Finovis

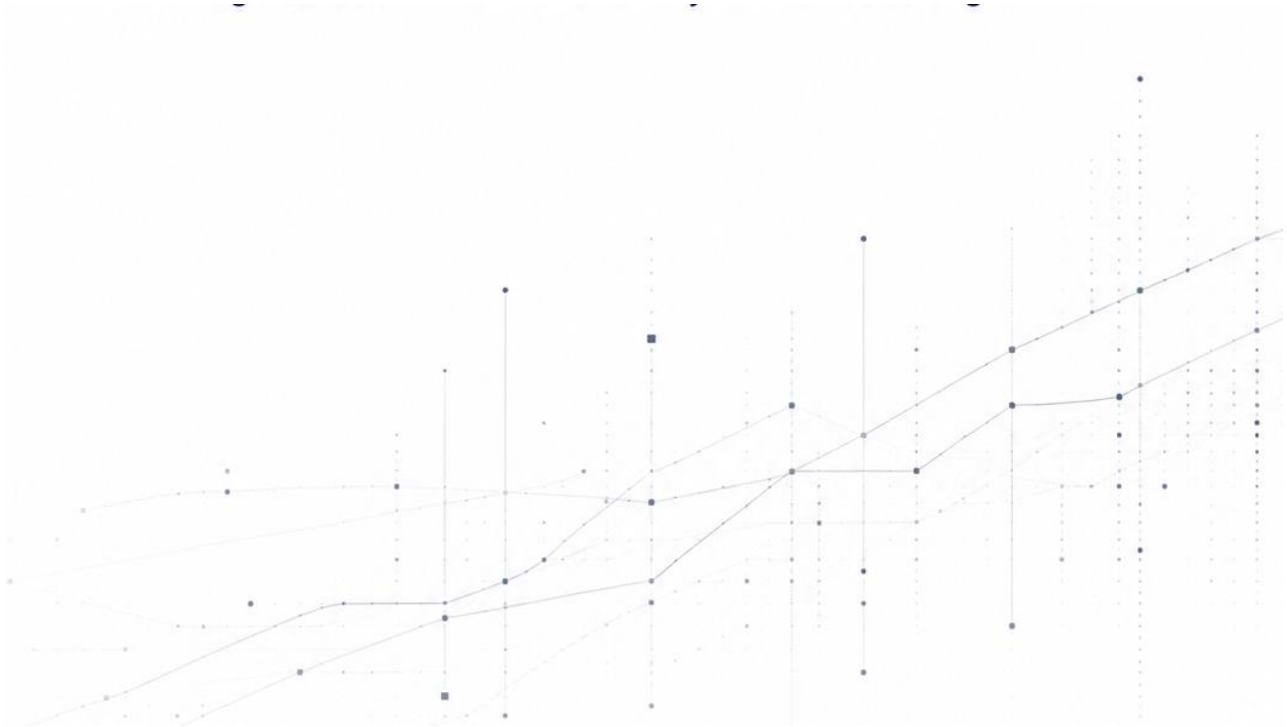
AI FOR THE FINANCIAL INDUSTRY

FINOVIS RESEARCH

QUANTITATIVE METHODS · RP-003

Robustness Before Performance

Evaluating Quantitative Models Beyond Backtesting



September 2026

finovis.ai

Finovis
AI FOR THE FINANCIAL INDUSTRY

Executive Summary

Historical backtesting is a fundamental tool in quantitative investment research, but attractive historical performance is not sufficient evidence of model robustness.

RP-003 examines why this distinction matters and asks what a more rigorous evaluation framework should consider before greater confidence is placed in historically observed quantitative-model behavior.

The paper argues that historical performance should be treated as conditional evidence. A backtest describes how a specified model behaved within a particular historical sample and under a defined research and simulation design. Its interpretation may be affected by sampling uncertainty, model selection, repeated testing, parameter optimization, specification dependence, temporal instability and implementation assumptions. None of these considerations makes backtesting uninformative. They affect how much evidential weight can reasonably be placed on the observed result.

A central conclusion is that the research process is part of the evidence. Performance selected from a broad universe of models, specifications or parameter configurations cannot necessarily be interpreted in the same way as the outcome of a more constrained research process. Data snooping, multiple testing, researcher flexibility, information leakage and repeated use of validation data can all affect the interpretation of an apparently successful historical result.

Out-of-sample evaluation can strengthen the evidence, but its value depends on informational independence rather than the label attached to the test. Historical hold-outs, pseudo-out-of-sample evaluation, rolling or walk-forward designs, prospective observation and live implementation provide different forms of evidence. Their relevance depends on what information was excluded from model development, whether evaluation outcomes subsequently influenced research decisions and how effectively the boundary between development and validation was preserved.

Robustness should also be challenged across model specification and time. Excessive dependence on narrow parameter configurations or particular specifications can weaken confidence in the interpretation of historical performance, although neither broad parameter stability nor model simplicity proves robustness. Similarly, performance variation through time or across market environments is informative but not self-explanatory. Observed degradation can be consistent with multiple mechanisms, including sampling variation, model-selection effects, changing relationships, structural change, market adaptation and altered implementation conditions. Degradation is evidence, not a causal diagnosis.

The paper further distinguishes predictive or statistical robustness from economic or implementation robustness. A statistically credible relationship may fail to produce an economically realizable trading outcome once transaction costs, turnover, achievable execution prices, liquidity, market impact or capacity are incorporated. Conversely, implementation failure does not necessarily invalidate the underlying statistical or predictive relationship.

Stress testing, scenario analysis, resampling and Monte Carlo methods can broaden the evaluation beyond a single historical path. Their contribution remains conditional on the assumptions and failure mechanisms they are designed to examine. Synthetic or resampled observations do not become genuinely unseen prospective evidence merely because additional paths have been generated, and greater computational sophistication does not by itself reduce model uncertainty.

These findings are integrated into the RP-003 Multidimensional Robustness Evidence Framework, which organizes model evaluation across six domains:

- Research-Process Validity
- Independent Validation
- Specification & Parameter Stability
- Temporal & Market-Environment Robustness
- Economic & Implementation Robustness
- Stress & Alternative Evidence

The framework does not assign a robustness score or prescribe a universal validation sequence. Instead, each domain is interpreted through four questions: What failure mechanism is being challenged? What evidence does the challenge add? Which assumptions does that evidence depend on? What residual uncertainty remains?

The resulting assessment is a cumulative evidential judgment. This does not mean that validation results are arithmetically additive or equally weighted. The evidential case for robustness is strengthened when materially different and appropriately designed challenges reduce credible alternative explanations for the observed historical performance while leaving residual uncertainty explicit.

The principal conclusion of RP-003 is therefore not that robustness can be proven.

It is that confidence in historically observed model behavior should depend on the quality, independence, relevance and limitations of the evidence supporting its interpretation.

> Historical performance begins the evaluation. It does not finish it.
>
> Backtesting creates evidence. Validation changes the quality and interpretation of that evidence. Neither historical performance nor subsequent validation eliminates uncertainty.

Key Findings

- Historical backtest performance is informative but insufficient evidence of quantitative-model robustness. Its interpretation depends on the historical sample, research process, model specification and implementation assumptions under which the result was produced, as well as on the design and independence of subsequent validation.
- The research process is part of the evidence. Data snooping, repeated testing, model selection, researcher flexibility, information leakage and repeated use of validation data can affect how an apparently successful historical result should be interpreted.
- Out-of-sample evidence is valuable only to the extent that it is meaningfully independent. A historical hold-out or out-of-sample label does not by itself establish independence, and historical out-of-sample evidence should not automatically be equated with genuinely unseen prospective evidence.
- Robustness should not depend excessively on narrow specifications or parameter optima. Stability across reasonable specifications and parameter perturbations can strengthen the evidential case, but neither model simplicity nor observed stability proves robustness.
- Robustness is conditional on time and market environment. Performance degradation can be informative, but observed degradation alone does not identify its cause. Sampling variation, model-selection effects, changing relationships, structural change, market adaptation and implementation conditions may provide competing explanations.
- Predictive robustness and economic robustness are related but distinct. A statistically credible relationship may fail to translate into an economically realizable trading outcome once transaction costs, turnover, liquidity, execution conditions, market impact or capacity are incorporated. Implementation failure does not necessarily invalidate the underlying statistical or predictive relationship.
- Stress testing, scenario analysis, resampling and Monte Carlo methods provide additional but conditional evidence. Their informational value depends on the assumptions and failure mechanisms they address, and simulated or resampled observations do not become genuinely unseen prospective evidence merely because additional paths are generated.
- No single test establishes robustness. The RP-003 Multidimensional Robustness Evidence Framework therefore evaluates robustness across six complementary domains: Research-Process Validity, Independent Validation, Specification & Parameter Stability, Temporal & Market-Environment Robustness, Economic & Implementation Robustness, and Stress & Alternative Evidence.
- Robustness is a cumulative evidential judgment rather than a mechanical score. Multiple weak or highly overlapping tests are not necessarily stronger evidence than one highly informative test that directly addresses a material failure mechanism.

- Validation can strengthen the evidence for robustness, but it cannot eliminate uncertainty or guarantee future performance.

Finovis
AI FOR THE FINANCIAL INDUSTRY

Table of Contents

1. Introduction

1.1 From Historical Performance to Evidence of Robustness

Backtesting is a central instrument in quantitative investment research. By applying a specified model or trading rule to historical market data, researchers can examine how a specified model performs in a historical simulation using observed market data and stated assumptions, evaluate relevant performance characteristics and identify weaknesses that may warrant further investigation. Historical simulation therefore provides meaningful evidence about model behavior. It does not, however, determine how much confidence should ultimately be placed in that evidence.

The distinction matters because a backtest is not an observation of model quality in isolation. Its results are generated within a particular historical sample, model specification, parameter configuration, research process and set of simulation assumptions. Reported performance measures are themselves estimated from finite observations and may be subject to material statistical uncertainty. For example, the sampling properties and interpretation of the Sharpe ratio depend on assumptions about the return-generating process, including serial dependence (Lo, 2002).

The research process can create an additional challenge. When the same historical information is repeatedly used to develop, compare or select models, the interpretation of the best-performing result changes. White (2000) formalizes this data-snooping problem in the context of model selection: searching across multiple specifications using the same data creates opportunities for apparently superior performance to emerge partly through sampling variation rather than persistent predictive structure.

Modern quantitative research environments may amplify this concern because greater computational capacity and model flexibility make it possible to examine increasingly large sets of specifications, parameters and candidate relationships. These concerns have motivated explicit research protocols for quantitative finance that address limited financial-market data, model flexibility and the risk of misapplying increasingly powerful quantitative methods (Arnott, Harvey & Markowitz, 2019).

These limitations do not imply that historical backtests are uninformative. Nor do they imply that a model whose subsequent performance differs from its historical record was necessarily overfit. Instead, they shift the analytical question. The relevant issue is not simply whether historical performance was attractive, but what that performance constitutes evidence of—and what additional evidence is required before greater confidence in model robustness is justified.

RP-003 addresses this distinction by treating historical performance as the starting point of model evaluation rather than its conclusion.

1.2 Research Question and Core Thesis

The central research question of RP-003 is:

> Why is historical backtest performance insufficient evidence of the robustness of a quantitative trading model, and what should a more rigorous model-evaluation framework examine?

The paper evaluates this question through the following Core Thesis:

> Historical backtest performance is informative but insufficient evidence of quantitative-model robustness because observed results may reflect not only persistent economic or statistical relationships, but also sampling variation, model selection, repeated testing, regime-specific relationships and implementation assumptions.
>
> Robustness should therefore be evaluated as a cumulative and falsification-oriented body of evidence: confidence in a model should depend on whether its observed behavior remains economically plausible and statistically credible when subjected to appropriately designed challenges across genuinely unseen data, alternative specifications, parameter perturbations, market environments, implementation conditions and stress scenarios.
>
> Such validation can strengthen the evidence for model robustness, but it cannot establish invariance or guarantee future performance.

Consistent with the analytical scope developed below, the phrase "genuinely unseen data" in the Core Thesis is not used as a synonym for all historical out-of-sample observations. RP-003 distinguishes historical hold-out, pseudo-out-of-sample and out-of-time evidence from genuinely prospective or live observations and evaluates their informational independence according to the surrounding validation design.

The thesis does not assume in advance that any particular validation technique is superior, nor that robustness can be demonstrated through a single statistical test. Instead, the paper investigates the informational contribution and limitations of different forms of evidence and examines how they alter the degree of confidence that can reasonably be placed in historically observed model behavior.

1.3 Scope and Analytical Boundaries

RP-003 examines the evidential robustness of quantitative investment and trading models whose behavior or performance is materially evaluated using historical financial-market data.

The analytical focus begins once an apparently successful historical model or strategy result has been observed. The paper does not primarily investigate how alpha should be discovered or how predictive signals should be constructed. Instead, it asks what evidence is required before historical model performance can support a stronger assessment of robustness.

The analysis covers research-process validity, sample integrity, out-of-sample evaluation, specification and parameter sensitivity, temporal and market-environment stability,

implementation realism, and selected forms of stress and alternative validation evidence. The paper is cross-market in principle, but empirical findings are interpreted within the actual asset class, dataset, period, model type and methodology examined by their source. Evidence from a particular market or strategy family is not automatically generalized to quantitative models as a whole.

RP-003 does not develop or test a proprietary trading strategy, conduct new FINOVIS backtests, prescribe universal validation thresholds, or evaluate the robustness of FINOVIS production models. It also does not attempt to establish a universal model-validation protocol. Machine learning and model complexity are considered only where they materially affect the robustness question, while execution and transaction-cost issues are examined only insofar as they affect the credibility of simulated trading outcomes.

The paper therefore occupies a deliberately defined analytical space between model discovery and production execution:

> An attractive historical result has been observed. What evidence is required before greater confidence in its robustness is justified?

1.4 Structure of the Paper

The analysis proceeds from the interpretation of historical performance toward progressively broader questions about the evidence supporting it.

Section 2 sets out the research methodology and evidence-appraisal approach. Section 3 examines what historical performance can legitimately establish, including statistical uncertainty and conceptual plausibility. Section 4 considers data snooping, multiple testing, model selection and sample-integrity problems that can affect the interpretation of a selected backtest. Section 5 evaluates out-of-sample evidence and the conditions required for meaningful informational independence.

Section 6 turns to specification sensitivity, parameter stability and model complexity. Section 7 examines temporal instability, structural change, market environments and the interpretation of performance degradation. Section 8 distinguishes predictive or statistical robustness from the economic robustness of an implementable trading strategy. Section 9 considers stress testing, resampling, simulation and other complementary forms of robustness evidence.

Section 10 integrates these dimensions into the RP-003 Multidimensional Robustness Evidence Framework. The framework does not produce a robustness score or universal pass/fail determination. Instead, it structures the relationship between potential failure mechanisms, the evidence supplied by different validation approaches, the assumptions on which that evidence depends and the uncertainty that remains.

Section 11 concludes by returning to the central research question and assessing what a rigorous validation process can—and cannot—justify about quantitative-model robustness.

2. Research Methodology

2.1 Critical Literature Review and Analytical Synthesis

RP-003 employs a Critical Literature Review and Analytical Synthesis to examine how the robustness of quantitative investment and trading models can be evaluated when historical financial-market data form a material part of the evidence base.

The designation is consistent with critical-review approaches that emphasize critical appraisal, analysis and conceptual synthesis rather than exhaustive systematic searching (Grant & Booth, 2009; Snyder, 2019). The term Analytical Synthesis describes the specific synthesis function of RP-003 and does not imply a standardized systematic-review or meta-analytic protocol.

The paper does not estimate a new trading model, conduct proprietary backtests, generate new simulations or statistically pool results across existing studies. Instead, it evaluates theoretical, methodological, empirical and selected institutional evidence concerning the limitations of historical model evaluation and the information contributed by different validation approaches.

The research is designed to critically evaluate, qualify and, where necessary, challenge the propositions developed in the paper rather than to assemble evidence in support of a predetermined conclusion. Findings that constrain or contradict stronger interpretations are therefore considered alongside evidence that supports them.

The resulting synthesis is conceptual rather than meta-analytic. Its purpose is to assess what different forms of validation evidence can reasonably establish about historically observed quantitative-model performance, under what assumptions those conclusions hold and what uncertainty remains.

2.2 Source Selection and Evidence Appraisal

The literature search follows a targeted and iterative rather than exhaustive process.

Research is organized around the principal validation dimensions examined in RP-003:

- research-process validity;
- sample integrity;
- out-of-sample and prospective evaluation;
- specification and parameter stability;
- temporal and market-environment robustness;
- implementation realism;
- stress, resampling and alternative evidence.

Initial searches are supplemented by backward citation tracing and, where useful, forward citation tracing to identify foundational work, later methodological developments, empirical applications and relevant counter-evidence.

An internal search and source-selection record is maintained to support traceability of material literature entering the analytical corpus. For material searches, the record captures where practicable the search source or platform, search date, principal search concepts, relevant citation tracing and material inclusion or exclusion decisions.

The process does not imply systematic-review completeness, exhaustive database coverage or a PRISMA-style screening protocol.

Sources are appraised claim by claim rather than through a mechanical hierarchy. Relevant considerations include:

- source authority;
- methodological quality;
- correspondence between the research object and the claim being evaluated;
- market and dataset;
- sample period;
- model or strategy type;
- performance measure;
- validation design;
- stated and implied limitations;
- external validity;
- publication status and currency where relevant.

Peer-reviewed methodological and empirical literature receives priority for material scientific claims where appropriate. Authoritative institutional guidance may be used when it directly informs questions of model validation, model risk or supervisory practice, but is not treated as empirical evidence of trading-strategy behavior unless it actually provides such evidence.

High-quality working papers from recognized researchers, universities or research institutions may be considered where they are directly relevant, materially recent or not yet superseded by a published version; their publication status is taken into account when determining evidential weight.

Practitioner publications and other secondary sources may assist with orientation, terminology, contextual explanation and identification of primary literature, but material scientific claims are traced wherever reasonably possible to original or authoritative sources.

For material claims, identification of a potentially relevant source is not sufficient. The underlying methodological context is reviewed before the source is mapped to a specific proposition. Particular attention is paid to avoiding generalization beyond the market, sample, methodology or research object actually examined.

2.3 Evidence Classification

RP-003 distinguishes explicitly between three forms of content.

External Evidence

External Evidence consists of empirical, statistical, theoretical or institutional findings derived from external sources.

Such findings are presented within the boundaries supported by the underlying research. Results from a specific market, model class or sample are not automatically generalized to quantitative models as a whole.

FINOVIS Analytical Synthesis

FINOVIS Analytical Synthesis refers to conceptual structures developed through comparative interpretation of the reviewed literature.

This includes taxonomies, distinctions between forms of robustness and the eventual RP-003 Multidimensional Robustness Evidence Framework.

These elements represent analytical synthesis rather than proprietary empirical FINOVIS findings. They do not describe confidential FINOVIS production models, research pipelines or validation thresholds.

Illustrative Examples

Illustrative Examples are hypothetical constructions used solely to clarify methodological concepts.

Where numerical values, parameter configurations or model responses are used illustratively, they are identified as hypothetical and do not represent empirical observations or FINOVIS strategy results.

This separation is intended to prevent conceptual interpretation, external evidence and hypothetical explanation from being mistaken for one another.

2.4 Methodological Limitations

The methodology has several limitations.

First, the literature review is targeted rather than exhaustive. Relevant research may therefore exist outside the final analytical corpus, and the search process does not establish complete coverage of all work relating to quantitative-model validation.

Second, the evidence base is heterogeneous. The reviewed literature includes technical trading rules, asset-pricing factors, forecasting models, alternative-beta strategies, general time-series methodology and institutional model-risk frameworks. Findings derived from these different research objects are not necessarily directly comparable.

Third, RP-003 does not conduct a new empirical study. It therefore does not establish through original testing which validation method or combination of methods would be superior for every quantitative model.

Fourth, the academic literature may itself be affected by publication, selection and replication problems of the type examined in this paper. Where material, RP-003 therefore incorporates methodological criticism, replication studies and contrary findings rather than relying exclusively on published positive results.

Finally, the RP-003 Multidimensional Robustness Evidence Framework is an analytical synthesis, not an empirically validated diagnostic instrument, regulatory standard or model-certification procedure. Its purpose is to organize the interpretation of evidence and residual uncertainty, not to establish a universal robustness score or guaranteed validation outcome.

3. Historical Performance as Conditional Evidence

3.1 What Backtesting Can Establish

A historical backtest provides a structured way to evaluate how a specified quantitative model behaves when applied to observed market data under defined simulation assumptions. Properly constructed, it can describe characteristics that are difficult to assess from model design alone, including the distribution and timing of simulated returns, risk and turnover characteristics, and model behavior during historical conditions represented in the evaluation sample (CFA Institute, 2026).

This makes backtesting an important component of quantitative research. It allows researchers to move from an abstract model specification to an observable set of simulated outcomes and provides a common basis for diagnosing model behavior, comparing prespecified alternatives and identifying questions that require further investigation. In this sense, historical performance is genuine evidence.

The evidential claim must nevertheless remain conditional. A backtest establishes neither what would necessarily have occurred in actual implementation nor how the model will behave in future observations. It describes model behavior within a particular historical record and under the assumptions embedded in the simulation. The relevant interpretation is therefore not:

> the model generated this performance

but more precisely:

> the specified model generated this simulated performance under these data, assumptions and evaluation conditions.

That distinction is foundational to RP-003. The objective is not to diminish the information contained in historical simulation, but to separate the observation of attractive historical outcomes from the stronger inference that those outcomes constitute evidence of model robustness.

3.2 Performance Measures and Statistical Uncertainty

The performance reported by a backtest is itself estimated from a finite sample. Average returns, volatility, risk-adjusted performance and related statistics are therefore subject to sampling uncertainty.

A point estimate summarizes what was observed in the historical sample, but it does not by itself communicate the estimation uncertainty surrounding that statistic or establish that the process generating it is stable beyond the sample examined.

The Sharpe ratio provides a useful illustration. Its inputs—expected return and volatility—are not known constants but quantities estimated from observed returns. Lo (2002) demonstrates that the statistical properties of estimated Sharpe ratios depend on assumptions about the return-generating process and that serial dependence can materially affect their interpretation and time aggregation. The implication is broader than the Sharpe ratio itself: the apparent precision of a historical performance statistic should not be confused with precision in the underlying inference.

This distinction becomes especially relevant when relatively small differences in estimated performance are used to support materially different conclusions about model quality. Historical evaluation should therefore consider not only the level of a reported performance measure but also the uncertainty surrounding it, the structure of the underlying observations and the assumptions required for its interpretation.

Statistical uncertainty does not make historical performance meaningless. It changes the form of the question from:

> How high was the reported performance?

to:

> How informative is that estimate, given the data from which it was obtained?

3.3 Conceptual and Economic Plausibility

Statistical evidence is not the only consideration in evaluating a quantitative model. The relationship between a model's stated purpose, its design, its assumptions and the behavior it is intended to capture can also affect how its historical results should be interpreted.

Conceptual scrutiny may include asking whether key assumptions are coherent with the intended use of the model, whether the model is being evaluated against an appropriate benchmark and whether its behavior is reasonably consistent with the mechanism or forecasting task it is intended to represent. In some applications, an economic or market rationale may contribute to this assessment. In others, particularly where models are highly statistical or complex, interpretability or comparison with alternative models may provide more useful forms of conceptual challenge.

This broader approach is consistent with the 2026 interagency *Supervisory Guidance on Model Risk Management*, which distinguishes conceptual soundness from outcome analysis and recognizes

that theoretical construction may be informative for some models while interpretability measures or benchmarking may be more practical for others (Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation & Office of the Comptroller of the Currency, 2026). The guidance is used here as an institutional model-validation reference, not as empirical evidence concerning quantitative trading strategies.

Conceptual plausibility must nevertheless not be treated as a substitute for empirical evidence. An attractive economic explanation does not demonstrate that a historical relationship is statistically robust. Conversely, the absence of a simple economic narrative does not by itself invalidate a statistically credible model.

Accordingly:

> A plausible story is not a substitute for validation, and the absence of a simple story is not a substitute for evidence of failure.

Conceptual and empirical evaluation should inform one another without being treated as interchangeable.

3.4 Performance Is Conditional on the Simulation Design

Historical performance cannot be separated from the simulation design that generated it.

At a minimum, reported results depend on the historical data used, the model specification, parameter choices, performance definition and the assumptions governing how model outputs are translated into simulated outcomes. Depending on the strategy, conclusions may also be sensitive to choices concerning data frequency, portfolio construction, rebalancing, leverage, transaction assumptions, benchmarks or other elements of the evaluation architecture.

The purpose of recognizing this conditionality is not to suggest that every modeling choice is arbitrary. Many choices may be economically necessary, theoretically justified or fixed before the evaluation begins. The important point is that the observed performance is conditional on them.

Arnott, Harvey & Markowitz (2019) emphasize the importance of research protocols in quantitative finance in an environment characterized by limited financial-market data and increasingly flexible analytical tools. For purposes of RP-003, this supports a broader analytical conclusion: methodological sophistication does not remove the need to understand how research design conditions the evidence produced by a backtest.

A backtest should therefore be read together with the specification of the experiment that produced it. Two reported performance figures need not represent comparable evidence merely because they use the same headline metric. Differences in model construction, data, assumptions or evaluation design may materially change what those figures mean.

The appropriate object of interpretation is consequently not performance in isolation but:

> performance conditional on the research and simulation design under which it was observed.

3.5 What an Attractive Backtest Cannot Establish

An attractive historical backtest can demonstrate that a particular model specification produced favorable simulated outcomes in the historical environment examined. It cannot, by itself, establish why those outcomes occurred or how much confidence should be placed in their persistence.

Historical performance alone does not determine whether the result is materially affected by model selection or repeated research choices. It does not establish that similar behavior persists under meaningfully independent evaluation, that the model is insensitive to reasonable specification changes, that the observed relationship remains stable through different market environments, or that simulated performance survives economically credible implementation assumptions. Nor can a favorable historical record establish invariance with respect to future market structure.

These are not reasons to discard backtesting. They are reasons to interpret it as conditional evidence requiring further challenge.

The analytical transition at the end of this section is therefore deliberate. Section 3 has asked what information an observed backtest contains and what limits apply to the interpretation of its headline performance. The next stage examines a different question: whether the research process that produced and selected the observed result can itself alter the strength of the evidence.

Historical performance is the starting observation.

> Robustness remains the proposition to be evaluated.

4. When the Research Process Creates Apparent Performance

4.1 Data Snooping and Repeated Data Use

The interpretation of a historical result depends not only on the final model but also on how the historical data were used before that model was selected.

White (2000) defines data snooping as the repeated use of a given dataset for inference or model selection. When researchers examine multiple models, specifications, signals or trading rules using the same historical information, the data influence not only the measurement of performance but also the process that determines which result ultimately receives attention.

This creates an important distinction between two research settings. In the first, the model and material evaluation choices are fixed before the evaluation result is observed. In the second, multiple candidate models or specifications are examined using the same historical information and the most attractive result is subsequently selected. Even where the selected models display

identical historical performance, the evidential interpretation of the two observations is not necessarily the same because the second result emerged from a broader search process.

White's Reality Check was developed specifically to address the inferential consequences of specification search and to test whether the best-performing model encountered in such a search demonstrates predictive superiority over a benchmark after accounting for data snooping (White, 2000). Sullivan, Timmermann & White (1999) applied this general problem to a large universe of technical trading rules, illustrating why an apparently successful rule should be evaluated in the context of the wider set from which it was selected rather than as an isolated observation.

The central issue is therefore not repeated computation itself. Historical data can legitimately be used for research, model comparison and diagnosis. The evidential problem arises when the same information influences the creation or selection of a result while the final result is interpreted as though no such search had taken place.

4.2 Multiple Testing and Strategy Selection

Data snooping becomes especially important when the research process involves many candidate hypotheses.

As a simple statistical benchmark, if many independent null hypotheses were tested under identical conventional significance criteria, the opportunity for at least some apparently unusual observations to arise by chance would increase with the number of tests. Quantitative research is usually more complicated than this simple setting because candidate strategies, parameters and signals may be statistically dependent, tests may be sequential rather than simultaneous, and the research universe itself may evolve as results are observed.

The underlying multiple-testing problem is nevertheless well established in financial research. Harvey, Liu & Zhu (2016), examining the cross-section of expected returns, argue that conventional significance criteria become difficult to interpret when large numbers of candidate factors have been investigated and develop a framework designed to account for multiple testing. Their empirical conclusions concern the particular asset-pricing literature they examine and should not be interpreted as a universal estimate of false discoveries across quantitative trading models.

Bailey & López de Prado (2014) address a related problem from the perspective of backtest evaluation. Their Deflated Sharpe Ratio explicitly incorporates selection bias arising from multiple trials together with the effect of non-normal returns. The methodological point relevant to RP-003 is not that every research process requires this particular statistic, but that the number and character of prior trials can affect how a selected performance measure should be interpreted.

Multiple-testing controls also involve trade-offs. Harvey & Liu (2020) emphasize both false discoveries and missed discoveries, showing that procedures designed to reduce false positives can also affect statistical power. A more demanding evidential standard is therefore not equivalent to mechanically rejecting more models.

The appropriate question is broader:

How much opportunity did the research process provide for an apparently successful result to emerge, and how should that opportunity affect the interpretation of the selected result?

4.3 Parameter Optimization and Researcher Flexibility

The relevant research universe may extend far beyond a set of formally distinct strategies.

A single strategy concept can generate a large family of candidate results through choices concerning parameters, estimation windows, signal definitions, filters, portfolio rules, data transformations, performance metrics and other elements of model design. If these choices are repeatedly adjusted after historical outcomes are observed, the resulting performance reflects both the underlying model concept and the path through which the final specification was selected.

Bailey & López de Prado (2014) describe how backtest optimization across alternative parameter combinations can generate performance inflation and backtest overfitting. This does not imply that parameter optimization is inherently invalid. Many quantitative models require calibration, and systematic model selection can be a legitimate part of research. The evidential concern arises when optimization freedom is large relative to the available information and the resulting search is not reflected in the interpretation of the selected performance.

Researcher flexibility also need not be fully automated to matter. Decisions to modify a filter, alter a sample, change a performance criterion or abandon one specification in favor of another can all become part of the effective research process if they are informed by observed historical results.

For RP-003, the relevant object is therefore the `research path`, not merely the final source code or mathematical specification.

The evidential interpretation of a selected historical result therefore cannot rest solely on the model's terminal form when substantial historical information also influenced how that form was reached.

4.4 Effective Multiplicity and Dependence

The number of research choices alone is not a sufficient description of the selection problem.

Candidate models are often related. Parameter variants may produce highly correlated returns; closely related factors may encode similar information; alternative specifications may differ only marginally; and sequential research decisions may depend on earlier results. Treating every variation as a fully independent statistical trial may therefore mischaracterize the structure of the research process.

This is one reason the literature does not reduce the problem to a universal adjustment based solely on the nominal number of tests. Hansen's (2005) test for Superior Predictive Ability, developed as an extension of White's Reality Check, is designed to reduce sensitivity to poor and

irrelevant alternatives within a model-comparison universe. This illustrates why the composition of the candidate set can matter for inference.

Dependence among candidate hypotheses creates a related but distinct problem. Giglio, Liao & Xiu (2021) develop a multiple-hypothesis-testing framework for linear asset-pricing models that accommodates features including cross-sectional dependence, omitted factors and missing data. Their empirical application is specific to linear asset-pricing and hedge-fund-performance settings, but the methodological lesson is relevant more generally: multiplicity is a joint inferential problem rather than simply a count of research outputs.

RP-003 therefore avoids treating the effective number of research trials as an observable universal quantity that can always be reconstructed precisely after the event. Instead, the concept is used diagnostically.

Relevant questions include how many materially different research paths were available, how dependent those paths were, how much discretion existed, and how strongly observed results influenced subsequent choices.

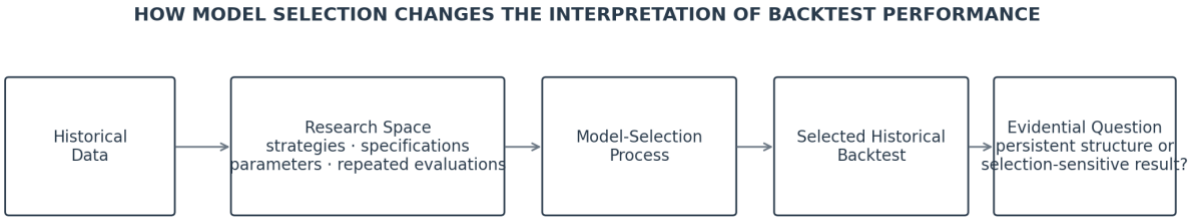


Figure 1. How Model Selection Changes the Interpretation of Backtest Performance

Historical Data → Research Space (multiple strategies, specifications, parameters and repeated evaluations) → Model-Selection Process → Selected Historical Backtest → Evidential question: persistent structure or selection-sensitive result?

Source: FINOVIS analytical synthesis based principally on White (2000) and Bailey et al. (2017); see Section 4.

4.5 Sample Integrity

Model selection is not the only way in which the research process can compromise the interpretation of historical evidence. The information available to a model or researcher must also correspond to what the evaluation design claims was available at the relevant point in time.

Several distinct problems fall within this broader concept of sample integrity.

Survivorship bias can arise when the historical sample is conditioned on entities that survived until a later date, thereby excluding observations that would have been part of the information set available to a contemporaneous investor. Brown et al. (1992), studying mutual-fund performance, demonstrate that survivorship-related sample truncation can generate an appearance of predictability. That finding is specific to the performance-study setting they examine, but it establishes why the construction of the historical universe can materially affect inference.

Look-ahead bias arises when the evaluation uses information that was not legitimately available at the point when the simulated decision was supposed to have been made. Duarte et al. (2026), in empirical options research, show that apparently attractive strategy results can arise when

information unavailable at portfolio formation is used to filter observations considered noisy or erroneous. Their finding concerns the options-research designs examined in that study and should not be treated as a quantitative estimate of look-ahead bias across other markets.

A related but broader problem is information leakage. Kaufman et al. (2012) describe leakage in data mining as the introduction of information about the prediction target that should not legitimately be available to the model. The source is general data-mining methodology rather than finance-specific empirical research. For purposes of RP-003, its methodological principle is applied analytically: an evaluation boundary can be weakened if information that should not be available at model construction or prediction enters through data preparation, feature construction, preprocessing or related research steps.

Sample integrity can also deteriorate through adaptive reuse. Dwork et al. (2015), in the general statistical literature on adaptive data analysis, show that a hold-out dataset can itself become subject to overfitting when its results repeatedly influence subsequent analyses. The paper is not evidence about the prevalence of such behavior in quantitative finance. It establishes the methodological principle that nominal separation of a dataset does not necessarily preserve informational independence when feedback from that dataset enters later research decisions.

For RP-003, these mechanisms should not be collapsed into a single label. Survivorship bias, look-ahead bias, target leakage and adaptive validation reuse describe different failures and may require different controls.

Their common evidential implication is narrower:

> a historical result cannot be interpreted correctly without understanding what information entered the research process, when it became available and how it influenced model construction or evaluation.

4.6 Interpreting Performance After Model Selection

Once historical performance has emerged from a model-selection process, its evidential interpretation cannot necessarily rest solely on the selected model's observed performance, a comparison with zero, or an informal benchmark. The selection process may itself become relevant to the inference.

White (2000), Sullivan, Timmermann & White (1999), Hansen (2005), Bailey & López de Prado (2014) and later multiple-testing approaches differ in their precise objectives and assumptions, but they share an important methodological concern: inference concerning a selected model can change when selection occurred across a broader universe of alternatives.

Bailey et al. (2017) address this problem from another perspective by proposing a framework for assessing the probability of backtest overfitting in investment simulations. Their methodology does not establish a universal probability that backtests are overfit. Rather, it illustrates that the relationship between in-sample model selection and subsequent evaluation can itself be treated as an object of analysis.

This leads to a central proposition of RP-003:

> the research process is part of the evidence.

A historical result should therefore be interpreted together with the material research choices that preceded it: the candidate universe, model-selection freedom, parameter search, repeated data use, dependence among alternatives and the integrity of the information boundaries used during evaluation.

This does not imply that every flexible research process invalidates its final model. Nor does it establish that a selected high-performing model is merely a statistical accident. The conclusion is more limited and more important.

Where extensive search or adaptive model selection has taken place, the evidential burden associated with the selected historical performance changes.

The relevant question is not merely:

> How well did the selected model perform?

It is:

> How much model-selection freedom existed for a seemingly successful result to emerge, and how should the multiplicity and dependence of those research choices affect its interpretation?

5. Out-of-Sample Validation and the Problem of Independence

5.1 In-Sample and Out-of-Sample Evidence

Out-of-sample evaluation is intended to answer a different question from in-sample model fitting.

In-sample data contribute directly to model estimation, calibration, specification or selection. Performance measured on those observations therefore reflects, to some degree, the information used to construct the model. Out-of-sample evaluation seeks to examine model behavior on observations that were excluded from defined parts of that development process.

For purposes of RP-003, out-of-sample is therefore not treated as a self-explanatory label. It means that the observations being evaluated were excluded from specified estimation, calibration or selection decisions under the stated research design.

This distinction matters because different designs exclude different information from different decisions. A dataset may be unused for parameter estimation but still influence model selection. It may be excluded during initial development but later inspected repeatedly. Alternatively, a historical test period may remain uninspected within the documented research process until all material model and evaluation choices have been fixed.

Forecast-evaluation research provides formal methods for assessing predictive performance. Diebold & Mariano (1995) develop tests of equal predictive accuracy between competing forecasts under relatively general forecast-error conditions. Building on this literature, Giacomini & White (2006) develop an explicitly out-of-sample framework for predictive-ability testing and forecast selection that also accommodates conditional evaluation objectives and the effects of parameter estimation.

These methods address particular inferential questions. Their existence does not imply that any observation described as out-of-sample automatically provides strong independent evidence.

For RP-003, the relevant question is therefore not simply:

> Was the performance measured out of sample?

It is:

> From which research decisions was the evaluation sample actually independent?

5.2 Why a Hold-Out Is Not Automatically Independent

A conventional hold-out design separates available historical observations into a development sample and a test sample. If the test sample remains untouched while the model and material evaluation rules are developed, it can provide evidence not directly used in constructing the selected model.

The strength of that evidence depends, however, on preserving the boundary.

Once a researcher observes hold-out performance and responds by changing the model, parameters, features, filters, strategy definition or evaluation criteria, information from the hold-out has entered the development process. Subsequent evaluation on the same observations is no longer informationally equivalent to the first untouched evaluation.

Dwork et al. (2015), in the broader literature on adaptive data analysis, formalize this problem by showing that repeated adaptive interaction with a hold-out can lead to overfitting to the hold-out itself. As established in Section 4, this is general statistical methodology rather than finance-specific evidence. Its relevance to RP-003 is the underlying principle: data separation and informational independence are not the same thing when evaluation feedback influences subsequent research decisions.

The investment-backtesting literature raises a related concern. Bailey et al. (2017) argue, in the context of investment simulations and their proposed probability-of-backtest-overfitting framework, that standard hold-out techniques can be unreliable when researchers repeatedly select among investment strategies using limited historical data. Their proposed combinatorially symmetric cross-validation methodology addresses a specific form of this problem; it does not establish that every conventional hold-out is invalid.

The relevant evidential distinction is therefore between:

> data that were formally separated

and

> data whose outcomes remained meaningfully independent of model development and selection.

A hold-out can provide valuable evidence. The label alone does not establish how independent that evidence remained.

5.3 Pseudo-Out-of-Sample and Out-of-Time Evaluation

Historical data can also be evaluated in chronological sequence so that model estimation at a given point uses only information that would have been available up to that point.

Terminology varies across research traditions. For purposes of RP-003, pseudo-out-of-sample evaluation refers to historical sequential designs that simulate forecasting or model evaluation using only information permitted by the stated historical information set at each evaluation point. The broader term out-of-time is used for evaluations in which a chronologically later historical segment is separated from earlier model-development observations. The precise evidential meaning depends on the design rather than on the label.

Their principal advantage is that they can preserve temporal ordering and approximate the information structure of sequential forecasting more closely than a randomly constructed historical split.

Out-of-sample forecast-evaluation research demonstrates that the design of such exercises is itself consequential. Tashman (2000), reviewing out-of-sample forecasting tests, identifies choices including the division between fit and test periods, fixed versus rolling forecast origins, coefficient updating or recalibration, fixed versus rolling estimation windows and the use of single or multiple evaluation periods.

Giacomini & White (2006) similarly show that out-of-sample predictive evaluation can address different objectives, including unconditional questions about average predictive performance and conditional questions about when one forecast may outperform another.

Historical pseudo-out-of-sample evidence therefore represents more than a mechanical split between earlier and later observations. Its interpretation depends on what was estimated at each point, what was re-estimated, what information entered the model, what evaluation objective was used and whether later results subsequently influenced further research.

Most importantly, an out-of-time historical observation remains historical evidence.

It may have been unavailable to the model at the simulated decision date while still being available to the present-day researcher designing the study. If the researcher has previously inspected, modeled or otherwise learned from that period, chronological ordering alone does not restore informational independence.

For this reason, RP-003 distinguishes historical out-of-sample evidence from genuinely prospective evidence.

5.4 Walk-Forward Validation

Walk-forward evaluation extends the logic of chronological testing by repeatedly advancing the evaluation origin through time.

At each stage, the evaluation origin advances through time and subsequent observations are used for assessment under a predefined historical information structure. Depending on the design, model coefficients or parameters may remain fixed, be updated, or be recalibrated using newly available observations.

The process is repeated as the evaluation origin advances. Depending on the design, the estimation sample may expand, remain fixed in length, or be recalibrated according to predefined rules.

Tashman (2000) analyzes rolling-origin evaluation as an important form of out-of-sample forecast assessment and shows that design choices concerning updating, recalibration, window structure and test periods influence the resulting evaluation.

Walk-forward designs can provide several useful forms of information. They can generate multiple sequential evaluation periods rather than allowing conclusions to depend entirely on one arbitrary historical split. They can also more closely represent research settings in which models are periodically re-estimated or recalibrated as new information becomes available.

They do not, however, remove the underlying problem of research feedback.

If the researcher repeatedly examines walk-forward results and modifies the model in response, the entire historical procedure can gradually become part of the development process. Likewise, a walk-forward test says little about future performance merely because its simulated chronology resembles live operation.

Walk-forward validation should therefore be interpreted as a historical sequential evaluation design, not as a substitute for genuinely prospective observation.

Its evidential strength depends on the rules governing model freezes, recalibration, evaluation horizons, information availability and researcher feedback.

5.5 Cross-Validation for Financial Time Series

Cross-validation provides another family of approaches for evaluating predictive models, but its application to time-dependent financial data requires particular care.

Standard cross-validation procedures developed for independent observations can disrupt temporal ordering or create dependencies between training and evaluation samples that are inappropriate for a given forecasting problem. The relevant issue is not that conventional cross-validation is universally invalid for time series, but that its validity depends on the statistical structure of the problem and the design of the procedure.

Bergmeir, Hyndman & Koo (2018) demonstrate this point explicitly. For purely autoregressive models, they show that standard K-fold cross-validation can be valid under conditions including uncorrelated model errors. Their result is deliberately narrow. For purposes of RP-003, it supports an analytical conclusion that neither blanket acceptance nor blanket rejection of K-fold cross-validation for time-dependent data is methodologically justified without examining the structure of the specific forecasting problem.

Time-series-specific procedures commonly preserve chronological information more explicitly. Rolling-origin approaches, for example, estimate the model using observations occurring before each evaluation point and assess forecasts on later observations. Such designs can be adapted for different forecast horizons, training-window lengths and re-estimation rules.

Investment-strategy evaluation introduces additional complications because the object being selected may be a complete trading strategy rather than a conventional forecasting model. Bailey et al. (2017), for example, propose combinatorially symmetric cross-validation specifically to estimate the probability of backtest overfitting in investment simulations. This methodology answers a particular model-selection question and should not be treated as a universally preferred cross-validation design for financial time series.

RP-003 therefore does not designate one form of cross-validation as intrinsically superior.

The relevant questions are instead:

- Does the procedure respect the information structure relevant to the model?
- What observations influence estimation, calibration and selection?
- How is temporal dependence handled?
- Are training and evaluation observations statistically dependent in ways material to the inference?
- Did evaluation outcomes subsequently influence the model being evaluated?

Cross-validation contributes evidence only to the extent that its design corresponds to the failure mechanism it is intended to challenge.

5.6 Prospective and Live Evidence as a Distinct Form of Validation Evidence

Historical validation procedures can increase the degree of separation between model development and evaluation, but they cannot recreate one feature of genuinely prospective observation: the future outcome did not yet exist when the model and evaluation rules were fixed.

For purposes of RP-003, prospective evidence begins when material model specifications, parameters and evaluation rules are defined before the observations on which they will be judged become available.

The distinction among historical hold-out, pseudo-out-of-sample or rolling historical evaluation, prospective evaluation and live implementation evidence is used in RP-003 as an analytical taxonomy for comparing forms of informational independence. It is not presented as a standardized hierarchy of validation methods.

This does not make prospective evidence infallible. Researchers may still alter models during the evaluation period, stop or extend an observation window after seeing results, change performance criteria, or use interim observations to guide subsequent decisions. Prospective status therefore strengthens one dimension of informational independence only when the surrounding research protocol preserves it.

Live implementation evidence is related but distinct. In addition to chronological prospective observation, it records model behavior in an actual operational environment rather than solely within a historical simulation. This can provide information unavailable from a backtest alone.

The distinction should not be overstated. A difference between historical and live results does not by itself identify why the difference occurred. Possible explanations—including structural change, statistical variation, implementation conditions or research-process effects—are examined separately in later sections. Section 5 is concerned only with the form and independence of the evidence, not with causal diagnosis of degradation.

The hierarchy is therefore deliberately avoided.

A carefully preserved historical hold-out may answer some questions more cleanly than a short or adaptively managed live record. A long pseudo-out-of-sample study may provide information across more historical environments than a limited prospective period. Prospective evidence can offer a stronger form of temporal non-exposure, while live evidence introduces actual implementation conditions. These forms of evidence are complementary rather than mechanically ranked.

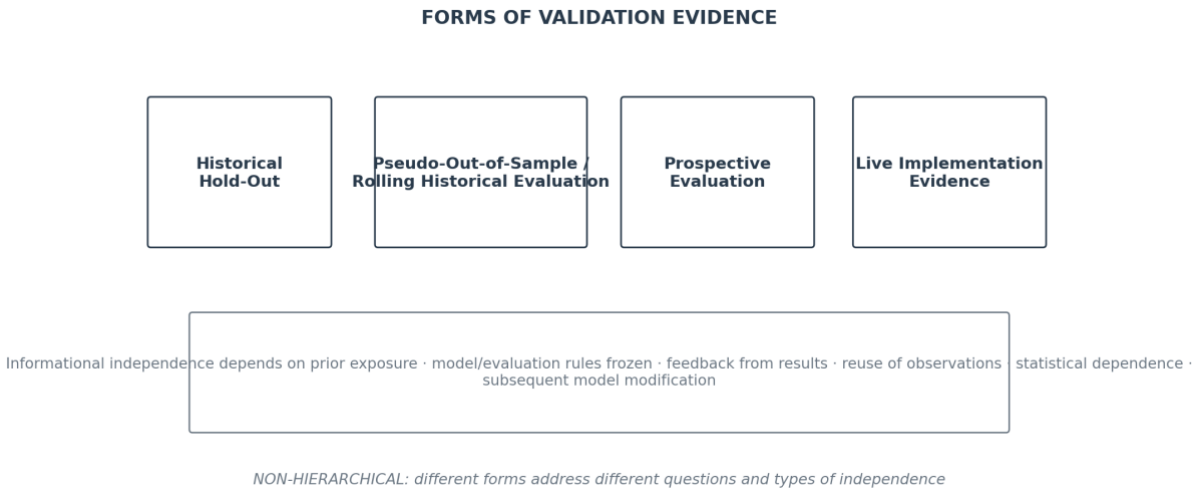


Figure 2. Forms of Validation Evidence and Conditions for Independence

Historical Hold-Out | Pseudo-Out-of-Sample / Rolling Historical Evaluation | Prospective Evaluation | Live Implementation Evidence

For each form, informational independence depends on prior exposure to the evaluation data, whether the model and evaluation rules were frozen, feedback from evaluation results, reuse of observations, statistical dependence and subsequent model modification.

The figure is non-hierarchical: different forms of validation evidence address different questions and may provide different degrees and types of independence.

Source: FINOVIS analytical synthesis based on the literature reviewed in Section 5. The classification is non-hierarchical and is not presented as a standardized validation taxonomy.

The principal conclusion of Section 5 is therefore not that robustness requires a particular validation label.

It is that validation evidence becomes informative to the extent that the research design creates and preserves meaningful independence between the evidence being evaluated and the decisions that produced the model.

Historical out-of-sample evidence can strengthen the case for robustness.

It should not automatically be equated with genuinely unseen prospective evidence.

6. Specification, Parameter Stability and Model Complexity

6.1 Narrow Optima

A quantitative model can appear attractive under one specification while producing materially different results under nearby alternatives.

This possibility matters because the selected parameter configuration is rarely the only conceivable representation of the underlying model idea. Lookback horizons, thresholds, weighting parameters, estimation windows, signal transformations and other choices may admit ranges of values that are defensible independently of the particular historical optimum ultimately selected.

If attractive performance is concentrated around a very narrow configuration, the historical result becomes more dependent on the exact specification selected. This does not establish that the model is overfit. A genuine relationship may itself be nonlinear, threshold-dependent or economically concentrated within a limited parameter region. The relevant evidential point is narrower: dependence on a precise local optimum increases the importance of understanding why that particular configuration performs differently from reasonable nearby alternatives.

The problem is especially important when the optimum was itself discovered through historical search. In that setting, parameter sensitivity interacts with the model-selection issues examined in Section 4. A sharp historical optimum may reflect economically meaningful structure, sampling variation, optimization against the historical sample, or some combination of these mechanisms.

Observed geometry alone does not identify which explanation is correct.

Accordingly, RP-003 treats a narrow optimum as a diagnostic observation requiring additional scrutiny, not as a universal indicator of model failure.

6.2 Parameter Perturbation and Sensitivity

Parameter sensitivity analysis asks how materially the model's evaluated behavior changes when selected inputs are varied within a defined range.

A basic exercise may hold the broader model specification constant while perturbing one parameter around its selected value. More extensive analysis may vary multiple parameters jointly, examine interactions or compare different regions of the admissible parameter space.

Wu et al. (2024), in a quantitative-trading application, examine the related concept of a parameter plateau and develop an optimization procedure intended to identify parameter regions with greater performance stability. Their methodology provides a specific example of parameter-stability analysis; it does not establish that plateau geometry is a universal or sufficient measure of model robustness.

The purpose of parameter perturbation is not simply to locate another high-performing configuration. It is to determine how much of the original historical conclusion depends on the exact parameter values selected.

For purposes of RP-003, lower sensitivity to modest, defensible perturbations can strengthen the evidential case that the observed result is not exclusively a feature of a single numerical configuration. Higher sensitivity can indicate that interpretation of the selected backtest depends materially on parameter choice and therefore warrants greater caution or further investigation.

Neither observation should be interpreted mechanically.

Sensitivity depends on how the parameter is represented, the range over which it is perturbed, the performance or behavior metric being examined and interactions with other model components. A movement that appears small in one parameterization may represent a materially different model in another. Similarly, a stable Sharpe ratio can coexist with instability in turnover, drawdown, exposure or other economically relevant characteristics.

Parameter sensitivity is therefore not a scale-invariant robustness statistic.

It is a diagnostic procedure whose meaning depends on the model and the perturbation experiment being performed.

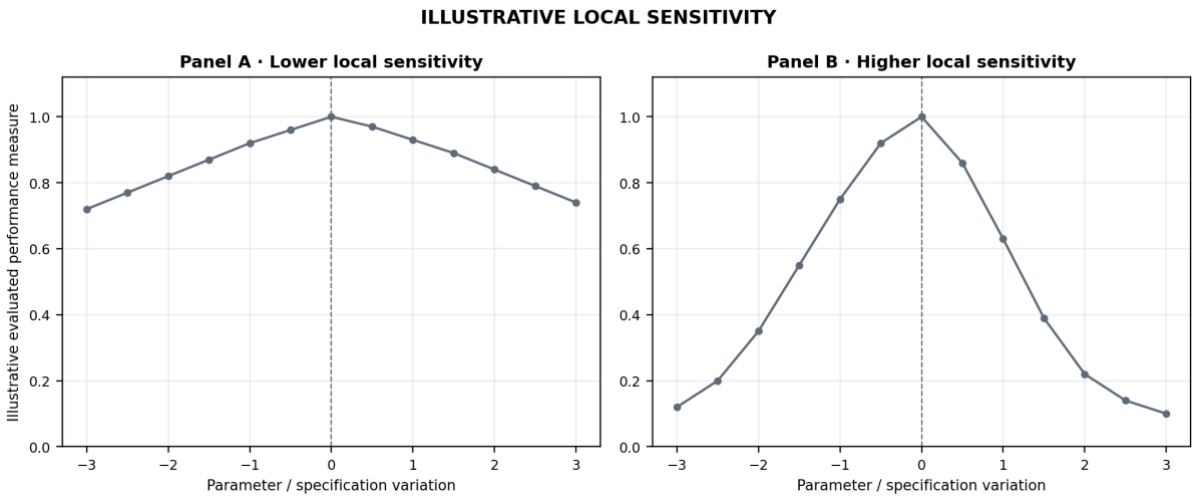


Figure 3. Illustrative Local Sensitivity to Parameter or Specification Variation

Panel A: Lower local sensitivity — the evaluated performance measure changes comparatively gradually around the selected configuration.

Panel B: Higher local sensitivity — the evaluated performance measure changes comparatively sharply around the selected configuration.

The panels are illustrative only. They do not classify either model as robust or overfit. Apparent sensitivity depends on parameterization, perturbation range, evaluation metric and the surrounding model structure.

Source: FINOVIS illustrative example. No empirical data represented.

6.3 Alternative Reasonable Specifications

Parameter perturbation examines variation within a given model structure. A broader robustness question asks whether the historical conclusion persists when other reasonable specifications of the underlying research idea are considered.

Depending on the model, such alternatives may include different but defensible signal definitions, estimation windows, normalization methods, portfolio-construction rules, benchmark choices, transformations or other assumptions that are not uniquely determined by theory or the intended economic mechanism.

The word reasonable is important.

A robustness exercise should not become another unrestricted search for whichever alternative specification happens to preserve attractive performance. Alternative specifications should instead correspond to substantively defensible modeling choices rather than alternatives selected because they reproduce the desired historical result.

Arnott, Harvey & Markowitz (2019) emphasize the importance of disciplined research protocols in quantitative finance when flexible analytical tools interact with limited financial-market data. For purposes of RP-003, this supports an analytical principle: specification robustness is most informative when the alternatives being examined represent genuine challenges to the original result rather than post hoc attempts to reproduce it.

Consistency across alternative reasonable specifications can strengthen confidence that a conclusion is not exclusively dependent on one modeling choice. Material disagreement across alternatives can reveal specification dependence that a single backtest conceals.

Again, the inference is diagnostic rather than binary.

A change in results does not automatically establish that the original specification is wrong. Different specifications may represent different economic hypotheses, information sets or decision rules. The analytical question is whether the historical conclusion remains credible once the range of reasonable modeling uncertainty is made visible.

6.4 Simplicity, Complexity and Flexibility

Model complexity is frequently associated with overfitting, but the relationship should not be reduced to the proposition that simpler models are inherently more robust.

Greater flexibility can allow a model to represent nonlinear relationships, interactions or other structures that a simpler specification cannot capture. At the same time, additional parameters, transformations, hyperparameters and modeling choices can enlarge the set of relationships the research process is capable of fitting.

This creates an evidential trade-off rather than a universal ranking.

In an empirical asset-pricing application, Gu, Kelly & Xiu (2020) examine machine-learning methods with different degrees of flexibility and emphasize regularization, temporally separated training, validation and testing samples, and hyperparameter tuning as mechanisms for controlling overfitting and assessing predictive performance. Their findings concern the specific return-prediction problem and model classes they study; they do not establish that greater model complexity is inherently inferior.

Evidence from commercially promoted investment strategies also illustrates why complexity may be relevant to validation. Suhonen, Lennkh & Perez (2017) examine 215 alternative-beta strategies developed by global investment banks across five asset classes. In their sample, Sharpe ratios deteriorated materially between backtested and live periods, and the reduction was larger for strategies classified as more complex under the study's complexity classification. The study establishes an empirical association within that strategy universe; it does not demonstrate that complexity alone caused the deterioration or that the reported magnitudes generalize to other quantitative models.

For RP-003, model complexity is therefore relevant principally through flexibility and evidential burden.

A model that permits many plausible specifications, tuning parameters or adaptive choices may require stronger evidence that its attractive historical behavior is not dependent on those degrees of freedom. Conversely, a simple model does not become robust merely because it has few parameters. A simple specification can still be based on a weak relationship, a contaminated sample or a narrow historical environment.

The relevant question is not:

> Is the model simple or complex?

It is:

> How much flexibility did the model and research process possess, and is the validation evidence commensurate with that flexibility?

6.5 Stability as Evidence, Not Universal Threshold

Specification and parameter stability can strengthen the evidential case for a quantitative model, but RP-003 does not define robustness through a universal stability threshold.

There is no general basis for requiring that every model preserve a fixed percentage of performance under a predetermined parameter change, remain profitable across every neighboring specification or exhibit a parameter region of a specified width.

Such thresholds would require assumptions about the model, parameterization, economic mechanism, data frequency, performance measure and intended use. A perturbation that is economically minor for one model may fundamentally alter another.

Stability should instead be interpreted relative to the particular failure mechanism being challenged.

If a model's central historical conclusion survives reasonable parameter perturbations and alternative specifications, that evidence can reduce concern that the result depends exclusively on a narrow choice. If the conclusion changes materially under small but meaningful variations, the sensitivity becomes evidence of specification dependence and increases the need for explanation or additional validation.

Neither outcome settles the robustness question.

Parameter and specification tests address one dimension of the evidence. They do not establish independence from model selection, temporal stability, implementation feasibility or resilience under other relevant challenges.

This leads to the central proposition of Section 6:

> Robustness evidence is stronger when a model's historical interpretation does not depend excessively on narrow specifications or parameter optima, but stability is evidence rather than a universal pass/fail criterion.

The same principle applies to complexity.

> Greater flexibility can increase the validation burden. It does not, by itself, establish overfitting.

7. Robustness Through Time and Changing Market Environments

7.1 Non-Stationarity and Structural Change

Historical model evaluation implicitly raises a temporal question: to what extent can relationships observed in one period be treated as informative about another?

A quantitative model may rely on parameters, relationships or conditional distributions that are sufficiently stable for historical estimation to remain informative over the period being considered. That stability should not, however, be assumed merely because a model produced an attractive aggregate backtest.

Statistical methods for structural-change analysis provide formal ways to investigate whether model parameters remain stable through time. Andrews (1993), for example, develops tests for parameter instability and structural change when the location of a potential change point is unknown. Bai & Perron (1998) extend structural-change analysis to regression models containing multiple potential break points and develop procedures for testing and estimating their number and location.

These methods address specific statistical formulations of instability. They do not imply that every change in investment-model performance corresponds to a discrete structural break, nor that all economically relevant change can be represented by a small number of identifiable break dates.

For purposes of RP-003, the broader expression temporal instability is used to describe the possibility that relationships relevant to model behavior may change through time. This analytical category may include, but is not limited to, formal violations of statistical stationarity or discrete structural breaks.

Depending on the research problem, such change may appear gradually, abruptly, recurrently or in ways that cannot be cleanly classified using a single structural-break framework.

The robustness question is consequently not whether a model is invariant through time.

It is whether the historical evidence remains credible when the possibility of temporal change is explicitly challenged.

7.2 Stability Across Subperiods

Aggregate historical performance can conceal substantial variation through time.

A model that appears attractive over a long sample may derive much of its performance from particular intervals, while other periods contribute little or produce materially different behavior. Subperiod analysis can therefore provide information that is lost when evaluation focuses exclusively on full-sample averages.

The purpose is not to require equal performance in every historical interval. Financial-market observations are noisy, sample lengths differ, and a model may reasonably perform differently as

the environment it encounters changes. Arbitrarily dividing a sample into many short periods can also create unstable estimates that are difficult to interpret.

Subperiod evaluation is most informative when the divisions correspond to a stated analytical purpose. These may include chronological stability, known institutional or market changes, pre-specified economic periods, or simply an examination of whether aggregate performance is disproportionately dependent on a limited part of the available history.

Where subperiod boundaries are motivated by institutional, economic or market events, their justification should be documented independently of the performance pattern they are subsequently used to interpret. Otherwise, subperiod analysis itself can become another source of research flexibility.

Giacomini & Rossi (2010) demonstrate why time variation can matter for forecast evaluation. Rather than examining only global average performance, they develop methods for studying the historical path of relative out-of-sample forecasting performance between competing models in the presence of possible instability. Their empirical application concerns monetary models of exchange-rate determination relative to a random-walk benchmark and should not be generalized directly to trading strategies.

For purposes of RP-003, the methodological contribution supports a broader analytical point: aggregate performance measures can conceal economically or statistically relevant variation through time.

Evidence that a model behaves credibly across materially different subperiods can strengthen the robustness case. Concentration of performance within a narrow interval can instead identify a dependency that requires explanation or additional validation.

Neither result is conclusive by itself.

7.3 Market Environments and Regime Dependence

Temporal variation is often described using the language of market regimes.

The term can be useful, but it is also potentially imprecise. A regime may refer to volatility conditions, liquidity states, monetary-policy environments, directional markets, correlation structures, macroeconomic conditions or a statistically inferred latent state. These definitions are not interchangeable.

RP-003 therefore does not adopt a universal market-regime taxonomy.

Where regime analysis is used, the environment should be defined empirically, economically or operationally in a manner appropriate to the research question. The classification should also be distinguished from the subsequent observation of model performance within that classification.

If market environments are defined or relabeled after model degradation has been observed, the resulting analysis may still be useful exploratorily, but it should not be interpreted as independent evidence that the subsequently defined regime caused or predicted the degradation.

Regime dependence is also not inherently evidence of model failure. Some quantitative relationships are designed to operate only under particular conditions. A model whose economic or statistical premise is conditional may reasonably exhibit conditional performance.

The relevant robustness question is therefore not:

> Does the model perform equally in every regime?

It is:

> Is the model's dependence on changing market conditions understood, consistent with its stated design and supported by evidence that was not constructed solely to explain the observed outcome?

A model can be conditionally useful without being universally invariant.

The evidential requirement is to make that conditionality visible rather than to infer it retrospectively whenever performance changes.

7.4 Interpreting Performance Degradation

A decline from historical to later performance is important evidence, but its interpretation is not unique.

McLean & Pontiff (2016) examine 97 variables previously reported to predict cross-sectional stock returns. In their sample, portfolio returns were lower out of sample and lower again after academic publication. The authors interpret the out-of-sample decline as providing an upper-bound estimate of data-mining effects and the additional post-publication deterioration as consistent with investors learning about published return predictors and trading on that information. The study concerns published cross-sectional stock-return predictors and does not establish a universal magnitude or cause of degradation for quantitative trading models.

Suhonen, Lennkh & Perez (2017) provide a different empirical setting. Across 215 commercially promoted alternative-beta strategies developed by global investment banks, they report material deterioration in Sharpe ratios from backtested to live periods and find differences associated with strategy characteristics, including complexity. As discussed in Section 6, the finding applies to the study's specific strategy universe and does not establish that complexity, structural change or any other single mechanism caused the observed deterioration.

These studies illustrate an important feature of performance degradation: similar observations can be consistent with different mechanisms.

A decline may reflect, among other possibilities:

- sampling variation or statistical uncertainty;
- research-process effects or overfitting;
- a changing relationship between predictors and outcomes;
- structural or institutional change;

- market adaptation after information becomes known;
- different market environments;
- altered implementation conditions;
- or combinations of these mechanisms.

The list is analytical rather than exhaustive, and observing degradation does not by itself allow the researcher to select among these explanations.

Observed Phenomenon	Possible Mechanism	Additional Evidence That May Be Relevant	What the Observation Alone Cannot Establish
Later performance below historical backtest	Sampling variation or estimation uncertainty	Confidence intervals, longer independent observation, distributional analysis	That the model was overfit
Later performance below historical backtest	Research-process selection or overfitting	Research-universe reconstruction, multiplicity analysis, untouched or prospectively preserved evidence	That temporal market structure changed
Performance changes across periods	Parameter or structural instability	Formal stability or structural-break tests, subperiod analysis	The economic cause of the instability
Performance concentrated in particular market conditions	Conditional relationship or environment dependence	Predefined or independently justified environment classifications, mechanism-specific analysis	That the identified "regime" caused the result
Post-publication or post-discovery deterioration	Market adaptation or changing economic relationship	Timing evidence, adoption or trading-activity evidence, and mechanism-specific empirical analysis	That publication or adoption was the sole cause
Backtest-to-live deterioration	Implementation effects, research effects, structural change or combinations	Execution and cost analysis, research-process audit, contemporaneous market evidence	Which mechanism dominates

Table 1. Performance Degradation: Competing Explanations and Evidential Limits

Source: FINOVIS analytical synthesis based on the literature reviewed in Sections 4–8. "Possible mechanism" denotes an explanation consistent with an observation, not an identified causal effect.

7.5 Degradation Is Evidence, Not a Causal Diagnosis

Performance degradation should therefore be treated as an observation that changes the evidential position of a model, not as a self-interpreting diagnosis.

A historically attractive model that subsequently performs materially differently has encountered evidence that the original historical interpretation may have been incomplete. The appropriate response is to investigate which assumptions or relationships are being challenged.

That investigation should distinguish at least three questions.

First, what changed observationally? This may involve the magnitude, timing, distribution or conditional pattern of performance.

Second, which mechanisms are consistent with that change? Multiple mechanisms may remain plausible.

Third, what additional evidence could discriminate among them?

Structural-change methods can identify statistical instability without necessarily explaining its economic origin. Regime classification can describe conditional differences without proving that the classification is causal. Backtest-to-live deterioration can reveal a gap in expected behavior

without establishing whether the gap originated in overfitting, market change, implementation or another mechanism.

The same caution applies in the opposite direction. Stable historical or subsequent performance does not establish permanent invariance. An absence of detected instability may depend on sample size, test power, the particular stability measure used and the range of environments represented in the evidence.

For purposes of RP-003, temporal and market-environment analysis therefore contributes to robustness evaluation by challenging the assumption that historically observed relationships can be treated as automatically stable.

Its purpose is not to demand identical performance through all conditions.

The central proposition is narrower:

> Robustness is conditional on time and market environment, and observed degradation alone does not identify which mechanism is responsible; causal interpretation requires additional evidence.

A model may remain credible despite variation through time.

But variation should be investigated rather than explained away.

8. Implementation Realism and Economic Robustness

8.1 From Simulated Gross Performance to Realizable Outcomes

A quantitative model can be statistically credible while the trading strategy built around it remains economically difficult to implement.

This distinction arises because a historical simulation typically converts model outputs into hypothetical positions and returns under assumptions about how trades could have been executed. Those assumptions determine the connection between a model's predictive or statistical behavior and the economic outcome represented by the backtest.

For purposes of RP-003, simulated gross performance refers to performance before the trading frictions specifically identified as implementation costs in the evaluation are deducted. Because gross-versus-net conventions can differ across studies and simulations, the relevant cost boundary should be stated explicitly.

A gross historical return therefore answers a narrower question than an implementable return. It describes simulated performance before some or all of the frictions associated with converting model decisions into executed positions are incorporated.

For a trading strategy, relevant frictions may include transaction costs, turnover, spreads, price impact, liquidity constraints, execution assumptions and the amount of capital the strategy is

expected to deploy. Their importance varies materially across markets, instruments, trading horizons and strategy designs.

This distinction is consistent with Perold's (1988) implementation-shortfall framework, which contrasts a hypothetical paper portfolio with the outcome associated with actual implementation. The reference is used here as a conceptual foundation for distinguishing simulated and implemented outcomes rather than as empirical evidence concerning contemporary quantitative trading costs.

Novy-Marx & Velikov (2016), studying a large set of equity anomalies, show that transaction costs reduce strategy profitability and that turnover is strongly related to whether anomaly portfolios retain statistically significant net spreads after cost-mitigation techniques are applied. Their findings concern the anomaly portfolios and cost methodology examined in that study; they do not provide a universal transaction-cost threshold for quantitative strategies.

The relevant RP-003 principle is broader but deliberately limited:

> a historical strategy result should not be interpreted as economically realizable merely because its underlying model appears statistically credible.

Implementation realism forms a separate layer of validation.

8.2 Transaction Costs and Turnover

Transaction costs can materially change the economic interpretation of a trading strategy.

Their effect is closely connected to turnover. A strategy that generates frequent or large portfolio changes exposes more of its gross return to trading frictions than an otherwise comparable strategy requiring less trading. Turnover is therefore not merely an operational statistic; in a trading-model backtest it can affect how much of the simulated performance could plausibly survive implementation.

Within their anomaly universe, Novy-Marx & Velikov (2016) report that most strategies with turnover below 50% per month generate significant net spreads when designed to mitigate transaction costs, whereas few strategies with higher turnover do. They also show that cost-mitigation rules can materially affect after-cost strategy performance.

The 50% figure is a feature of the study's empirical classification and should not be interpreted as a universal turnover threshold.

Trading costs depend on the instruments traded, order sizes, liquidity, portfolio construction, execution methodology, market structure and historical period. A turnover level that is economically manageable for one strategy can be prohibitive for another.

The robustness question is therefore not:

> Is turnover below a fixed threshold?

It is:

Has the historical performance been evaluated using cost assumptions that are credible for the actual trading activity implied by the model?

Transaction-cost assumptions should also be internally consistent with the simulated strategy. A model whose apparent edge is small relative to plausible trading frictions faces a different evidential burden from a model whose simulated economic margin is substantially larger.

This remains an evidential question rather than a universal profitability rule.

8.3 Slippage and Execution Assumptions

Backtests must make assumptions about the prices at which hypothetical trades are treated as having occurred.

RP-003 uses the term slippage descriptively for differences between an execution price assumed in a simulation and an economically credible price at which the modeled trade could have been executed. It is not treated as a single standardized cost measure.

The materiality of execution-price assumptions depends on the strategy's trading frequency, order size, liquidity environment, signal horizon, portfolio construction and other characteristics of the intended implementation.

They may become especially consequential where a strategy trades frequently, reacts rapidly to signals, executes large orders or operates in instruments where reference prices do not adequately represent achievable transaction prices for the required size.

Korajczyk & Sadka (2004), studying momentum strategies in U.S. equities, explicitly distinguish proportional trading costs associated with spreads from non-proportional costs arising through price impact. Using intraday data and estimated trading-cost models, they find that after-cost momentum profitability depends on portfolio construction and that estimated price impact causes abnormal returns to decline as portfolio size increases.

RP-003 does not use these findings to prescribe a particular execution model for other strategies or markets.

The methodological implication is narrower: a simulation that assumes execution at reference prices without adequately reflecting material trading frictions may overstate the economic meaning of its gross historical performance.

This subsection deliberately does not address execution algorithms, order routing, latency architecture or trading-system engineering. Those are separate implementation questions.

For robustness evaluation, the relevant issue is whether the execution assumptions embedded in the historical simulation are sufficiently realistic for the economic conclusion being drawn from it.

8.4 Liquidity and Capacity

Implementation realism also changes with strategy scale.

A strategy that appears economically viable at a small hypothetical portfolio size may not produce the same net outcome when substantially more capital is required to trade the same opportunities. Position sizes may become large relative to available liquidity, execution costs may increase and the trades themselves may affect achievable prices.

Korajczyk & Sadka (2004) provide direct empirical evidence of this mechanism within the momentum strategies they study. Their price-impact estimates imply declining abnormal returns as portfolio size increases, and they calculate strategy-specific break-even fund sizes at which estimated abnormal performance is eliminated by trading costs.

Novy-Marx & Velikov (2016) similarly examine the capacity of anomaly strategies to absorb additional capital. In their sample, the extent to which additional capital reduces profitability is related to turnover, with substantial differences across strategy families.

The numerical capacities estimated in these studies are specific to their markets, periods, portfolio constructions and transaction-cost models. They should not be transferred mechanically to other quantitative strategies.

Capacity should therefore be interpreted as a strategy- and market-dependent economic constraint, not as a fixed characteristic inferred from gross backtest performance.

A robustness assessment should ask whether the scale assumed by the economic interpretation of a strategy is compatible with the liquidity and market-impact assumptions embedded in the simulation.

This creates another important distinction:

> a strategy can be economically viable at one scale and materially less viable at another without any change in the mathematical form of its predictive model.

8.5 Predictive Robustness versus Economic Robustness

The evidence reviewed above motivates a distinction between two related but non-identical robustness questions.

Predictive or statistical robustness concerns whether the model's statistical relationship, predictive behavior or other relevant empirical properties remain credible when challenged through independent data, alternative specifications, parameter variation, temporal instability and other appropriate forms of validation.

Economic or implementation robustness concerns whether a trading strategy constructed from that model remains economically credible after realistic assumptions about trading frictions, turnover, liquidity, capacity and implementation conditions are incorporated.

The two forms of robustness can fail separately.

A statistically credible predictive relationship may generate trades whose expected economic benefit is too small relative to implementation costs. Conversely, disappointing realized trading

outcomes do not necessarily demonstrate that the model's underlying predictive relationship has disappeared.

The distinction can be summarized as follows.

Dimension	Predictive / Statistical Robustness	Economic / Implementation Robustness
Primary Question	Do the model's statistical relationship, predictive behavior or other relevant empirical properties remain credible under appropriate validation challenges?	Can the model's outputs be translated into economically credible trading outcomes?
Principal Concerns	Sampling uncertainty, model selection, evaluation independence, specification dependence, temporal instability	Transaction costs, turnover, achievable prices, liquidity, market impact, capacity
Typical Failure	The historical relationship or relevant empirical behavior weakens or disappears under appropriate statistical challenge	The relationship may persist, but simulated gross performance does not survive credible implementation assumptions
Relevant Evidence	Independent evaluation, specification tests, parameter sensitivity, temporal analysis, statistical stress tests	Net-of-cost simulations, turnover analysis, execution assumptions, liquidity and capacity analysis
What Success Does Not Establish	That the strategy is economically implementable	That the underlying predictive relationship is statistically robust
What Failure Does Not Automatically Establish	That implementation frictions caused the observed statistical or predictive weakness	That the underlying statistical or predictive relationship is false

Table 2. Predictive Robustness versus Economic Robustness

Source: FINOVIS analytical synthesis based on Sections 3–8 and the literature cited therein.

The distinction is analytical rather than a claim that the two dimensions are independent. Implementation characteristics can influence model design, and model characteristics can determine transaction intensity and capacity.

The purpose is instead to prevent two logically different questions from being collapsed into one.

> A statistically credible signal is not automatically an investable strategy, and an implementation-constrained strategy is not automatically evidence of a statistically invalid signal.

8.6 Implementation Failure Does Not Necessarily Invalidate Signal

The distinction becomes particularly important when historical gross performance fails to survive realistic implementation assumptions.

As an illustrative hypothetical example, suppose a model identifies a relationship that remains statistically credible under independent evaluation, but the expected benefit from acting on its predictions is small relative to transaction costs or market impact.

The example is conceptual and does not represent an empirical FINOVIS strategy result.

The resulting trading strategy may be economically unattractive even though the underlying statistical relationship remains observable.

In that situation, the failed inference is not necessarily:

> the predictive relationship does not exist.

It may instead be:

> the simulated gross performance overstated the economically realizable outcome associated with exploiting that relationship under the implementation conditions examined.

The converse also matters. A model whose simulated returns remain strong after assumed transaction costs has not thereby established statistical robustness. Cost-adjusted profitability cannot repair weaknesses created by data snooping, contaminated evaluation, unstable parameters or insufficiently independent validation.

Implementation analysis therefore adds a distinct form of evidence.

Korajczyk & Sadka (2004) and Novy-Marx & Velikov (2016) demonstrate, in different equity-strategy settings, that trading costs, turnover, market impact and scale can materially alter the profitability implied by gross historical results. Their studies do not imply that every quantitative strategy faces the same constraints or that implementation effects explain every backtest-to-live deterioration.

For purposes of RP-003, the central proposition is therefore:

> Implementation failure does not necessarily invalidate the underlying statistical or predictive relationship; it may instead invalidate the inference that simulated gross performance represents an economically realizable trading outcome.

Statistical or predictive robustness and economic or implementation robustness are related but distinct.

A rigorous model evaluation should examine both.

9. Stress Testing and Alternative Robustness Evidence

9.1 Why Additional Challenges Are Needed

The validation methods examined in earlier sections address different potential weaknesses in historical model evidence.

Research-process analysis can reveal whether selection affected the interpretation of performance. Out-of-sample evaluation can increase separation between development and evaluation. Parameter and specification analysis can expose narrow dependence on modeling choices. Temporal analysis can investigate whether relationships remain credible through time, while implementation analysis examines whether simulated performance survives economically realistic trading assumptions.

None of these challenges exhausts the robustness question.

A quantitative model may remain vulnerable to assumptions or conditions that are only weakly represented in the observed historical sample. Historical data may contain relatively few observations of particular market states, extreme movements, unusual dependency structures or combinations of conditions relevant to the model. Other uncertainties may concern the statistical properties of the sample itself rather than a specific historical episode.

Additional robustness procedures can therefore be useful when they are designed to challenge a clearly identified assumption or failure mechanism.

For purposes of RP-003, these procedures are grouped broadly into stress testing, scenario analysis, resampling and Monte Carlo methods. The categories overlap in practice and are not presented as a universal taxonomy. Their evidential value depends on what is changed, what is preserved, what assumptions generate the alternative observations and what conclusion the procedure is intended to support.

A more sophisticated challenge does not automatically produce stronger evidence.

The relevant question is:

> What additional uncertainty or failure mechanism does the procedure expose that was not adequately examined by the original historical evaluation?

9.2 Stress Testing

Stress testing evaluates model or strategy behavior under deliberately adverse or altered conditions.

The defining feature is not a particular numerical technique but the purposeful challenge to assumptions that may be important to the historical result. Depending on the model, a stress test might alter market movements, volatility, correlations, transaction costs, liquidity assumptions, parameter values, execution conditions or other inputs relevant to the inference being examined.

Stress testing is widely used in financial risk management. The Basel Committee's stress-testing principles describe adverse scenarios as sets of economic and financial conditions designed to stress financial performance under severe conditions and emphasize that stress-testing frameworks should be linked to clear objectives, methodology, processes and documentation (Basel Committee on Banking Supervision, 2018). The guidance concerns banking risk management and supervision rather than empirical validation of quantitative trading strategies; RP-003 uses it only as an institutional reference for the general logic of structured stress testing.

For model robustness, the value of a stress test depends on the connection between the imposed challenge and a plausible failure mechanism.

Increasing transaction costs may challenge economic margin. Altering parameter values may challenge specification dependence. Changing correlations may examine sensitivity to dependency assumptions. Applying unusually adverse market movements may reveal nonlinear exposures that were not obvious in average historical performance.

The test becomes less informative when stressed inputs are selected arbitrarily or when the result is interpreted without reference to the assumptions generating the stress.

Observed model behavior under a particular stress can provide evidence about sensitivity to that specified challenge under the stated design.

It does not establish robustness to conditions that were not tested.

Nor does a severe stress scenario automatically represent a probable future state.

Stress testing should therefore be interpreted as targeted evidence about sensitivity to specified adverse conditions, not as a prediction of what the future will contain.

9.3 Scenario Analysis

Scenario analysis examines model behavior under a defined configuration of conditions considered jointly.

This distinguishes it conceptually from a simple one-variable sensitivity test. A scenario may combine changes in several market, economic or implementation variables where their joint configuration is relevant to the question being investigated.

For example, an analysis might consider how a strategy behaves when elevated volatility coincides with reduced liquidity and higher assumed transaction costs. Another scenario might examine a shift in correlations together with materially different directional market behavior. These examples are illustrative only and do not prescribe specific scenarios for quantitative-model validation.

The evidential quality of scenario analysis depends heavily on how the scenario is constructed.

Consistent with the general scenario-design principles described by the Basel Committee (2018), scenario variables should be internally coherent and linked to the risks or vulnerabilities the exercise is intended to examine. The banking guidance is used here only as an institutional reference for scenario construction, not as empirical evidence of trading-model robustness.

A historically observed scenario can provide information about model behavior during a documented environment, but its relevance remains conditioned on the historical circumstances that produced it. A hypothetical scenario can examine conditions not observed in the sample, but its result is necessarily conditional on the assumptions used to construct those conditions.

Scenario analysis therefore should not be interpreted as independent evidence merely because the scenario is numerically different from the historical record.

For purposes of RP-003, an informative scenario should have a stated rationale, an explicit connection to a model vulnerability or assumption and sufficient internal coherence for the resulting model behavior to be interpretable.

Scenario analysis can answer:

> How would the evaluated model respond under this specified combination of conditions?

It cannot establish:

> How likely is this scenario to occur?

unless probability estimation is separately justified.

Nor can successful performance under several chosen scenarios establish invariance to other conditions that were not represented.

9.4 Resampling and Bootstrap

Resampling methods approach robustness from a different direction. Rather than imposing a specific economic scenario, they use the observed data to examine the variability of statistics or model outcomes under repeated reconstitution of the available sample.

Efron (1979) introduced the bootstrap as a general method for estimating the sampling distribution of a statistic from observed data. In its basic form, the method is developed for random samples drawn from an underlying distribution.

Financial time series create additional difficulties because observations may be serially dependent. Resampling individual observations as though they were independent can destroy dependency structures relevant to the statistic being evaluated.

Politis & Romano (1994) develop the `stationary bootstrap` for weakly dependent stationary observations, while Lahiri (2003) surveys bootstrap and related resampling methods designed for temporal and other dependent data. These contributions demonstrate why the resampling design must correspond to the dependence structure of the underlying problem rather than being transferred mechanically from an independent-data setting.

For quantitative-model evaluation, resampling may be useful for investigating the sampling variability of performance statistics, the sensitivity of conclusions to particular historical observations or the range of outcomes consistent with a specified empirical resampling scheme.

Its limitations are equally important.

Empirical bootstrap procedures generate resamples from the observed data under a specified resampling scheme, while model-based bootstrap procedures generate observations using an estimated probabilistic model. In either case, the resulting evidence remains conditional on the original data and on the assumptions embedded in the resampling design or generating model.

It therefore does not become genuinely independent evidence merely because a computer generates a new sequence of returns.

If a historical sample does not contain important structures relevant to future model behavior, purely empirical resampling cannot necessarily manufacture those structures. Conversely, a resampling scheme that modifies dependence assumptions can generate alternative sequences, but the evidential interpretation then depends on the validity of the imposed resampling design.

For purposes of RP-003:

> resampling can provide evidence about uncertainty conditional on the observed sample and the resampling assumptions; it does not convert historical information into genuinely unseen prospective evidence.

The choice of bootstrap procedure, block structure, dependence assumptions and statistic being evaluated is therefore part of the evidence rather than a purely technical implementation detail.

9.5 Monte Carlo

Monte Carlo simulation uses probabilistic models and repeated random generation to study distributions of possible outcomes under specified assumptions.

The method is fundamental in financial engineering and risk measurement. Glasserman (2003) develops Monte Carlo methods for a wide range of financial applications, including path generation, derivatives valuation and risk measurement.

For RP-003, Monte Carlo is relevant not because those applications establish trading-model robustness, but because simulation provides a general mechanism for examining model behavior across many realizations generated under an explicitly specified stochastic structure.

This can be valuable where the robustness question concerns uncertainty that is difficult to examine from a single historical path.

A simulation might, for example, investigate the distribution of a model's performance under alternative return paths generated from a specified process, assess sensitivity to assumed parameters or examine the frequency and magnitude of adverse outcomes within that simulation environment.

The strength of the exercise depends directly on the quality and relevance of the generating assumptions.

For purposes of RP-003, this motivates a distinction between simulation uncertainty and model uncertainty. Increasing the number of Monte Carlo replications can improve numerical precision conditional on the assumed generating model; it does not, by computation alone, establish that the generating model itself is appropriate.

This distinction is fundamental.

> Monte Carlo can generate many observations from an assumed world. It does not establish that the assumed world is an adequate representation of the actual one.

Simulation results should therefore distinguish at least conceptually between uncertainty generated *within* the assumed model and uncertainty about whether the model used to generate the simulations is itself appropriate.

Monte Carlo evidence is consequently conditional evidence.

Its value increases when the generating process, calibration, dependence assumptions and relationship to the targeted failure mechanism are made explicit.

9.6 What Stress Tests Can—and Cannot—Establish

Stress testing, scenario analysis, resampling and Monte Carlo methods can substantially broaden model evaluation beyond a single realized historical path.

They can reveal sensitivities that an aggregate backtest obscures. They can examine outcomes under deliberately altered assumptions. They can help characterize sampling uncertainty and expose dependence on particular observations or statistical structures. They can also explore model behavior under hypothetical conditions that were poorly represented or absent from the historical sample.

These contributions are valuable precisely because the methods answer different questions.

They should not be collapsed into a single category of "additional testing" whose evidential weight is assumed to increase mechanically with the number of procedures performed.

Several highly similar simulations may interrogate essentially the same assumption. Conversely, one well-designed stress test may directly challenge a material model vulnerability that numerous weakly related tests leave untouched.

The interpretation should therefore return to the same questions that guide the broader RP-003 framework:

- What failure mechanism is being challenged?
- What new evidence does the procedure add?
- Which assumptions does that evidence depend on?
- What uncertainty remains after the test?

Successful stress testing cannot establish that a model will remain stable under every future condition.

Scenario analysis cannot establish the probability of a scenario merely by constructing it.

Resampling cannot create informational independence from the historical evidence on which it depends.

Monte Carlo cannot eliminate model uncertainty by increasing the number of simulated paths.

An adverse result under one of these procedures also requires interpretation. The observed sensitivity may arise because the challenged assumption is materially relevant, because the imposed scenario is poorly matched to the intended question, because the resampling design fails to preserve relevant data structure or because the simulation model is misspecified.

The procedures therefore add evidence without converting validation into proof.

For purposes of RP-003, their role is complementary:

- alternative robustness methods strengthen model evaluation when they expose the model to relevant challenges that are not adequately represented

> by the original backtest, while keeping the assumptions and evidential limits of those challenges explicit.

This leads directly to the synthesis developed in Section 10.

No single stress, scenario, resampling procedure or simulation establishes robustness.

> Robustness remains a cumulative evidential judgment.

10. From Validation Tests to a Robustness Framework

10.1 Why No Single Test Establishes Robustness

The preceding sections have examined several distinct challenges to the interpretation of historical quantitative-model performance.

Research-process analysis asks whether selection and repeated data use affected the observed result. Out-of-sample evaluation examines the degree of informational independence between development and validation. Specification and parameter analysis challenges dependence on particular modeling choices. Temporal analysis examines whether historically observed relationships remain credible through changing conditions. Implementation analysis asks whether simulated performance represents an economically realizable trading outcome. Stress, scenario and resampling methods expose the model to additional assumptions and alternative representations of uncertainty.

Each challenge can add information.

None establishes robustness by itself.

This follows from the fact that different validation procedures address different potential failure mechanisms. A model may perform well on a historical hold-out while remaining highly sensitive to parameter choice. It may exhibit parameter stability while relying on unrealistic implementation assumptions. It may retain statistically credible predictive behavior while failing economically after trading costs. It may survive several stress scenarios while the original research process remains heavily affected by model selection.

A favorable result under one validation challenge therefore does not automatically compensate for weakness identified in another.

The same principle applies to the quantity of validation performed. Multiple procedures can produce highly overlapping evidence if they depend on the same historical observations, assumptions or model characteristics.

For purposes of RP-003:

> Multiple weak or highly overlapping tests are not necessarily stronger evidence than one highly informative test that directly addresses a material failure mechanism.

Validation should therefore be interpreted through the informational contribution of each challenge rather than through the number of tests completed.

This perspective is consistent with the broader principle reflected in the 2026 interagency *Supervisory Guidance on Model Risk Management* that validation approaches should depend on model characteristics and use, that validation should identify limitations and that material model risk may remain even after rigorous validation (Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation & Office of the Comptroller of the Currency, 2026).

The guidance concerns model risk management in banking organizations and applies its own regulatory definition of a model, which is not identical to the broader analytical object examined in RP-003. It is therefore used here only as an institutional reference for general validation principles, not as a validation standard applicable to all quantitative investment or trading models considered in this paper.

RP-003 goes further in a different direction: it develops an analytical framework specifically for interpreting the robustness evidence associated with quantitative investment and trading models evaluated materially through historical financial-market data.

10.2 Validation Layers and Failure Mechanisms

A useful validation framework should begin with the potential failure mechanism being challenged rather than with a preferred catalogue of statistical procedures.

The same technique can contribute different evidence depending on how it is designed. Conversely, different techniques may address substantially the same vulnerability.

For RP-003, the evidence reviewed in Sections 3–9 is organized into six validation domains.

Research-Process Validity

This domain examines whether the historical result should be interpreted differently because of data snooping, multiple testing, model selection, parameter optimization, researcher flexibility, repeated data use or weaknesses in sample integrity.

The failure mechanism being challenged is selection-sensitive historical performance.

Independent Validation

This domain examines the extent to which evaluation evidence was informationally separated from the decisions that produced the model.

Relevant distinctions include historical hold-out evidence, pseudo-out-of-sample or rolling historical evaluation, prospective evaluation and live implementation evidence.

The failure mechanism being challenged is performance that appears convincing because evaluation information has already influenced development or selection.

Specification & Parameter Stability

This domain examines whether the interpretation of historical performance depends excessively on narrow parameter values, particular specifications or highly localized modeling choices.

The failure mechanism being challenged is specification-sensitive evidence.

Temporal & Market-Environment Robustness

This domain examines whether model behavior remains credible when temporal instability, structural change, subperiod dependence and changing market environments are considered.

The failure mechanism being challenged is historical evidence whose relevance depends materially on a limited temporal or environmental configuration.

Economic & Implementation Robustness

This domain examines whether simulated model outputs remain economically credible after relevant transaction costs, turnover, achievable execution prices, liquidity and capacity assumptions are considered.

The failure mechanism being challenged is a gap between simulated gross performance and economically realizable trading outcomes.

Stress & Alternative Evidence

This domain examines model behavior under deliberately altered assumptions, scenarios, resampling designs or simulation environments.

The failure mechanism being challenged depends on the procedure: it may involve sensitivity to adverse conditions, sampling uncertainty, dependence assumptions, particular historical observations or other vulnerabilities not adequately represented in the original evaluation.

The six domains are analytical categories rather than independent statistical factors. Their boundaries can overlap because the underlying vulnerabilities themselves can interact.

A parameter choice may have been selected through extensive historical search. Temporal degradation may reveal specification dependence. Implementation assumptions may interact with particular market environments. A stress test may challenge both parameter stability and economic viability.

The purpose of the domains is therefore organizational rather than ontological.

They structure the evidence without implying that model weaknesses occur in isolated compartments.

10.3 Complementary Evidence Rather Than a Mechanical Checklist

A robustness framework can become misleading if it is converted into a checklist in which each completed procedure contributes another implicit point toward a final robustness score.

RP-003 deliberately rejects that interpretation.

Different validation results can vary substantially in informational value. A test that directly challenges a central assumption may matter more than several tests directed at peripheral concerns. Evidence generated from materially different information sources, evaluation periods or assumptions may add something different from repeated analysis of the same historical observations. Conversely, apparently diverse tests may contribute little additional information when they share the same assumptions or failure mechanism.

The interpretation also depends on the model itself.

Not every quantitative model requires identical validation procedures, identical stress conditions or identical evidence. The relevant challenges depend on model purpose, structure, research process, data, trading horizon, intended implementation and the conclusions being drawn from historical results.

Accordingly, the framework does not ask whether every possible test has been completed.

It asks four common questions of each material validation domain:

- What failure mechanism is being challenged?
- What evidence does the challenge add?
- Which assumptions does that evidence depend on?
- What residual uncertainty remains?

These questions shift attention from procedural completion to evidential interpretation.

They also prevent a favorable validation result from being read more broadly than the design permits. A parameter-sensitivity exercise showing limited dependence under the specified perturbations adds evidence about parameter dependence. It does not establish sample integrity. A preserved historical hold-out can add evidence about evaluation independence. It does not establish implementation realism. A realistic cost analysis can strengthen the economic interpretation of a strategy. It does not correct prior data snooping.

Validation evidence is therefore complementary when different challenges address materially different vulnerabilities.

It is not automatically cumulative simply because more tests have been performed.

10.4 Evidential Burden of Model Flexibility

The amount and type of robustness evidence required to interpret a historical result cannot be separated from the flexibility of the model and the research process that produced it.

Greater flexibility can arise through many channels: larger candidate-model universes, additional parameters, alternative transformations, hyperparameter choices, adaptive research decisions, flexible feature selection or other degrees of freedom.

Flexibility is not itself evidence that a model is invalid.

As established in Section 6, complex or flexible models may represent relationships that simpler specifications cannot capture. The relevant concern is instead that a more flexible research environment can create more ways for historical observations to influence the final model.

For purposes of RP-003, this creates an evidential burden, not an automatic rejection rule.

Where substantial model-selection freedom exists, stronger separation between development and evaluation may become particularly informative. Where parameterization is extensive, specification and perturbation analysis may become more important. Where implementation assumptions materially determine the reported outcome, economic validation becomes correspondingly important.

This does not imply a mechanical formula in which model complexity determines a predetermined number of required tests.

Nor does it imply that a simple model deserves a presumption of robustness.

The relationship is qualitative:

> the greater the material flexibility that can shape a historical result, the more important it becomes that validation evidence directly challenges the ways in which that flexibility could have influenced the result.

The burden should therefore follow the potential failure mechanism rather than an abstract complexity label.

10.5 RP-003 Multidimensional Robustness Evidence Framework

The preceding analysis can now be integrated into the RP-003 Multidimensional Robustness Evidence Framework.

The framework constitutes the principal analytical contribution of RP-003. It is derived from the literature and analytical distinctions developed in Sections 3–9 and should be understood as FINOVIS Analytical Synthesis rather than as an empirically validated diagnostic instrument.

The framework begins with one observation:

> OBSERVED HISTORICAL PERFORMANCE

This is the evidence produced by the historical evaluation. It is informative, but its interpretation remains conditional.

The framework then asks how that evidence changes when challenged across six dimensions:

1. Research-Process Validity
2. Independent Validation
3. Specification & Parameter Stability
4. Temporal & Market-Environment Robustness
5. Economic & Implementation Robustness
6. Stress & Alternative Evidence

Each domain is examined through the same four questions:

- Failure Mechanism Challenged
- Evidence Added
- Assumptions
- Residual Uncertainty

The output is not a numerical score.

It is a:

CUMULATIVE EVIDENTIAL JUDGMENT

The judgment concerns how much confidence the combined evidence reasonably supports in the historically observed behavior of the model.

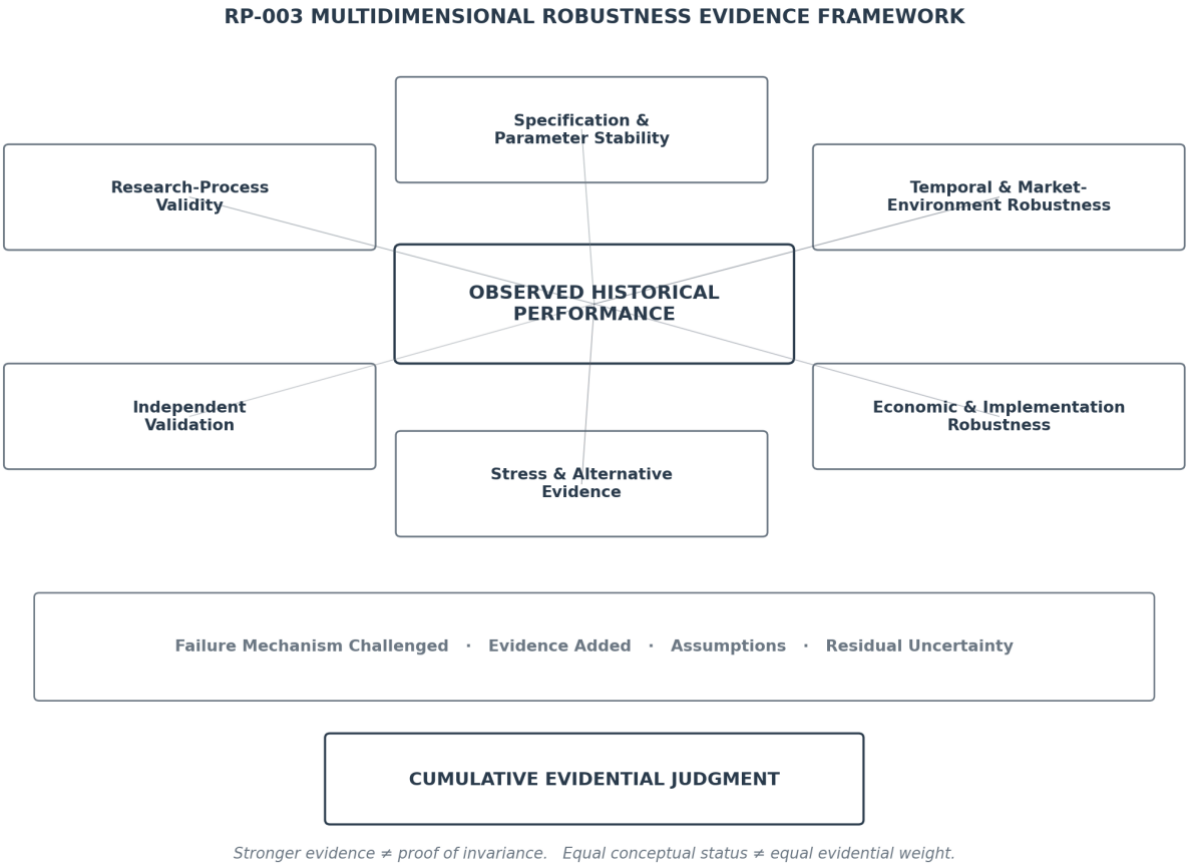


Figure 4. RP-003 Multidimensional Robustness Evidence Framework

Center / analytical starting point:

OBSERVED HISTORICAL PERFORMANCE

Six equal validation domains surrounding the starting observation:

Research-Process Validity

Independent Validation

Specification & Parameter Stability

Temporal & Market-Environment Robustness

Economic & Implementation Robustness

Stress & Alternative Evidence

Each domain is evaluated through four common questions:

Failure Mechanism Challenged

Evidence Added

Assumptions

Residual Uncertainty

Analytical output:

CUMULATIVE EVIDENTIAL JUDGMENT

Stronger evidence ≠ proof of invariance.

Equal conceptual status ≠ equal evidential weight.

Source: FINOVIS analytical synthesis based on Sections 3–9 and the literature cited therein. The framework is not a scoring model, standardized validation protocol, regulatory standard or pass/fail certification procedure.

The framework should not be interpreted spatially as a sequence through which every model must progress.

The six domains are intended to have equal conceptual status. Equal conceptual status does not imply equal evidential weight. The relevance and informational contribution of each domain depend on the model, research process, intended use and failure mechanisms being examined.

Nor should evidence across the domains be mechanically aggregated.

A material weakness in one dimension cannot necessarily be neutralized by favorable evidence in unrelated dimensions. Equally, uncertainty in one domain does not automatically invalidate evidence obtained elsewhere.

The framework instead asks whether the overall body of evidence becomes more or less persuasive as distinct explanations for apparently attractive historical performance are subjected to appropriately designed challenge.

This is what RP-003 means by a cumulative evidential judgment.

Cumulative in this context means that relevant evidence is interpreted jointly and in relation to the failure mechanisms it addresses; it does not mean that validation results are arithmetically additive or that evidence from different domains can be reduced to a common score.

10.6 What Robustness Evidence Does Not Establish

A rigorous validation process can materially change the quality and interpretation of the evidence supporting a quantitative model.

It can reveal that an attractive result was highly dependent on research choices. It can provide evidence that model behavior persists under more independent evaluation. It can expose parameter sensitivity, temporal instability or unrealistic implementation assumptions. It can show, under the stated validation design, whether particular stresses or alternative assumptions materially alter the evaluated behavior of the model.

What it cannot do is remove uncertainty entirely.

No finite historical record contains every future market condition. No out-of-sample design guarantees that the future will resemble its evaluation period. No parameter-sensitivity exercise explores every plausible specification. No structural-change test establishes permanent stability. No implementation model reproduces every future trading condition. No stress scenario or simulation contains every possible state of the world.

Stronger robustness evidence therefore does not establish invariance.

It does not guarantee future performance.

It does not prove that a model will remain useful under all future conditions.

Nor does a weakness identified by one validation procedure automatically establish that the model has no statistical or economic value.

The framework is instead intended to make the structure of the evidence and its remaining uncertainty more explicit.

This distinction can be summarized in a final interpretive principle:

> Backtesting creates evidence. Validation changes the quality and interpretation of that evidence. Neither historical performance nor subsequent validation eliminates uncertainty.

The objective of robustness assessment is therefore not to transform uncertainty into certainty.

It is to determine whether confidence in historically observed model behavior remains justified after the evidence has been subjected to relevant, appropriately designed and transparently interpreted challenges.

11. Conclusion

11.1 Answer to the Central Research Question

RP-003 began with a central question:

> Why is historical backtest performance insufficient evidence of the robustness of a quantitative trading model, and what should a more rigorous model-evaluation framework examine?

The analysis leads to a clear answer.

Historical backtest performance is informative because it documents how a specified model behaved within a defined historical simulation. It can reveal economically and statistically relevant characteristics of the model and provide an essential starting point for further evaluation.

It is insufficient as a robustness verdict because the observed result remains conditional on the data, model specification, parameter choices, research process, evaluation design and implementation assumptions through which it was produced.

An attractive historical result may therefore reflect a combination of persistent economic or statistical structure, sampling variation, model selection, repeated testing, specification dependence, temporal conditions and implementation assumptions. Historical performance alone cannot determine the contribution of those mechanisms.

A more rigorous evaluation should consequently examine not only the magnitude of historical performance but the quality and independence of the evidence supporting its interpretation.

For RP-003, this requires challenges across six dimensions:

- Research-Process Validity
- Independent Validation
- Specification & Parameter Stability
- Temporal & Market-Environment Robustness
- Economic & Implementation Robustness
- Stress & Alternative Evidence

These dimensions do not form a mechanical checklist or hierarchy. They organize different questions about whether historically observed model behavior remains credible beyond the conditions under which it was observed and whether statistically credible behavior can translate into economically realizable outcomes.

Robustness is therefore best understood as a cumulative evidential judgment rather than the result of a single backtest, statistic or validation procedure.

11.2 Principal Analytical Conclusions

Several conclusions follow from the analysis.

First, historical performance is conditional evidence. A backtest should be interpreted together with the data, assumptions and research design that produced it rather than as an isolated measure of model quality.

Second, the research process is part of the evidence. Data snooping, multiple testing, parameter optimization, researcher flexibility and repeated use of historical information can materially alter the interpretation of a selected result.

Third, out-of-sample status does not by itself establish informational independence. The value of evaluation evidence depends on what information was excluded from which research decisions, whether evaluation outcomes influenced subsequent model development and how effectively the boundary between development and validation was preserved.

Fourth, parameter and specification stability provide diagnostic evidence rather than universal robustness thresholds. Dependence on narrow configurations can weaken confidence in the interpretation of a historical result, but neither broad stability nor model simplicity proves robustness.

Fifth, robustness is conditional on time and market environment. Temporal instability, structural change and changing market conditions can alter model behavior, but observed degradation alone does not identify the mechanism responsible. Causal interpretation requires additional evidence.

Sixth, statistical credibility and economic realizability are distinct. A predictive relationship may remain statistically credible while the trading strategy built around it becomes unattractive after transaction costs, liquidity constraints, price impact or capacity effects are considered.

Seventh, stress tests, scenario analysis, resampling and Monte Carlo methods contribute conditional and complementary evidence. Their value depends on the failure mechanism being challenged and the assumptions embedded in the procedure. Synthetic or resampled observations do not become genuinely unseen prospective evidence merely because additional data paths have been generated.

Taken together, these conclusions support a broader principle:

> Robustness is not established by accumulating favorable test results. The evidential case for robustness is strengthened when materially different and appropriately designed challenges reduce credible alternative explanations for the observed historical performance while leaving residual uncertainty explicit.

11.3 Implications for Quantitative Model Evaluation

The practical implication is that quantitative-model evaluation should move beyond the question of whether a backtest appears attractive.

A more relevant set of questions includes:

> How was the result produced?
>
> Which information influenced model development and selection?
>
> How uncertain is the reported performance?
>
> Does the interpretation depend excessively on particular specifications or parameters?
>
> Does the evidence remain credible through time and across relevant market environments?
>
> Can the simulated result survive economically realistic implementation assumptions?
>
> What additional vulnerabilities are exposed by stress, scenario, resampling or simulation methods?
>
> What uncertainty remains after those challenges?

The answers will differ across models.

A simple strategy and a highly flexible machine-learning model need not require identical validation procedures. A low-turnover strategy and a capacity-constrained trading model may face different implementation questions. A long historical sample and a short prospective record can contribute different forms of evidence.

The purpose of the RP-003 Multidimensional Robustness Evidence Framework is therefore not to prescribe a universal sequence or threshold.

It is to ensure that model evaluation remains aligned with the failure mechanisms that could plausibly explain an apparently attractive historical result.

This also changes the interpretation of favorable validation outcomes.

Evidence that a model remains credible under independent evaluation, reasonable specification changes, temporal challenges, realistic implementation assumptions and relevant stress conditions can materially strengthen confidence in its historical behavior.

It does not convert that confidence into certainty.

11.4 Limits of Robustness Assessment

Robustness assessment has an unavoidable boundary.

Future financial-market conditions are not fully observable in advance. Historical data provide only a finite representation of possible environments. Prospective evidence grows only as new observations occur. Statistical procedures depend on assumptions. Implementation conditions can change. Models, markets and research processes can interact in ways that are only partially represented by prior evidence.

Validation can therefore interrogate sources of uncertainty, organize the resulting evidence and make residual uncertainty more explicit.

It cannot eliminate it.

A model that performs credibly across multiple validation dimensions may warrant greater confidence than one supported primarily by an attractive backtest. That conclusion remains conditional on the quality, independence and relevance of the evidence examined.

Similarly, weakness under one challenge should not automatically be converted into a universal rejection of the model. The significance of an adverse result depends on the failure mechanism tested, the design of the evaluation and the inference being drawn.

The appropriate endpoint is consequently not a declaration that a model is permanently robust.

It is a reasoned judgment about whether the available evidence supports greater confidence in historically observed model behavior and where material uncertainty remains.

The central conclusion of RP-003 can therefore be stated simply:

> Historical performance begins the evaluation. It does not finish it.

And the broader interpretive principle remains:

> Backtesting creates evidence. Validation changes the quality and interpretation of that evidence. Neither historical performance nor subsequent validation eliminates uncertainty.

References

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica, 61*(4), 821–856. doi: 10.2307/2951764.

Arnott, R., Harvey, C. R., & Markowitz, H. (2019). A backtesting protocol in the era of machine learning. *The Journal of Financial Data Science, 1*(1), 64–74. doi: 10.3905/jfds.2019.1.064.

Bai, J., & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica, 66*(1), 47–78. doi: 10.2307/2998540.

Bailey, D. H., & López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management, 40*(5), 94–107. doi: 10.3905/jpm.2014.40.5.094.

Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2017). The probability of backtest overfitting. *The Journal of Computational Finance, 20*(4), 39–69. doi: 10.21314/JCF.2016.322.

Basel Committee on Banking Supervision. (2018). *Stress testing principles*. Bank for International Settlements.

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis, 120*, 70–83. doi: 10.1016/j.csda.2017.11.003.

Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, & Office of the Comptroller of the Currency. (2026). *Supervisory Guidance on Model Risk Management*. April 17, 2026. Attachment to Federal Reserve Supervisory Letter SR 26-2, *Revised Guidance on Model Risk Management*.

Brown, S. J., Goetzmann, W., Ibbotson, R. G., & Ross, S. A. (1992). Survivorship bias in performance studies. *The Review of Financial Studies, 5*(4), 553–580. doi: 10.1093/rfs/5.4.553.

CFA Institute. (2026). *Backtesting & Simulation*. 2026 CFA Program Level II Curriculum, Portfolio Management. CFA Institute.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 253–263. doi: 10.1080/07350015.1995.10524599.

Duarte, J., Jones, C. S., Khorram, M., Mo, H., & Wang, J. L. (2026). Too good to be true: Look-ahead bias in empirical options research. *The Review of Financial Studies*, hhag061. Advance online publication. doi: 10.1093/rfs/hhag061.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science, 349*(6248), 636–638. doi: 10.1126/science.aaa9375.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26. doi: 10.1214/aos/1176344552.

Giacomini, R., & Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics, 25*(4), 595–620. doi: 10.1002/jae.1177.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica, 74*(6), 1545–1578. doi: 10.1111/j.1468-0262.2006.00718.x.

Giglio, S., Liao, Y., & Xiu, D. (2021). Thousands of alpha tests. *The Review of Financial Studies, 34*(7), 3456–3496. doi: 10.1093/rfs/hhaa111.

Glasserman, P. (2003). *Monte Carlo methods in financial engineering*. Springer. doi: 10.1007/978-0-387-21617-1.

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal, 26*(2), 91–108. doi: 10.1111/j.1471-1842.2009.00848.x.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies, 33*(5), 2223–2273. doi: 10.1093/rfs/hhaa009.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics, 23*(4), 365–380. doi: 10.1198/073500105000000063.

Harvey, C. R., & Liu, Y. (2020). False (and missed) discoveries in financial economics. *The Journal of Finance, 75*(5), 2503–2553. doi: 10.1111/jofi.12951.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *The Review of Financial Studies, 29*(1), 5–68. doi: 10.1093/rfs/hhv059.

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data, 6*(4), 15:1–15:21. doi: 10.1145/2382577.2382579.

Korajczyk, R. A., & Sadka, R. (2004). Are momentum profits robust to trading costs? *The Journal of Finance, 59*(3), 1039–1082. doi: 10.1111/j.1540-6261.2004.00656.x.

Lahiri, S. N. (2003). *Resampling methods for dependent data*. Springer. doi: 10.1007/978-1-4757-3803-2.

Lo, A. W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal, 58*(4), 36–52. doi: 10.2469/faj.v58.n4.2453.

McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance, 71*(1), 5–32. doi: 10.1111/jofi.12365.

Novy-Marx, R., & Velikov, M. (2016). A taxonomy of anomalies and their trading costs. *The Review of Financial Studies, 29*(1), 104–147. doi: 10.1093/rfs/hhv063.

Perold, A. F. (1988). The implementation shortfall: Paper versus reality. *The Journal of Portfolio Management, 14*(3), 4–9. doi: 10.3905/jpm.1988.409150.

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association, 89*(428), 1303–1313. doi: 10.1080/01621459.1994.10476870.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333–339. doi: 10.1016/j.jbusres.2019.07.039.

Suhonen, A., Lennkh, M., & Perez, F. (2017). Quantifying backtest overfitting in alternative beta strategies. *The Journal of Portfolio Management, 43*(2), 90–104. doi: 10.3905/jpm.2017.43.2.090.

Sullivan, R., Timmermann, A., & White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap. *The Journal of Finance, 54*(5), 1647–1691. doi: 10.1111/0022-1082.00163.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting, 16*(4), 437–450. doi: 10.1016/S0169-2070(00)00065-0.

White, H. (2000). A reality check for data snooping. *Econometrica, 68*(5), 1097–1126. doi: 10.1111/1468-0262.00152.

Wu, J. M.-T., Lin, W.-Y., Huang, K.-W., & Wu, M.-E. (2024). On the design of searching algorithm for parameter plateau in quantitative trading strategies using particle swarm optimization. *Knowledge-Based Systems, 293*, Article 111630. doi: 10.1016/j.knosys.2024.111630.

Research Disclaimer

This publication has been prepared by Finovis Software Solutions LLC-FZ ("FINOVIS") for general informational and research purposes only. It is intended for professional and sophisticated readers with an interest in quantitative investment methods, model evaluation and financial-market research. The description of the intended readership is editorial in nature and does not classify any reader as a professional investor, professional client, counterparty or equivalent category under any applicable regulatory framework.

This publication is thematic and methodological in nature. It does not constitute investment advice, an investment recommendation, a personal recommendation, financial advice or any other form of advice tailored to the circumstances of a particular person. It does not provide a recommendation or suggestion concerning the purchase, sale, holding, valuation or expected performance of any specific financial instrument, issuer, investment product or investment strategy. Nothing contained in this publication constitutes or should be construed as an offer, solicitation or invitation to acquire, dispose of or enter into any financial instrument, investment product, trading strategy or transaction.

The analysis is based on academic literature, institutional publications and other external sources considered relevant to the subject matter. External sources are identified where applicable. FINOVIS has sought to select and interpret external sources considered relevant and appropriate for the purposes of this publication. However, FINOVIS has not independently verified every underlying third-party dataset, calculation, empirical result or methodological assumption and makes no representation or warranty as to the completeness or accuracy of the underlying third-party materials. Findings from individual studies should be interpreted within the markets, periods, datasets, methodologies and other limitations applicable to those studies.

The concepts, classifications and analytical frameworks presented in this publication, including the RP-003 Multidimensional Robustness Evidence Framework, represent analytical synthesis developed for the purposes of this paper. They are not presented as an empirically validated diagnostic instrument, standardized model-validation protocol, regulatory standard, scoring system or certification procedure. They do not disclose, describe or represent any proprietary FINOVIS trading model, strategy, signal, algorithm, source code, production system, research pipeline, validation threshold or other confidential methodology.

References to backtests, historical simulations, out-of-sample evaluation, prospective evidence, stress testing, resampling, Monte Carlo methods or other forms of model evaluation are provided solely to examine methodological and evidential questions. Historical, simulated, hypothetical or model-based results do not guarantee future outcomes. A model or strategy that has performed favorably under historical or simulated conditions may perform differently under future market conditions, and validation cannot eliminate uncertainty or establish permanent robustness.

Financial markets and quantitative trading strategies involve substantial risk. Market conditions, liquidity, volatility, transaction costs, execution conditions, model behavior and other relevant

factors may change over time. Actual outcomes, including losses, may differ materially from historical, simulated, hypothetical or model-based results.

FINOVIS has not assessed the suitability or appropriateness of any financial instrument, investment product, quantitative model, strategy or transaction for any particular reader. Nothing in this publication should be interpreted as a statement that any such instrument, product, model, strategy or transaction is suitable, profitable or appropriate for any particular investor or market participant.

Any examples, scenarios or illustrations presented in the publication are used solely to clarify methodological concepts unless expressly identified as external empirical evidence. Illustrative material does not represent actual FINOVIS strategy performance or proprietary FINOVIS research results.

Readers should conduct their own independent assessment and, where appropriate, obtain advice from qualified professional advisers before making investment, financial, legal, tax, regulatory or other decisions. No decision should be made solely in reliance on this publication.

The information and analysis reflect the sources and methodological understanding available at the time of preparation. FINOVIS undertakes no obligation to update, revise or supplement this publication as a result of subsequent research, market developments, regulatory changes or other events.

To the fullest extent permitted by applicable law, FINOVIS and its affiliates, officers, employees, representatives and contributors accept no liability for any direct or indirect loss, damage, cost or expense arising from reliance on this publication or from any investment, trading, financial or other decision made or action taken on the basis of information contained in it.

Nothing in this publication creates any advisory, fiduciary, contractual or other professional relationship between FINOVIS and the reader.

FINOVIS RESEARCH

QUANTITATIVE METHODS · RP-003

Robustness Before Performance: Evaluating Quantitative Models Beyond Backtesting

Finovis Software Solutions LLC-FZ

finovis.ai