Large Language Quant: Cross-Domain Mathematical Trading Strategy Generation with an Agentic LLM

Andrew Clayman
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
ac25767@essex.ac.uk

Michael Kampouridis
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
mkampo@essex.ac.uk

Tasos Papastylianou
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
tasos.papastylianou@essex.ac.uk

***Abstract*—We present an agentic large language model (LLM) system that draws on the breadth of an LLM's training across many disciplines to import advanced mathematical concepts from fields outside quantitative finance, including information theory, rough path signatures, information geometry, differential geometry, optimal transport, fractal analysis, and statistical physics, and turn them into systematic trading signals. Through a structured prompt loop, the LLM proposes a candidate trading strategy grounded in one such concept, implements it as a Python-based signal generator, runs a backtest, and refines the implementation in response to the results, producing 8 strategies across 7 mathematical domains. We evaluate the resulting strategies in two stages. First, each strategy is run independently across 19 assets (cryptocurrency, forex, equity indices, commodities) and 4 disjoint out-of-sample (OOS) test periods spanning 2013–2023; 7 of 8 LLM-generated strategies achieve positive mean Sharpe ratios, and 3 of the top 5 strategies overall are LLM-generated. Second, we deploy a portfolio-level genetic algorithm (GA) that simultaneously chooses which strategy runs on each asset and the capital weight allocated to it, optimising the Sharpe ratio of the combined portfolio. We run the GA under three pool variants (*Full* = LLM + benchmarks + cash; *LLM-Only* = LLM + cash; *Benchmark-Only* = benchmarks + cash) so that the contribution of the LLM-generated strategies can be isolated. Across 6 rolling walk-forward periods (2018–2024), with 50 independent GA runs per (period, variant) cell, the Full GA significantly outperforms the Benchmark-Only GA on aggregate Sharpe ($\Delta = +0.21$, Mann-Whitney U $p = 0.042$) and Calmar ratio ($p = 0.005$), and achieves significantly shallower maximum drawdowns in 5 of 6 periods. Given the choice of the Full pool, the GA allocates 51% of its capital on average to LLM-generated strategies and 43% to conventional benchmarks. The benefit is regime-dependent: LLM strategies excel in volatility-driven and transition regimes (2018, 2022, 2023) and lag in strong-momentum bull markets (2020, 2024) where buy-and-hold dominates, supporting a complementary rather than dominant role for cross-domain LLM-generated alpha.**

***Index Terms*—large language model, agentic AI, generative AI, algorithmic trading, cross-domain reasoning, genetic algorithm, portfolio optimisation**

I. Introduction

Large language models (LLMs) encode statistical patterns across vast corpora spanning mathematics, physics, and finance, enabling them to draw connections across domains that might be more challenging to uncover by humans. In quantitative finance, LLMs have been applied to sentiment analysis [1], [2] and, more recently, to generating formulaic alpha factors, mathematical expressions that map price and volume data to expected-return signals, by recombining standard financial operators [3], [4]. A less-explored direction is designing systems that draw on the breadth of an LLM's training across many disciplines to pull advanced mathematical concepts from fields such as information geometry, optimal transport, and stochastic analysis, and to turn them into trading signals, a task that has traditionally required expertise spanning several disciplines.

The design of systematic trading strategies has traditionally required deep expertise in both finance and mathematics, limiting the strategy space to well-known techniques such as moving-average crossovers, momentum, and mean-reversion bands. While advanced mathematical tools such as information theory, rough path signatures, optimal transport, and information geometry have rich theoretical foundations, they are not commonly found in the systematic trading literature, where simpler technical indicators dominate.

Contributions. We propose an agentic LLM system, Large Language Quant (LLQ), that iteratively generates, implements, and evaluates trading strategies expressed as Python code,[1] drawing on mathematical domains rarely applied in systematic trading. We evaluate the resulting strategies under a two-stage protocol designed to address both individual viability and aggregate portfolio value. *Stage 1* runs each LLM-generated strategy alongside conventional technical benchmarks across 19 assets and 4 disjoint OOS test periods spanning 2013–2023, to see how each strategy performs across a range of market regimes. *Stage 2* embeds the LLM-generated strategies and conventional benchmarks into a portfolio-level genetic algorithm that simultaneously chooses which strategy runs on each asset and how much capital it receives. To isolate the contribution of the LLM-generated strategies, we run the GA under three pool variants: a *Full* pool (LLM + benchmarks + cash), an *LLM-Only* pool (LLM + cash), and a *Benchmark-Only* pool (benchmarks + cash); the cash option lets the GA decline to trade an asset rather than being forced to assign

[1]The generation prompt and the generated strategy code are available at https://github.com/andrewclayman/LLQ_CIFEr.

one of the active strategies (Section III-C). Across 6 rolling walk-forward periods (2018–2024), we run the GA 50 times for each (period, variant) cell, yielding distributions of out-of-sample performance on which we can perform statistical testing. A panel of complementary tests (Mann-Whitney U with Holm-Bonferroni correction, bootstrap confidence intervals, Cohen's d, Kruskal-Wallis, and multi-metric t-tests) quantifies the aggregate value added by LLM-generated strategies relative to a benchmark-only GA, while a revealed-preference analysis examines how the Full GA allocates capital across the strategy pool when it has access to both LLM and benchmark strategies.

The contributions of this work are: (i) an agentic system that brings advanced mathematical domains into working trading strategies; and (ii) an empirical account of where the LLM-generated strategies add value, where they do not, and how a downstream optimisation layer can combine them with conventional benchmarks.

The remainder of the paper is organised as follows. Section II reviews related work on LLMs in finance, automated strategy discovery, and statistical evaluation in algorithmic trading. Section III presents the LLQ system, the strategy pool, the portfolio GA, and the evaluation design. Section IV reports per-strategy OOS results and Section V reports portfolio-level GA results with the full set of statistical tests. Section VI discusses findings, limitations, and future directions.

II. Related Work

LLMs in finance. Application of LLMs to quantitative finance has evolved rapidly from domain-specific language understanding to active strategy generation. BloombergGPT [2] demonstrated domain-specific pretraining, FinGPT [5] provided an open-source fine-tuning framework for financial text, and Lopez-Lira and Tang [1] showed that ChatGPT-derived sentiment signals contain predictive information for stock returns. More recently, researchers have moved toward LLMs that generate formulaic alpha factors: Alpha-GPT [4] translates natural-language investment hypotheses into alpha expressions, and Kou et al. [3] employ multi-agent LLM systems to iteratively generate and filter alpha candidates. These systems generate recombinations of conventional technical indicators rather than drawing on advanced mathematical frameworks. In the broader scientific domain, FunSearch [6] demonstrated that LLMs paired with evolutionary search can produce genuinely novel mathematical constructions, motivating our use of LLMs to bridge advanced mathematical theory and practical trading systems.

Automated strategy discovery and portfolio optimisation. Automating strategy discovery predates LLMs. Allen and Karjalainen [7] used genetic algorithms to discover technical trading rules but found that automatically generated rules struggle to outperform buy-and-hold after costs on S&P 500 data, while Neely et al. [8] applied genetic programming to foreign exchange with more promising results. These evolutionary approaches recombine known technical operators via combinatorial search; we instead use the GA to optimise portfolio-level deployment of an existing LLM-generated strategy pool, addressing strategy assignment and capital allocation jointly. The portfolio-level GA design follows the broader literature on evolutionary methods in portfolio construction [9], [10], but operates over a discrete strategy-asset assignment combined with a continuous weight simplex.

Information-theoretic measures in finance. Information-theoretic tools have a rich history in financial analysis: transfer entropy [11] extends Granger causality to nonlinear settings, while permutation entropy [12] and approximate entropy [13] quantify time-series complexity. These measures are rarely used as actionable trading signals within systematic strategies, a gap the LLQ system targets directly.

Backtest overfitting is a well-documented risk [14], [15], motivating our use of distributions of OOS performance from repeated runs rather than single-point estimates in Stage 2.

III. Methodology

A. The LLQ Agentic System

The Large Language Quant (LLQ) system is an iterative strategy-design agent built on Claude Opus 4.6 (Anthropic, 2025), accessed via Claude Code, an agentic coding tool that grants the LLM direct access to the local filesystem and shell. Within this setup the LLM behaves much like a human developer: it writes Python files, runs backtests, inspects the resulting metrics, and edits its own code, all within a single interactive session per strategy. Each generated signal-generator file was reviewed by the authors, and all OOS metrics in Sections IV and V were obtained through an author-controlled backtesting harness implemented using the `vectorbt` Python library rather than read from the agent's session logs.

Each strategy is generated in a separate session. The authors provide a structured prompt containing: (i) a directive to discover a novel trading strategy grounded in an advanced mathematical topic not commonly used in systematic trading; (ii) the `compute_signals(df)` interface specification (input: a pandas DataFrame of daily OHLCV (Open, High, Low, Close, Volume) bars; output: boolean entry/exit columns indicating, for each bar, whether to open or close a position); (iii) development data file paths and a backtesting harness built on `vectorbt` [16]; and (iv) the objective: an aggregated score $\tilde{S} - 0.5\,\tilde{D} - 0.2\,\tilde{T}$, where $\tilde{S}$, $\tilde{D}$, $\tilde{T}$ are the median Sharpe, median maximum drawdown, and median turnover penalty across disjoint development segments, computed per segment to encourage robustness across regimes. No specific mathematical domain is prescribed; the LLM chooses the domain, drafts a candidate strategy, and implements it autonomously. It iterates within the session by drafting a strategy, implementing the module, running the backtest, analysing failure modes, and refining the implementation. Each session is capped at a budget of 60 backtests and is required to try at least 5 distinct strategy ideas (meaningfully different trading concepts) within the first 25% of the budget, so that the search broadens before it deepens and the LLM does not overfit a single idea by running it through too many tuning rounds [14]. Strict data isolation is enforced:

TABLE I
THE 8 LLM-GENERATED STRATEGIES AND 6 CONVENTIONAL BENCHMARKS.

#	Strategy	Domain
LLM-Generated		
1	Spectral Entropy	Information Theory
2	Rough Path Signatures	Stochastic Analysis
3	Information Geometry	Differential Geometry
4	Wasserstein Distance	Optimal Transport
5	Rate Distortion	Information Theory
6	Fisher Information	Information Geometry
7	Volatility Phase Trans.	Statistical Physics
8	Hurst Mean Reversion	Fractal Analysis
Benchmark		
B1	Mean Reversion Bands	Technical
B2	RSI Divergence	Technical
B3	MA Crossover	Technical
B4	Bollinger Breakout	Technical
B5	Donchian Breakout	Technical
B6	Buy & Hold	Passive

the LLM has access only to development data, with test sets hidden during generation.

B. Strategy Pool

The LLQ system produced 8 strategies across 7 mathematical domains (Table I). Each strategy is a self-contained Python module mapping raw OHLCV bars to boolean entry/exit signals; all 8 independently converged on long-only positions despite no such constraint. Signals use rolling windows without lookahead and are shifted by one bar before execution.

All numerical parameters within each strategy (rolling windows, thresholds, multipliers, confirmation periods) were selected autonomously by the LLM during the iterative development process and held fixed for OOS evaluation. The novelty contribution lies in the LLM's choice of mathematical concept and signal construction; parameter values were not human-tuned. As a robustness check (Appendix A), we ran a per-strategy grid search around each LLM-chosen value on the development set: the best grid configuration improves mean Sharpe by only $+0.11$ on dev and $+0.11$ out-of-sample, indicating the LLM-selected parameters are close to optimal within the search neighbourhood and that the qualitative findings are not driven by parameter selection artefacts. In the descriptions that follow, a *bar* denotes a single daily OHLCV candlestick (the time resolution of our data), and all rolling windows are expressed in bars (trading days) unless otherwise noted.

Spectral Entropy [17] detects when price movements lose any dominant cyclical pattern, meaning the market has no exploitable periodicity. This disorder often coincides with regime transitions: the old pattern has broken down but a new trend has not yet been priced in, creating a window for early entry. To measure this, the price over the most recent 60 bars is decomposed into its frequency components via the Fourier transform. If one frequency dominates (e.g. a regular cycle) entropy is low; if all frequencies contribute equally (no pattern) entropy is high. Formally, the power spectrum is normalised into a probability distribution $\{p_i\}$, where p_i is the fraction of total power at frequency i, and entropy is computed as

$$H = -\sum_i p_i \log p_i \tag{1}$$

An adaptive threshold (the mean μ_H plus $0.80\,\sigma_H$, where σ_H is the standard deviation of the trailing 15 entropy values) calibrates "high" relative to recent history. The strategy enters long when entropy exceeds this threshold and the current bar's close is below the previous bar's close (buying into disorder after a one-bar dip), and exits when entropy falls below the threshold or on the first up-close after entry.

Rough Path Signatures [18], [19] builds on the observation that before a sustained price move the trading range often widens while returns have not yet committed to a direction: volatility leads, price follows. Rough path theory provides tools for summarising the behaviour of complex paths; the *signature* is a set of statistics that capture a path's shape. The strategy tracks two cumulative series over a 20-bar window. The first, x_i, is the cumulative log-return up to bar i; the second, y_i, is the cumulative sum, also up to bar i, of the per-bar normalised range $(\text{High} - \text{Low})/\text{Close}$, which serves as a volatility proxy. Both series are z-normalised over the window and plotted against each other to produce a 2D path. If volatility rises first and returns follow, this path traces a loop enclosing a large area; if the two move in lockstep the path is roughly a straight line with little enclosed area; if they are independent the path scribbles randomly and signed contributions to the area cancel out. The level-2 signature of a 2D path consists of its first-order increments together with the four iterated integrals S^{ij}; the antisymmetric combination $A = |S^{12} - S^{21}|$ is (twice) the Lévy area enclosed by the path, where

$$S^{12} = \sum_i x_i(y_{i+1} - y_i), \quad S^{21} = \sum_i y_i(x_{i+1} - x_i) \tag{2}$$

The strategy operates in two modes. When $A > 1.5$ (range systematically leads returns, indicating that volatility expansion precedes directional movement) it enters long with 14-bar momentum above 0.8%, following the emerging trend. When $A < 0.3$ (the two series move in lockstep or independently, so no sequential pattern is available to exploit) it switches to mean reversion: entering long when price falls more than 2% below the 20-bar moving average and exiting when price rises 2% above it. In the intermediate region $0.3 \leq A \leq 1.5$ no new positions are opened, but any open position is exited on negative momentum.

Information Geometry [20] asks whether the market's behaviour is changing gradually or abruptly. When the statistical profile of returns evolves smoothly it is safer to hold positions; when it shifts suddenly, existing positions are at risk. To quantify this, returns over rolling 20-bar windows are discretised into 10-bin histograms (fixed bin range $[-0.05, +0.05]$ with add-one smoothing). Each histogram is treated as a point on a *statistical manifold* (the space of probability distributions equipped with the Fisher–Rao geometry), and the distance

between two consecutive histograms p and q, the natural notion of distance on this manifold, is

$$d_{FR}(p,q) = 2\arccos\Big(\sum_i \sqrt{p_i\, q_i}\Big) \tag{3}$$

Over 18 consecutive 20-bar windows the histograms trace a path through this manifold. *Path curvature* is the deviation of that path from the corresponding *geodesic*, the shortest path on the manifold between its endpoints: $\kappa = (\text{path length} - \text{direct distance})/\text{path length}$, where path length is the sum of step-by-step Fisher–Rao distances along the trajectory and direct distance is the Fisher–Rao distance from the first to the last histogram. High curvature means the distribution is zigzagging erratically through the manifold; low curvature means it is evolving smoothly. The strategy enters long during low-curvature periods (below the 75th percentile of a rolling 30-value curvature history) with positive 10-bar momentum above 0.1%, and exits when curvature exceeds this adaptive threshold or momentum turns negative.

Wasserstein Distance [21] measures whether the market is behaving normally or has shifted into an unusual regime. When recent returns look similar to a longer-term baseline, the conditions that established the current trend are still intact and positions are more likely to persist; when they diverge sharply, the market environment has changed and existing positions may no longer be justified. The strategy uses the 1-Wasserstein (earth mover's) distance, which measures how much "work" is needed to reshape one return distribution into another, comparing a recent 48-bar window (≈10 weeks) against a longer-term 240-bar baseline (≈1 year):

$$W_1(F,G) = \int_0^1 \left|F^{-1}(u) - G^{-1}(u)\right| du \tag{4}$$

where F and G are the empirical cumulative distribution functions of the recent and baseline return windows respectively. The asymmetric windows are deliberate: the short window captures the current state while the long window provides a stable reference. The strategy enters long when W_1 remains below the 75th percentile of a rolling 100-value distance history for 8 consecutive bars with positive momentum, and exits when W_1 exceeds this adaptive threshold or momentum turns negative. Note that in the 1-year Stage 2 OOS windows this strategy fired only one trade across the first OOS period and none in the remaining periods, leaving the Stage 2 GA to treat it as a no-trade option; we therefore drop it from the Stage 2 pool and let the Cash option (Section III-C) play that role directly.

Rate Distortion [22] identifies when price action is repetitive and therefore predictable, the kind of structured behaviour that tends to persist in the short term. Conversely, when price movements become noisy and unstructured, there is no pattern to exploit. The underlying idea comes from rate–distortion theory:

$$R(D) = \min_{p(\hat{x}|x):\, E[d(x,\hat{x})]\le D} I(X;\hat{X}) \tag{5}$$

where $R(D)$ is the minimum information (in bits) needed to represent a signal X with average distortion at most D, and $I(X;\hat{X})$ is the mutual information between the original and reconstructed signal. If price action can be compressed into a short description (low $R(D)$) it is structured; if it resists compression (high $R(D)$) it is noisy. The strategy approximates this by taking the last 48 bars of close prices, downsampling to $[4, 8, 16, 32]$ points, interpolating back to the original length, and measuring the reconstruction error (normalised mean squared error) at each resolution. A price pattern that can be reconstructed accurately from just a few points is compressible; one that loses its shape is noisy. The area under the resulting error curve is converted to a compressibility score between 0 and 1. The strategy enters long when the score exceeds 0.7 (highly structured) for 3 consecutive bars with positive 24-bar momentum, and exits when compressibility falls below 0.4 (noise-like regime) or momentum reverses.

Fisher Information [20] distinguishes between stable markets, where returns behave consistently and directional bets are favoured, and unstable markets, where the return distribution is shifting rapidly and positions are risky. Fisher information measures how sensitive a distribution is to small changes in its parameters; in trading terms, this captures how easily the market's behaviour could tip into something different. Assuming returns follow a Gaussian distribution with mean μ and variance σ^2, the Fisher information matrix is

$$\mathcal{I} = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix} \tag{6}$$

The trace $\mathrm{tr}(\mathcal{I}) = 1/\sigma^2 + 1/(2\sigma^4)$ summarises how sensitive the distribution is overall: low variance yields a high trace (the distribution is tightly concentrated and sensitive to shifts), while high variance yields a low trace. To account for fat tails, the trace is scaled by a kurtosis adjustment $\mathrm{tr}(\mathcal{I}) \times (1 + 0.1\,\max(0,\,\kappa - 3))$, where κ is the empirical kurtosis. The strategy estimates this from the most recent 24 bars of returns. A low value means returns are stable and predictable; a high value means the market's statistical character is in flux. The strategy enters long when the trace is below the 75th percentile for 4 consecutive bars with positive 24-bar momentum, and exits on regime transition (a trace spike).

Volatility Phase Transition models the market as a system that transitions between *ordered* (low-volatility, trending) and *disordered* (high-volatility, chaotic) phases, and tries to detect early warning signs before transitions occur. Rolling volatility σ_t, the standard deviation of returns over 24 bars, serves as the regime indicator. Adaptive percentile thresholds from a trailing 200-bar history of σ_t classify the current state:

$$\text{ordered: } \sigma_t < P_{40}(\sigma), \quad \text{disordered: } \sigma_t > P_{75}(\sigma) \tag{7}$$

where $P_k(\sigma)$ denotes the k-th percentile of the trailing volatility history. Between these thresholds the strategy monitors the lag-1 autocorrelation of recent volatility (over 48 bars) as a *critical slowing* indicator: when this autocorrelation exceeds 0.7, volatility shocks are persisting longer than usual, warning that a phase transition into disorder is approaching. The strategy enters long after 4 consecutive bars in the ordered phase with positive 24-bar momentum, and exits when critical slowing

is detected, the disordered phase is reached, or momentum reverses.

Hurst Mean Reversion [23] identifies whether the market is trending or mean-reverting by estimating the Hurst exponent H via rescaled-range (R/S) analysis over a 30-bar window, fitting the scaling relationship

$$E\left[\frac{R(n)}{S(n)}\right] = C\, n^H \tag{8}$$

For each sub-window of length $n \in \{2, 4, 8, 16\}$, the simple returns within the sub-window are mean-adjusted and cumulatively summed to produce a series of *cumulative deviates* $\{Y_t\}$, which measures how far returns have drifted from their local mean. $R(n)$ is the max–min range of this cumulative-deviates series within the sub-window, and $S(n)$ is its standard deviation; the rescaled range R/S therefore expresses how far the price wanders relative to the variability of that wandering. If $H > 0.5$ the range grows faster than a random walk would predict, so moves tend to continue in the same direction (persistence); if $H < 0.5$ it grows slower and moves tend to reverse (anti-persistence); $H = 0.5$ corresponds to a random walk with no exploitable pattern. The strategy enters long when $H > 0.55$ and the 3-bar mean return is positive, and exits when $H < 0.45$ or the 3-bar mean return falls below -0.5%.

Active benchmarks (B1–B5). Five conventional indicator-based strategies [24] serve as active benchmarks (Table I). Parameters were grid-searched on non-overlapping development segments disjoint from all OOS test periods, with the optimal configuration selected by maximising the same composite score $\tilde{S} - 0.5\,\tilde{D} - 0.2\,\tilde{T}$ used as the LLQ objective (Section III-A), then held fixed for all OOS evaluation. A sensitivity check restricting the grid search to the earliest segment (2017–2018) yielded comparable OOS performance (median Sharpe difference $+0.08$), confirming benchmark parameter selection is not driven by lookahead. **Mean Reversion Bands** enters at the lower Bollinger band, exits at the mean. **RSI Divergence** uses the Relative Strength Index (RSI), a momentum oscillator scaled to $[0, 100]$; it enters long on oversold RSI with bullish price divergence and exits on overbought. **MA Crossover** (fast=15, slow=80) enters on bullish cross, exits on bearish cross. **Bollinger Breakout** (period=50, $k = 2.5\sigma$) enters above the upper band, exits at the mean. **Donchian Breakout** (entry=50-day high, exit=20-day low) enters long on a close above the 50-day high and exits on a close below the 20-day low, a classical trend-following rule.

Passive baseline (B6). Buy & Hold enters at the start of each evaluation window and holds throughout, providing a passive baseline. Because B6 generates no signal it is not evaluated as an active strategy in Stage 1; it appears as a structural allocation option within the Stage 2 GA, allowing the optimiser to assign passive holding to any asset where active signals offer no advantage.

C. Portfolio-Level Genetic Algorithm

The portfolio GA simultaneously optimises (i) which strategy to run on each of the 19 assets and (ii) the capital weight allocated to each asset. The chromosome is a pair $(\mathbf{s}, \mathbf{w})$, where $\mathbf{s} \in \{0, 1, \ldots, K-1\}^{19}$ assigns a strategy index s_i to each asset i, and $\mathbf{w} \in \Delta^{19}$ is the capital-weight vector drawn from the 19-dimensional probability simplex $\Delta^{19} = \left\{\mathbf{w} \in \mathbb{R}^{19}_{\geq 0} : \sum_{i=1}^{19} w_i = 1\right\}$, so weights are non-negative and sum to one.

Fitness. Fitness is the annualised Sharpe ratio of the combined weighted portfolio's training-period daily returns:

$$\text{fit}(\mathbf{s}, \mathbf{w}) = \frac{\bar{r}}{\sigma_r}\sqrt{252}, \quad r_t = \sum_{i=1}^{19} w_i\, r_{i,t}^{(s_i)} \tag{9}$$

where $r_{i,t}^{(s_i)}$ is the daily return of strategy s_i on asset i at time t. This is *not* a per-asset Sharpe average: diversification effects (inter-asset covariance) are reflected in the combined-portfolio Sharpe, so two negatively correlated strategies can produce a high-Sharpe portfolio even if each is mediocre alone.

Operators. The GA uses tournament selection and asset-level crossover that treats each asset's (strategy, weight) as a coupled gene, swapped together to preserve any learned strategy-weight pairings recombined from parents. The crossover operator itself (one-point, two-point, or uniform) is treated as a tunable hyperparameter (see below). Mutation has two independent forms: weight mutation (Gaussian noise followed by simplex projection) and strategy mutation (re-sample the strategy index for an asset uniformly from the variant's allowed pool). The strategy mutation deliberately breaks the existing strategy-weight pairing for that asset, providing exploration that crossover alone cannot supply once the population has converged on a local optimum; the weight at that asset is preserved on the assumption that the asset's appropriate capital share is largely independent of which strategy operates on it. Elitism preserves the top k chromosomes between generations, where k is tuned per period in the range $\{1, \ldots, 5\}$ (see below).

Hyperparameter tuning. GA hyperparameters were tuned per period via Optuna's Tree-structured Parzen Estimator (TPE) [25]. TPE is a Bayesian-style sequential sampler that, after each batch of completed trials, splits them at a quantile of the validation objective: it then fits one probability distribution over the hyperparameter values that fell in the top-performing group and another over the remaining trials, and proposes new configurations where the ratio of the two distributions is highest, concentrating the search on regions that look promising. We pair TPE with *median pruning*, which terminates a trial whose intermediate validation Sharpe falls below the median of completed trials at the same step, so that the search budget is not wasted on configurations that are clearly underperforming. Tuning ran for 50 trials with 3 GA runs per trial, on a 2-year-train / 1-year-validate split carved from each period's training years (the OOS test year is never seen during tuning). The tuned space comprises population size ($[100, 500]$, step 50), generations ($[50, 500]$, step 50), crossover rate ($[0.4, 0.9]$), crossover type ({one-point, two-point, uniform}), strategy mutation rate ($[0.05, 0.30]$), weight mutation rate ($[0.05, 0.40]$), weight mutation sigma

TABLE II
STAGE 2 WALK-FORWARD PERIODS (3YR TRAIN / 1YR OOS TEST). TUNING IS PERFORMED ON A 2YR-TRAIN / 1YR-VALIDATE SPLIT CARVED FROM EACH PERIOD'S TRAINING WINDOW; OOS TEST DATA IS NEVER SEEN DURING TRAINING OR TUNING.

Period	Training	Validate (tuning)	OOS Test
P1	2015–2017	2017	2018
P2	2017–2019	2019	2020
P3	2018–2020	2020	2021
P4	2019–2021	2021	2022
P5	2020–2022	2022	2023
P6	2021–2023	2023	2024

($[0.01, 0.15]$), and elite count ($\{1, \ldots, 5\}$). Tournament size scales with population as $\max(2, \lfloor 0.01 \cdot \text{pop} \rfloor)$ rather than being tuned independently, on the rationale that selection pressure should track population size. Across the 6 stage-2 periods, Optuna selected one-point crossover in 5 periods and uniform crossover in 1 (P2). Tuning was performed on the Full GA only, and the resulting per-period parameters were applied to all three GA variants to ensure a fair comparison.

Variants. We compare three GA variants with identical operators and tuned hyperparameters: **Full GA** (7 LLM strategies + 6 benchmarks + Cash), **LLM-Only GA** (7 LLM strategies + Cash), and **Benchmark-Only GA** (6 benchmarks + Cash). *Cash* is included as a strategy option to give the GA a "none of the above" choice for any asset on which no strategy in the pool offers a useful signal. Without it, the GA would be forced to assign one of the active strategies to every asset and would inject noise into the portfolio when no good fit exists; the Cash option lets the GA decline to trade that asset instead. Mechanically, Cash returns zero on every bar, so capital routed to a Cash-assigned asset sits idle rather than being redistributed. Cash is not a competing alpha source, and we report its share separately from LLM-generated and benchmark allocations in the revealed-preference analysis (Section V-C).

D. Evaluation Design

Stage 1 (per-strategy) evaluates each strategy independently on each asset across 4 OOS test periods spanning 2013–2023, chosen to cover a range of market regimes (post-crisis recovery, pre-COVID and COVID dislocation, stimulus bull, rate-hiking environment); strategy parameters are held fixed at the values chosen during development. We rank strategies by mean OOS Sharpe across the applicable asset-period combinations. ETH and BCH did not trade through the full 2013–2015 window, so the first test period covers 17 of the 19 assets.

Stage 2 (portfolio GA) uses 6 rolling walk-forward periods with a 3-year training window and 1-year OOS test (Table II). For each period and each variant we run the GA 50 times with different random seeds, producing a distribution of OOS Sharpe ratios per (period, variant) cell that supports formal statistical inference. The 50-run protocol is critical: a single GA run is one realisation of a stochastic optimiser, and single-point OOS Sharpes are insufficient to claim outperformance [14], [15].

Assets and costs. Both stages evaluate on 19 assets across four classes: 3 cryptocurrencies (BTC, ETH, BCH), 5 forex pairs (EUR/USD, GBP/USD, USD/JPY, USD/CHF, NZD/USD), 8 equity indices (DAX, CAC 40, STOXX 50, FTSE 100, DJ 30, S&P 500, NASDAQ, Nikkei 225), and 3 commodities (Gold, Silver, Brent Crude). Daily OHLCV bars are sourced from Pepperstone, used as the single broker throughout to ensure consistent pricing. Transaction costs are applied per asset using the 95th percentile of broker-quoted bid–ask spreads from a 4-month tick sample (October 2025–January 2026), held constant across all periods. The tick sample post-dates the test windows, but the same spread enters every variant equally, so any drift between current and historical spreads contributes a level effect rather than a comparative bias.

Performance metrics. Beyond the annualised Sharpe ratio $\bar{r}/\sigma_r \sqrt{252}$, where $\bar{r}$ and σ_r are the mean and standard deviation of daily returns, we report three further metrics. The *Sortino ratio* $\bar{r}/\sigma_d \sqrt{252}$ replaces σ_r with the downside deviation σ_d, the standard deviation of returns below zero, so it penalises only adverse volatility. The *maximum drawdown* $\text{MDD} = \max_t \big(\max_{s \le t} V_s - V_t\big) / \max_{s \le t} V_s$, where V_t is portfolio equity at time t, is the largest peak-to-trough loss observed over the evaluation window. The *Calmar ratio* is the annualised return divided by $|\text{MDD}|$ and expresses reward per unit of worst-case drawdown. Across all metrics the Sharpe and Sortino ratios are annualised with risk-free rate set to zero, avoiding the ambiguity of selecting a representative risk-free rate across periods spanning 2013–2024.

E. Statistical Methods

Stage 2 inference combines several standard statistical tests so that we do not rely on any single distributional assumption and so that we control the family-wise error rate across the multiple pairwise comparisons. We first screen each of the 18 (period, variant) Sharpe distributions with the Shapiro–Wilk test [26]; only 8 of the 18 are consistent with normality at $p \ge 0.05$, motivating a non-parametric primary test rather than a parametric t-test. The Mann-Whitney U test [27] is therefore the primary pairwise comparison between GA variants, applied per period to the (Full vs Bench) and (LLM-Only vs Bench) Sharpe distributions and to the corresponding aggregate-pooled distributions across the 6 periods; across the 12 per-period pairwise tests we apply Holm–Bonferroni step-down correction [28] to control the family-wise error rate at $\alpha = 0.05$. Cohen's d [29] reports effect size separately from sample size (large at $|d| > 0.8$), and non-parametric bootstrap 95% confidence intervals [30] on the mean Sharpe difference (5,000 resamples) corroborate the U-test result without a parametric sampling-distribution assumption. The Kruskal–Wallis test [31] serves as a three-way omnibus check on the aggregate-pooled distributions. Finally, per-period Mann-Whitney U tests are repeated on Sortino, Calmar, maximum drawdown, and total return to verify that the qualitative findings on Sharpe are not artefacts of optimising for Sharpe alone.

IV. STAGE 1: PER-STRATEGY OOS EVALUATION

Table III ranks all 8 LLM-generated strategies alongside the 5 active benchmarks (B1–B5) by mean OOS Sharpe ratio across 19 assets and 4 test periods spanning 2013–2023. Seven

TABLE III
STAGE 1 OOS PERFORMANCE RANKED BY MEAN SHARPE RATIO ACROSS 19 ASSETS AND 4 TEST PERIODS 2013–2023. $\bar{S}$: MEAN SHARPE; $\bar{R}$: MEAN ANNUALISED RETURN (%); $\overline{\text{DD}}$: MEAN MAXIMUM DRAWDOWN (%); $\bar{N}$: MEAN TRADES PER ASSET-PERIOD. †BENCHMARK.

#	Strategy	$\bar{S}$	$\bar{R}$	$\overline{\text{DD}}$	$\bar{N}$
1	Mean Rev. Bands†	+0.614	+5.4	−9.7	7
2	RSI Divergence†	+0.534	+3.1	−8.8	2
3	Spectral Entropy	+0.411	+4.7	−8.7	38
4	Hurst Mean Reversion	+0.334	+9.0	−16.8	20
5	Rough Path Sigs.	+0.310	+4.2	−14.3	19
6	Volatility Phase Trans.	+0.181	+1.7	−6.1	5
7	Info. Geometry	+0.147	+2.1	−11.8	16
8	Fisher Information	+0.095	+0.0	−10.8	8
9	Rate Distortion	+0.021	+1.1	−13.0	8
10	MA Crossover†	−0.022	+1.9	−15.9	3
11	Bollinger Breakout†	−0.045	+7.4	−12.1	2
12	Donchian Breakout†	−0.120	+4.4	−13.6	3
13	Wasserstein Distance	−0.200	−0.7	−12.7	10

of 8 LLM-generated strategies achieve positive mean Sharpe ratios, compared with only 2 of 5 active benchmarks, and 3 of the top 5 strategies overall are LLM-generated.

The top 3 LLM-generated strategies (Spectral Entropy, Hurst Mean Reversion, Rough Path Signatures) rank 3rd–5th overall, placing above 3 of the 5 benchmarks. Mean Reversion Bands and RSI Divergence rank 1st and 2nd among per-strategy results because these are well-tuned mean-reverting indicators on a sample with several mean-reverting periods (2013–2015 post-crisis, 2019–2020 pre-COVID). LLM-generated strategies span return-risk profiles: Hurst Mean Reversion pairs the highest mean return (+9.0%) with the deepest drawdown (−16.8%), while Spectral Entropy pairs a +4.7% return with the shallowest drawdown of the top five (−8.7%). Volatility Phase Transition's positive mean ($\bar{S} = +0.181$) masks a negative median (−0.04), indicating that its headline ranking is driven by strong performance on a subset of asset–period combinations rather than broad effectiveness; it is well-suited to a small set of regimes rather than as a default strategy.

Asset-class effect. Strategy effectiveness varies strongly by asset class, and the sign of the lag-1 return autocorrelation ρ_1 largely explains the pattern: positive ρ_1 rewards trend-following signals, negative ρ_1 rewards mean-reversion, and near-zero ρ_1 leaves no exploitable serial dependence for any signal-based timing. Cryptocurrency (annualised volatility 85.9%, $\rho_1 = +0.04$) and equity indices ($\rho_1 = -0.09$) are the most productive classes, with 12 of 13 and 11 of 13 strategies positive respectively, the former rewarding trend-following and the latter mean-reverting signals. Forex (lowest volatility 9.4%, near-zero $\rho_1 = -0.00$) is most challenging: only 4 of 13 strategies achieve positive mean Sharpe, manifesting as flat performance (−0.3% mean return, −4.4% shallow drawdown) rather than large losses, consistent with lack of exploitable serial dependence rather than elevated risk. On its best-suited asset class, several LLM-generated strategies are competitive: Hurst Mean Reversion outperforms all benchmarks on cryptocurrency (+0.88), and Rough Path Signatures (+0.65) plus Spectral Entropy (+0.56) outperform 3 of 5 benchmarks on equity indices.

Top individual results. Beyond averages, Table IV lists the highest Sharpe ratios achieved on any single (asset, period) cell. Eight of the top 12 results are LLM-generated, and the LLM entries cluster on equity indices in 2021 and 2023, exactly the cells where negative-autocorrelation structure aligns with these strategies' assumptions. The only benchmark to crack the top tier is Mean Reversion Bands, which dominates on low-volatility forex (EUR/USD and USD/JPY) where its small but consistent return stream produces high Sharpe at very shallow drawdowns. This regime separation, LLM strategies on mean-reverting equity indices and Mean Reversion Bands on low-volatility forex, reinforces the asset-class effect above: top performance reflects alignment between strategy assumption and asset statistical structure rather than uniform superiority.

TABLE IV
TOP INDIVIDUAL STAGE 1 OOS RESULTS ACROSS ALL STRATEGIES (HIGHEST SHARPE RATIOS ON A SINGLE ASSET–PERIOD CELL). †BENCHMARK.

Strategy	Asset	Period	S	DD%
Mean Rev. Bands†	EUR/USD	19–20	2.96	−0.6
Mean Rev. Bands†	USD/JPY	2021	2.86	−0.2
Spectral Entropy	S&P 500	2023	2.38	−2.2
Rough Path Sigs.	STOXX 50	2021	2.37	−3.0
Fisher Information	DJ 30	2023	2.35	−0.7
Spectral Entropy	Nikkei 225	2023	2.27	−2.3
Info. Geometry	Nikkei 225	2023	2.26	−3.5
Mean Rev. Bands†	STOXX 50	2021	2.22	−2.3
Rough Path Sigs.	CAC 40	2021	2.20	−5.3
Info. Geometry	S&P 500	2023	2.17	−3.4
Rate Distortion	DJ 30	2023	2.14	−3.8
Fisher Information	Nikkei 225	19–20	2.12	−1.9

V. STAGE 2: PORTFOLIO GA RESULTS

A. Per-Period Sharpe and Effect Sizes

Table V reports per-period Sharpe distributions for each variant, mean differences, Cohen's d, bootstrap 95% confidence intervals, and Mann-Whitney U adjusted p-values (Holm-Bonferroni over the 12 pairwise tests).

The Full GA significantly beats Benchmark-Only in 3 of 6 periods (2018, 2022, 2023) with effect sizes ranging from large to very large ($d = +2.24$ to $+7.22$). Benchmark-Only beats Full in 2 periods (2020 COVID V-recovery, 2024 trending bull) with moderate-to-large effects, and 2021 is a tie. The LLM-Only variant shows a similar but slightly weaker pattern, with the same set of 3 wins, 2 losses, 1 tie, but with a substantial loss against Benchmark-Only in 2024 ($d = -7.34$) where Buy & Hold-led benchmarks dominate the AI-driven trending year.

B. Aggregate and Multi-Metric Robustness

Pooling all 50 runs across the 6 periods (300 runs per variant), the Full GA achieves significantly higher mean Sharpe than the Benchmark-Only GA ($\Delta = +0.21$, Mann-Whitney U $p = 0.042$). The LLM-Only GA's aggregate advantage is marginal ($\Delta = +0.16$, $p = 0.098$). The 3-group Kruskal-Wallis test is not significant at $\alpha = 0.05$ ($H = 4.56$, $p = 0.10$), reflecting limited power with only 6 period-level clusters; the per-period results retain full power with $n = 50$ per cell.

The advantage is not confined to Sharpe (Table VI). On the aggregate-pooled distributions, the Full GA's improvement over

TABLE V

STAGE 2 PER-PERIOD OOS SHARPE STATISTICS (50 GA RUNS PER CELL). SHARPE VALUES ARE MEAN±STD. Δ: MEAN DIFFERENCE VS BENCHMARK-ONLY. d: COHEN'S d (LARGE AT $|d| > 0.8$). ADJ. p: HOLM-BONFERRONI ADJUSTED MANN-WHITNEY U p-VALUE OVER THE 12 PAIRWISE TESTS; "NS" DENOTES $p \geq 0.05$.

		Sharpe (mean±std)			Full vs Bench			LLM-Only vs Bench		
P	**Year**	**Full**	**LLM**	**Bench**	Δ	d	adj. p	Δ	d	adj. p
P1	2018	−0.26±0.36	−0.67±0.21	−0.99±0.28	+0.73	+2.24	<0.001	+0.32	+1.26	<0.001
P2	2020	+0.52±0.21	+1.14±0.17	+1.28±0.08	−0.75	−4.67	<0.001	−0.13	−1.00	<0.001
P3	2021	+0.24±0.19	+0.34±0.17	+0.32±0.29	−0.08	−0.31	0.24 ns	+0.02	+0.09	0.76 ns
P4	2022	−0.85±0.19	−0.82±0.21	−1.41±0.17	+0.56	+3.05	<0.001	+0.60	+3.10	<0.001
P5	2023	+1.72±0.12	+1.82±0.13	+0.70±0.16	+1.02	+7.22	<0.001	+1.12	+7.50	<0.001
P6	2024	+1.34±0.25	+0.61±0.13	+1.55±0.12	−0.21	−1.06	<0.001	−0.94	−7.34	<0.001

TABLE VI

STAGE 2 MULTI-METRIC FULL VS BENCHMARK-ONLY COMPARISON. AGGREGATE COLUMNS POOL ALL 50 RUNS ACROSS 6 PERIODS (300 RUNS PER VARIANT); THE AGGREGATE p IS MANN-WHITNEY U ON THE POOLED DISTRIBUTION AND IS THE PRIMARY INFERENCE FOR EACH METRIC. "WINS" COUNTS PERIODS WHERE THE FULL GA MEAN EXCEEDS THE BENCHMARK-ONLY GA (LESS NEGATIVE FOR MAX DRAWDOWN).

	Aggregate			**Per-period**	
Metric	**Full**	**Bench**	p	Δ	**Wins**
Sharpe	+0.45	+0.24	0.042	+0.21	3/6
Sortino	+0.54	+0.28	0.053	+0.26	3/6
Calmar	+0.72	+0.49	0.005	+0.22	4/6
Max DD	−4.1%	−4.7%	<0.001	+0.7%	5/6
Total Ret	−0.2%	−0.2%	0.12	+0.0%	3/6

the Benchmark-Only GA is significant on Sharpe ($p = 0.042$), Calmar ($p = 0.005$), and maximum drawdown ($p < 0.001$), and is on the boundary on Sortino ($p = 0.053$). On total return there is no aggregate difference. The per-period "Wins" column shows the same direction of effect appears in multiple periods rather than being driven by a single year: shallower maximum drawdown in 5 of 6 (the strongest signal across metrics) and better Calmar in 4 of 6. The advantage is concentrated on risk-adjusted and risk-controlled metrics rather than raw return, suggesting the LLM-augmented portfolio's contribution lies in dampening drawdowns rather than generating incremental upside.

C. Revealed Preference: GA Capital Allocation

When the Full GA is given a free choice across LLM-generated and benchmark strategies, the evolved best chromosomes allocate, on average across the 6 periods, 50.9% of portfolio weight to LLM-generated strategies, 43.0% to benchmarks, and 6.0% to Cash. The most-favoured single strategy is Volatility Phase Transition at 22.7% of total weight, followed by RSI Divergence (19.2%, benchmark), Mean Reversion Bands (17.7%, benchmark), and Spectral Entropy (10.5%, LLM-generated). Two of the top 4 are LLM-generated. The chromosome-level allocation reflects what worked on the training window, not what works on the unseen OOS test year, and the two can diverge sharply when the test period departs from the recent training regime. The clearest example is 2024: the Full GA allocates 76% of capital to LLM-generated strategies after training on 2021–2023 (a mix of bear and AI-driven recovery), but 2024's strong directional bull regime favours passive Buy & Hold and the Benchmark-Only variant outperforms. Period-level allocation mix is benchmark-leaning in 2018 and 2022 (64% and 68%) and LLM-leaning in 2020, 2021, 2023, and 2024 (58%, 57%, 47%, 76%); the 2024 case underscores that an absence of dynamic regime adaptation is a clear limitation of the current setup.

VI. DISCUSSION AND CONCLUSION

Across the per-strategy evaluation on the 2013–2023 OOS periods (Stage 1), which cover a range of market regimes, 7 of 8 of the LLM-generated strategies achieve positive mean Sharpe and the top 3 of them rank 3rd–5th overall, ahead of 3 of the 5 active benchmarks. At the portfolio level across 2018–2024 (Stage 2), the Full GA significantly outperforms the Benchmark-Only GA on aggregate Sharpe ($\Delta = +0.21$, $p = 0.042$), Calmar ($p = 0.005$), and maximum drawdown ($p < 0.001$). The LLM-Only GA, which has no access to conventional benchmarks, performs comparably to but not significantly better than the Benchmark-Only GA, suggesting that the LLM-generated strategies hold their own as a standalone family but realise their main incremental gain when combined with the conventional benchmarks in the Full pool.

The benefit is regime-dependent. The Full GA wins in volatility-driven and transition regimes (2018 trade-tension volatility, 2022 rate-hiking bear, 2023 AI-driven mixed regime) and loses in strong-momentum regimes (2020 COVID V-recovery, 2024 trending bull) where Buy & Hold-led benchmarks dominate. This is consistent with the adaptive-markets hypothesis [32]: no single approach dominates across all conditions, and these LLM-generated strategies appear to capture signals tied to volatility regime change and statistical-property shifts rather than persistent trend exploitation. Where these signal sources align with prevailing market dynamics they add value; elsewhere, simple momentum or passive holding wins.

The nature of the value added is more visible on the risk-controlled metrics than on raw return. On the aggregate-pooled distributions, the Full GA's improvement over Benchmark-Only is significant on Calmar and maximum drawdown ($p = 0.005$ and $p < 0.001$ respectively), with the same directional advantage appearing in 5 of 6 periods on max drawdown and 4 of 6 on Calmar, while aggregate total return is statistically indistinguishable between the two variants. The LLM-augmented portfolio's contribution is to dampen drawdowns rather than to generate incremental upside, a

defensive characteristic that is practically useful for portfolio construction.

Limitations. First, 8 strategies represent a small sample from a vast space of mathematical domains, and the LLM's generation process is non-deterministic, so specific strategies depend on prompt design. Second, the GA optimises Sharpe; alternative objective functions (Sortino, Calmar) might yield different allocations, although the multi-metric robustness results suggest the qualitative findings are stable. Third, strategy-internal parameters were held fixed across all OOS evaluation rather than re-tuned per walk-forward window. The grid-search robustness check (Appendix A) indicates the LLM-chosen values are close to development-set optimal, but per-window re-tuning could adapt parameters to changing regimes and remains a deployment variant for future work. Fourth, the Stage 2 horizon (2018–2024) is short for definitive claims about regime-specific behaviour; longer rolling horizons would strengthen inference.

Future directions. Several avenues merit investigation. *Regime-conditioned strategy selection* could build on the observation that LLM-generated strategies do best in volatility-transition regimes by adding a regime-detection layer that shifts weight toward LLM strategies when market statistics indicate a transition is under way; the absence of any such adaptation is a clear limitation, most visible in the 2024 result. *Targeted deployment* based on asset-level statistical properties (volatility, autocorrelation, trend strength) could pre-screen which (asset, strategy) pairings to even consider, since some LLM strategies are sharply more effective on specific asset classes (e.g. Hurst Mean Reversion on cryptocurrency). *Larger LLM strategy pools* generated through repeated prompt-conditioned sessions could expand coverage across mathematical domains, supplying the GA with a richer search space.

Conclusion. Across 6 walk-forward OOS periods with multiple statistical tests, an LLM-augmented portfolio achieves significantly better risk-adjusted performance and shallower drawdowns than a portfolio restricted to conventional benchmarks, while the LLM-generated strategies as a standalone set perform on par with well-tuned indicator strategies. The benefit is concentrated on risk-controlled metrics rather than raw return, and depends on market regime: the LLM-generated strategies are most useful in volatility-driven and transition regimes, and complement rather than replace the established indicator family.

APPENDIX A
PARAMETER SENSITIVITY ROBUSTNESS CHECK

To assess whether the LLM's parameter choices are near-optimal, we ran a per-strategy grid search over each strategy's 2–3 principal parameters, defined as a tight neighbourhood around the LLM-chosen values. Each grid point was scored by mean Sharpe across all assets on the development set (universal-parameter regime), and the best-performing configuration on the development set was then evaluated on the same 4 OOS periods used for the LLM-chosen values (Table VII).

TABLE VII
PER-STRATEGY GRID SEARCH VS. LLM-CHOSEN PARAMETERS. $\bar{S}_{\text{LLM}}$ AND $\bar{S}_{\text{GRID}}$ ARE MEAN SHARPE UNDER THE LLM'S PARAMETERS AND THE BEST DEVELOPMENT-SET PARAMETERS RESPECTIVELY; Δ IS THE UPLIFT.

	Dev			OOS		
Strategy	$\bar{S}_{\text{LLM}}$	$\bar{S}_{\text{grid}}$	Δ	$\bar{S}_{\text{LLM}}$	$\bar{S}_{\text{grid}}$	Δ
Spectral Entropy	$+0.03$	$+0.26$	$+0.22$	$+0.41$	$+0.26$	-0.15
Rough Path Sigs.	$+0.20$	$+0.37$	$+0.17$	$+0.31$	$+0.38$	$+0.07$
Info. Geometry	$+0.29$	$+0.35$	$+0.06$	$+0.15$	$+0.26$	$+0.12$
Wasserstein Distance	$+0.15$	$+0.33$	$+0.18$	-0.20	$+0.25$	$+0.45$
Rate Distortion	$+0.33$	$+0.43$	$+0.09$	$+0.02$	$+0.24$	$+0.22$
Fisher Information	$+0.17$	$+0.27$	$+0.10$	$+0.10$	$+0.26$	$+0.17$
Volatility Phase Trans.	$+0.29$	$+0.30$	$+0.02$	$+0.18$	$+0.15$	-0.04
Hurst Mean Rev.	$+0.38$	$+0.40$	$+0.02$	$+0.33$	$+0.34$	$+0.01$
Mean	$+0.23$	$+0.34$	$+0.11$	$+0.16$	$+0.27$	$+0.11$

Of 8 strategies, 6 show positive OOS uplift under the development-tuned parameters, but the mean uplift of $+0.11$ Sharpe is modest, comparable to one standard error of a Sharpe ratio estimated over a single OOS period. Spectral Entropy is the one strategy where dev-tuning materially hurts OOS performance ($\Delta = -0.15$), likely reflecting overfitting of the spectral threshold to the dev-period volatility regime. We conclude that the LLM-selected parameters are close to grid-optimal on the development set, and OOS performance is not bottlenecked by parameter choice within each strategy's chosen structure.

REFERENCES

[1] A. Lopez-Lira and Y. Tang, "Can ChatGPT forecast stock price movements? Return predictability and large language models," *SSRN Working Paper*, 2023.

[2] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "BloombergGPT: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[3] Z. Kou, H. Yu, J. Luo, J. Peng, X. Li, C. Liu, J. Dai, L. Chen, S. Han, and Y. Guo, "Automate strategy finding with LLM in quant investment," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 18 517–18 533.

[4] S. Wang, H. Yuan, L. Zhou, L. Ni, H.-Y. Shum, and J. Guo, "Alpha-GPT: Human-AI interactive alpha mining for quantitative investment," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2025, pp. 196–206.

[5] H. Yang, X.-Y. Liu, and C. D. Wang, "FinGPT: Open-source financial large language models," in *International Symposium on Large Language Models for Financial Services (FinLLM) at IJCAI 2023*, 2023.

[6] B. Romera-Paredes, M. Barekatain, A. Novikov, M. Balog, M. P. Kumar, E. Dupont, F. J. R. Ruiz, J. S. Ellenberg, P. Wang, O. Fawzi, P. Kohli, and A. Fawzi, "Mathematical discoveries from program search with large language models," *Nature*, vol. 625, no. 7995, pp. 468–475, 2024.

[7] F. Allen and R. Karjalainen, "Using genetic algorithms to find technical trading rules," *Journal of Financial Economics*, vol. 51, no. 2, pp. 245–271, 1999.

[8] C. Neely, P. Weller, and R. Dittmar, "Is technical analysis in the foreign exchange market profitable? A genetic programming approach," *Journal of Financial and Quantitative Analysis*, vol. 32, no. 4, pp. 405–426, 1997.

[9] D. Maringer, *Portfolio Management with Heuristic Optimization*. Springer, 2005.

[10] T.-J. Chang, N. Meade, J. E. Beasley, and Y. M. Sharaiha, "Heuristics for cardinality constrained portfolio optimisation," *Computers & Operations Research*, vol. 27, no. 13, pp. 1271–1302, 2000.

[11] T. Schreiber, "Measuring information transfer," *Physical Review Letters*, vol. 85, no. 2, pp. 461–464, 2000.

[12] C. Bandt and B. Pompe, "Permutation entropy: A natural complexity measure for time series," *Physical Review Letters*, vol. 88, no. 17, p. 174102, 2002.

[13] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proceedings of the National Academy of Sciences*, vol. 88, no. 6, pp. 2297–2301, 1991.
[14] D. H. Bailey, J. M. Borwein, M. López de Prado, and Q. J. Zhu, "Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance," *Notices of the American Mathematical Society*, vol. 61, no. 5, pp. 458–471, 2014.
[15] M. López de Prado, *Advances in Financial Machine Learning*. John Wiley & Sons, 2018.
[16] O. Polakow, "vectorbt: Lightning-fast backtesting engine to find your trading edge," 2021. [Online]. Available: https://github.com/polakowo/vectorbt
[17] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
[18] T. J. Lyons, M. Caruana, and T. Lévy, *Differential Equations Driven by Rough Paths*. Springer, 2007.
[19] I. Chevyrev and A. Kormilitzin, "A primer on the signature method in machine learning," in *Signature Methods in Finance: An Introduction with Computational Applications*. Springer, 2025, pp. 3–64.
[20] S.-i. Amari, *Information Geometry and Its Applications*. Springer, 2016.
[21] C. Villani, *Optimal Transport: Old and New*. Springer, 2009.
[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.
[23] H. E. Hurst, "Long-term storage capacity of reservoirs," *Transactions of the American Society of Civil Engineers*, vol. 116, pp. 770–799, 1951.
[24] J. J. Murphy, *Technical Analysis of the Financial Markets*. New York Institute of Finance, 1999.
[25] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.
[26] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965.
[27] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.
[28] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.
[29] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, 1988.
[30] B. Efron, "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.
[31] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.
[32] A. W. Lo, "The adaptive markets hypothesis: Market efficiency from an evolutionary perspective," *The Journal of Portfolio Management*, vol. 30, no. 5, pp. 15–29, 2004.