Structured Equity Research with Large Language Models: A Benchmark Study of GPT-5.5 and Claude Sonnet 4.6 Across Tickers, Horizons, and Trials

Sam Naghshineh

Founder, StuxMux.com

info@stuxmux.com

Abstract

Large language models (LLMs) are being deployed across the equity research workflow, from idea generation to written analyst notes, on the implicit assumption that they produce stable, rule-following output when given a structured analytical framework. This paper tests that assumption directly. We benchmark two frontier models, GPT-5.5 and Claude Sonnet 4.6, on a structured forward-analysis task spanning twelve U.S. equities and four investment horizons (1 to 3 months, 6 to 12 months, 1 to 3 years, and 3 to 5 years), with a within-model variance subset of three trials per cell on three tickers. The framework imposes scenario probability weighting, a quality-adjusted valuation rule, an expected-value alpha computation, a reward-to-risk ratio, a confidence gate that prohibits Neutral ratings at high confidence, and a mega-cap benchmark override designed to address index contamination. Across the 96 cross-model observations, exact rating agreement is 72.9%, but agreement collapses to 25% at the 1 to 3 month horizon and remains above 83% at every horizon of six months or longer. GPT-5.5 issued nine Neutral ratings against Claude's zero on identical inputs, evidence of a "neutral regression" prior that overrides explicit framework rules. Within-model trials reveal a stability tradeoff: Claude is more directionally consistent (75% identical-rating cells versus GPT's 58.3%), while GPT is more numerically consistent on confidence and expected-value metrics. The 1 to 3 month horizon shows rating instability across trials in every observed cell. We frame these findings as evidence that LLMs are not yet ready for autonomous structured equity research and propose mitigations including horizon-conditioned deployment, multi-model voting at the cell level, and supervisor-agent orchestration of the kind implemented at platforms such as stuxmux.com. The iterative prompt-engineering process required to produce the framework is itself reported as a finding about LLM brittleness on structured financial tasks.

Keywords: large language models, equity research, instruction following, ensemble methods, reliability, benchmark study

1. Introduction

The integration of large language models into equity research has moved from experiment to production in roughly twenty-four months. Sell-side desks, asset managers, and retail platforms now use LLMs to summarize filings, draft analyst notes, score management commentary, and (increasingly) to generate forward-looking views on individual securities. Vendors describe these systems as "AI analysts," and a growing literature treats LLM output as a tractable substitute for, or supplement to, human judgment on questions of fundamental value.

This paper asks a narrower question that has not yet been answered convincingly: when two frontier LLMs are given identical inputs, identical tools, and an identical structured analytical framework, do they produce consistent, rule-following output? And when one model is run multiple times on the same inputs, do its own outputs agree with themselves?

These questions matter because the operational case for LLM-based research rests on two implicit promises. The first is reproducibility: a framework specified in the prompt should produce the same answer when the inputs are the same. The second is fidelity: the model should follow the framework's rules even when its own pretrained priors point elsewhere. We find that both promises break down in identifiable, characterizable ways.

We benchmark GPT-5.5 and Claude Sonnet 4.6, both run with web search enabled, against a forward-analysis framework that produces a structured rating (Underweight, Underperform, Neutral, Outperform, Overweight) together with a probability-weighted expected-value alpha, a downside scenario, a reward-to-risk ratio, and a small set of decision gates including a confidence gate that prohibits Neutral output at high confidence levels. The universe consists of twelve tickers selected to span mega-cap technology, defensive healthcare, cyclical energy, financially weaker mid-cap, speculative growth, and zero-yield compounders. Each ticker is evaluated at four horizons. A three-ticker subset (AAPL, UNH, XOM) is run three times per cell to measure within-model variance.

The headline findings preview the paper. Cross-model exact rating agreement is 72.9% on average, but collapses to 25% at the shortest horizon and stays above 83% at every horizon of six months or longer. GPT-5.5 produces nine Neutral ratings out of 48 cells; Claude Sonnet 4.6 produces zero, despite the framework's confidence gate making Neutral the wrong choice when stated confidence exceeds the threshold. On three-trial repetition, Claude is more directionally stable but more numerically variable; GPT is the reverse. Every 1 to 3 month cell exhibits rating instability across trials.

We interpret these results as a cautionary critique of single-model deployments. The paper contributes (i) a horizon-conditioned reliability map for two frontier models on a realistic structured equity task, (ii) characterization of two distinct failure modes (neutral regression and short-horizon instability), (iii) evidence of a directional-versus-numerical stability tradeoff with implications for ensemble design, and (iv) a methodological note on a mega-cap benchmark contamination override that both models applied correctly when the framework specified it. We close with mitigation recommendations centered on multi-model orchestration.

2. Related Work

2.1 LLM Evaluation, Instruction Following, and Prompt Brittleness

A substantial body of work documents that LLM behavior is sensitive to prompt structure, context length, and the position of relevant information within long inputs. Liu et al. (2024) show that performance on retrieval-style tasks degrades when relevant information is placed in the middle of long contexts; the U-shape persists even for explicitly long-context models. Zhou et al. (2023) introduce IFEval, a benchmark that measures instruction-following on verifiable rules and finds significant gaps between flagship models on simple format constraints. Sclar et al. (2024) demonstrate that small, semantically irrelevant prompt perturbations (such as whitespace or list delimiter choices) move accuracy by tens of points. These results are consistent with our finding that small differences in framework version produced large differences in model output during the development described in Section 4.1.

A second strand of LLM evaluation is concerned with reasoning fidelity. Wei et al. (2022) show that chain-of-thought prompting elicits step-by-step reasoning but does not, on its own, guarantee that the reasoning is faithful to the final answer. Turpin et al. (2023) document systematic unfaithfulness, where models confabulate plausible reasoning chains that do not correspond to the actual driver of their answer. The tension this paper exposes

between explicit framework rules and pretrained priors is closely related: a model that produces a well-formed reasoning trace can still "round" the rating toward its prior, as we observe with GPT-5.5's Neutral ratings.

Recent work on numerical reliability is also relevant. Nori et al. (2023) and others document that LLM numerical outputs vary across runs even at low temperature. The OpenAI and Anthropic technical reports for the relevant model generations note that web-search tool use and retrieval introduce additional non-determinism not captured by sampling temperature alone (OpenAI, 2025; Anthropic, 2025). We accept this non-determinism in our experimental design because the alternative (forbidding web search) would compromise the realism of the equity research task.

2.2 Analyst Rating Reliability and Behavioral Finance

The benchmark for our work is the older literature on sell-side analyst dispersion and bias. Welch (2000) shows that analyst recommendations exhibit substantial herding around the consensus, a finding extended by Hong and Kubik (2003) who document career-incentive-driven bias in forecast revisions. Trueman (1994) provides the early theoretical model in which analysts rationally bias forecasts toward prior consensus when their reputation depends on perceived ability. Clement (1999) measures forecast accuracy as a function of analyst experience and brokerage size, finding meaningful but small differentials. Womack (1996) shows that analyst recommendation changes contain price-relevant information, while Loh and Stulz (2011) establish that "influential" recommendations are a small subset of all recommendations issued.

LLM "analysts" can be read as the next iteration of this literature. They share two of the structural features that drive sell-side bias (training data that overrepresents consensus narratives; reward signals that penalize confident wrongness) and lack two important constraints (career risk that disciplines extreme calls in either direction; explicit accountability for outcomes). The neutral regression we document in GPT-5.5 is recognizable as an automated form of the herding behavior Welch (2000) describes, with the consensus being not the analyst median but the model's pretrained prior.

2.3 Systematic Investing and Rule-Based Decision Frameworks

The motivation for imposing a structured framework on an LLM analyst is that explicit rules reduce judgmental noise. Kahneman, Sibony, and Sunstein (2021) make the general case that algorithmic and rule-based decisions outperform unaided expert judgment in noisy domains. Tetlock and Gardner (2015) document the value of structured decomposition in geopolitical forecasting. Heuer (1999) catalogs structured analytic techniques for intelligence analysis; many of his recommendations (explicit hypothesis enumeration, weighted evidence assessment) map directly onto the scenario weighting and confidence-dimension scoring used here.

In quantitative finance, the analog is the factor literature. Fama and French (1993) introduce the size and value factors; Asness, Frazzini, and Pedersen (2019) and Frazzini and Pedersen (2014) develop the quality-minus-junk and betting-against-beta factors that motivate our framework's quality_score adjustment. Ang (2014) provides a textbook treatment of factor-based asset management. Bessembinder (2018) shows that lifetime stock returns are positively skewed and concentrated in a small number of names, a result that motivates our mega-cap benchmark override and the framework's emphasis on long-horizon scenario probability rather than short-horizon point estimates.

The framework tested in this paper sits between these traditions. It is rule-based in the Kahneman sense and factor-aware in the Fama-French sense, but its execution layer is a stochastic LLM. The empirical question is how much of the rule-based discipline survives execution.

3. Methodology

3.1 The Analytical Framework

We do not reproduce the prompt verbatim. The framework is described conceptually here, with sufficient detail for replication of the experimental design.

The framework asks the model to produce, for a given ticker and horizon, a forward analysis consisting of three probability-weighted scenarios (Bull, Base, Bear) with explicit total return estimates inclusive of capital return; six confidence dimensions each scored on a 0 to 20 scale (covering, broadly, business durability, financial strength, valuation margin of safety, catalyst clarity, sentiment positioning, and macro fit); a quality_score derived from the business durability and financial strength dimensions; an explicit ev_alpha computed as the probability-weighted expected total return relative to a horizon-appropriate benchmark; a downside scenario value used to compute the reward-to-risk ratio RR = EV_stock / |Downside_stock|; and a final rating drawn from a five-point ordinal scale.

Several design features are worth flagging because they bear directly on the empirical findings.

Confidence gate. The framework includes an explicit rule (the confidence_gate_triggered field) that high stated confidence is incompatible with a Neutral rating. The intuition is operational: if an analyst genuinely has high confidence in a forward view, that view must be directional. The rule is meant to prevent rhetorical hedging.

Quality-adjusted valuation. The framework's valuation_margin_of_safety score is reduced or eliminated when the quality_score is high. This reflects the empirical observation, formalized by Frazzini and Pedersen (2014), that high-quality businesses persistently trade at premium multiples and that mechanically penalizing them on valuation systematically underweights compounders. The rule introduces a known degree of freedom: two models with different priors on what counts as "high quality" can apply the rule consistently and still arrive at different ratings.

Capital return adjustment with horizon prorating. Total return scenarios incorporate capital return (dividends plus net buybacks) prorated to the horizon. For a 1 to 3 month view, four-year capital return is divided down accordingly; for a 3 to 5 year view, the full forward yield compounds. The prorating explicitly avoids the common error of overstating short-horizon dividend contribution.

Mega-cap benchmark override. When the stock under analysis exceeds 5% of QQQ weight, the benchmark for ev_alpha is replaced with SPY. The rule addresses index contamination: a stock that is itself a major driver of QQQ cannot meaningfully be evaluated against QQQ without circularity. We treat this as a methodological contribution and discuss it in Section 6.5.

Self-audit. The framework requires a final consistency check: the rationale must be consistent with the numerical scoring, and the rating must be consistent with both ev_alpha and the confidence gate. Fields are produced in a structured order such that a downstream auditor can verify rule compliance directly.

The framework was developed iteratively over eleven versions. Each version targeted specific failure modes observed in the previous version's output: early versions allowed Neutral output at high confidence (removed in version 4); valuation penalties were too binary (replaced with a quality-conditioned schedule in version 6); short-horizon scenarios collapsed onto recent price action (mitigated in version 8 by requiring a dated catalyst and a measurable expectation gap); and the benchmark override was introduced in version 9 after early AAPL outputs produced spurious negative alpha against QQQ. We treat the iterative development process itself as a finding: producing a stable structured framework on a realistic equity task required eleven iterations across two model families. A practitioner deploying any single version would inherit the residual brittleness of that version.

3.2 Models and Tools

Both models were run with web search enabled, on identical prompts, in single-shot mode for the cross-model comparison and in three-trial mode for the variance subset. We use GPT-5.5 and Claude Sonnet 4.6 as the two

contemporaneous frontier models with comparable reasoning and tool-use capability. Web search was enabled because the task requires current price, valuation, and catalyst information; the cost is non-determinism beyond what sampling temperature alone introduces, since each trial may retrieve a slightly different evidence set. We accept this cost because the alternative compromises external validity.

To homogenize the dataset, AAPL was re-run on the latest framework version after the mega-cap benchmark override was introduced. All other tickers were single-shot on the latest version.

3.3 Ticker Selection

The twelve-ticker universe is constructed to probe distinct framework components.

Ticker	Role
AAPL	Mega-cap tech; benchmark override probe
MSFT	Mega-cap tech; quality compounder
GOOGL	Mega-cap tech; antitrust catalyst
BRK.B	Zero-yield capital return; quality
JPM	Cyclical financial; macro probe
XOM	Cyclical energy; macro and capital return
COST	Quality-defensive; valuation premium probe
F	Financially weaker mid-cap
PLTR	Speculative growth; valuation extreme
SHOP	Speculative growth; quality-vs-valuation tension
UNH	Out-of-favor defensive; sentiment probe
CVS	Value trap probe

Table 1. Ticker universe and the framework feature each is intended to probe. The selection is biased toward U.S. large-cap by design and is not a tradable universe.

3.4 Variance Subset

Three tickers (AAPL, UNH, XOM) were selected for the three-trial within-model variance test. The selection spans technology, defensive healthcare, and cyclical energy, with the explicit goal of measuring whether variance behavior generalizes across sector and factor exposures. Each ticker was run three times per horizon per model, producing 72 observations (3 trials x 4 horizons x 2 models x 3 tickers).

3.5 Statistical Methods

Ratings are encoded on a -2 to +2 scale (Underweight, Underperform, Neutral, Outperform, Overweight). We report exact agreement rate (proportion of cells in which both models produce the same rating), rating distance (absolute difference on the -2 to +2 scale), and signed difference (Claude minus GPT) to measure directional bias. Within-cell variance is reported as standard deviation across the three trials for ratings, confidence, ev_alpha, and the framework's raw_total score. Krippendorff's alpha is used as an inter-rater reliability statistic on the cross-model ratings; McNemar's test is reported on the paired Neutral-versus-non-Neutral comparison given its asymmetry.

4. Results

4.1 Cross-Model Agreement

Across the 48 ticker-horizon cells common to both models, exact rating agreement is 35 out of 48, or 72.9%. The distribution of rating distances is as follows.

Rating distance	Cells	Share
0 (exact)	35	72.9%
1	9	18.8%
2	2	4.2%
3	2	4.2%
Total	**48**	**100.0%**

Table 2. Distribution of cross-model rating distances on the -2 to +2 scale. The two distance-3 cells correspond to PLTR at the 1 to 3 month and 6 to 12 month horizons.

Mean rating score is 0.77 for Claude Sonnet 4.6 and 0.50 for GPT-5.5. The mean signed difference (Claude minus GPT) is +0.27, indicating a structural bullishness gap. Claude is more bullish than GPT in 10 cells; GPT is more bullish in 3; the remaining 35 cells are exact matches. The asymmetry is statistically meaningful and not the result of a few outliers.

The full rating distributions are shown in Table 3.

Rating	GPT-5.5 (n=48)	Claude Sonnet 4.6 (n=48)
Overweight (+2)	1	3
Outperform (+1)	30	38
Neutral (0)	9	0
Underperform (-1)	8	7
Underweight (-2)	0	0

Table 3. Rating distribution by model across all 48 cells. The most striking line is Neutral: 9 from GPT, 0 from Claude. Under the framework's confidence gate, Neutral is permitted only at low stated confidence; the GPT cells in question all reported confidence above the gate threshold.

Mean stated confidence is 70.3% for GPT (standard deviation 5.2) and 66.7% for Claude (standard deviation 2.5). GPT's confidence is on average higher but markedly more dispersed; Claude's is lower but tightly clustered in a narrow band.

A visual companion to Table 2 (not produced here) would show a heatmap of rating distance by ticker and horizon, with darker cells concentrated in the leftmost (1 to 3 month) column and on the COST and PLTR rows.

4.2 Horizon Analysis

The headline empirical finding of this paper is the horizon-conditional reliability profile.

Horizon	Agreement (n=12)

1 to 3 months	3/12 = 25.0%
6 to 12 months	11/12 = 91.7%
1 to 3 years	11/12 = 91.7%
3 to 5 years	10/12 = 83.3%

Table 4. Cross-model exact rating agreement by horizon. Agreement at the shortest horizon is roughly one-quarter of the rate at all longer horizons. The collapse is monotonic in horizon length up to the medium term and degrades only mildly at the longest horizon.

The 25% short-horizon agreement is an order of magnitude worse than the long-horizon agreement and is a strong signal that LLM reasoning on structured forward equity tasks is not horizon-invariant. We discuss possible mechanisms in Section 5.2.

4.3 Per-Ticker Agreement

Ticker	Agreement (n=4)
AAPL	75%
BRK.B	50%
COST	50%
CVS	75%
F	75%
GOOGL	100%
JPM	75%
MSFT	100%
PLTR	50%
SHOP	75%
UNH	75%
XOM	75%

Table 5. Cross-model exact agreement rate by ticker, averaged across the four horizons. The 50% names (BRK.B, COST, PLTR) cluster around different framework components: BRK.B on the zero-yield capital return rule, COST on the quality-adjusted valuation rule, and PLTR on the high-confidence directional rating gate at short horizons.

4.4 Within-Model Variance

The three-trial variance subset (n=72) is the second core empirical contribution of this paper.

Across the 24 model-ticker-horizon cells (2 models x 3 tickers x 4 horizons), 16 produced identical ratings on all three trials (66.7% identical-rating consistency). Broken down by model: Claude is consistent in 9 of 12 cells (75%); GPT is consistent in 7 of 12 (58.3%).

The numerical variance profile inverts the directional one.

Metric	Claude Sonnet 4.6	GPT-5.5

Identical-rating cells	9/12 (75%)	7/12 (58.3%)
Mean within-cell confidence std	1.91	1.47
Mean within-cell ev_alpha std	1.93	1.38
Mean within-cell raw_total std	2.03	1.76
Maximum within-cell raw_total range	9	13

Table 6. Within-model variance summary across the three-trial subset (3 tickers, 4 horizons, 3 trials per cell). Claude is more directionally stable (rating-level) but more numerically variable; GPT is the reverse. The maximum raw_total range is the only measure on which GPT shows wider tails despite tighter average dispersion.

The horizon profile of variance mirrors the cross-model agreement profile. The 1 to 3 month horizon shows rating instability across trials in every observed cell, for both models. Specifically:

• AAPL at 1 to 3 months. Claude went Outperform, Underperform, Underperform across the three trials. GPT went Neutral, Outperform, Outperform.

• UNH at 1 to 3 months. Both models exhibit at least one rating change across trials, with Claude crossing the Neutral boundary in one trial.

• XOM at 1 to 3 months. Both models exhibit at least one rating change across trials.

At 6 to 12 months and longer, almost every cell is unanimous across trials.

A visual companion (not produced here) would plot within-cell ev_alpha standard deviation against horizon, with the 1 to 3 month bar markedly taller than the others for both models.

4.5 Case Studies

4.5.1 PLTR at 1 to 3 months: the distance-3 disagreement. Claude rates Overweight (+2); GPT rates Underperform (-1). The underlying scoring divergence is not in the data inputs (both models retrieved similar price levels and similar consensus estimates) but in the quality_score. Claude assigns PLTR a high business_durability score on the basis of government contract stickiness and developer ecosystem; GPT assigns a low durability score on the basis of customer concentration risk. The high quality_score in Claude's output then triggers the framework's quality-adjusted valuation rule, eliminating the valuation penalty that would otherwise apply to a stock at PLTR's multiple. GPT's low quality_score leaves the valuation penalty in place. The framework was applied correctly by both models; the priors that fed into the durability score were not.

4.5.2 PLTR at 6 to 12 months. The same pattern as 4.5.1, with the same distance-3 split.

4.5.3 COST at 1 to 3 years and 3 to 5 years. Both horizons produce a distance-2 disagreement: Claude Outperform (+1), GPT Underperform (-1). The underlying mechanism is the same as PLTR but inverted. Both models score COST as a high-quality business; both apply the framework's quality-adjusted valuation rule. The disagreement is over how much the rule should reduce the penalty. Claude applies the full elimination available under the rule when quality_score exceeds the threshold; GPT applies a partial reduction. The framework permits both interpretations because it specifies the eligibility for adjustment but not the magnitude with full precision. The COST cells are the cleanest demonstration in the dataset that explicit rules cannot eliminate priors entirely when the rules contain any degree of freedom.

4.5.4 AAPL at 1 to 3 months across trials. Claude's three trials produced Outperform, Underperform, Underperform. The first trial leaned on a dated product-cycle catalyst; the second and third leaned on near-term valuation compression

and a deteriorating channel-check signal. GPT's three trials produced Neutral, Outperform, Outperform. The Neutral trial is the framework's confidence-gate failure: stated confidence was 71%, above the gate threshold, yet the rating remained Neutral. The trial-to-trial divergence within a single model on a single ticker at a single horizon is, in our view, the single most operationally important finding for practitioners considering production deployment.

4.6 Confidence Calibration

The full picture on confidence is captured by two numbers per model and one structural observation. GPT's mean confidence is 70.3% with standard deviation 5.2; Claude's is 66.7% with standard deviation 2.5. Claude's confidence lives almost entirely in a 64 to 70 band; GPT's spans 60 to 80. A visual companion would show two histograms, one tightly peaked (Claude) and one flatter and right-shifted (GPT). The structural observation is that GPT's higher mean confidence is not associated with more directional ratings; on the contrary, it is associated with more Neutral ratings, which is the opposite of what the framework's confidence gate prescribes.

4.7 EV_alpha and Reward-to-Risk

Mean ev_alpha (in basis points relative to the appropriate benchmark) is broadly comparable between models across horizons longer than three months. At the 1 to 3 month horizon, both models cluster ev_alpha near zero, consistent with the short-horizon scenario compression we discuss in Section 5.2. The within-trial variance pattern reported in Table 6 (GPT std 1.38, Claude std 1.93 on ev_alpha) is the most actionable single statistic for ensemble design: it implies that a numerical-aggregation strategy weighted toward GPT's expected-value estimates, combined with a directional-aggregation strategy weighted toward Claude's ratings, would dominate any single-model deployment on these particular cells.

Reward-to-risk distributions are broadly similar in mean and median across models, with Claude's higher long-horizon ratings producing a slightly higher mean RR. We do not interpret this as evidence of a structural difference in risk modeling so much as a downstream consequence of the bullishness gap reported in Section 4.1.

5. Discussion

5.1 Neutral Regression as Pretrained Prior Interference

GPT-5.5 produced nine Neutral ratings; Claude Sonnet 4.6 produced zero. The framework explicitly forbids Neutral output at high stated confidence via the confidence_gate_triggered rule. In the GPT cases in question, stated confidence exceeded the gate threshold; the rule should have forced a directional rating; the model rated Neutral anyway.

We call this the neutral regression bias. It is an automated analog of the herding behavior Welch (2000) and Hong and Kubik (2003) document in human analysts, and an example of the unfaithful reasoning Turpin et al. (2023) describe more generally. The mechanism is most likely that GPT's training emphasizes hedged, low-risk output more than Claude's, and that this prior is sufficiently strong to override an explicit rule when the rule conflicts with it. The framework's audit fields catch the violation after the fact (a downstream system can flag any cell where confidence_gate_triggered is true and rating equals Neutral), but they do not prevent it during generation.

This finding has direct deployment implications. Any single-model production system using GPT-5.5 on this kind of structured task should treat Neutral output as suspect and route those cells to either an alternative model or a human reviewer.

5.2 The Short-Horizon Failure Mode

The 1 to 3 month horizon shows 25% cross-model agreement and 100% within-model rating instability across trials. This is a structural failure, not a sampling artifact.

We hypothesize three contributing mechanisms.

First, base rates for short-horizon equity returns are noisier. The signal-to-noise ratio in any forward forecast scales unfavorably with horizon shortness; under-three-month moves are dominated by flow, positioning, and idiosyncratic news rather than the fundamental factors the framework is built around. The framework is therefore being asked to do the wrong job.

Second, scenario priors are weaker at short horizons. The Bull/Base/Bear decomposition has well-developed priors for medium-term cash-flow paths but compresses awkwardly to a one-to-three-month window. Small changes in the model's near-term price-action interpretation flip scenario probabilities sharply.

Third, the framework's catalyst rule (which requires both a dated catalyst and a measurable expectation gap) is much harder to satisfy at short horizons. Either the catalyst is too far out (reducing weight) or the expectation gap is unmeasurable (the consensus has not yet anchored). The rule produces unstable scoring on insufficient evidence.

The practical implication is that LLMs as currently deployed are not appropriate for short-horizon equity analysis on this kind of structured task. They are appropriate, with caveats, at six months and longer. This is a horizon-conditioned reliability map that we believe should inform deployment decisions directly.

5.3 The Variance-Stability Tradeoff and Implications for Ensemble Design

Claude is more directionally stable but more numerically variable. GPT is the reverse. This tradeoff is the most novel methodological finding of the paper and we believe it is fundamental rather than accidental.

The tradeoff has direct implications for ensemble design. A naive ensemble that averages across models inherits the worst of both: Claude's numerical variance contaminates the average ev_alpha, while GPT's directional instability contaminates the rating vote. A better-designed ensemble decouples the two. Ratings are aggregated by majority vote across models and trials; numerical fields (ev_alpha, RR, confidence) are aggregated by weighted average where weights favor the more numerically stable model.

A more sophisticated mitigation, and the one we recommend, is supervisor-agent orchestration. A supervisor agent observes which model is more reliable on which task type (mega-cap vs. small-cap, long-horizon vs. short-horizon, quality-name vs. cyclical) and routes work accordingly. The supervisor can also flag cells where rating distance is large for human review and can cross-check rule compliance using the framework's audit fields. Platforms such as stuxmux.com implement this kind of multi-model ensemble with a supervisor agent layer; we mention this as one architectural example of the mitigation pattern, not as an endorsement of any specific implementation. The general principle, multi-model orchestration with task-conditioned routing, is what we believe is required.

5.4 Quality-Valuation Tension and the COST and PLTR Disagreements

The COST and PLTR cells are the cleanest demonstrations in the dataset that even highly-specified frameworks cannot fully eliminate model priors. The framework's quality-adjusted valuation rule is rigorous in stating the eligibility for adjustment (quality_score above threshold) but admits two reasonable interpretations of magnitude (full elimination versus partial reduction). On COST, both models score the business as eligible; the disagreement is over magnitude. On PLTR, both models apply the rule consistently; the disagreement is over the durability score that feeds the eligibility check.

A practitioner reading these results might be tempted to tighten the framework further. We caution against this. Each iteration of the framework that removes a degree of freedom also removes an analyst's substantive judgment from the system. At some point, the framework becomes a rule machine rather than an analytical tool. The right response, in our view, is not to over-specify but to (i) measure where models disagree, (ii) flag those cells, and (iii) bring them to human or supervisor-agent review.

5.5 The Mega-Cap Benchmark Override as Methodological Contribution

A small but consequential element of the framework is the rule that replaces QQQ with SPY as the alpha benchmark when the stock under analysis exceeds 5% of QQQ weight. This is a response to index contamination: a stock that is itself a major component of the benchmark cannot be evaluated against that benchmark without double counting.

Both models applied the rule correctly for AAPL once the framework specified it. In the prior framework version that did not include the rule, both models produced spuriously negative alpha against QQQ for AAPL. The cleanup was mechanical once the rule was made explicit.

The point connects to Bessembinder's (2018) wealth creation literature, which shows that a small number of stocks drive aggregate market returns; those same stocks are by construction the largest index components, and their relative-to-index alpha is mechanically suppressed when they are allowed to be self-referential in the benchmark. We believe this is a generalizable methodological contribution applicable to any factor or analyst process that uses a market-cap-weighted benchmark and includes the largest index constituents in the investable universe. For practitioners, the operational rule is simple: if the stock is more than 5% of the benchmark, change the benchmark.

5.6 Limitations

Several limitations qualify our findings. Nine of the twelve tickers were single-shot, and within-model variance behavior was measured only on the AAPL/UNH/XOM subset; the variance findings should be read as suggestive at the population level. Web search introduces non-determinism that we have not isolated from sampling-temperature non-determinism. The ticker selection is biased toward U.S. large-cap, and the findings may not extend to small-cap, international, or fixed-income analogs. The framework itself drifted over eleven versions during development; the cross-model dataset is on the latest version (with AAPL re-run to homogenize), but the eleven-version path is itself an unobserved degree of freedom in the experimental design. Most importantly, this paper does not validate the LLM ratings against subsequent realized returns; we report inter-rater reliability and within-model consistency, not predictive accuracy. Out-of-sample return validation is the natural next study and we identify it as the most important open question in the literature.

6. Mitigation Recommendations

We close the discussion with a set of operational recommendations for practitioners deploying LLMs on structured equity research tasks.

Multi-model voting with distance-conditioned escalation. Run at least two frontier models on every cell. At distance-0 cells (exact agreement), accept the output as the system rating. At distance-1 cells, flag for review and treat the average as a soft signal. At distance-2 or distance-3 cells, do not accept either model; route to human override or to a supervisor agent for adjudication. The empirical justification for this rule is the observed concentration of failures at high-distance cells, which constitute a small minority of the dataset (4 cells out of 48 at distance 2 or 3) and are therefore tractable for human review.

Horizon-conditioned deployment. Restrict LLM use to horizons of six months or longer on this kind of structured task. At 1 to 3 months, either supplement with traditional quantitative methods (factor models, momentum, options-implied measures) or route to human review. The 25% short-horizon agreement and 100% rating-instability findings are sufficient evidence in our view to recommend against autonomous LLM deployment at the shortest horizon.

Run-level voting. Run a minimum of three trials per cell. Take majority vote on ratings; take the average on numerical outputs. The 16-of-24 within-model consistency rate suggests that three trials are usually enough to identify the modal rating, but not always; for cells where three trials produce three different ratings, escalate to a fourth model or to human review.

Supervisor-agent architectures. Deploy a supervisor agent that selects between models on a per-task basis and applies the audit checks the framework supports. The supervisor should at minimum verify confidence_gate_triggered consistency, apply distance-conditioned escalation, and maintain a per-model reliability map that updates as data accumulates. The architecture used by platforms such as stuxmux.com is one concrete example of this pattern: multiple frontier models run in parallel, a supervisor layer aggregates and adjudicates, and task routing is informed by accumulated reliability data.

Mandatory reasoning trace and audit fields. The framework's intermediate fields (ev_alpha, RR, confidence_gate_triggered, scenario weights, quality_score) make it possible to audit rule compliance directly. Any production deployment should require these fields and run automated checks on them. Cells where a rule was violated (Neutral at high confidence; valuation penalty applied where quality_score eliminates it; benchmark not overridden where required) should be flagged automatically.

Self-consistency forcing functions. During the iterative development of the framework, the introduction of explicit self-audit consistency checks materially reduced contradictions between numerical scoring and final rating. We recommend including these checks as required output fields in any production prompt. The cost is small (a few additional tokens) and the benefit is substantial in the form of catchable rule violations.

7. Conclusion

We benchmarked GPT-5.5 and Claude Sonnet 4.6 on a structured forward-analysis task spanning twelve U.S. equities and four horizons, with a within-model variance subset on three tickers. Cross-model exact rating agreement is 72.9% on average but collapses to 25% at the 1 to 3 month horizon. GPT produced nine Neutral ratings against Claude's zero on identical inputs, evidence of a pretrained prior overriding an explicit rule. Within-model trials reveal a stability tradeoff: Claude is more directionally consistent (75% identical-rating cells versus GPT's 58.3%), while GPT is more numerically consistent on confidence and expected-value metrics. The 1 to 3 month horizon shows rating instability across trials in every observed cell.

The headline interpretation is that LLMs are not yet ready for autonomous structured equity research. They are usable, with caveats, at horizons of six months or longer; they are unreliable at the shortest horizon; and they exhibit characterizable failure modes (neutral regression, quality-valuation interpretation drift, short-horizon instability) that are diagnosable and partially addressable through orchestration. Multi-model ensembles with distance-conditioned escalation, horizon-conditioned deployment, run-level voting, supervisor-agent architectures, and mandatory audit fields together constitute a viable mitigation stack.

The methodological contribution is not the framework itself but the iterative process required to produce it: eleven versions, each addressing failure modes discovered in the prior version. We argue that the iterative process is itself a finding about LLM brittleness on structured financial tasks. A practitioner deploying any single version would inherit

the residual brittleness of that version.

Future work should focus on three priorities. First, out-of-sample return validation against realized prices, which would convert this reliability study into a predictive-accuracy study. Second, expansion of the model panel to three or more frontier models, which would test whether the variance-stability tradeoff generalizes or is specific to the GPT-Claude pair. Third, longer time-horizon validation, since the 3 to 5 year ratings cannot be evaluated for predictive accuracy without a multi-year hold-out period. We hope this paper contributes a useful baseline for that work and a cautionary note for practitioners considering autonomous LLM deployment in equity research today.

References

Ang, A. (2014). *Asset Management: A Systematic Approach to Factor Investing*. Oxford University Press.

Anthropic. (2025). *Claude Sonnet 4.6 Model Card and Technical Report*. Anthropic Technical Reports.

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

Bessembinder, H. (2018). Do stocks outperform Treasury bills? *Journal of Financial Economics*, 129(3), 440-457.

Clement, M. B. (1999). Analyst forecast accuracy: Do ability, resources, and portfolio complexity matter? *Journal of Accounting and Economics*, 27(3), 285-303.

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

Frazzini, A., & Pedersen, L. H. (2014). Betting against beta. *Journal of Financial Economics*, 111(1), 1-25.

Heuer, R. J. (1999). *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, CIA.

Hong, H., & Kubik, J. D. (2003). Analyzing the analysts: Career concerns and biased earnings forecasts. *Journal of Finance*, 58(1), 313-351.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgment*. Little, Brown Spark.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3), 411-433.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173.

Loh, R. K., & Stulz, R. M. (2011). When are analyst recommendation changes influential? *Review of Financial Studies*, 24(2), 593-627.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153-157.

Nasdaq. (2024). *Nasdaq-100 Index Methodology*. Nasdaq Index Research.

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375.

OpenAI. (2025). *GPT-5.5 System Card*. OpenAI Technical Reports.

Petajisto, A. (2011). The index premium and its hidden cost for index funds. *Journal of Empirical Finance*, 18(2), 271-288.

S&P Dow Jones Indices. (2024). *S&P U.S. Indices Methodology*. S&P Global.

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design. *International Conference on Learning Representations*.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. Crown.

Trueman, B. (1994). Analyst forecasts and herding behavior. *Review of Financial Studies*, 7(1), 97-124.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35.

Welch, I. (2000). Herding among security analysts. *Journal of Financial Economics*, 58(3), 369-396.

Womack, K. L. (1996). Do brokerage analysts' recommendations have investment value? *Journal of Finance*, 51(1), 137-167.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., & Hou, L. (2023). Instruction-following evaluation for large language models. arXiv:2311.07911.

Author note: empirical observations reported in this paper are drawn from a benchmark study conducted by the author. The platform stuxmux.com is referenced as one example of a multi-model orchestration architecture; the author is the founder of stuxmux.com, which constitutes the relevant disclosure. The framework described here is the property of the author and was developed iteratively over eleven versions during the study period.