analysis. Across the full discovery exercise, $M = 365$ unique signals produce valid test statistics, so the conventional $|t| > 1.96$ threshold materially understates the evidentiary bar. We apply three correction procedures from Harvey et al. (2016), each controlling a different error rate, and complement them with the Bayesian framework of Harvey (2017).

Table 5 reports the results. At the 5% level, 30 of the 38 horse-race survivors remain significant under BHY, 25 under Holm, and 24 under Bonferroni. Even at the 1% level, BHY retains 27 and Bonferroni retains 16. Under a skeptical Bayesian prior, 28 of the 38 survivors have posterior $\Pr(H_0 \mid \text{data}) < 5\%$. These results matter for two reasons. First, they show that the horse-race survivors are not merely a consequence of testing many formulas. Second, they clarify where the main attrition does *not* occur. Most of the surviving signals remain statistically credible after correcting for multiple comparisons. The more economically consequential attrition arises later, once the signals are evaluated against benchmark factor models and closely related published anomalies. Multiple-testing survival is therefore a necessary condition for further attention, not evidence of novelty by itself.

6.2 Multi-Model Factor Alphas

Statistical credibility does not imply incremental pricing information. A signal may be significant in a univariate Fama–MacBeth regression and yet largely proxy for exposures already captured by standard factor models. To assess this issue, we evaluate each long–short portfolio under the Fama–French five-factor and six-factor models, and the Hou–Xue–Zhang q-factor model. We also report a market-equity-weighted Fama–MacBeth specification that further downweights smaller firms within the non-microcap universe.

Table 6 and Figure 4 show that factor absorption is the central source of attrition. Of the 38 horse-race survivors, 13 remain significant in the weighted Fama–MacBeth specification, 8 retain FF5 alpha significance, 6 retain FF6 significance, and 5 retain q-factor significance. The pattern is consistent with the view that much of the initial predictability loads on profitability-, and investment-related return variation already captured by benchmark models. In other words, the agent often discovers economically sensible formulas whose return spreads are real in sample but whose pricing content is substantially aligned with known factor structures.

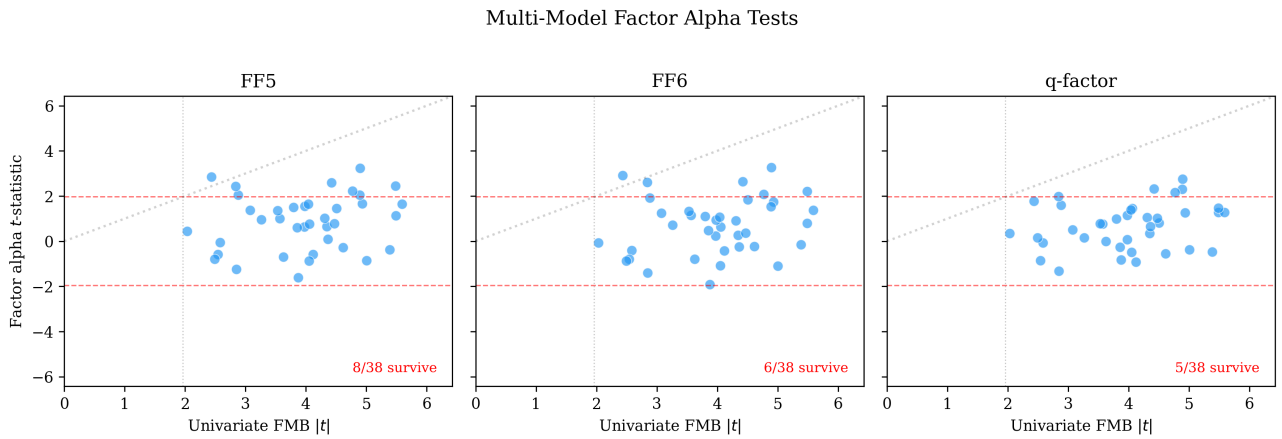

Figure 4: Multi-model factor alpha tests. Each panel plots the univariate FMB $|t|$-statistic (horizontal axis) against the factor model alpha t-statistic (vertical axis) for all 38 horse-race survivors. The dashed red lines indicate $|t| = 1.96$. Points below the 45-degree line have alphas smaller than their univariate t-statistics, indicating that factor exposures absorb part of the signal's return predictability. The count of survivors retaining $|t^{\alpha}| > 1.96$ is shown in each panel.

6.3 Spanning Tests Against Known Anomalies

Factor models address systematic risk exposures, but they do not answer whether a signal is new relative to the anomaly literature. A signal can generate factor alpha and still be a close variant of a published predictor. We therefore benchmark the horse-race survivors against the 209-signal Chen and Zimmermann (2022) open-source library. For each survivor, we compute its average cross-sectional Spearman rank correlation with each published anomaly, select the 10 closest neighbors, and estimate a multivariate Fama–MacBeth regression that includes the new signal together with those controls. Because the controls are chosen to be the signal's nearest published alternatives, the resulting conditional t-statistic should be interpreted as a targeted novelty screen rather than as another generic horse race.

Table 7 and Figure 5 show substantial attenuation but not complete absorption. Of the 34 survivors with stable conditional estimates, 9 retain significance at $|t^{cond}| > 1.96$ and 6 remain above $|t^{cond}| > 3.00$. The average decline in absolute t-statistics is large, indicating that many proposals recombine ideas already represented in the anomaly zoo. The closest matches are most often variants of accruals–value and asset-growth signals, again suggesting that the agent frequently converges on familiar themes through new functional forms.

At the same time, a small set of signals remains meaningfully distinct from its nearest published neighbors. `LeveragedPPEGrowth` has the highest conditional statistic in the paper, with $t^{cond} = 4.75$ despite a nontrivial correlation with its closest Chen–Zimmermann match.

`UltimateAlphaComposite` is the second strongest, with $t^{cond} = 3.81$ and a relatively low correlation with its nearest published analogue. These cases suggest that the agent can occasionally generate signals that are not well described as simple relabelings of existing anomalies.

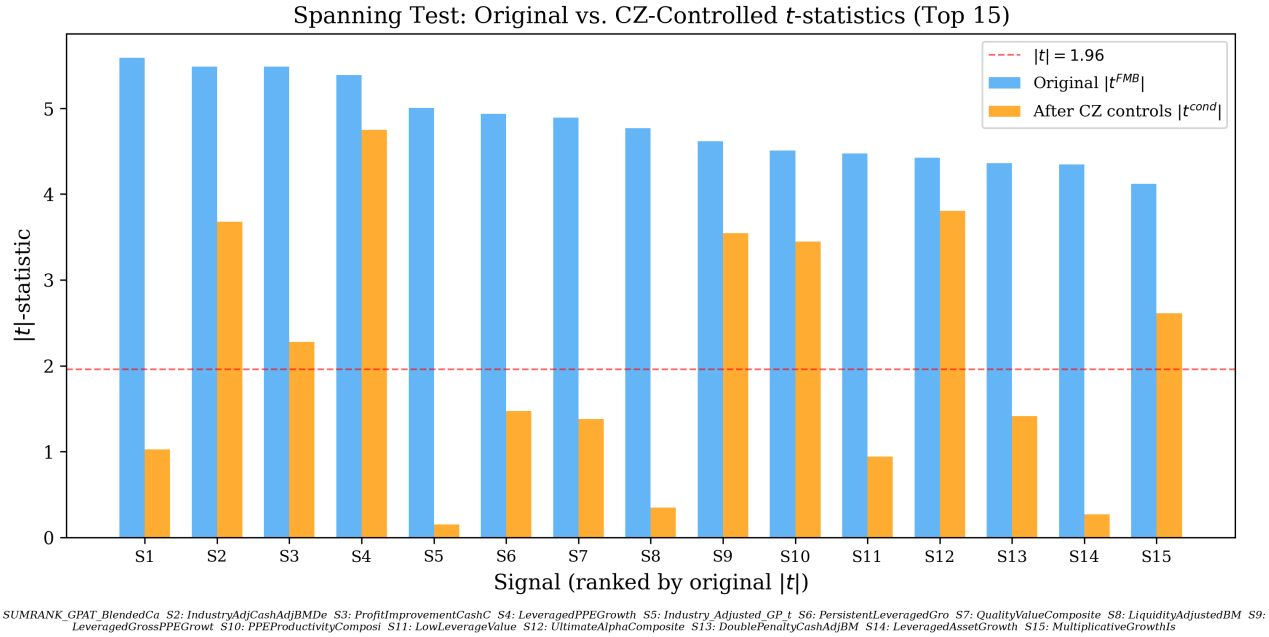

Figure 5: Spanning test results for the top 15 signals ranked by original $|t^{FMB}|$. Blue bars show the original univariate t-statistic; orange bars show the conditional t-statistic after controlling for the 10 most-correlated anomalies from the Chen–Zimmermann (2022) library. The dashed red line indicates $|t| = 1.96$. Signal labels (S1–S15) are mapped to full names in the legend below the figure.

6.4 Subsample Robustness and Decay

A signal that is significant in the full sample may be concentrated in a single era or may weaken materially once it becomes closer to real time. We therefore evaluate each survivor across seven subsamples: the full sample, the pre-anomaly era, the post-publication era, pre-2000, post-2000, post-2010, and a sample that excludes NBER recession months. Appendix table 7 and Figure 6 summarize the results.

Twelve of the 38 horse-race survivors remain significant in all seven subsamples, and 22 are significant in at least six. The post-2010 window is the most demanding: only 16 of 38 survive in that period. Given that the discovery loop uses the full 1963–2024 sample, the post-2010 results provide the closest available proxy to out-of-sample stability within the current design. The comparison between earlier and later periods also points to broad attenuation. Mean absolute t-statistics decline from 4.64 before 2000 to 3.01 after 2000, and 26 of the 38 survivors

are stronger in the earlier sample.

We formalize this pattern by computing rolling 60-month Fama–MacBeth statistics and testing for linear time trends. The resulting classification is mixed but not encouraging: 15 signals are stable, 17 are decaying, and 6 are strengthening. This pattern closely parallels the broader anomaly literature and reinforces a central theme of the paper. The framework is capable of producing many apparently strong signals, but temporal robustness meaningfully narrows the set of claims that should be taken seriously.

Figure 6: Subsample robustness heatmap. Each row is a horse-race survivor (sorted by number of robust subsamples, then by full-sample $|t|$). Each column is a subsample. Color intensity indicates $|t^{FMB}|$ (green = significant, red = weak). Signals that are uniformly green across all columns are the most temporally robust.

6.5 Permutation Tests

The preceding exercises benchmark the signals against known statistical and economic alternatives. A final concern is that the measured predictability could still reflect accidental structure in the panel rather than the hypothesized link between signals and returns. To address this issue, we conduct time-series and cross-sectional permutation tests with 1,000 iterations each. The time-series procedure shuffles month labels on returns while preserving cross-sectional signal assignments; the cross-sectional procedure permutes signal values across firms within each month while preserving the time-series structure. In both cases, the permutation p-value is the fraction of permuted absolute t-statistics that exceed the observed value.

The permutation evidence is reassuring. Of the 38 horse-race survivors, 33 have time-series permutation $p < 0.05$, and all 38 have cross-sectional permutation $p < 0.05$. Stronger signals are correspondingly harder to reproduce under random reassignments. These results do not establish economic novelty, but they do indicate that the measured predictability is unlikely to be a mechanical artifact of the panel's temporal or cross-sectional structure.

6.6 Summary

Table 8 summarizes the attrition across the full validation framework. The main message is not that the initial discovery results disappear, nor that they survive intact. Rather, the evidence yields a clear hierarchy. Multiple-testing adjustments eliminate relatively few of the horse-race survivors. The decisive attrition occurs when the signals are confronted with benchmark factor models and with the closest anomalies in the published literature. Temporal robustness further narrows the set of credible claims, while permutation tests indicate that the remaining predictability is not spurious.

Taken together, the evidence supports a disciplined interpretation of AI-guided discovery in asset pricing. The agent generates many statistically interesting and economically interpretable candidates, but most of these candidates map onto known return patterns once subjected to modern validation. A smaller group of signals, most notably `UltimateAlphaComposite` and `LeveragedPPEGrowth`, remains difficult to explain away along several dimensions at once. These signals constitute the paper's strongest empirical claims.

7 How AI Reasons: Evidence from the Discovery Trace

A central advantage of the framework is that it records not only final discoveries but also the sequence of proposals, outcomes, and revisions that produced them, including the model's internal chain-of-thought reasoning preserved across all seven generations. This section uses those traces to study how the AI research agent updates hypotheses in response to empirical feedback. Our objective is descriptive rather than philosophical: we do not infer internal cognition from text outputs, but we analyze observable proposal behavior, recurring revision motifs, and the relationship between those motifs and subsequent empirical success.

The analysis proceeds in four parts. We first present two detailed case studies that illustrate the agent's reasoning at its most informative, one showing a paradigm shift in how investment and financing interact, the other showing the deliberate composition of orthogonal alpha channels. We then classify proposal text into a taxonomy of recurrent patterns and assess how proposed signals relate to the existing anomaly literature. Finally, we document recurring failure modes. Two additional case studies appear in Appendix C.

7.1 Case Study A: The Leverage-Stock Breakthrough

The single strongest signal to emerge from the CZ spanning test is `LeveragedPPEGrowth` (conditional $t = 4.75$, robust in 7/7 subsamples, stable decay classification). Its discovery illustrates how the agent shifted from one conceptual paradigm to another in response to accumulated evidence.

Generations 0–1: The flow×flow paradigm. Through the first two generations of the Investment×Financing pair, every successful signal used the same template: multiply an investment *flow* (asset growth, capex change) by a financing *flow* (debt issuance, equity issuance). The reasoning was natural—firms simultaneously expanding and raising external capital are the clearest empire-builders.

Generation 2: The conceptual shift. In Generation 2, the agent proposed a qualitatively different interaction. Its reasoning trace articulates the distinction 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, capturing a distinct mechanism: highly leveraged firms face higher costs of financial distress, so aggressive investment on an already-strained balance sheet signals managerial recklessness."

The resulting signal, `LeveragedAssetGrowth`, immediately surpassed all prior flow×flow variants ($t^{FMB} = 4.35$). The Generation 3 reflection recognized the shift as systematic: the financing *stock* conditions the investment effect more powerfully than financing *flows*.

Generation 3: Substitution and refinement. The agent then substituted the strongest standalone investment measure, PP&E growth, for generic asset growth, reasoning that PP&E growth captures deliberate capacity expansion rather than passive balance sheet changes. `LeveragedPPEGrowth` achieved $t^{FMB} = 5.39$ and ultimately produced the highest CZ spanning t-statistic in the study ($t^{cond} = 4.75$), with its closest published anomaly (`InvestPPEInv`) at a correlation of only 0.64.

The signal is robust in all 7 subsamples and classified as temporally stable. Its FF5 alpha, however, is not significant ($t = -0.37$): the return predictability is explained by known factor exposures. The CZ spanning and factor model tests thus tell complementary stories that the signal is *novel in construction* (no published anomaly replicates the leverage×PP&E-growth interaction) but *explained in content* (its returns map onto known investment and leverage premia). The agent discovered a new way of accessing existing risk compensation, not a new source of it.

7.2 Case Study B: Compositing Orthogonal Alpha Channels

The most robust signal in the study across all test dimensions is `UltimateAlphaComposite`—a Gen 6 signal that survives FF5 ($t^{\alpha} = 2.60$), FF6 ($t^{\alpha} = 2.64$), the q-factor model ($t^{\alpha} = 2.32$), CZ spanning ($t^{cond} = 3.81$, with the closest CZ anomaly at $\rho = 0.34$), and all 7 subsamples. Its construction illustrates how the agent synthesized six generations of accumulated evidence into a single composite.

The two channels. By Generation 5, the Investment×Financing pair had produced two distinct families of successful signals. The first family conditioned investment growth on leverage

levels—the stock×flow paradigm from Case Study A. The second family paired capex *growth* with debt *issuance flows*, capturing a different mechanism: firms that simultaneously accelerate capital spending and tap debt markets are making concurrent timing decisions that the market penalizes.

The agent's reasoning trace identifies the key insight:

> *"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."*

The deliberate composition. In Generation 6, the agent proposed combining these two channels via `SUMRANK`, explicitly reasoning about their orthogonality:

> *"These two signals capture economically distinct mechanisms—structural balance sheet fragility versus active concurrent market-timing decisions—that should be largely uncorrelated, making their SUMRANK composite a diversified alpha source superior to either alone."*

The resulting composite performs as the reasoning predicts. Its low correlation with the closest CZ anomaly ($\rho = 0.34$ with `CompositeDebtIssuance`) confirms that the signal is not a repackaging of any single published predictor. Its survival under factor models indicates that the diversification across channels does provide incremental information beyond known factors. What makes this case study distinctive is not the final performance alone but the observable reasoning chain: the agent identified two mechanisms, diagnosed their orthogonality, and composed them deliberately. Whether this constitutes genuine economic understanding or sophisticated pattern matching is a question we leave open. What we can verify is that the reasoning is specific, the orthogonality claim is empirically supported, and the resulting signal passes tests that most LLM-discovered signals fail.

7.3 Reasoning Taxonomy

Beyond the case studies, we classify the text associated with each of the 280 proposals into a taxonomy of recurrent reasoning patterns. Appendix table 9 and Figure 7 report the results.

Categories are non-exclusive; a proposal may employ multiple strategies simultaneously, so column totals exceed the number of unique signals.

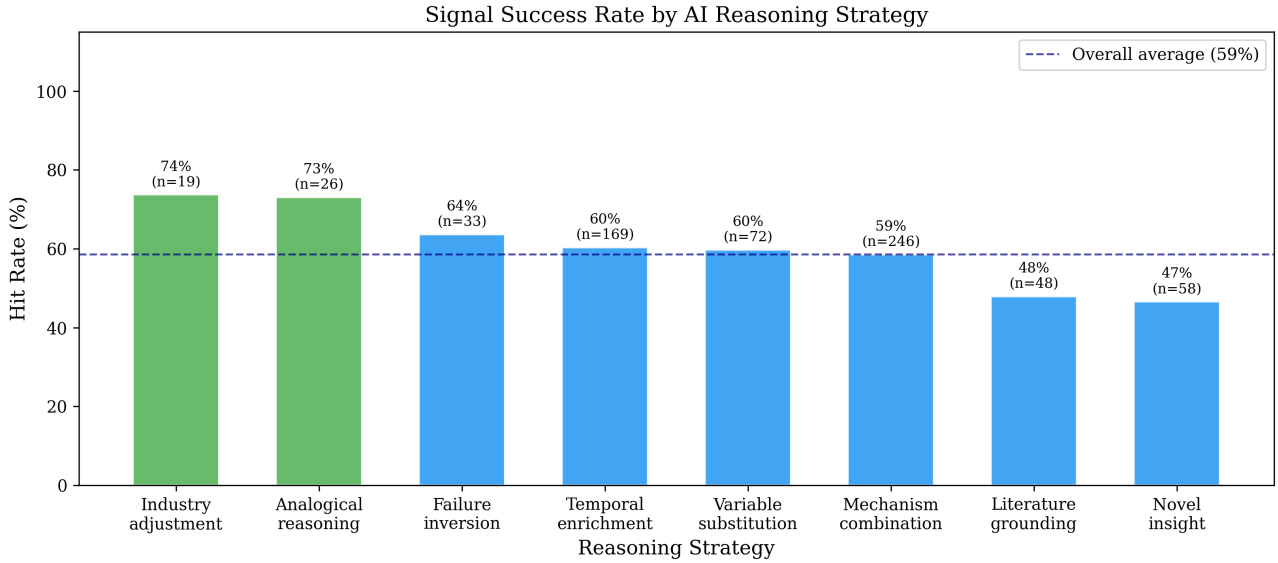

Figure 7: Signal success rate by AI reasoning strategy, sorted by hit rate. The dashed line indicates the overall average. Each bar is labeled with its hit rate and sample size.

Three features stand out. First, *industry adjustment* achieves the highest per-category hit rate (74%), though it is deployed less frequently than mechanism combination or temporal enrichment. Its usage grows over generations, suggesting that the reflection loop increasingly redirects search toward industry-relative measurement once early results reveal its value. Second, *mechanism combination*, the most common strategy, achieves a 59% hit rate that mirrors the overall average, consistent with its role as the default cross-theme approach. Third, *novel economic insight* and *literature grounding* achieve lower hit rates (47% and 48%), suggesting that the agent's comparative advantage lies in structured recombination rather than in generating ideas from theoretical first principles.

7.4 Novelty Classification

We next map each proposed signal to the existing literature on cross-sectional return predictors. Appendix table 8 and Figure 8 report the results.

Novel combinations, defined as known building blocks combined in previously undocumented ways, constitute 71% of all proposals and achieve a 62% hit rate. Variants of known signals achieve 48%. Direct rediscoveries are rare (11 of 280) and have the lowest success rate (18%).

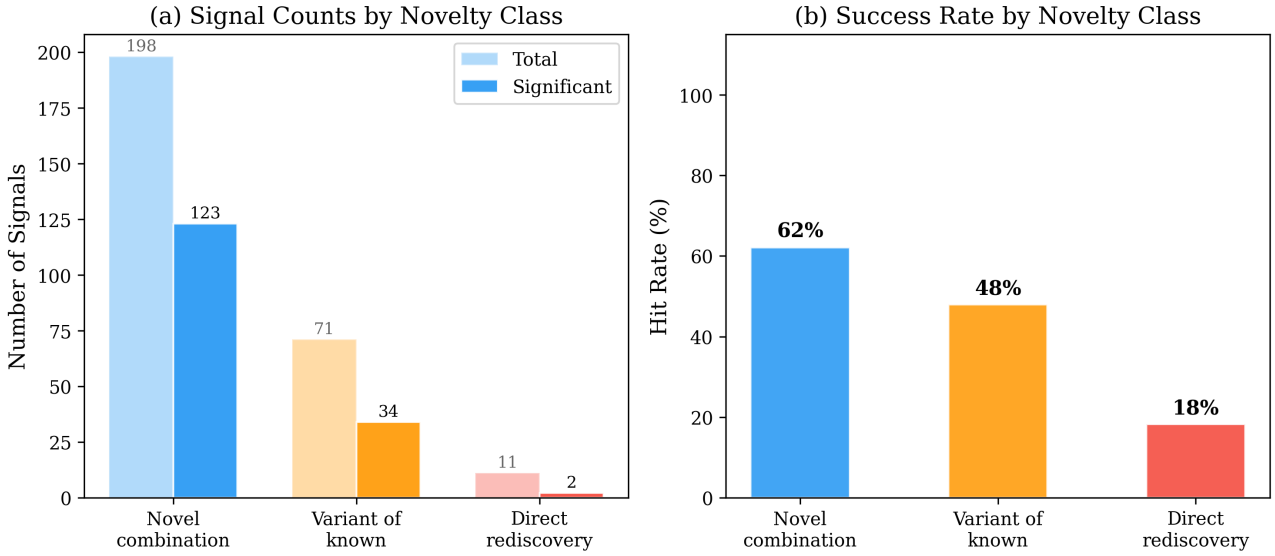

Figure 8: Novelty classification: counts and hit rates. Panel (a): total and significant signal counts by novelty class. Panel (b): hit rate by novelty class.

The 44 percentage-point gap between novel combinations and direct rediscoveries is the clearest evidence that the framework's value lies in combinatorial creativity rather than in reproducing the existing literature.

7.5 Failure Modes

The discovery trace is also informative about where the framework performs less well. Three recurring failure modes help define the boundaries of the approach.

1. **Noise amplification.** Temporal operators such as `GROWTH` and `DELTA` can amplify noise when applied to already-volatile ratios. This failure mode is partially self-correcting: the agent learns to prefer `MA` (smoothing) over raw `GROWTH` in later generations.

2. **Factor subsumption.** As documented in Section 6.2, this is the dominant failure mode: 30 of 38 survivors lose significance under at least one factor model. The agent's reasoning traces show partial awareness of this risk, but it lacks direct access to factor model results during the discovery loop.

3. **Anchoring on prior successes.** In later generations, the agent often proposes local variants of previously successful structures. This behavior produces additional successful proposals but narrows exploration of more distant mechanisms. The prompt partially offsets this tendency by requiring at least one novel-direction proposal per generation.

These failure modes map onto well-known challenges in the asset pricing literature. Noise amplification parallels the concern about unstable anomalies; factor subsumption is precisely the issue that Hou et al. (2020) document at scale; and anchoring on prior successes echoes the "factor zoo" critique (Harvey et al., 2016). The fact that an AI research agent encounters the same challenges as human researchers suggests that these are properties of the underlying empirical landscape, not artifacts of any particular discovery method.

8 Discussion

The results should be interpreted as evidence about a discovery architecture and its empirical limits rather than as a definitive catalog of new anomalies. The paper studies whether a human-designed, AI-executed framework can generate, test, and revise interpretable accounting-based hypotheses inside a fixed empirical environment, and, equally important, what fraction of those hypotheses survives when subjected to the full battery of tests that the modern asset pricing literature considers standard.

8.1 What the LLM Can and Cannot Do

The empirical evidence points to a clear division. The LLM excels at hypothesis search: it rapidly converges on combinations of accounting variables that produce statistically significant cross-sectional predictability. The interpretability constraint imposed by the expression language is a feature of the design rather than a limitation. By restricting the agent to symbolic expressions built from accounting quantities, the framework ensures that every proposal can be inspected, critiqued, and subjected to the full suite of asset pricing tests. Black-box signals cannot be tested against the CZ anomaly library or decomposed into factor loadings in the same way. Of 280 proposals, 159 clear the conventional screen, and hit rates nearly double from Generation 0 to Generation 5. The reasoning traces add a dimension that has no analogue in black-box machine learning or in conventional anomaly discovery. The case studies in Section 7 show an agent that diagnoses wrong-sign results, articulates structural insights about leverage stocks versus financing flows, identifies orthogonal alpha channels, and independently reconstructs known factor premia. Whether this constitutes "understanding" is a question we leave to others. What we can document is that the trace is specific, verifiable, and in several cases contains economic insights, such as the stock×flow conditioning principle, that are absent

from the published literature.

What the LLM cannot do, at least in the current design, is generate many signals that are predominantly novel relative to known factors and published anomalies. Under the FF5 model, only 8 of 38 survivors retain significant alpha. The CZ spanning test finds that 9 of 34 testable survivors carry incremental information beyond the 209 published anomalies, but even these survivors are substantially attenuated relative to their univariate significance. The LLM is an efficient search engine for the cross-sectional signal space, but most of what it finds maps onto factor premia that the literature already recognizes. This result is not a failure of the framework. It is a characterization of the empirical landscape. The space of interpretable accounting-based signals that predict cross-sectional returns is heavily populated by known effects. The value of the LLM lies not in escaping that landscape but in navigating it more efficiently and transparently than manual search.

8.2 Inference and the Role of the Testing Framework

The discovery loop constitutes an in-sample evolutionary search: the agent proposes signals, observes their performance on the same dataset, and refines its next proposals accordingly. This design raises the standard concerns about data mining, adaptive search, and multiple testing (Harvey et al., 2016; Hou et al., 2020). We address these concerns through the testing framework of Section 6 rather than through a conventional train–test split. The rationale is that the discovery process is the object of study: splitting the sample would weaken the evidence on how the agent reasons and revises. Instead, we apply the full battery of post-discovery corrections recommended by the methodological literature. This approach does not eliminate the in-sample concern. A researcher who is skeptical of all in-sample results, regardless of the corrections applied, should treat the surviving signals as promising candidates rather than settled anomalies.

8.3 Generalizability

The framework is portable to other empirical settings that share three features: a structured set of candidate primitives, a transparent evaluation pipeline, and enough data to discriminate signal from noise. Potential applications include credit risk modeling, corporate event studies, macroeconomic forecasting, and accounting measurement. In each case, the central requirement is that candidate hypotheses can be formalized in a constrained language and evaluated under

a stable protocol. The present design does not accommodate signals that require unstructured text, high-frequency data, cross-firm network effects, or real-time information. Those extensions define a natural research frontier but require fundamentally different hypothesis languages and evaluation pipelines.

8.4 Limitations

Several limitations define the boundaries of the present study and the natural next steps. First, the discovery exercise is conducted in-sample. The agent sees performance summaries from the same CRSP/Compustat panel it is implicitly optimizing over. While the multiple testing corrections, factor model tests, and subsample splits provide substantial discipline, they are not a substitute for a true out-of-sample holdout. A natural extension is to freeze the discovery protocol in an initial sample period and evaluate the resulting signals in later periods or alternative markets.

Second, the agent operates with domain knowledge acquired during pretraining on a large corpus that includes the finance literature. Some successful proposals may therefore reflect learned concepts rather than fully de novo discovery. Our novelty analysis partly addresses this concern by showing that direct rediscoveries are less successful than novel combinations (18% vs. 62% hit rate), but the issue remains conceptually important. Third, the main experiments use a single frontier model (Claude Opus 4.6) and a fixed symbolic language with 66 variables and 24 operators. The results do not establish how sensitive the framework is to model choice, prompt design, or the boundaries of the hypothesis space. Broader operator libraries, alternative LLMs, and multi-model ensembles are natural extensions.

8.5 Conclusion

We develop a human-designed, AI-executed framework for hypothesis discovery in empirical asset pricing. Within a constrained language of 66 accounting primitives and 24 operators, evaluated on a microcap-excluded CRSP/Compustat universe spanning 1963–2024, an AI research agent generates 280 cross-theme signals, 159 of which clear a conventional significance screen and 38 of which survive multivariate horse races. We then subject these survivors to the full battery of modern asset pricing tests: multiple testing corrections, multi-model factor spanning, novelty tests against 209 published anomalies, subsample robustness analysis, and permutation tests. Most signals are absorbed by known factors or are variants of published anomalies. But a

small set, most notably `UltimateAlphaComposite` and `LeveragedPPEGrowth`, survives across multiple dimensions simultaneously, carrying genuinely incremental predictive content.

The reasoning traces reveal an agent that diagnoses empirical surprises, extracts structural principles from accumulated evidence, and composes orthogonal information channels, all observable in its chain-of-thought reasoning preserved across seven generations of iterative refinement. More broadly, the paper suggests that the most productive role for AI in empirical finance is not to replace economic judgment or to discover entirely unknown phenomena, but to operate inside carefully designed research architectures that make the search faster, more systematic, and more transparent.

References

Agrawal, A. K., McHale, J., and Oettl, A. (2026). AI in science. Working Paper 34953, National Bureau of Economic Research.

Asness, C. S., Frazzini, A., and Pedersen, L. H. (2019). Quality minus junk. *Review of Accounting Studies*, 24(1):34–112.

Bernard, D., Blankespoor, E., de Kok, T., and Toynbee, S. (2026). Using GPT to measure business complexity. *The Accounting Review*. Published online January 14, 2026.

Chen, A. and Zimmermann, T. (2022). Open source cross-sectional asset pricing. *¿ Critical Finance Review*, 11(02):207–264.

Chen, L., Pelger, M., and Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, 70(2):714–750.

Cooper, M. J., Gulen, H., and Schill, M. J. (2008). Asset growth and the cross-section of stock returns. *The Journal of Finance*, 63(4):1609–1651.

Cui, C., Wang, W., Zhang, M., Chen, G., Luo, Z., and Ooi, B. C. (2021). AlphaEvolve: A learning framework to discover novel alphas in quantitative investment. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2208–2216. Association for Computing Machinery.

Dechow, P. M. and Dichev, I. D. (2002). The quality of accruals and earnings: The role of accrual estimation errors. *The Accounting Review*, 77(Supplement):35–59.

Dong, M. M., Stratopoulos, T. C., and Wang, V. X. (2024). A scoping review of ChatGPT research in accounting and finance. *International Journal of Accounting Information Systems*, 55:100715.

Eisfeldt, A. L. and Papanikolaou, D. (2013). Organization capital and the cross-section of expected returns. *The Journal of Finance*, 68(4):1365–1406.

Fairfield, P. M., Whisenant, J. S., and Yohn, T. L. (2003). Accrued earnings and growth: Implications for future profitability and market mispricing. *The Accounting Review*, 78(1):353–371.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *The Journal of Finance*, 72(4):1399–1440.

Harvey, C. R. and Liu, Y. (2020). False (and missed) discoveries in financial economics. *The Journal of Finance*, 75(5):2503–2553.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ...and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Harvey, C. R., Sancetta, A., and Zhao, Y. (2026). What threshold should be applied to tests of factor models? Working Paper 34898, National Bureau of Economic Research.

Hou, K., Xue, C., and Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies*, 28(3):650–705.

Hou, K., Xue, C., and Zhang, L. (2020). Replicating anomalies. *The Review of Financial Studies*, 33(5):2019–2133.

Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Kelly, B. T., Kuznetsov, B., Malamud, S., and Xu, T. A. (2025). Artificial intelligence asset pricing models. Working Paper 33351, National Bureau of Economic Research.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.

Kou, Z., Yu, H., Luo, J., Peng, J., Li, X., Liu, C., Dai, J., Chen, L., Han, S., and Guo, Y. (2025). Automate strategy finding with LLM in quant investment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18517–18533, Suzhou, China. Association for Computational Linguistics.

Lev, B. and Sougiannis, T. (1996). The capitalization, amortization, and value-relevance of r&d. *Journal of Accounting and Economics*, 21(1):107–138.

Lopez-Lira, A. and Tang, Y. (2023). Can ChatGPT forecast stock price movements? return predictability and large language models.

Ludwig, J. and Mullainathan, S. (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827.

Manning, B. S., Zhu, K., and Horton, J. J. (2024). Automated social science: Language models as scientist and subjects.

McLean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, 71(1):5–32.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, 108(1):1–28.

Novy-Marx, R. and Velikov, M. (2024). Assaying anomalies. Working paper, SSRN.

Novy-Marx, R. and Velikov, M. (2026). Artificial intelligence–powered (finance) scholarship. *Journal of Economic Literature*, 64(1):5–37.

Peters, R. H. and Taylor, L. A. (2017). Intangible capital and the investment-q relation. *Journal of Financial Economics*, 123(2):251–272.

Petersen, M. A. (2008). Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of financial studies*, 22(1):435–480.

Piotroski, J. D. (2000). Value investing: The use of historical financial statement information to separate winners from losers. *Journal of Accounting Research*, 38:1–41. Supplement.

Piotroski, J. D. and So, E. C. (2012). Identifying expectation errors in value/glamour strategies: A fundamental analysis approach. *The Review of Financial Studies*, 25(9):2841–2875.

Pontiff, J. and Woodgate, A. (2008). Share issuance and cross-sectional returns. *The Journal of Finance*, 63(2):921–945.

Richardson, S. (2006). Over-investment of free cash flow. *Review of Accounting Studies*, 11(2–3):159–189.

Richardson, S. A., Sloan, R. G., Soliman, M. T., and Tuna, I. (2005). Accrual reliability, earnings persistence and stock prices. *Journal of Accounting and Economics*, 39(3):437–485.

Shi, H., Song, W., Zhang, X., Shi, J., Luo, C., Ao, X., Arian, H., and Seco, L. A. (2025a). AlphaForge: A framework to mine and dynamically combine formulaic alpha factors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12524–12532. AAAI Press.

Shi, Y., Duan, Y., and Li, J. (2025b). Navigating the alpha jungle: An LLM-powered MCTS framework for formulaic factor mining.

Shumway, T. (1997). The delisting bias in CRSP data. *The Journal of Finance*, 52(1):327–340.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting review*, pages 289–315.

Tang, Z., Chen, Z., Yang, J., Mai, J., Zheng, Y., Wang, K., Chen, J., and Lin, L. (2025). AlphaAgent: LLM-driven alpha mining with regularized exploration to counteract alpha

decay. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2813–2822, Toronto, ON, Canada. Association for Computing Machinery.

Titman, S., Wei, K. C. J., and Xie, F. (2004). Capital investments and stock returns. *Journal of Financial and Quantitative Analysis*, 39(4):677–700.

Udrescu, S.-M. and Tegmark, M. (2020). AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631.

Wang, H., Fu, T., Du, Y., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Wang, S., Yuan, H., Zhou, L., Ni, L., Shum, H.-Y., and Guo, J. (2025). Alpha-GPT: Human-AI interactive alpha mining for quantitative investment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 196–206, Suzhou, China. Association for Computational Linguistics.

Weng, Z., Zhang, S., Wang, T., and Xia, Y. (2026). AlphaLogics: A market logic-driven multi-agent system for scalable and interpretable alpha factor generation.

Yu, S., Xue, H., Ao, X., Pan, F., He, J., Tu, D., and He, Q. (2023). Generating synergistic formulaic alpha collections via reinforcement learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5476–5486. Association for Computing Machinery.

Zhang, T., Li, Y., Jin, Y., and Li, J. (2020). AutoAlpha: an efficient hierarchical evolutionary algorithm for mining alpha factors in quantitative investment.

Table 1: Main Experimental Conditions

Pair	Economic Mechanism	Gens
Profitability × Investment	Cash cow (high profit, low growth)	7
Valuation × Quality	Quality value (cheap, high quality)	7
Investment × Financing	Empire building (high invest, external finance)	7
Profitability × Intangibles	R&D productivity (profitable innovators)	7
Profitability × Valuation	Value-quality wedge (profitable cheap stocks)	7
Accruals × Quality	Earnings quality premium (cash vs. accrual)	7
Valuation × Financing	Cheap stocks with low external financing	7
Investment × Accruals	Investment-accrual interaction	7

Eight cross-theme pairs, each explored over 7 generations (Gen 0–6) with 5 proposals per generation, yielding up to $8 \times 7 \times 5 = 280$ signal proposals. Each signal is a symbolic expression combining Compustat and CRSP variables via a fixed operator library (30+ operators including RATIO, GROWTH, RANK, IND_ADJ, etc.). Financial firms (SIC 6000–6999) are excluded for 5 of 8 pairs whose signals involve operating cash-flow or investment variables. Microcap stocks (below the 10th NYSE market-equity percentile) are excluded throughout, following the main specification of Hou, Xue, and Zhang (2020). All signals are evaluated on CRSP/Compustat data spanning July 1963 to December 2024.

Table 2: Cross-Theme Signal Discovery: Aggregate Results

Theme Pair	Tested	Sig.	Wrong	Weak	Err	Hit Rate	Best $\lvert t \rvert$
Profitability × Investment	34	20	1	13	1	59%	5.49
Valuation × Quality	34	24	1	9	1	71%	5.60
Investment × Financing	34	24	1	9	1	71%	5.39
Profitability × Intangibles	33	17	2	14	2	52%	6.16
Profitability × Valuation	34	25	0	9	1	74%	6.26
Accruals × Quality	33	13	1	19	2	39%	3.07
Valuation × Financing	34	22	0	12	1	65%	6.23
Investment × Accruals	34	14	2	18	1	41%	3.72
Total	**270**	**159**	**8**	**103**	**10**	**59%**	

A proposal is *successful* if its univariate Fama–MacBeth $|t| > 1.96$ with the predicted sign. "Wrong sign" indicates $|t| > 1.96$ but opposite to the hypothesis. "Weak" indicates $|t| \leq 1.96$. "Err" indicates expression execution failure (e.g., unknown operator). Hit rate = successful / tested (excluding errors). All evaluations use the microcap-excluded universe (NYSE ME > 10th percentile), NYSE quintile breakpoints, value-weighted portfolios, and Fama–MacBeth t-statistics with Newey–West standard errors (6 lags).

Table 3: Horse Race Results: Independent Survivors per Theme Pair

Theme Pair	Input	Survivors	Top Survivor (cond. t)
Profitability × Investment	20	4	`CashCowSpread` ($t = 2.55$)
Valuation × Quality	24	13	`GrossProfitYield` ($t = 4.55$)
Investment × Financing	24	11	`MultiplicativeGrowthIssuanceInteraction` ($t = -2.99$)
Profitability × Intangibles	17	3	`GPAT_RDME_Composite` ($t = 3.82$)
Profitability × Valuation	25	2	`Industry_Adjusted_GP_to_Price` ($t = 3.43$)
Accruals × Quality	13	1	`InvGrowth_SmoothedGP_Composite` ($t = 3.98$)
Valuation × Financing	22	3	`DoublePenaltyCashAdjBM` ($t = -3.02$)
Investment × Accruals	14	1	`AccrualGrowthComposite` ($t = 2.61$)
Total	**159**	**38**	

"Input" is the number of successful signals entering the horse race after removing near-duplicates ($|\rho| > 0.95$). Stepwise backward elimination: at each step, the signal with the lowest conditional $|t|$ is dropped from a simultaneous multivariate Fama–MacBeth regression; the process repeats until all remaining signals have conditional $|t| > 1.96$. "Top Survivor" is the signal with the highest $|t|$ in the final regression. All evaluations use microcap-excluded returns.

Table 4: Cross-Theme vs. Single-Theme Signal Discovery

	Single-Theme (Programmatic)	Cross-Theme (AI)
Signals tested	99	270
Hit rate	34%	59%
Best $\lvert t^{FMB} \rvert$	5.60	6.26
Best Sharpe	0.40	0.54
Horse-race survivors	—	38

Single-theme: 99 programmatic signals across 8 individual themes (profitability, valuation, investment, etc.) constructed from the finance literature, with no LLM involvement. Cross-theme: AI-guided iterative discovery across 8 theme *pairs* over 7 generations with extended thinking and reflection. Both evaluated identically on the microcap-excluded CRSP/Compustat universe (1963–2024), NYSE quintile value-weighted portfolios, Fama–MacBeth with Newey–West SE (6 lags).

Table 5: Multiple Testing Corrections and Bayesian Inference

	$\alpha = 5\%$		$\alpha = 1\%$	
Method	$\|t\|$ hurdle	Survivors	$\|t\|$ 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 inference (Harvey 2017): $Pr(H_0\|data) < 5\%$				
$\pi_0 = 5\%$ (skeptical): 28/38		$\pi_0 = 20\%$ (moderate): 30/38		$\pi_0 = 50\%$ (agnostic): 37/38

Panel A: multiple testing corrections applied to $M = 365$ unique signals with valid t-statistics (10 proposals with expression execution errors excluded from M). Bonferroni and Holm control the family-wise error rate (FWER); BHY (Benjamini–Hochberg–Yekutieli) controls the false discovery rate (FDR). Thresholds and adjusted p-values computed per Harvey, Liu, and Zhu (2016). "Survivors" counts how many of the 38 horse-race survivors also pass each correction. Panel B: Bayesian minimum Bayes factor (MBF $= e^{-t^2/2}$) combined with prior odds per Harvey (2017). π_0 is the prior probability that the effect is true. Counts show survivors with posterior $\Pr(H_0|\text{data}) < 5\%$ under each prior.

Table 6: Multi-Model Factor Alpha Tests

Signal	t^{FMB}	t^{WLS}	$t^{\alpha,FF5}$	$t^{\alpha,FF6}$	$t^{\alpha,q4}$
GPAT_RDME_Composite	4.90	3.51	3.24	3.26	2.75
ProfitVsDebtExpansion	2.44	1.07	2.85	2.90	1.77
UltimateAlphaComposite	4.42	2.22	2.60	2.64	2.32
ProfitImprovementCashCow	5.49	2.70	2.45	2.21	1.27
GP_Trend_RDME_Composite	2.84	2.08	2.42	2.61	1.99
LiquidityAdjustedBM	4.77	3.59	2.23	2.08	2.16
QualityValueComposite	4.89	3.29	2.05	1.54	2.30
CapexGrowthTimesDebtIssuance	2.88	1.36	2.04	1.92	1.60
PersistentLeveragedGrowth	4.93	1.19	1.66	1.75	1.25
SUMRANK_GPAT_BlendedCashAdjV	5.59	4.14	1.64	1.37	1.28
Survivors ($\|t\| > 1.96$) of 38 horse-race survivors		13	8	6	5

Top 10 signals ranked by $|t^{\alpha,FF5}|$. t^{FMB}: univariate Fama–MacBeth t-statistic with Newey–West SE (6 lags). t^{WLS}: market-equity-weighted Fama–MacBeth (Hou, Xue, and Zhang 2020). t^{α}: long–short (Q5−Q1) quintile portfolio alpha t-statistic under each factor model. FF5: Fama–French five-factor (Mkt-RF, SMB, HML, RMW, CMA). FF6: FF5 plus momentum. q4: Hou–Xue–Zhang q-factor model (Mkt, ME, I/A, ROE). q5: q-factor plus expected growth. A signal subsumed by FF5 or q5 is capturing known profitability/investment premia rather than a genuinely novel effect. All evaluations use microcap-excluded returns.

Table 7: Spanning Tests Against 209 Chen–Zimmermann Anomalies

Signal	t^{orig}	t^{cond}	Closest CZ	ρ
LeveragedPPEGrowth	5.39	4.75*	InvestPPEInv	0.64
UltimateAlphaComposite	4.42	3.81*	CompositeDebtIs	0.34
IndustryAdjCashAdjBMDebtRet	5.49	3.68*	AccrualsBM	0.85
LeveragedGrossPPEGrowth	4.61	3.54*	InvestPPEInv	0.62
CashCowSpread	3.86	3.47*	GP	0.68
PPEProductivityComposite	4.51	3.45*	InvestPPEInv	0.57
MultiplicativeGrowthIssuance	4.12	2.61*	XFIN	0.32
ProfitImprovementCashCow	5.49	2.28*	AssetGrowth	0.43
GPlessDepCapxToEV	3.53	2.07*	GP	0.65
AccrualGrowthComposite	2.85	1.89	AssetGrowth	0.76
Survive $\lvert t^{cond}\rvert > 1.96$: 9/34		Survive $\lvert t^{cond}\rvert > 3.00$: 6/34		

Top 10 signals ranked by $|t^{cond}|$. t^{orig}: univariate Fama–MacBeth t-statistic. t^{cond}: conditional t-statistic from a multivariate Fama–MacBeth regression controlling for the 10 most-correlated anomalies from the Chen and Zimmermann (2022) open-source library of 209 published cross-sectional predictors. Controls are selected by average cross-sectional Spearman rank correlation (sampled every 12 months). ρ: correlation with the single closest CZ anomaly. *: $|t^{cond}| > 1.96$. 4 signals have missing t^{cond} due to near-singular regression matrices (high multicollinearity with controls).

Table 8: Summary of Attrition Across Rigorous Tests

Filter	Pass	of 38
BHY FDR 5%	30	79%
Bonferroni 5%	24	63%
WLS FMB $\|t\| > 1.96$	13	34%
FF5 alpha $\|t\| > 1.96$	8	21%
q-factor alpha $\|t\| > 1.96$	5	13%
CZ spanning $\|t^{cond}\| > 1.96$	9	24%
CZ spanning $\|t^{cond}\| > 3.00$	6	16%
All 7 subsamples $\|t\| > 1.96$	12	32%
TS permutation $p < 0.05$	33	87%

Each row is an independent filter applied to the 38 horse-race survivors. Rows are not nested: a signal can pass CZ spanning but fail FF5, or vice versa. The tests target different threats to validity (multiple testing, factor exposure, novelty, temporal stability, spurious correlation). Signals appearing on multiple "pass" lists simultaneously represent the strongest empirical claims.

Appendix A Variable and Operator Catalog

Appendix table 1 lists the 66 Compustat annual variables available to the AI research agent. Appendix table 2 lists the 24 operators in the symbolic expression language.

Appendix Table 1: Compustat Variable Catalog (66 Items)

Mnemonic	Name	Description
Balance Sheet (33 items)		
`aco`	Current Assets Other	Other current assets
`act`	Current Assets	Current assets; used in working capital and accruals computations
`ajex`	Adjustment Factor	Cumulative adjustment factor for stock splits
`ao`	Assets Other	Other assets; captures non-standard items
`ap`	Accounts Payable	Accounts payable; trade credit, used in accruals
`at`	Total Assets	Total assets; the most common denominator for scaling
`ceq`	Common Equity	Book value of common equity; denominator for ROE and equity-scaled signals
`che`	Cash and Short-Term Investments	Cash and equivalents; high cash can signal precautionary savings or low investment
`csho`	Common Shares Outstanding	Shares outstanding; per-share computations
`dlc`	Debt in Current Liabilities	Short-term debt; total debt = dltt + dlc
`dltt`	Long-Term Debt	Long-term debt; key leverage component
`emp`	Employees	Number of employees (thousands); hiring rate, productivity
`gdwl`	Goodwill	Goodwill from acquisitions; high goodwill may indicate overpriced acquisitions
`intan`	Intangible Assets	Reported intangible assets; intangibility = intan/at
`invt`	Inventories	Total inventories; inventory growth predicts low returns
`itcb`	Investment Tax Credit	Investment tax credit balance sheet; fallback component of txditc
`lct`	Current Liabilities	Current liabilities; current ratio = act/lct
`lo`	Liabilities Other	Other liabilities; captures off-balance-sheet type items
`lt`	Total Liabilities	Total liabilities; leverage = lt/at
`mib`	Minority Interest	Minority interest on balance sheet

Continued on next page

Appendix table 1 continued

Mnemonic	Name	Description
ppegt	Gross Property, Plant & Equipment	Gross PP&E before depreciation; capital intensity measure
ppent	Net Property, Plant & Equipment	Net PP&E; tangibility = ppent/at, fixed asset turnover = sale/ppent
prcc_f	Price Close (Fiscal Year End)	Stock price at fiscal year end; Compustat-based market cap = prcc_f * csho
pstk	Preferred Stock - Carrying	Preferred stock carrying value; last fallback
pstkl	Preferred Stock - Liquidating	Preferred stock liquidating value; fallback for pstkrv
pstkrv	Preferred Stock - Redemption	Preferred stock redemption value; used in book equity computation
re	Retained Earnings	Cumulative retained earnings; Altman Z-score component, earned/contributed capital
rect	Receivables Total	Accounts receivable; used in accruals decomposition and turnover ratios
seq	Stockholders' Equity	Total stockholders' equity including preferred; used in book equity computation
txdi	Deferred Taxes	Deferred income taxes; fallback component of txditc
txditc	Deferred Taxes and ITC	Deferred taxes and investment tax credit; book equity component
txp	Income Taxes Payable	Current income taxes payable; used in balance-sheet accruals
xacc	Accrued Expenses	Accrued expenses; accrued liabilities
Income Statement (16 items)		
cogs	Cost of Goods Sold	Direct production costs; gross profit = revt - cogs
dp	Depreciation and Amortization	D&A from income statement; used in accruals computation
ebitda	EBITDA	Earnings before interest, taxes, D&A; enterprise multiple = (ME+DLTT+DLC-CHE)/EBITDA
ib	Income Before Extraordinary	Clean earnings measure; used in ROA, accruals, earnings quality
ni	Net Income	Bottom-line earnings; used in ROE, payout ratios
oiadp	Operating Income After D&A	EBIT proxy; operating earnings after depreciation
oibdp	Operating Income Before D&A	EBITDA proxy; used in enterprise multiples
pi	Pre-Tax Income	Pre-tax income; effective tax rate = txt/pi

Continued on next page

Appendix table 1 continued

Mnemonic	Name	Description
`revt`	Revenue Total	Total revenue; top-line growth and profitability numerator
`sale`	Net Sales	Net sales/turnover; slightly different from revt for some firms
`spi`	Special Items	Non-recurring items; earnings quality indicator
`txt`	Total Income Taxes	Income tax expense; tax burden signals
`xad`	Advertising Expense	Advertising spending; brand capital anomaly, often missing
`xint`	Interest Expense	Interest on debt; used in operating profitability (FF 2015)
`xrd`	R&D Expense	R&D spending; R&D intensity is positively priced (intangibles anomaly)
`xsga`	SG&A Expense	Selling, general & admin; operating profit = GP - xsga
Cash Flow & Financing (17 items)		
`aoloch`	Change in Other Assets/Liabilities	Net change in other operating items; accruals decomposition
`apalch`	Change in Accounts Payable	Increase (decrease) in AP; accruals decomposition component
`aqc`	Acquisitions	Cash spent on acquisitions; M&A-driven vs organic growth
`capx`	Capital Expenditures	Capital spending; investment intensity, free cash flow = oancf - capx
`dltis`	LT Debt Issuance	Long-term debt issuance; net debt issuance anomaly
`dltr`	LT Debt Retirement	Long-term debt retirement; net debt change = dltis - dltr
`dpc`	Depreciation (CF Statement)	D&A from cash flow statement; may differ from income statement dp
`dvc`	Common Dividends	Cash dividends on common stock; dividend yield, payout ratio
`dvt`	Total Dividends	Total dividends including preferred; total payout measure
`fopt`	Funds from Operations	Funds from operations; fallback for oancf in pre-1988 data

Continued on next page

Appendix table 1 continued

Mnemonic	Name	Description
`invch`	Change in Inventories	Decrease (increase) in inventories; accruals decomposition component
`oancf`	Operating Cash Flow	Net cash from operations; cash ROA = oancf/at, accruals = ib - oancf
`prstkc`	Purchase of Stock	Share repurchases; net payout = dvc + prstkc - sstk
`recch`	Change in Receivables	Decrease (increase) in receivables; accruals decomposition component
`sstk`	Sale of Stock	Equity issuance; net issuance predicts low returns
`txach`	Change in Taxes Payable	Increase (decrease) in taxes payable; accruals decomposition
`wcap`	Working Capital	Working capital from balance sheet; alternative accruals base

All 66 Compustat annual items available to the AI research agent. Each variable is annotated with its financial statement category, a plain-English name, and common scaling denominators (not shown). BS = Balance Sheet, IS = Income Statement, CF = Cash Flow and Financing.

Appendix Table 2: Operator Catalog

Operator	Arity	Params	Description
Tier 1: Algebra & Scaling			
`ADD`, `SUB`, `MUL`	2	—	Arithmetic
`RATIO`	2	—	a/b (division)
`SCALE`	2	—	a/b (alias)
Tier 2: Composite			
`SUMRANK`	2	—	Within-year rank(a) + rank(b) − 1
Tier 3: Temporal			
`GROWTH`	1	lag	$(x_t - x_{t-\text{lag}})/\lvert x_{t-\text{lag}}\rvert$
`DELTA`	1	lag	$x_t - x_{t-\text{lag}}$
`LAG`	1	d	x_{t-d}
`MA`	1	k	k-year moving average
`VOL`	1	k	k-year rolling std. dev.
`TREND`	1	k	Linear slope over k years
`ACCEL`	1	—	Second difference
Tier 4: Cross-Sectional & Industry			
`IND_ADJ`	1	—	x − industry median(x)
`IND_ZSCORE`	1	—	$(x - \mu_{\text{ind}})/\sigma_{\text{ind}}$
`ZSCORE`	1	—	Cross-sectional z-score
`PERCENTILE`	1	—	Cross-sectional percentile rank
Tier 5: Nonlinear & Robustness			
`LOG`, `ABS`, `NEG`, `INV`	1	—	Math transforms
`SIGN`	1	—	Sign function
`INDICATOR`	1	thresh	$\mathbf{1}[x > \text{thresh}]$
`WINSOR`	1	—	1%–99% winsorization

All operators compose recursively up to nesting depth 6. Industry classifications use two-digit SIC codes. Temporal operators compute within-firm across fiscal years.

Appendix B Additional Tables

This appendix collects supplementary tables referenced in the main text. Appendix table 3 lists all 38 horse-race survivors with their symbolic expressions, theme pairs, and key metrics. Appendix table 4 reports Bayesian inference results for each survivor. Appendix tables 5–9 provide generation-by-generation statistics, top signals by Sharpe ratio, subsample robustness, novelty classification, and reasoning taxonomy results.

Appendix Table 3: All 38 Horse-Race Survivors: Definitions and Metrics

Signal	Pair	t^{FMB}	Cond. t	Sharpe
LowLeverageValue	Val×Qual	4.47	6.43	0.30
SUB(RATIO(ceq, MUL(prcc_f, csho)), RATIO(ADD(dltt, dlc), at))				
GrossMarginTimesBM	Val×Qual	2.49	5.66	0.01
MUL(RATIO(SUB(revt, cogs), revt), RATIO(ceq, MUL(prcc_f, csho)))				
SmoothedCashAdjBM_ShareCountAd	Val×Fin	4.05	5.49	0.20
SUB(MA(RATIO(ADD(ceq, che), MUL(prcc_f, csho)), 3), GROWTH(csho, 1))				
LeveragedPPEGrowth	Inv×Fin	5.39	5.36	0.10
NEG(MUL(RATIO(ADD(dltt, dlc), at), GROWTH(ppent, 1)))				
InventoryGrowthTimesDebtGrowth	Inv×Fin	3.97	4.84	0.14
NEG(MUL(GROWTH(invt, 1), GROWTH(ADD(dltt, dlc), 1)))				
RnDAdjustedBM	Val×Qual	3.87	-4.78	0.07
RATIO(ADD(ceq, xrd), MUL(prcc_f, csho))				
UltimateAlphaComposite	Inv×Fin	4.42	-4.69	0.41
SUMRANK(NEG(MUL(WINSOR(RATIO(ADD(dltt, dlc), ceq)), MA(GROWTH(at, 1), 3))), NEG(MUL(GROWTH(capx, 1), RATIO(SUB(dltis, dltr), at))))				
PersistentLeveragedGrowth	Inv×Fin	4.93	4.59	0.30
NEG(MUL(RATIO(ADD(dltt, dlc), at), MA(GROWTH(at, 1), 3)))				
GrossProfitYield	Val×Qual	3.63	4.55	0.36
RATIO(SUB(revt, cogs), MUL(prcc_f, csho))				
LeveragedGrossPPEGrowth	Inv×Fin	4.61	-4.50	-0.01
NEG(MUL(RATIO(ADD(dltt, dlc), at), GROWTH(ppegt, 1)))				
CapexGrowthTimesDebtIssuance	Inv×Fin	2.88	4.45	0.39
NEG(MUL(GROWTH(capx, 1), RATIO(SUB(dltis, dltr), at)))				
InvGrowth_SmoothedGP_Composite	Acc×Qual	3.07	3.98	0.25
SUMRANK(NEG(GROWTH(invt, 1)), MA(RATIO(SUB(revt, cogs), at), 3))				
GPAT_RDME_Composite	Prof×Int	4.90	3.82	0.47
SUMRANK(RATIO(SUB(revt, cogs), at), RATIO(xrd, MUL(prcc_f, csho)))				
StableGP_RDME_Composite	Prof×Int	4.31	3.76	0.27
SUMRANK(NEG(VOL(RATIO(SUB(revt, cogs), at), 5)), RATIO(xrd, MUL(prcc_f, csho)))				
MarketLeveragedGrowth	Inv×Fin	3.26	-3.75	0.22
NEG(MUL(RATIO(ADD(dltt, dlc), MUL(prcc_f, csho)), GROWTH(at, 1)))				
ProfitImprovementCashCow	Prof×Inv	5.49	3.73	0.46

Continued on next page

Table 3 continued

Signal	Pair	t^{FMB}	Cond. t	Sharpe
`SUMRANK(SCALE(DELTA(SUB(revt, cogs), 1), at), NEG(GROWTH(at, 1)))`				
`ConditionalQualityBM`	Val×Qual	4.06	-3.56	0.38
`MUL(PERCENTILE(RATIO(ceq, MUL(prcc_f, csho))), PERCENTILE(RATIO(SUB(revt, cogs), at)))`				
`TangibleBookToMarket`	Val×Qual	2.54	3.55	0.04
`RATIO(SUB(ceq, gdwl), MUL(prcc_f, csho))`				
`IndustryAdjCashAdjBMDebtRet`	Val×Fin	5.49	3.48	0.30
`IND_ADJ(ADD(RATIO(ADD(ceq, che), MUL(prcc_f, csho)), RATIO(SUB(dltr, dltis), at)))`				
`Industry_Adjusted_GP_to_Price`	Prof×Val	5.01	3.43	0.26
`IND_ADJ(RATIO(SUB(revt, cogs), MUL(prcc_f, csho)))`				
`LiquidityAdjustedBM`	Val×Qual	4.77	-3.24	0.32
`MUL(RATIO(ceq, MUL(prcc_f, csho)), RATIO(act, lct))`				
`DoublePenaltyCashAdjBM`	Val×Fin	4.36	-3.02	0.29
`SUB(RATIO(ADD(ceq, che), MUL(prcc_f, csho)), ADD(GROWTH(csho, 1), RATIO(SUB(dltis, dltr), at)))`				
`MultiplicativeGrowthIssuanceIn`	Inv×Fin	4.12	-2.99	0.03
`NEG(MUL(GROWTH(at, 1), RATIO(ADD(SUB(sstk, prstkc), SUB(dltis, dltr)), at)))`				
`PPEProductivityComposite`	Prof×Inv	4.51	2.95	0.36
`SUMRANK(RATIO(SUB(revt, cogs), ppent), NEG(GROWTH(ppent, 1)))`				
`GPlessDepCapxToEV`	Val×Qual	3.53	2.88	0.51
`RATIO(SUB(SUB(SUB(revt, cogs), dp), capx), ADD(MUL(prcc_f, csho), SUB(ADD(dltt, dlc), che)))`				
`AcquisitionShareOfIssuance`	Inv×Fin	2.58	2.87	-0.22
`NEG(RATIO(aqc, ADD(dltis, sstk)))`				
`GP_Trend_RDME_Composite`	Prof×Int	2.84	-2.86	0.21
`SUMRANK(TREND(RATIO(SUB(revt, cogs), at), 5), RATIO(xrd, MUL(prcc_f, csho)))`				
`ConditionalLeveragedGrowth`	Inv×Fin	3.57	2.82	0.12
`NEG(MUL(PERCENTILE(GROWTH(at, 1)), PERCENTILE(RATIO(ADD(dltt, dlc), at))))`				
`AccrualGrowthComposite`	Inv×Acc	2.85	2.61	-0.06
`SUMRANK(NEG(RATIO(SUB(ib, oancf), at)), NEG(GROWTH(at, 1)))`				
`LeveragedAssetGrowth`	Inv×Fin	4.35	2.57	0.26
`NEG(MUL(RATIO(ADD(dltt, dlc), at), GROWTH(at, 1)))`				

Continued on next page

Table 3 continued

Signal	Pair	t^{FMB}	Cond. t	Sharpe
`CashCowSpread`	Prof×Inv	3.86	2.55	0.23
`SUB(RATIO(SUB(revt, cogs), at), GROWTH(at, 1))`				
`ConditionalSP_Quality`	Val×Qual	3.98	-2.48	0.54
`MUL(PERCENTILE(RATIO(sale, MUL(prcc_f, csho))), PERCENTILE(SUB(RATIO(SUB(revt, cogs), at), RATIO(ADD(dltt, dlc), at))))`				
`QualityValueComposite`	Val×Qual	4.89	-2.40	0.48
`SUMRANK(RATIO(ceq, MUL(prcc_f, csho)), SUB(RATIO(SUB(revt, cogs), at), VOL(RATIO(ib, at), 5)))`				
`SUMRANK_GPAT_BlendedCashAdjVal`	Prof×Val	5.59	2.38	0.53
`SUMRANK(RATIO(SUB(revt, cogs), at), ADD(ZSCORE(RATIO(ADD(ceq, che), MUL(prcc_f, csho))), ZSCORE(RATIO(sale, MUL(prcc_f, csho)))))`				
`NegativeAccrualsToPrice`	Val×Qual	2.03	2.22	0.21
`RATIO(SUB(oancf, ib), MUL(prcc_f, csho))`				
`GPtoTotalClaims`	Val×Qual	4.04	-2.03	0.47
`RATIO(SUB(revt, cogs), ADD(MUL(prcc_f, csho), lt))`				
`GPlessDepToTotalClaims`	Val×Qual	3.80	-2.01	0.46
`RATIO(SUB(SUB(revt, cogs), dp), ADD(MUL(prcc_f, csho), lt))`				
`ProfitVsDebtExpansion`	Prof×Inv	2.44	-2.01	0.40
`SUB(RATIO(SUB(revt, cogs), at), SCALE(DELTA(ADD(dltt, dlc), 1), at))`				

Sorted by conditional $|t|$ from the within-pair horse race (descending). Each signal's formula is shown below its name row. t^{FMB}: univariate Fama–MacBeth t-statistic (Newey–West, 6 lags). Cond. t: conditional t controlling for all other survivors in the same theme pair. Sharpe: annualized long–short (Q5−Q1) Sharpe ratio. Expressions use the symbolic language defined in Section 3.1. Variable mnemonics follow Compustat conventions (see Table 1). Key operators: RATIO = division, SUMRANK = composite of within-year ranks, GROWTH(x,k) = k-year growth rate, MA(x,k) = k-year moving average, NEG = negation, IND_ADJ = industry-median adjustment, MUL = multiplication.

Appendix Table 4: Bayesian Inference for All Horse-Race Survivors

			$\Pr(H_0\|\text{data})$		
Signal	$\|t\|$	MBF	$\pi_0 = 5\%$	$\pi_0 = 20\%$	$\pi_0 = 50\%$
SUMRANK_GPAT_BlendedCashAdjV	5.59	1.7e-07	<0.001	<0.001	<0.001
IndustryAdjCashAdjBMDebtRet	5.49	2.9e-07	<0.001	<0.001	<0.001
ProfitImprovementCashCow	5.49	2.9e-07	<0.001	<0.001	<0.001
LeveragedPPEGrowth	5.39	5.0e-07	<0.001	<0.001	<0.001
Industry_Adjusted_GP_to_Pric	5.01	3.6e-06	<0.001	<0.001	<0.001
PersistentLeveragedGrowth	4.93	5.2e-06	<0.001	<0.001	<0.001
GPAT_RDME_Composite	4.90	6.3e-06	<0.001	<0.001	<0.001
QualityValueComposite	4.89	6.4e-06	<0.001	<0.001	<0.001
LiquidityAdjustedBM	4.77	1.2e-05	<0.001	<0.001	<0.001
LeveragedGrossPPEGrowth	4.61	2.4e-05	<0.001	<0.001	<0.001
PPEProductivityComposite	4.51	3.9e-05	<0.001	<0.001	<0.001
LowLeverageValue	4.47	4.5e-05	<0.001	<0.001	<0.001
UltimateAlphaComposite	4.42	5.7e-05	0.001	<0.001	<0.001
DoublePenaltyCashAdjBM	4.36	7.4e-05	0.001	<0.001	<0.001
LeveragedAssetGrowth	4.35	7.9e-05	0.001	<0.001	<0.001
StableGP_RDME_Composite	4.31	9.4e-05	0.002	<0.001	<0.001
MultiplicativeGrowthIssuance	4.12	2.1e-04	0.004	<0.001	<0.001
ConditionalQualityBM	4.06	2.6e-04	0.005	0.001	<0.001
SmoothedCashAdjBM_ShareCount	4.05	2.7e-04	0.005	0.001	<0.001
GPtoTotalClaims	4.04	2.9e-04	0.005	0.001	<0.001
ConditionalSP_Quality	3.98	3.6e-04	0.007	0.001	<0.001
InventoryGrowthTimesDebtGrow	3.97	3.7e-04	0.007	0.001	<0.001
RnDAdjustedBM	3.87	5.5e-04	0.010	0.002	<0.001
CashCowSpread	3.86	5.9e-04	0.011	0.002	<0.001
GPlessDepToTotalClaims	3.80	7.4e-04	0.014	0.003	<0.001
GrossProfitYield	3.63	0.0014	0.026	0.006	0.001
ConditionalLeveragedGrowth	3.57	0.0017	0.032	0.007	0.002
GPlessDepCapxToEV	3.53	0.0020	0.036	0.008	0.002
MarketLeveragedGrowth	3.26	0.0049	0.085	0.019	0.005
InvGrowth_SmoothedGP_Composi	3.07	0.0089	0.144	0.034	0.009
CapexGrowthTimesDebtIssuance	2.88	0.0158	0.231	0.059	0.016
AccrualGrowthComposite	2.85	0.0174	0.249	0.065	0.017
GP_Trend_RDME_Composite	2.84	0.0178	0.252	0.066	0.017

Continued on next page

Appendix table 4 continued

			$\Pr(H_0\|\text{data})$		
Signal	$\|t\|$	MBF	$\pi_0 = 5\%$	$\pi_0 = 20\%$	$\pi_0 = 50\%$
`AcquisitionShareOfIssuance`	2.58	0.0357	0.404	0.125	0.035
`TangibleBookToMarket`	2.54	0.0397	0.430	0.137	0.038
`GrossMarginTimesBM`	2.49	0.0449	0.460	0.152	0.043
`ProfitVsDebtExpansion`	2.44	0.0516	0.495	0.171	0.049
`NegativeAccrualsToPrice`	2.03	0.1264	0.706	0.336	0.112
$\Pr(H_0\|\text{data}) < 5\%$:			28/38	30/38	37/38

Sorted by $|t|$ descending. MBF $= e^{-t^2/2}$ is the minimum Bayes factor—the maximum evidence the data provide against the null under any alternative prior (Harvey, 2017). π_0 is the prior probability that the effect is true. $\Pr(H_0|\text{data})$ is the posterior probability that the null is true, computed as $\text{MBF} \times (1-\pi_0)/\pi_0 \;/\; (1 + \text{MBF} \times (1-\pi_0)/\pi_0)$. Under the skeptical prior ($\pi_0 = 5\%$), the framework effectively requires $|t| \gtrsim 3.4$ for the posterior to drop below 5%.

Appendix Table 5: Signal Quality by Generation

Gen	Tested	Sig.	Hit Rate	Best $\lvert t \rvert$
0	34	15	44%	4.47
1	36	13	36%	5.01
2	40	20	50%	6.14
3	40	22	55%	6.26
4	40	30	75%	5.97
5	40	31	78%	6.23
6	40	28	70%	6.16

Aggregated across all 8 theme pairs within each generation. Gen 0 is the initial hypothesis generation (no prior feedback); Gen 1–6 are reflection-based evolutionary iterations where the LLM receives all prior results and its own extended thinking chain before proposing new signals. Hit rate excludes error signals. The improvement from Gen 0 to Gen 4–5 reflects the LLM learning which variable combinations, operators, and economic mechanisms produce significant cross-sectional return predictability.

Appendix Table 6: Top 10 Discovered Signals by Sharpe Ratio

	Signal	Pair	Gen	t^{FMB}	t^{α}	Sharpe
1	`ConditionalSP_Quality`	Val×Qual	4	3.98	2.52	0.54
2	`SUMRANK_GPAT_BlendedCashAdjValue`	Prof×Val	6	5.59	2.77	0.53
3	`GPlessDepCapxToEV`	Val×Qual	5	3.53	2.42	0.51
4	`QualityValueComposite`	Val×Qual	2	4.89	2.13	0.48
5	`GPAT_RDME_Composite`	Prof×Int	2	4.90	3.92	0.47
6	`GPtoTotalClaims`	Val×Qual	4	4.04	2.78	0.47
7	`ProfitImprovementCashCow`	Prof×Inv	5	5.49	3.28	0.46
8	`GPlessDepToTotalClaims`	Val×Qual	5	3.80	2.76	0.46
9	`UltimateAlphaComposite`	Inv×Fin	6	4.42	2.60	0.41
10	`ProfitVsDebtExpansion`	Prof×Inv	5	2.44	4.09	0.40

All signals are horse-race survivors (conditional $|t| > 1.96$ controlling for all other survivors in the same theme pair). t^{FMB}: univariate Fama–MacBeth t-statistic (Newey–West, 6 lags). t^{α}: Carhart four-factor (Mkt-RF, SMB, HML, Mom) long–short quintile alpha t-statistic. Sharpe: annualized Sharpe ratio of the Q5−Q1 value-weighted portfolio. NYSE breakpoints, microcap-excluded universe, CRSP/Compustat 1963–2024.

Appendix Table 7: Subsample Robustness and Decay Analysis

Signal	Full	Pre-00	Post-00	Post-10	Ex-Rec	#Rob	Δ
LeveragedPPEGrowth	**9.94**	**8.06**	**5.78**	**4.22**	**9.42**	7	S
IndustryAdjCashAdjBMDebtR	**8.11**	**5.91**	**5.61**	**3.63**	**7.49**	7	S
LiquidityAdjustedBM	**8.03**	**6.61**	**4.93**	**2.85**	**7.16**	7	S
LeveragedAssetGrowth	**7.99**	**5.84**	**5.66**	**3.80**	**7.26**	7	S
LeveragedGrossPPEGrowth	**7.77**	**6.00**	**4.96**	**4.06**	**7.09**	7	S
PersistentLeveragedGrowth	**7.70**	**5.40**	**5.91**	**4.14**	**7.07**	7	+
AcquisitionShareOfIssuanc	**7.02**	**4.58**	**5.76**	**4.82**	**6.69**	7	+
ConditionalLeveragedGrowt	**6.67**	**4.28**	**5.89**	**5.25**	**5.89**	7	+
MarketLeveragedGrowth	**6.57**	**4.29**	**5.50**	**4.15**	**6.10**	7	+
UltimateAlphaComposite	**6.07**	**4.13**	**4.79**	**3.17**	**5.21**	7	S
RnDAdjustedBM	**5.85**	**4.08**	**4.28**	**2.97**	**5.54**	7	S
InventoryGrowthTimesDebtG	**3.60**	**2.45**	**2.52**	**1.97**	**2.79**	7	S
Robust in all 7: 12/38 ≥6: 22/38 Stable: 15 Decaying: 17 Strengthening: 6							

Top 12 signals ranked by number of robust subsamples, then $|t^{FMB}|$. t-statistics in bold where $|t| > 1.96$. Columns: Full (1963–2024), Pre-00 (1963–2000), Post-00 (2000–2024), Post-10 (2011–2024), Ex-Rec (excluding NBER recession months). #Rob: count of 7 subsamples (full, pre-anomaly 1963–1990, post-publication 1991+, pre-2000, post-2000, post-2010, ex-recession) in which $|t| > 1.96$. Δ: decay classification from 60-month rolling Fama–MacBeth t-statistics with OLS trend test (S=stable $|t_{trend}| \leq 1.96$, D=decaying $t_{trend} < -1.96$, +=strengthening $t_{trend} > 1.96$). All with microcap exclusion.

Appendix Table 8: Novelty Classification of Proposed Signals

Category	Count	Successful	Hit Rate
Novel combination	198	123	62%
Variant of known	71	34	48%
Direct rediscovery	11	2	18%

"Direct rediscovery": expression closely matches a published anomaly (e.g., GP/AT $\approx$ Novy-Marx 2013). "Variant of known": same economic mechanism but different functional form or variable choice. "Novel combination": known building blocks (variables, operators) combined in a previously undocumented way. Classification based on mapping signal expressions and hypothesis text to published anomaly lists from the Chen–Zimmermann (2022) database.

Appendix Table 9: AI Reasoning Taxonomy × Outcome

Category	Signals	Successful	Hit Rate
Mechanism combination	246	144	59%
Temporal enrichment	169	102	60%
Variable substitution	72	43	60%
Novel economic insight	58	27	47%
Failure inversion	33	21	64%
Industry adjustment	19	14	74%
Literature grounding	48	23	48%
Analogical reasoning	26	19	73%
Total	**671**	**393**	**59%**

Each signal is coded into one or more reasoning categories based on the hypothesis and reasoning text from the LLM trace (a signal may belong to multiple categories, so column totals exceed the number of unique signals). "Mechanism combination": combining variables from two economic themes. "Temporal enrichment": adding MA, TREND, VOL, or GROWTH operators. "Industry adjustment": using IND_ADJ or IND_ZSCORE. "Failure inversion": flipping the sign or modifying a previously failed signal. Hit rate = successful / total (including errors).

Appendix C Additional Case Studies

This appendix presents four additional case studies from the discovery traces, complementing the two main case studies in Section 7.

C.1 The R&D Intensity Reversal

In the Profitability×Intangibles pair, the agent proposed two Generation 0 signals based on R&D *productivity*—gross profit per R&D dollar—expecting that firms extracting more output per research dollar would earn higher returns. Both came back significantly wrong-signed ($t \approx -2.0$).

The Generation 1 thinking trace captures the diagnosis: firms with *low* GP/xrd (high R&D intensity relative to profits) earn higher returns, consistent with the intangibles literature rather than the agent's initial productivity framing. Rather than simply negating the expression, the agent restructured it in Generation 2 as `GPAT_RDME_Composite`: a `SUMRANK` of gross profitability and R&D-to-market-equity, combining both channels additively rather than as a ratio where one dominates. The composite achieved $t^{FMB} = 4.90$ and FF5 α $t = 3.24$.

C.2 The Persistence Discovery

After the leverage-stock breakthrough (Section 7.1) dominated Generations 2–3 of the Investment×Financing pair, Generation 4 introduced temporal smoothing. The agent reasoned that firms sustaining high asset growth over three years are "persistent empire-builders whose overinvestment is structural rather than transitory." The resulting `PersistentLeveragedGrowth` used a 3-year moving average of asset growth interacted with leverage, jumping from $t = 4.35$ (single-year) to $t = 4.93$ (MA-smoothed).

This signal is one of only 6 classified as *strengthening* in the decay analysis—its predictive power is increasing over time, with post-2010 $t = 4.14$ compared to pre-2000 $t = 5.40$. It is robust in all 7 subsamples. However, its FF5 alpha is marginal ($t = 1.66$) and CZ spanning fails ($t^{cond} = 1.47$), suggesting that the persistence insight, while temporally robust, is partially captured by existing investment anomalies. The distinction between transient and persistent overinvestment is not formalized in any published factor model, making this a candidate for future theoretical development.

C.3 Three Failures and a Meta-Rule

The Investment×Financing pair saw three consecutive failures across Generations 0–2 when the agent used capex *intensity* (RATIO(capx, at)) as the investment component. Each attempt produced weak or insignificant results. By Generation 3, the agent's reflection trace articulated a structural rule: "investment GROWTH always outperforms investment LEVEL—every successful signal uses GROWTH while every signal using RATIO(capx, at) fails."

This meta-learning led to `CapexGrowthTimesDebtIssuance`, which replaced capex intensity with capex *growth*. The signal achieved $t^{FMB} = 2.88$ and FF5 α $t = 2.04$. However, it is the weakest of the case study

signals—it fails BHY correction, is significant in only 3 of 7 subsamples, and has near-zero conditional t in the CZ spanning test. The meta-learning story is compelling as an illustration of the discovery process, but the signal itself does not survive rigorous scrutiny.

C.4 Sales-to-Price Beats Book-to-Market

In the Valuation×Quality pair, the agent tried several valuation–quality composites across Generations 2–3. `GPtoEV_StabilityComposite`, which combined GP/EV with earnings stability, failed ($t = 0.53$). The agent diagnosed the problem: earnings volatility and GP/EV both capture quality, making them non-orthogonal as composite ingredients.

Generation 4 proposed `ConditionalSP_Quality`, using *sales-to-price* as the valuation base: "SP and the quality score are more independent since SP uses sales while quality uses gross margin and leverage." The signal achieved the highest Sharpe ratio in the study (0.54) and FF4 α $t = 2.52$. The rigorous tests temper the enthusiasm. The signal is decaying (significant trend $t < -1.96$) and significant in only 4 of 7 subsamples. The orthogonality diagnosis, that SP is more independent of quality measures than BM, is intellectually valuable and verifiable, but the resulting signal's economic content maps onto known value and profitability premia that are weakening over time.

Appendix D Robustness Details

D.1 Bayesian Inference Details

Appendix table 4 reports the minimum Bayes factor (MBF) and Bayesianized posterior probability $\Pr(H_0|\text{data})$ for each horse-race survivor under three prior specifications, following Harvey (2017). Under the skeptical prior ($\pi_0 = 5\%$), the Bayesian framework effectively requires $|t| \gtrsim 3.4$ for the posterior to drop below 5%. Twenty-eight of 38 survivors clear this bar, confirming that the signals are credible even under priors that place 19-to-1 odds against the effect being real.

D.2 Placebo Null Distribution

We construct an empirical null distribution using 113 placebo signals from the Chen and Zimmermann (2022) repository. These placebos are randomly constructed characteristics (quarterly variants, lagged transformations) with no a priori expected predictive power. We evaluate each placebo using the same simplified Fama–MacBeth estimator used in the permutation tests, on the microcap-excluded return panel.

The placebo null distribution has mean $|t| = 2.83$, 95th percentile $|t| = 7.95$, and 99th percentile $|t| = 9.43$. These values are substantially higher than the theoretical χ^2 null ($|t_{95}| \approx 1.96$), reflecting the fact that CZ placebos are not pure noise—they include quarterly variants of real signals that may retain residual cross-sectional structure. Zero of the 38 horse-race survivors exceed the placebo 95th percentile, indicating that the CZ placebos define an extremely conservative benchmark. We therefore treat the placebo null as a supplementary diagnostic rather than a primary significance filter.

D.3 Generation-by-Generation Pair-Level Breakdown

Appendix table 5 aggregates hit rates across all 8 pairs. For completeness, the online code repository provides the full 8×7 matrix of per-pair, per-generation hit rates and best $|t|$-statistics, along with the per-generation reflection summaries extracted from the LLM traces.