AI & QUANTITATIVE STRATEGY RESEARCH · JUNE 2026

One Prompt Won’t Solve the Market

LLMs, the “Medallion” Meme, and a Realistic Playbook for AI-Assisted Strategy Research

We took the viral “make medallion fund” prompt at its word, built the textbook strategy it implies, and measured what actually happens. Spoiler: a single dollar becomes two cents. This report explains why — and lays out where frontier language models genuinely earn their keep in a quant research workflow.

PREPARED FOR

Portfolio Managers

Hedge fund quantitative desk

PREPARED BY

Perplexity Computer

for Jiri Pik / RocketEdge

Executive Summary

- The meme is funny because it's wrong. A widely shared LinkedIn post claimed a frontier model recreated Renaissance Technologies' Medallion Fund from the single prompt "make medallion fund," complete with a chart drowning in overlaid trendlines. We ran the implied strategy for real. It does not work.
- Our replication: $1 → $0.02. A textbook daily cross-sectional mean-reversion strategy on 20 US large caps was roughly flat gross (Sharpe 0.24) and catastrophic net of realistic costs (Sharpe −3.04, −97.7% max drawdown). Medallion's pace over the same window would have turned $1 into roughly $195.
- The edge is execution and data, not a describable recipe. Medallion's ~66% gross / ~39% net annualized returns (1988–2018) come from decades of curated tick data, microstructure-aware execution, capacity discipline, and institutional learning — none of which an LLM can emit in a code block.
- LLMs add value as inputs, not as portfolio managers. Peer-reviewed work shows LLM headline-sentiment scores predict out-of-sample returns; broad backtests show autonomous LLM strategies decay badly. Use the model for feature extraction and research velocity, not for running the book.
- Realistic hit rate: ~15% of generated strategies survive costs. That is comparable to skilled human idea generation — but produced in hours, not weeks. The defensible edge is testing 100× more hypotheses with rigorous guardrails, not magical alpha.
- Govern the new risk surface. Model risk, crowding (everyone prompting the same models), IP leakage into public endpoints, and explainability for IC and compliance are now first-order concerns. A workflow with human-in-the-loop gates and deflated-Sharpe validation is the actual playbook.

Bottom line. Frontier LLMs are a genuine force multiplier for quantitative research, but they multiply the quality of your process, your data, and your execution — they do not substitute for them. There is no prompt for an edge you do not otherwise possess.

SECTION 1

The Meme and What We Tested

The artifact at the center of this report is a satirical LinkedIn post in which a fictional frontier model — cheekily branded "Claude Fable 5" — is asked to "make medallion fund" and obligingly produces an equity curve buried under dozens of overlaid trendlines and Gann fans. The joke lands because it compresses two real anxieties of the moment: that frontier LLMs can conjure proprietary alpha on command, and that they will happily overfit anything you point them at. Both deserve a serious answer, so we gave the prompt a fair, literal trial.

We implemented the strategy the prompt most plausibly implies — the canonical "Medallion-adjacent" textbook recipe: a daily cross-sectional mean-reversion book that goes long the bottom quintile and short the top quintile of prior-day returns, equal-weighted, dollar-neutral, with gross exposure of 1. The universe is 20 US large caps; the window is January 2016 through June 2026 on daily closes; transaction costs are modeled at a deliberately modest 10 basis points per side.

Backtest results

Variant	CAGR	Ann. vol	Sharpe	Max DD	Avg daily turnover
Gross (zero costs)	+2.15%	11.58%	0.24	−25.6%	151% of book
Net (10 bps / side)	−30.21%	11.59%	−3.04	−97.7%	151% of book
Medallion reference (net of fees)	~+66% / yr*	—	—	—	—

*Gross figure for the Medallion reference; ~39% net of fees historically. Medallion data from Quartr[1] and Cornell Capital[2].

Growth of $1 over the window tells the story bluntly: at Medallion's historical pace, $1 compounds to roughly $195; our naive net toy strategy decays to about $0.02.[1]

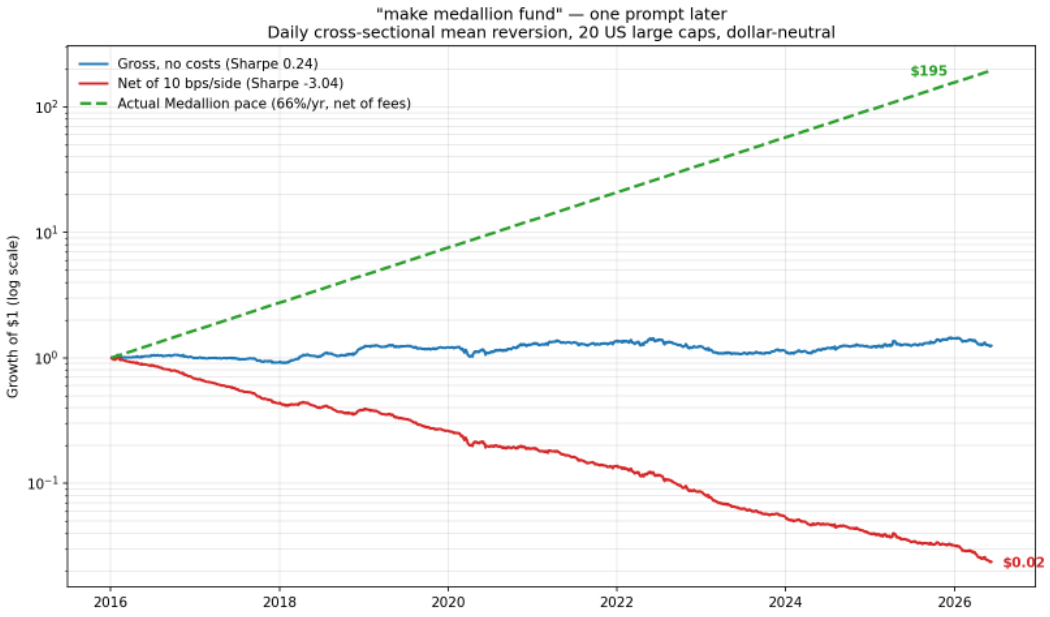

Figure 1. Growth of $1, log scale. The gross signal (blue) is essentially flat; realistic costs at 151% daily turnover drag the net curve (red) toward zero, while Medallion's historical pace (green, dashed) compounds to ~$195. Same idea, three very different destinies.

Interpretation

The result is not a bug; it is the entire lesson. The naive signal is approximately flat gross of costs, which means it carries almost no exploitable edge to begin with. Layer on 151% average daily turnover at 10 bps per side and the friction alone is more than enough to convert a flat line into a slow-motion liquidation. The prompt produces plausible, runnable code — and a portfolio that loses money with quiet determination. Medallion's edge was never the describable strategy; it lived in execution, data, and the breadth of thousands of weak signals combined with discipline.

1. Quartr — Renaissance Technologies and the Medallion Fund.
https://quartr.com/insights/edge/renaissance-technologies-and-the-medallion-fund
2. Cornell Capital — The Medallion Fund: The Ultimate Counterexample.
https://www.cornell-capital.com/blog/medallion-fund-the-ultimate-counterexample/

SECTION 2

Why the Prompt Fails: The Anatomy of Medallion's Edge

To see why no prompt can replicate Medallion, it helps to decompose where the fund's edge actually resides and ask, component by component, whether a language model could supply it. The honest answer is "no" for every structural source of advantage. An LLM is a superb writer of strategy code; it owns none of the assets that make the strategy work.

Medallion by the numbers

Metric	Figure	Note
Gross annualized return (1988–2018)	~66%	Before fees
Net annualized return (1988–2018)	~39%	After 5-and-44 fee structure
Cumulative net to 2022	90,000×+	Compounded, employee capital
$100 in 1988 → 2018 (Cornell)	$398.7M	"Stretches explanation to the limit"
Average leverage	~12.5×	Up to ~20×
Capacity	~$10–15B	Profits distributed to maintain cap
Hit rate on individual bets	barely >50%	Across millions of trades; Kelly sizing

Sources: Zuckerman via Quartr[1]; Cornell Capital[2]; Institutional Investor[3]; Renaissance Technologies overview[4].

Can an LLM replicate it?

Edge component	What it really is	Can an LLM replicate it?
Proprietary data moat	Decades of curated, cleaned tick data & alternative datasets	No — the data is the asset; a model has only its training corpus
Execution & microstructure	Alpha roughly breakeven before costs; edge is in how orders are placed	No — requires venue access, latency, queue position, fills
Capacity discipline	Hard cap ~$10–15B; profits distributed to protect returns	No — a business & governance choice, not code
Talent & secrecy	Extreme retention; ideas never leave the building	No — public models share knowledge with everyone
Risk management & sizing	Kelly-style sizing over thousands of weak, decorrelated signals	Partly — can scaffold the math; cannot supply the signals or data

Read the right-hand column top to bottom and the verdict is plain. The edge is a stack of institutional assets — data, infrastructure, governance, people — that compound over decades. A model can describe a mean-reversion strategy in seconds, and that is precisely why the description is worthless: anything a public model can emit on demand is, by definition, not proprietary.

1. Quartr — Renaissance Technologies and the Medallion Fund. https://quartr.com/insights/edge/renaissance-technologies-and-the-medallion-fund
2. Cornell Capital — The Medallion Fund: The Ultimate Counterexample. https://www.cornell-capital.com/blog/medallion-fund-the-ultimate-counterexample/
3. Institutional Investor — Famed Medallion Fund 'Stretches Explanation to the Limit'. https://www.institutionalinvestor.com/article/2bswymr8cih3jeaslxc00/portfolio/famed-medallion-fund-stretches-explanation-to-the-limit-professor-claims
4. Wikipedia — Renaissance Technologies. https://en.wikipedia.org/wiki/Renaissance_Technologies

SECTION 3

What "Profitable" Realistically Looks Like

If the prompt-as-portfolio-manager fantasy is dead, the useful question becomes: where do frontier LLMs add measurable value in a quant workflow today, and where do they reliably disappoint? The evidence base is now large enough to answer with specifics rather than vibes.

Where LLMs add value — and where they fail

Use case	Verdict	Evidence / rationale
Signal extraction from text, news, earnings calls	Adds value	Headline-sentiment scores predict out-of-sample daily returns, stronger in small caps and after negative news[6]
Research velocity (hypothesis generation)	Adds value	500 generated strategies tested in ~2 hours vs weeks manually[7]
Code generation & data cleaning	Adds value	Code-assistant use documented at Man Group, Two Sigma[9]
Literature mining / RAG over archives	Adds value	Retrieval over research archives a leading adoption pattern[9]
Autonomous trading / running the book	Fails	Reported advantages deteriorate under broader, longer evaluation[5]
Regime detection	Fails	Too conservative in bull markets, too aggressive in bears[5]
Novel alpha from memorized patterns	Fails	Textbook patterns (MACD, RSI) are worst out-of-sample[7]

The pattern is consistent: LLMs are valuable as an input to the research process — a feature extractor, a tireless junior researcher, a code assistant — and unreliable as the output, the autonomous decision-maker. Lopez-Lira and Tang[6] make the cleanest version of the point: the model adds value by reading the news, not by managing the money.

Realistic performance expectations

Dimension	Illustrative expectation	Caveat
Strategies surviving costs	~15% of generated ideas	Comparable to skilled human hit-rate
Time to test 500 ideas	~2 hours vs weeks	Velocity, not edge, is the gain
Sentiment signal (text input)	Small but significant IC uplift	Strongest in small caps / post-bad-news
Best-performing prompt type	Novel-constraint, hypothesis-first	Textbook-pattern prompts worst

All figures in this table are illustrative estimates synthesized from the cited studies, not guarantees. Information-coefficient improvements from text signals are real but modest; the durable advantage is throughput.

6. Lopez-Lira & Tang — Can ChatGPT Forecast Stock Price Movements? (SSRN 4412788). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4412788
7. StratCraft — LLM-Generated Strategies: A Backtesting Reality Check. https://stratcraft.ai/news/llm-generated-strategies-backtesting-reality
5. FINSABER (arXiv:2505.07078) — Re-evaluating LLM trading strategies. https://arxiv.org/html/2505.07078v3
9. Resonanz Capital — How Hedge Funds Are Really Using Generative AI. https://resonanzcapital.com/insights/how-hedge-funds-are-really-using-generative-ai-and-why-it-matters-for-manager-selection

SECTION 4

How to Make the Toy Strategy Less Unprofitable

Illustrative, not advice. The toy strategy fails for two diagnosable reasons: the raw signal is weak, and the turnover is ruinous. Neither is fatal in principle. Below are the levers a researcher would pull to drag the net curve back toward the gross line — framed as hypotheses to test, with directional, illustrative magnitudes only.

- Cut turnover. At 151% daily turnover, costs dominate everything. Smoothing the signal (e.g. exponentially-weighted prior returns) and extending the holding period from one day to several can plausibly halve turnover — the single highest-leverage fix, since cost scales linearly with trading.
- Choose less-efficient universes. Short-horizon reversal persists more reliably in small caps and less-liquid markets than in 20 mega-caps where reversal is heavily arbed. The trade-off is higher per-share cost and capacity limits.
- Move to intraday / microstructure horizons. Reversal effects are often strongest at intraday and overnight horizons — but this raises the bar on execution infrastructure dramatically, which is exactly where Medallion's real edge lives.
- Ensemble many weak signals. Medallion's lesson: thousands of weak, decorrelated signals sized by a Kelly-style rule beat one "clever" signal. Diversification across signals is the closest a small shop can come to the approach.
- Earn execution alpha. Provide liquidity (post limit orders, capture spread) rather than taking it. For a turnover-heavy book, flipping from spread-payer to spread-earner can swing the cost term from a large negative to roughly neutral.
- Overlay text-derived sentiment. Per Lopez-Lira & Tang, an LLM sentiment overlay can tilt the book toward names with favorable news, lifting the information coefficient[6] — a concrete, evidence-backed use of the model as an input.

Illustrative lever impact (directional only)

Lever	Primary effect	Illustrative direction
Signal smoothing + longer hold	Turnover ↓	Largest cost reduction
Small-cap / inefficient universe	Gross signal ↑	Higher per-trade cost
Liquidity provision (limit orders)	Cost term → ~neutral	Execution alpha
Weak-signal ensemble	Sharpe stability ↑	Diversification
LLM sentiment overlay	Information coefficient ↑	Small but significant

Arrows indicate hypothesized direction of effect, not quantified returns. No combination of these levers is claimed to reach Medallion-like performance; the realistic ceiling for a transparent, replicable strategy is a modest positive Sharpe after costs.

6. Lopez-Lira & Tang — Can ChatGPT Forecast Stock Price Movements? (SSRN 4412788). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4412788

SECTION 5

A Practical Playbook: Using Claude (or Any Frontier LLM) for Strategy Research

Used well, a frontier model behaves like an exceptionally fast, occasionally overconfident junior researcher. The desk's job is to point it at the right problems and gate everything it produces through rigorous validation. The patterns and guardrails below separate the funds that benefit from those that simply generate expensive overfitting.

Prompt patterns that work

- Novel constraints over textbook patterns. Strategies built from generic indicators (MACD, RSI) are the worst out-of-sample because they are the most data-snooped. Prompts that impose novel constraints — specific microstructure assumptions, unusual holding periods, capacity limits — produce the best survivors.[7]
- Hypothesis-first, not strategy-first. Ask the model to articulate an economic rationale before any code. A signal with no story is a coincidence waiting to be priced out.
- State microstructure assumptions explicitly. Force the model to declare its cost, turnover, and execution assumptions up front, so they can be stress-tested rather than buried in defaults.

Guardrails and validation

- Bias defense. Actively screen every generated strategy for look-ahead, survivorship, and data-snooping bias — the three failure modes FINSABER documents[5].
- Walk-forward + deflated Sharpe. Validate out-of-sample with walk-forward analysis and report the deflated Sharpe ratio to penalize the multiple-testing inherent in generating hundreds of candidates.
- Human-in-the-loop gates. No strategy reaches paper trading without a human reviewing rationale, costs, and capacity. The model proposes; the desk disposes.
- Data security. Never paste proprietary data, positions, or signals into public model endpoints. Use enterprise/on-prem deployments with no-training guarantees for anything sensitive.
- Explainability for IC and compliance. Every deployed signal needs a human-legible explanation. "The model said so" does not survive an investment-committee review or a regulator.

Workflow: who does what, and the gate to pass

Stage	LLM role	Human role	Validation gate
Ideation	Generate hypotheses with economic rationale	Curate, reject incoherent ideas	Plausible economic story?
Prototyping	Write backtest code, clean data	Review code for leakage	No look-ahead in pipeline?
Backtesting	Run variations, summarize results	Set cost & turnover assumptions	Survives realistic costs?
Validation	Draft walk-forward harness	Interpret deflated Sharpe	Deflated Sharpe > threshold?
Deployment	Document rationale for IC	Final sign-off, sizing	Explainable & compliant?

7. StratCraft — LLM-Generated Strategies: A Backtesting Reality Check. https://stratcraft.ai/news/llm-generated-strategies-backtesting-reality
5. FINSABER (arXiv:2505.07078) — Re-evaluating LLM trading strategies. https://arxiv.org/html/2505.07078v3

SECTION 6

Risk & Governance Considerations for the Fund

Bringing frontier models into the research stack opens a new risk surface that sits alongside the familiar market and operational risks. The leading industry analyses converge on four areas that warrant explicit governance.

- Model risk. LLMs are non-deterministic, can hallucinate plausible-looking code or citations, and drift across versions. Treat any model output as an untrusted draft subject to the same model-risk-management discipline applied to any quantitative model.
- Herding and crowding risk. If every fund prompts the same handful of frontier models for ideas, those models will surface the same "novel" strategies to everyone — converting them into crowded trades the moment they are deployed. Commoditized ideas decay fastest.
- Compliance and surveillance. GenAI is already used for compliance surveillance, but the tooling itself must be governed[9]. Document model usage, retain prompts and outputs, and ensure deployed signals are explainable to regulators.
- IP and data leakage. Proprietary signals, positions, and datasets pasted into public endpoints may be retained or used in training. Black-box behavior and data-leakage risk are the headline concerns flagged in manager-selection due diligence[9].

The macro context raises the stakes. Industry surveys report that front-office GenAI adoption has shifted from "if" to "when" among leading managers[8], and that nearly every manager now uses generative AI somewhere in the workflow[10]. Widespread adoption is precisely what makes the crowding risk above non-hypothetical.

9. Resonanz Capital — How Hedge Funds Are Really Using Generative AI. https://resonanzcapital.com/insights/how-hedge-funds-are-really-using-generative-ai-and-why-it-matters-for-manager-selection

8. AIMA — Front-office GenAI adoption shifts from 'if' to 'when'. https://www.aima.org/article/press-release-front-office-gen-ai-adoption-shifts-from-if-to-when-for-leading-fund-managers-aima-research-finds.html

10. Marex — Generative AI in Hedge Funds: From Experimentation to Everyday Use. https://www.marex.com/news/2025/12/generative-ai-in-hedge-funds-from-experimentation-to-everyday-use

SECTION 7

Alternative Perspectives & Provoking Questions

A report that only confirms the comfortable conclusion is its own kind of overfitting. Here are three contrarian readings worth arguing about at the next research offsite.

1. The alpha is now in execution and data exclusivity, not ideas.

If a frontier model can articulate any strategy on demand, then strategy ideas have a marginal cost approaching zero — and zero-cost goods do not carry economic rent. On this view, the entire defensible edge migrates to the two things models cannot supply: exclusive data and superior execution. That is exactly the shape of Medallion's moat, and it suggests the AI era will reward infrastructure investment over cleverness.

> Provoking question: If ideas are free, is your research budget mis-allocated toward generating them rather than executing on them?

2. LLM commoditization makes mean reversion the crowded trade.

The flip side of cheap ideas: when thousands of desks prompt the same models, the models recommend the same canonical strategies — mean reversion, momentum, pairs — to all of them simultaneously. The very accessibility that makes these strategies easy to generate is what makes them crowded and short-lived. The model is, in effect, a synchronization device for the herd.

> Provoking question: Are the strategies your model proposes most enthusiastically the ones you should most distrust?

3. Research velocity is the real edge — 100× more hypotheses per quarter.

The optimistic reading does not rely on any single strategy being good. If a desk can generate and rigorously test 100 times more hypotheses per quarter, then even a constant ~15% survival rate yields far more validated signals in absolute terms. The edge is not a smarter idea; it is an industrialized, well-governed idea factory — provided the validation gates keep pace with the generation.

> Provoking question: Is your validation pipeline fast and rigorous enough to exploit a 100× jump in idea throughput — or will it become the bottleneck?

Synthesis. These three views are not mutually exclusive. Taken together they imply a barbell: invest in proprietary data and execution on one end, and in a fast, disciplined AI-assisted research factory on the other — while treating any idea the crowd can also generate as a depreciating asset.

SECTION 8

Sources

All external claims in this report are footnoted to the numbered sources below. Footnote markers throughout the document correspond to these entries. URLs are clickable.

1. Quartr — Renaissance Technologies and the Medallion Fund.
   https://quartr.com/insights/edge/renaissance-technologies-and-the-medallion-fund
2. Cornell Capital — The Medallion Fund: The Ultimate Counterexample.
   https://www.cornell-capital.com/blog/medallion-fund-the-ultimate-counterexample/
3. Institutional Investor — Famed Medallion Fund 'Stretches Explanation to the Limit'.
   https://www.institutionalinvestor.com/article/2bswymr8cih3jeaslxc00/portfolio/famed-medallion-fund-stretches-explanation-to-the-limit-professor-claims
4. Wikipedia — Renaissance Technologies.
   https://en.wikipedia.org/wiki/Renaissance_Technologies
5. FINSABER (arXiv:2505.07078) — Re-evaluating LLM trading strategies.
   https://arxiv.org/html/2505.07078v3
6. Lopez-Lira & Tang — Can ChatGPT Forecast Stock Price Movements? (SSRN 4412788).
   https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4412788
7. StratCraft — LLM-Generated Strategies: A Backtesting Reality Check.
   https://stratcraft.ai/news/llm-generated-strategies-backtesting-reality
8. AIMA — Front-office GenAI adoption shifts from 'if' to 'when'.
   https://www.aima.org/article/press-release-front-office-gen-ai-adoption-shifts-from-if-to-when-for-leading-fund-managers-aima-research-finds.html
9. Resonanz Capital — How Hedge Funds Are Really Using Generative AI.
   https://resonanzcapital.com/insights/how-hedge-funds-are-really-using-generative-ai-and-why-it-matters-for-manager-selection
10. Marex — Generative AI in Hedge Funds: From Experimentation to Everyday Use.
    https://www.marex.com/news/2025/12/generative-ai-in-hedge-funds-from-experimentation-to-everyday-use