OpenQuant: Reasoning Algorithms & Thesis

Alex Reid[*1]
Tsinghua University
alex.reid@sc.tsinghua.edu.cn

Cameron Ramsey [*2]
Queen's University Belfast
cramsey11@qub.ac.uk

Edward Liu [*3]
New York University Abu Dhabi
el4371@nyu.edu

Kamai Jackson-Wade [*4]
University of Warwick
kamai.jackson-wade@warwick.ac.uk

[1]Schwarzman College, Tsinghua University
[2]School of Mathematics and Physics, Queen's University Belfast
[3]New York University Abu Dhabi
[4]School of Engineering, University of Warwick

Abstract

We introduce **OpenQuant**, a reasoning framework for evaluating quantitative trading strategies under benchmark dominant constraints. Unlike conventional evaluation pipelines that optimise a singular metric (e.g. Sharpe Ratio), OpenQuant treats strategy selection as a constrained multi-determinant decision problem.
The OpenQuant framework separates feasibility from optimisation: strategies must first satisfy strict dominance and risk constraints relative to a benchmark before being admitted into a nonlinear scoring functional. Performance metrics are mapped through calibrated sigmoid transforms to prevent metric explosion and enforce diminishing marginal reward.
We formalise the evaluation problem, define admissibility conditions, and present a computational engine that implements this reasoning algorithm. The system provides interpretable ranking, benchmark-relative accountability, and modular extensibility. We argue that such structured evaluation constitutes a form of algorithmic reasoning under uncertainty and offers a reproducible alternative to discretionary strategy selection.

1 Introduction

The evaluation of quantitative trading strategies is often reduced to the optimisation of a single performance statistic, most commonly the Sharpe ratio. While convenient, such scalarisation is structurally fragile: it obscures trade-offs between return, risk, drawdown behaviour, and benchmark-relative performance. In practice, capital allocators evaluate strategies through layered reasoning, imposing hard risk constraints before considering relative outperformance and robustness.

This paper proposes **OpenQuant**, a formal framework that encodes this layered reasoning into an explicit algorithm. We treat strategy evaluation as a constrained multi-objective optimisation problem defined over a space of empirical performance functionals. The system operates in two stages:

[*]Equal contribution, names listed in alphabetical order. Alex proposed the original concept and spearheaded the research direction; Kamai developed the computational modeling for the metric system; Edward engineered the agentic machine learning system; and Cameron led the development of the mathematical framework.

1. **Feasibility gating:** Strategies must dominate a benchmark under predefined admissibility constraints.
2. **Nonlinear aggregation:** Admissible strategies are evaluated via a bounded composite scoring functional constructed from calibrated metric transforms.

This separation between admissibility and optimisation reflects institutional practice and prevents degenerate solutions that maximise one metric at the expense of structural risk.

We introduce **OpenQuant**, a principled reasoning architecture for evaluating strategies under uncertainty built using two-stage reasoning framework that addresses this problem by separating feasibility from optimisation. In the first stage, strategies are tested against a set of hard admissibility gates derived from benchmark dominance and risk constraints - strategies that fail any gate are immediately disqualified. In the second stage, admitted strategies are evaluated by a **nonlinear composite scoring** function, whose components are mapped through calibrated sigmoid transforms, which enforce marginal reward and prevent metric explosion.

The resulting scores are further normalised against a same-ticker population via **z-score rescaling**, so that rankings reflect relative performance within a comparable set rather than numerical coincidences.

OpenQuant also addresses the upstream problem of pitch integrity: before a strategy can be evaluated, the data underlying its backtest must be validated for fabrication and methodological soundness.

We implement **Eva**, a multi-agent validation pipeline comprising of a fabrication detector, a coding errors auditor, and a data-quality assessor that gates strategies into one of three outcomes:

1. **Rejection** for suspected data manipulation
2. **Request** for clarification on methodological gaps
3. **Advancement** to final scoring

This pipeline is implemented as a set of concurrent LLM-backed (**ClaudeAI**) agents operating over structured pitch drafts.

First, we formalise the strategy evaluation problem as a constrained multi-determinant optimisation, and define precise admissibility conditions rooted in benchmark dominant theory. Second, we introduce a calibrated sigmoid scoring functional with theoretically motivated component weights. Third, we present a full computational implementation — spanning data validation, methodology auditing, and normalised ranking, and demonstrate that structured algorithmic reasoning offers a reproducible and interpretable alternative to *discretionary* strategy selection.

2 Motivations

The Sharpe ratio has become the standard for strategy comparison in both academic and industrial settings (Sharpe, 1966). Its appeal is intuitive, by normalising excess return by volatility, you yield a dimensionless measure of return per unit of risk.
However, this normalisation conflates fundamentally distinct phenomena. A strategy with low volatility and low return may achieve the same Sharpe ratio as one with high volatility and high return; their risk profiles and practicality are, however, entirely different. Furthermore, Sharpe ratio optimisation provides no guarantee of benchmark outperformance: a strategy can achieve a Sharpe of 2.0 while still losing to a passive buy-and-hold portfolio on a **cumulative** return basis.

A structural inefficiency in modern quantitative finance is geographic and institutional concentration. Top-tier quant firms draw from a narrow pipeline of talent, predominantly researchers at elite universities in a handful of cities, leaving substantial intellectual capital untapped. OpenQuant represents a rebuttal, this concentration is not a function of ability, but of access.

Let $\mathcal{H}_{\text{firm}}$ denote the set of human-hours of analytical effort contributed annually by a quant shop, and let $\mathcal{H}_{\text{OQ}}$ denote the equivalent quantity for an open-submission platform such as OpenQuant. By construction:

$$|\mathcal{H}_{\text{OQ}}| \gg |\mathcal{H}_{\text{firm}}| \tag{1}$$

since $\mathcal{H}_{\text{OQ}}$ aggregates contributions from a geographically and institutionally unbounded population. More formally, let $\mathcal{T}_{\text{firm}} \subset \mathcal{S}$ and $\mathcal{T}_{\text{OQ}} \subset \mathcal{S}$ be the sets of strategies accessible to a firm and to OpenQuant respectively. Under reasonable assumptions about the distribution of alpha-generating insight across the global population:

$$|\mathcal{T}_{\text{OQ}}| \gg |\mathcal{T}_{\text{firm}}|, \qquad \mathbb{E}[\alpha^* \mid s \in \mathcal{T}_{\text{OQ}}] \geq \mathbb{E}[\alpha^* \mid s \in \mathcal{T}_{\text{firm}}] \tag{2}$$

where $\alpha^*(s)$ denotes the true out-of-sample Jensen's alpha of strategy s. The inequality in expected alpha follows from the observation that a larger, more diverse strategy pool samples more of the tail of the alpha distribution, the rare, non-consensus insights that generate excess returns.

Cost Efficiency: Let $C(s)$ denote the cost of sourcing, evaluating, and onboarding strategy s. For a traditional quant firm, this cost is dominated by researcher salaries and infrastructure:

$$C_{\text{firm}}(s) = c_{\text{salary}} + c_{\text{infra}} + c_{\text{overhead}} \tag{3}$$

Under the OpenQuant model, strategy sourcing is decentralised: contributors bear their own development costs, and the platform incurs only evaluation costs c_{eval}:

$$C_{\text{OQ}}(s) = c_{\text{eval}} \ll C_{\text{firm}}(s) \tag{4}$$

The risk-adjusted return on evaluation spend is therefore:

$$\rho = \frac{\mathbb{E}[\alpha^*(s) \mid \mathcal{A}(s) = 1]}{C_{\text{OQ}}(s)} \gg \frac{\mathbb{E}[\alpha^*(s) \mid \mathcal{A}(s) = 1]}{C_{\text{firm}}(s)} \tag{5}$$

The admissibility gates and composite scoring functional of Section 3 serve a critical role in this model: by filtering the large incoming strategy pool to only those satisfying hard dominance and risk constraints, they transform a high-volume, low-cost intake process into a high-quality, curated output set. The mathematical framework is therefore not merely an evaluation tool, but the mechanism by which scale is converted into quality.

The financial industry widely acknowledges that active strategy management is only justified when it outperforms a passive alternative on a risk-adjusted basis (Jensen, 1968), yet evaluation systems frequently compare strategies only against each other, rather than against the relevant benchmark. This creates a selection environment in which the "best" strategy in a pool of uniformly underperforming candidates is elevated to a false prominence. OpenQuant addresses this by treating benchmark dominance as a hard **prerequisite**: a strategy that fails to deliver positive excess return, positive Jensen's alpha, and a Sharpe ratio exceeding zero is *disqualified entirely*, regardless of how it ranks against peers.

Quantitative strategies are almost universally evaluated on historical data, introducing a well-documented family of methodological hazards: look-ahead bias (where future information leaks into training features), survivorship bias (where the universe of tickers is defined ex post using only companies that persisted), data snooping (where the space of potential strategies is implicitly searched until an apparently strong backtest is found), and overfitting to in-sample noise (Bailey and López de Prado, 2014; López de Prado, 2018). These errors systematically inflate reported performance metrics, and no ranking system operating on reported metrics alone can detect them. OpenQuant introduces a validation stage that explicitly targets these failure modes using structured LLM-based auditing.

Even a methodologically sound strategy with genuine alpha may score poorly in absolute terms if the benchmark's period was 'difficult' and inversely, may appear strong simply because the comparison period was bull-market-dominant.

3 Mathematical Framework

3.1 Problem Formulation

Let $\mathcal{S}$ be a finite set of candidate trading strategies. Each strategy $s \in \mathcal{S}$ is characterised by a feature vector $\mathbf{f}(s) \in \mathbb{R}^d$, whose components encode performance and risk metrics derived from

historical backtesting, and by a benchmark reference $b(s)$ representing the buy-and-hold return series for the same ticker and period. The evaluation problem is to construct a total preorder $\preceq$ on $\mathcal{S}$ that is consistent with the following prerequisites:

1. **Benchmark dominance**: any s that fails to strictly outperform $b(s)$ on all core risk-adjusted criteria should be ranked below all s' that do.
2. **Multi-criterion consistency**: the ranking should reflect trade-offs across all relevant performance dimensions, not a single scalar proxy.
3. **Population relativity**: rankings within a homogeneous group (same ticker, same benchmark period) should reflect relative excellence, not absolute numeric coincidence.
4. **Interpretability**: the scoring functional should decompose into interpretable components with well-defined economic meaning.

3.2 Admissibility Gates

A strategy s is admissible ($\mathcal{A}(s) = 1$) only if it passes all five gates:

$$\alpha(s) > 0, \quad r(s) - r_b(s) > 0, \quad \mathrm{SR}(s) > 0, \quad \mathrm{PF}(s) > 1, \quad \mathrm{MDD}(s) > -0.50 \tag{6}$$

These jointly enforce benchmark dominance, positive risk-adjusted returns, profitable trading, and catastrophic drawdown prevention. Any strategy with $\mathcal{A}(s) = 0$ is excluded from ranking with score zero.

3.3 Component Scoring

Each metric is mapped to $[0, 1]$ via a sigmoid transform centred at a decision-relevant threshold μ_k:

$$\phi_k(x) = \sigma(\lambda_k(x - \mu_k)), \qquad \sigma(z) = \frac{1}{1 + e^{-z}} \tag{7}$$

This is bounded, strictly monotone, and concentrates scoring discrimination at μ_k, implementing diminishing marginal reward automatically. Twelve components across three categories are defined in Table 1.

Table 1: OpenQuant scoring components.

Component	**Category**	μ_k	λ_k	w_k	**Metric**
Excess Return	Benchmark	0.05	20	3.0	$r(s) - r_b(s)$
Alpha	Benchmark	0.02	30	3.0	$\alpha(s)$
Information Ratio	Benchmark	0.50	3	2.5	$\mathrm{IR}(s)$
Up/Down Capture	Benchmark	1.00	3	2.0	$u(s)/d(s)$
Drawdown vs. Benchmark	Benchmark	0.00	15	2.0	$\mathrm{MDD}(s) - \mathrm{MDD}_b(s)$
Sharpe Ratio	Risk	1.00	2	2.5	$\mathrm{SR}(s)$
Sortino Ratio	Risk	1.50	1.8	2.0	$\mathrm{SoR}(s)$
Calmar Ratio	Risk	1.00	2	1.5	$\mathrm{CR}(s)$
Profit Factor	Trades	1.50	2	1.5	$\mathrm{PF}(s)$
Win/Loss Ratio	Trades	1.50	1.5	1.0	$\lvert\bar{w}\rvert/\lvert\bar{l}\rvert$
Expectancy	Trades	0.005	200	1.5	$\mathbb{E}[\text{trade P\&L}]$
Trade Consistency	Trades	50	0.04	0.5	$N_{\text{trades}}(s)$

3.4 Composite Score

The composite score is a weighted average of component scores, scaled to $[0, 100]$:

$$\Phi(s) = \frac{100}{W} \sum_{k=1}^{12} w_k \, \phi_k(f_k(s)), \qquad W = \sum_{k=1}^{12} w_k = 25 \tag{8}$$

Benchmark metrics carry the highest aggregate weight (12.5), followed by risk metrics (6.0) and trade metrics (4.5).

3.5 Population-Normalised Scoring

When $|\mathcal{P}_{\text{ticker}}| \geq 3$ strategies exist for the same ticker, component scores are z-score normalised and re-sigmoidised:

$$\tilde{\phi}_k(s) = \sigma\left(\frac{\phi_k(f_k(s)) - \bar{p}_k}{\hat{\sigma}_k}\right), \qquad \tilde{\Phi}(s) = \frac{100}{W}\sum_{k=1}^{12} w_k\,\tilde{\phi}_k(s) \tag{9}$$

This double-sigmoid maps an average strategy to 0.5 per component, is symmetric around the population mean, and compresses outliers. Adding any new strategy triggers a full pool rescore.

3.6 Allocation Mapping

A capital allocation $A(s)$ is derived from $\Phi(s)$ and time horizon $h(s)$:

$$A(s) = \min(20{,}000,\ \text{round}_{100}(A_0(\Phi(s)) \cdot \gamma(h(s)))) \tag{10}$$

$$A_0(\Phi) = \begin{cases} 0 & \Phi < 55 \\ 1{,}000 & 55 \leq \Phi < 65 \\ 2{,}500 & 65 \leq \Phi < 75 \\ 5{,}000 & 75 \leq \Phi < 85 \\ 10{,}000 & 85 \leq \Phi < 93 \\ 15{,}000 & \Phi \geq 93 \end{cases}, \qquad \gamma(h) = \begin{cases} 0.8 & h = \text{days} \\ 1.0 & h = \text{weeks} \\ 1.2 & h = \text{months} \\ 1.4 & h = \text{years} \end{cases} \tag{11}$$

The horizon multiplier discounts shorter-horizon strategies to reflect greater uncertainty and liquidity costs.

4 The OpenQuant System

4.1 System Overview

The OpenQuant system comprises two interacting subsystems: a *pitch intake and validation pipeline* that accepts raw strategy submissions and gates them into one of three validation outcomes, and a *scoring and ranking engine* that implements the mathematical framework of Section 3. Both subsystems operate on a shared data model—the `PitchDraft` and `Strategy` dataclasses—and produce structured, auditable outputs.

4.2 Pitch Intake Pipeline

Strategy submissions arrive as *pitch drafts*, structured documents containing the trading thesis, ticker universe, data source URLs, methodology description, and attached CSV/TSV price data. Drafts are progressively enriched through a combination of structured field updates and heuristic extraction from free-text user input. Required fields include thesis, time horizon, methodology summary, tickers, and source URLs; a draft missing any required field cannot advance to evaluation.

Data Ingestion: Uploaded tabular files are parsed to extract a close price series, from which returns, Sharpe ratio, and maximum drawdown are computed deterministically. A data quality score $q \in [0, 1]$ is derived as:

$$q = \text{clamp}(1 - m_f - 0.5 \cdot d_f) \tag{12}$$

where m_f is the fraction of missing cells and d_f is the fraction of duplicate rows. At least 30 valid price observations are required; fewer constitutes a hard rejection.

Deterministic Scoring: The intake pipeline computes a preliminary composite score using the functional defined in Equation 3.4, with six components: Sharpe ratio ($w = 0.30$), maximum drawdown ($w = 0.20$), annualised volatility as a risk proxy ($w = 0.15$), data quality ($w = 0.15$), methodology score ($w = 0.10$), and data source match rate ($w = 0.10$). The methodology score is itself a weighted keyword-presence function rewarding explicit mentions of out-of-sample validation, walk-forward testing, risk management, and related methodological safeguards.

4.3 Multi-Agent Validation

Beyond deterministic data checks, the intake pipeline deploys two concurrent LLM-backed validation agents to detect failure modes that resist rule-based detection.

Fabrication Detector: The fabrication detection agent receives a structured payload containing the pitch metadata, data summary statistics, and a sample of the uploaded records, and is prompted to identify evidence of data manipulation along five dimensions: impossible values (negative volume, constant prices where variation is expected), unnaturally smooth return series inconsistent with market microstructure, timestamp anomalies suggesting post-hoc editing, source-data mismatches between declared URLs and submitted files, and systematic tampering patterns. The agent outputs a structured JSON verdict with per-finding severity codes in **low, medium, high, critical**.

Coding Errors Auditor: The coding errors agent targets unintentional methodological mistakes that inflate backtest quality: look-ahead bias from future-data leakage, target leakage in feature construction, survivorship bias in the ticker universe, overfitting and data snooping risk, weak or absent out-of-sample validation protocols, and unrealistic transaction cost assumptions. Unlike the fabrication detector, this agent is explicitly oriented toward good-faith errors rather than fraud intent.

Both agents are instantiated from the same base model (**Google's Gemini**) with temperature set to zero, ensuring deterministic outputs. They execute concurrently via a thread pool, and their outputs are merged into a unified flag set before final outcome determination.

4.4 Validation Outcome Determination

The validation pipeline maps each pitch to one of three outcomes:

1. Blocked $\rightarrow$ fabrication
2. Needs clarification
3. Ready for next stage

The outcome is determined by the following logic:

1. If the fabrication detector returns a critical-severity flag or a "fabrication" verdict, the pitch is immediately blocked and the submitter is notified without further explanation.
2. Otherwise, if any agent returns a medium- or higher-severity flag, or if hard rejection conditions are present (insufficient data, missing required fields), the pitch enters a clarification loop in which the submitter is presented with targeted questions derived from the specific flags raised.
3. If no actionable issues are detected, the pitch advances to final review with a capital allocation recommendation computed per Section 3

This three-outcome structure implements a form of *iterative triage*: submitters are not simply accepted or rejected, but guided toward compliance through structured feedback. The system generates up to four clarification questions per pitch, de-duplicated across agents and prioritised by flag severity.

4.5 Scoring and Ranking Engine

The ranking engine accepts a JSON array of fully specified strategy objects and produces a ranked output with full score decompositions. The processing pipeline is: (1) schema validation and sanity checking; (2) hard gate application; (3) composite score computation (absolute or population-normalised depending on pool size); (4) within-ticker and overall ranking; (5) report generation in either human-readable text or structured JSON.

The text report provides a multi-level view: a global ranking table, per-ticker sub-rankings, and detailed per-strategy breakdowns including bar-chart visualisations of per-component scores, enabling practitioners to identify at a glance whether a strategy's overall ranking reflects balanced excellence or a single dominant metric compensating for weaknesses elsewhere.

4.6 Extensibility

Both the scoring component list and the sigmoid parameters are defined as a single declarative configuration structure, allowing practitioners to introduce domain-specific metrics, adjust thresholds to reflect different market regimes, or reweight categories without modifying any algorithmic logic. Similarly, the multi-agent validation pipeline is designed to admit additional specialised agents (e.g., a regime-sensitivity auditor or a liquidity risk assessor) by registering them in the concurrent executor and merging their flag outputs into the existing aggregation logic.

5 Conclusion

We have presented **OpenQuant**, a two-stage reasoning framework for the structured evaluation of quantitative trading strategies. The framework addresses three distinct failure modes in conventional evaluation practice: single-metric scalarisation, which obscures structural risk; the absence of hard benchmark dominance requirements, which permits uniformly underperforming strategies to appear competitive; and the lack of upstream data integrity validation, which allows methodological errors and fabricated backtests to distort rankings.

The admissibility gate structure enforces benchmark dominance as a non-negotiable prerequisite, ensuring that no strategy reaches the scoring stage without demonstrating positive excess return, positive Jensen's alpha, positive risk-adjusted return, profitable gross trading, and controlled drawdown. The sigmoid scoring functional then maps the remaining strategy space into a calibrated, bounded score in $[0, 100]$, with population normalisation ensuring that rankings reflect relative excellence within a comparable peer group rather than absolute metric coincidence.

Critically, the mathematical framework is not merely an evaluation tool. In the OpenQuant model, where strategy sourcing is decentralised and $|\mathcal{T}_{\text{OQ}}| \gg |\mathcal{T}_{\text{firm}}|$, the admissibility and scoring pipeline is the mechanism by which a high-volume, geographically unbounded submission pool is converted into a curated, capital-allocable output set. The multi-agent validation layer, comprising a fabrication detector and a coding errors auditor operating concurrently over structured pitch drafts, further ensures that the scoring inputs are methodologically sound before any allocation is recommended.

Taken together, these components constitute a form of *algorithmic reasoning under uncertainty*: a reproducible, auditable, and institutionally grounded alternative to discretionary strategy selection. We release the full implementation as an open system, and we hope it serves both as a practical evaluation tool and as a foundation for further research into structured, multi-criterion decision-making in quantitative finance.

References

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

M. C. Jensen. The performance of mutual funds in the period 1945–1964. *The Journal of Finance*, 23(2):389–416, 1968. doi: 10.1111/j.1540-6261.1968.tb00815.x.

M. López de Prado. *Advances in Financial Machine Learning*. Wiley, Hoboken, NJ, 2018. ISBN 978-1-119-48208-6.

W. F. Sharpe. Mutual fund performance. *The Journal of Business*, 39(1):119–138, 1966. doi: 10.1086/294846.