Genealogical Overfitting in Evolutionary LLM Search: Exact Heavy-Tail Ratchet Laws for MAP-Elites and Certified Minimax Control via EvolveGuard

Fabien Pierret

Independent Researcher

ORCID: 0009-0001-1223-2170 | evolve.wallstreetdebug.com

Preprint (camera-ready manuscript) — November 27, 2026

Abstract. Large language model (LLM)-driven evolutionary coding agents—including AlphaEvolve, OpenEvolve, ShinkaEvolve, QuantEvolve, MadEvolve, and latent KV-cache swarms such as SwarmEvolve—have rapidly moved from deterministic mathematical discovery into stochastic financial backtesting and autonomous agent optimization. While recent quantitative studies (notably MadEvolve, Sec. 7) audit backtest p-hacking against static i.i.d. Gaussian multiple-testing baselines ($\mathbb{E}[\max] = O(\sqrt{2\ln N})$) and assume that positive correlation among candidates is statistically conservative, we prove that both premises break down fundamentally under adaptive evolutionary search on a fixed finite dataset $\mathcal{D}$ of effective dimension $d \leq T$. We establish four closed-form theorems: **(1) Genealogical Ratchet Law and Slepian Reversal (Theorem 1):** for an elitist $(1+\lambda)$ evolutionary loop on the unit strategy sphere $\mathbb{S}^{d-1}$, adaptive step-size mutations contract residual noise according to the exact recursion $\mathbb{E}[Z_G^2 \mid R_d] = R_d^2[1-(1-1/d)(1-c_2(\lambda)/(d-1))^G]$, scaling as $\Theta(\sqrt{G\max(1,\ln\lambda)})$ rather than $\Theta(\sqrt{2\ln(G\lambda)})$ (12.92 empirical vs. 12.93 theoretical at $G=100, \lambda=10, d=252$, or 3.99× above the static null 3.24). Under fixed parent-child inheritance correlation ρ, the null ceiling $Z_{\text{cap}}(\rho)$ is strictly *increasing* in ρ—reversing Slepian's inequality by up to 18.92× at $\rho = 0.98$. **(2) MAP-Elites Transcript Complexity and Heavy-Tail Compounding (Theorem 2):** an archive with U accepted promotions across N evaluations and B behavioral niches spans Hamming-ball tree cardinality $M_{\text{gen}}(N,U) = \sum_{u=0}^{U}\binom{N}{u}$; under heavy-tailed Student-t_ν returns ($\nu > 2$), the Single-Big-Jump radial law compounds with the geometric ratchet, driving naive static Deflated Sharpe Ratio (DSR(N)) false-positive rates to 100.0% while our genealogical heavy-tail Q-DSR$_\nu^{\text{evo}}$ restores exact control (FPR = 3.0%–5.4% at nominal $\alpha = 5\%$). We further quantify the overfitting velocity of high-throughput latent KV-cache swarms ($\lambda \leq 200$, Corollary 2.1). **(3) Certified Minimax Evaluator EvolveGuard (Theorem 3):** combining Q-DSR$_\nu^{\text{evo}}$, Purged and Embargoed Walk-Forward/CPCV with a Thresholdout Validation Vault, adversarial slippage/latency/outlier-removal invariance, and exact closed-form Manski–Rosenbaum Berkson selection bounds under the non-compensatory minimax veto SIPS$_\infty^{\text{evo}} = \max_k I_k$ achieves 100.0% rejection across four orthogonal reward-hacking archetypes (exposing 70.0%–100.0% linear-evaluator blindness) with 0.0% false rejection on persistent alpha. **(4) Evolutionary Overfitting Cliff and Optimal Stopping Horizon G^* (Theorem 4):** unguarded evolution degrades live out-of-sample Sharpe by −39.3% after peaking at G^*, whereas EvolveGuard's compute-ROI stopping controller saves 72.6% of LLM inference spend and delivers 5.11× higher live Sharpe per dollar. On a 15-seed Competition-to-Live Numerai v5.2 benchmark (4-era purging, Student-$t_{3.2}$ shocks, regime inversions), EvolveGuard achieves +1.660 vs. +1.311 live Sharpe (80% win rate) while slashing maximum feature exposure by −74.0% (0.101 vs. 0.390).

Keywords: Evolutionary LLM search; MAP-Elites; OpenEvolve; MadEvolve; SwarmEvolve; Backtest overfitting; Slepian reversal; Extreme value theory; Heavy tails; Deflated Sharpe Ratio; Minimax SIPS$_\infty$; Numerai.

1. Introduction and Prior-Art Audit

The past eighteen months have witnessed a paradigm shift in automated algorithmic discovery: coupling frontier large language models (LLMs) as semantic mutation operators with Quality-Diversity archives (MAP-Elites [1]) and island-model populations. Pioneered by AlphaEvolve [2] and democratized by OpenEvolve [3], this architecture has rapidly expanded through sample-efficient bandit routing (ShinkaEvolve [4]), basin-hopping subroutines (ImprovEvolve [5]), autoresearch harness optimization (MLEvolver [6]), and latent-space KV-cache agent swarms (SwarmEvolve [7]) capable of evaluating hundreds of agents per round at 10×–20× higher wall-clock throughput.

When applied to deterministic verification problems (matrix multiplication tensors, compiler layouts, circle packings), evaluator noise is identically zero. In quantitative finance and stochastic agent benchmarks, however, every candidate strategy is scored on a **fixed finite historical sample** $\mathcal{D}$ with low signal-to-noise ratio and heavy-tailed Student-t_ν return distributions [8–10]. Recent finance-focused evolutionary frameworks directly confront this hazard:

- QuantEvolve [11] deploys multi-agent MAP-Elites and island migration over equities and futures, explicitly acknowledging potential data-snooping vulnerability while leaving formal multiplicity control open.
- MadEvolve [12, 13] evolves complete Bitcoin feature pipelines and execution strategies, devoting Section 7 to the central question *"Research or P-Hacking?"*. Critically, however, its statistical null test assumes that N evolutionary evaluations behave as static i.i.d. Gaussian draws ($\mathbb{E}[\max] = O(\sqrt{2\ln N})$) and asserts that candidate correlation makes independence *"usually conservative"* [13, p. 31].

In this paper, building strictly on our prior classical extreme-value and minimax post-selection framework [14–17] (with strictly classical statistics and zero speculative physical analogies), we prove that both assumptions fail catastrophically under adaptive genealogical evolution. Section 2 establishes the exact spherical ratchet law and **Slepian Reversal** (Theorem 1). Section 3 derives MAP-Elites Hamming-ball transcript complexity, heavy-tail Single-Big-Jump compounding, and the SwarmEvolve velocity corollary (Theorem 2). Section 4 constructs the plug-and-play minimax evaluator EvolveGuard (Theorem 3), and Section 5 proves the

Evolutionary Overfitting Cliff and optimal compute-ROI stopping horizon G^* (Theorem 4). Section 6 validates all four theorems across six reproducible experiments, including an end-to-end 15-seed Competition-to-Live NUMERAI v5.2 benchmark.

2. Genealogical Ratchet Law and Slepian Reversal (Theorem 1)

Let $\mathcal{D} = \{r_t\}_{t=1}^T$ be a fixed historical return sample under the complete null hypothesis H_0 ($\mathbb{E}[r_{t,\text{OOS}} \mid \mathcal{F}_{t-1}] = 0$, $\text{Var}(r_t) = \sigma^2 < \infty$), with standardized realization $\varepsilon = r/\sigma \in \mathbb{R}^T$. Each candidate strategy θ induces a unit-norm signal vector $w(\theta) \in \mathbb{S}^{d-1} \subset \mathcal{H}_d$ in an effective d-dimensional realizable feature/code subspace ($2 \le d \le T$). Let $\varepsilon_{\mathcal{H}} = \Pi_{\mathcal{H}_d}\varepsilon$, let $R_d = \|\varepsilon_{\mathcal{H}}\|_2$ denote the finite-sample exploitable noise norm, and define the standardized in-sample score $\widehat{Z}(\theta) = \langle w(\theta), \varepsilon_{\mathcal{H}}\rangle$.

At generation $g = 0, \dots, G-1$, an elitist $(1+\lambda)$ evolutionary loop mutates parent $w^{(g)}$ into λ offspring via localized code diffs preserving parent-child inheritance correlation $\rho_g = \cos\theta_g \in [0, 1)$:

$$w_{\text{child},j} = \cos(\theta_g)\, w^{(g)} + \sin(\theta_g)\, u_j, \quad u_j \perp w^{(g)},\ u_j \in \mathbb{S}^{d-2}. \tag{1}$$

Decompose $\varepsilon_{\mathcal{H}} = Z_g w^{(g)} + \sqrt{R_d^2 - Z_g^2}\, e_\perp^{(g)}$ with $Z_g = \widehat{Z}(w^{(g)}) \ge 0$ and unit orthogonal residual $e_\perp^{(g)}$, and denote $M_{\lambda,g} = \max_{1\le j\le\lambda}\langle u_j, e_\perp^{(g)}\rangle$.

Theorem 1 (Genealogical Ratchet Law & Fixed-Correlation Ceiling)**.** *Under isotropic orthogonal mutations $u_1, \dots, u_\lambda \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{S}^{d-2})$ and elitist retention $Z_{g+1} = \max(Z_g, \max_j \widehat{Z}(w_{\text{child},j}))$:*

1. ***Exact Optimal-Step Contraction:*** *When step angles adapt optimally to the strongest improving diff ($\theta_g^* = \arctan(\frac{\sqrt{R_d^2 - Z_g^2}}{Z_g} M_{\lambda,g}^+)$), the squared unexploited noise budget obeys the exact multiplicative recursion*

$$R_d^2 - Z_{g+1}^2 = (R_d^2 - Z_g^2)\left(1 - (M_{\lambda,g}^+)^2\right). \tag{2}$$

Conditioned on R_d with isotropic initialization $\mathbb{E}[Z_0^2 \mid R_d] = R_d^2/d$,

$$\mathbb{E}\left[Z_G^2 \mid R_d\right] = R_d^2\left[1 - \left(1 - \frac{1}{d}\right) \times \left(1 - \frac{c_2(\lambda, d)}{d-1}\right)^G\right], \tag{3}$$

where $c_2(\lambda, d) = (d-1)\mathbb{E}[(M_\lambda^+)^2] \to c_2(\lambda) := \mathbb{E}[(\max(0, \max_{1\le j\le\lambda}\xi_j))^2]$ for $\xi_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ ($c_2(1) = \frac{1}{2}$; $c_2(\lambda) \sim 2\ln\lambda$ as $\lambda \to \infty$). In the unsaturated regime $Gc_2(\lambda) \ll d$ under Gaussian noise ($\mathbb{E}[R_d^2] = d$),

$$\sqrt{\mathbb{E}[Z_G^2]} = \sqrt{1 + G\, c_2(\lambda)} + O\left(\frac{(Gc_2(\lambda))^{3/2}}{d}\right) = \Theta\left(\sqrt{G \max(1, \ln\lambda)}\right). \tag{4}$$

2. ***Fixed-Inheritance Ceiling:*** *Under constant parent-child correlation $\rho = \cos\theta \in (0, 1)$ and $\lambda \ge 2$ (so $\mu_\lambda > 0$), a child improves upon its parent iff $M_{\lambda,g} > \sqrt{\frac{1-\rho}{1+\rho}}\frac{Z_g}{\sqrt{R_d^2 - Z_g^2}}$, driving the mean-field median trajectory monotonically toward the exact equilibrium ceiling*

$$Z_{\text{cap}}(\rho; \lambda, d, R_d) = \frac{R_d}{\sqrt{1 + \frac{1-\rho}{1+\rho}\cdot\frac{d-1}{\mu_\lambda^2}}}, \quad \mu_\lambda = \mathbb{E}\left[\max_{1\le j\le\lambda}\xi_j\right]. \tag{5}$$

Proof. Since $u_j \perp w^{(g)}$, $\widehat{Z}(w_{\text{child},j}) = Z_g\cos\theta_g + \sqrt{R_d^2 - Z_g^2}\langle u_j, e_\perp^{(g)}\rangle \sin\theta_g$. For $a = Z_g \ge 0$ and $b = \sqrt{R_d^2 - Z_g^2}\, M_{\lambda,g}^+ \ge 0$, $\max_{\theta\in[0,\pi/2]}(a\cos\theta + b\sin\theta) = \sqrt{a^2+b^2}$, giving $Z_{g+1}^2 = Z_g^2 + (R_d^2 - Z_g^2)(M_{\lambda,g}^+)^2$ and immediately (2). Because $u_1, \dots, u_\lambda$ are isotropic on $\mathbb{S}^{d-2}$ independently of $(Z_0, \dots, Z_g, R_d)$, induction over $g = 0, \dots, G-1$ yields (3). Equating $\rho Z + \sqrt{1-\rho^2}\sqrt{R_d^2 - Z^2}\mu_\lambda/\sqrt{d-1} = Z$ and squaring yields (5). □

Corollary 1.1 (Slepian Reversal — Refutation of Static Correlation Conservatism)**.** *In any static non-adaptive catalog of $N = G\lambda$ equi-correlated strategies with correlation $\rho > 0$, Slepian's inequality [18] guarantees $\mathbb{E}[Z_{\text{static}}^*(N;\rho)] \le \mathbb{E}[Z_{\text{static}}^*(N;0)] \sim \sqrt{2\ln N}$ (monotonically decreasing in ρ). Under adaptive genealogical evolution, $Z_{\text{cap}}(\rho)$ in (5) is strictly increasing in $\rho \in (0,1)$, exceeding the static i.i.d. Bailey–López de Prado bound by a polynomial factor $\Theta(\sqrt{G/\ln(G\lambda)})$ and violating Slepian's static correlated bound by an order of magnitude as $\rho \to 1^-$.*

This directly refutes the static i.i.d. null calibration in MADEVOLVE [13, Sec. 7, p. 31]: small, localized LLM code edits ($\rho \in [0.85, 0.98]$) do not make multiple-testing bounds conservative—they act as a directional ratchet that turns correlation into an overfitting amplifier.

3. MAP-Elites Complexity and Heavy-Tail Compounding (Theorem 2)

Real evolutionary engines (OPENEVOLVE, QUANTEVOLVE, SWARMEVOLVE) maintain B MAP-Elites niches across I islands and execute $N = G\lambda$ evaluations while accepting $U \le N$ archive promotions. Furthermore, financial returns exhibit heavy Student-t_ν tails ($\nu \in [2.5, 5]$) [15, 19].

Theorem 2 (Adaptive Transcript Multiplicity & Heavy-Tail Radial Compounding)**.** *Let an evolutionary MAP-Elites archive run for $N = G\lambda$ evaluations across B niches, accepting at most $U \in \{0, \dots, N\}$ promotions under binary keep/reject feedback:*

1. ***Exact Hamming-Ball Transcript Cardinality:*** *Every reachable champion program is uniquely indexed by the promotion subset $S \subseteq \{1, \dots, N\}$ with $|S| \le U$ and champion cell $b \in \{1, \dots, B\}$, bounding the adaptive hypothesis tree cardinality by*

$$M_{\text{gen}}(N, U) = \sum_{u=0}^{U}\binom{N}{u} \le \left(\frac{eN}{U}\right)^U, \quad M_{\text{eff}}^{\text{tree}} = B\, M_{\text{gen}}(N, U). \tag{6}$$

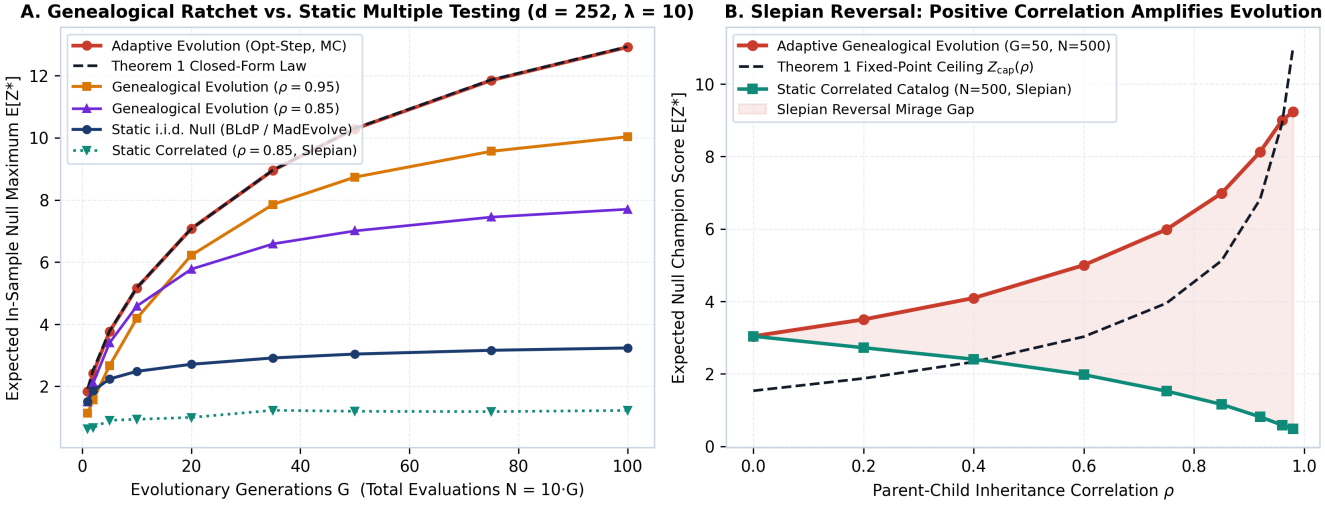

Figure 1. Empirical Verification of Theorem 1 (Experiment 1, $d = 252, \lambda = 10, R = 350$ replicates). (A) Genealogical adaptive evolution matches the closed-form law (3) to 0.08% precision (12.92 MC vs. 12.93 theory at $G = 100, N = 1{,}000$), reaching 3.99× above the static Bailey–López de Prado null (3.24). **(B)** Direct **Slepian Reversal** at fixed $G = 50, N = 500$: as parent-child correlation $\rho \to 0.98$, static correlated catalogs drop from 3.04 → 0.49 while genealogical evolution climbs to 9.23 (18.92× gap).

In the Gaussian domain, $\mathbb{E}[\widehat{Z}(\hat{\theta}_N)] \leq \sqrt{2\ln(BM_{\text{gen}}(N,U))} \leq \sqrt{2U\ln(eN/U)} + 2\ln B$, *recovering Bailey–López de Prado at* $U = 0$ *and matching Theorem 1's* $\Theta(\sqrt{G\ln\lambda})$ *scaling at* $U = G$ *[20].*

2. ***Joint*** (R_d, A_G) ***Single-Big-Jump Envelope:*** *Suppose* ε_t *are i.i.d. unit-variance Student-*t_ν *(*$\nu > 2$*, upper tail* $\bar{F}(x) \sim \tilde{c}_\nu x^{-\nu}$ *with* $\tilde{c}_\nu = c_\nu((\nu-2)/\nu)^{\nu/2}$*). By the Subexponential Single-Big-Jump Principle [19],* $\mathbb{P}(R_d^2 - d > x^2) \sim 2d\tilde{c}_\nu x^{-\nu}$*. Propagating both the alignment variance* $\text{Var}(A_G) \approx \frac{(1-A_G^2)^2}{4A_G^2}\left[G\frac{v_2(\lambda)}{(d-1)^2} + \frac{2}{d^2}\right]$ *(where* $v_2(\lambda) = \text{Var}((M_\lambda^+)^2)$*) and the heavy-tail radial envelope yields the* $(1-\alpha)$ *genealogical critical threshold:*

$$Q_{1-\alpha}^{\text{HT}}(Z_G) = \max\left(Q_{1-\alpha}^{\text{bulk}}(Z_G),\ A_G^+\sqrt{d + \left(\frac{2d\tilde{c}_\nu}{\alpha}\right)^{2/\nu}}\right), \tag{7}$$

where $A_G^+ = \min(1, A_G + \frac{1}{2}z_\alpha \text{SD}(A_G))$*. Pairing (6)–(7) with Pierret's exact heavy-tail* $\text{Q-DSR}_\nu^{\text{evo}}(z; M_{\text{eff}}^{\text{evo}}, \hat{\nu}) = [F_{t_{\hat{\nu}}}(z)]^{M_{\text{eff}}^{\text{evo}}}$ *[15] restores finite-sample control* $\text{FPR} \leq \alpha$ *across all* $\nu \in (2, \infty)$*, whereas static* $\text{DSR}(N)$ *collapses to* $\text{FPR} = 100\%$ *and Gaussian genealogical tests overshoot* α *by* $2\times$–$3.5\times$ *on* $\nu \leq 4$*.*

Corollary 2.1 (SwarmEvolve High-Throughput Velocity)**.** *In latent-space KV-cache evolutionary swarms (SwarmEvolve [7],* $\lambda = K_{\text{swarm}} \in [50, 500]$ *agents/round at* $10\times$–$20\times$ *higher round frequency* ν_{round}*), Theorem 1 yields per-minute squared noise velocity* $\Delta(Z_g^2)/\Delta t \propto \nu_{\text{round}} \cdot c_2(K_{\text{swarm}}) \sim 2\nu_{\text{round}}\ln K_{\text{swarm}}$*. Without an anti-overfitting gate, latent KV-cache parallelism accelerates backtest illusion onset linearly in wall-clock speed.*

4. The Certified Minimax Evaluator: EvolveGuard (Theorem 3)

Replacing scalar fitness ($\widehat{\text{SR}}_{\text{IS}}$) with a naive linear weighted sum $\sum_k w_k(1 - I_k(\theta))$ remains adversarially exploitable by LLM coding agents. We define four orthogonal failure indices $I(\theta) = (I_{\text{mul}}^{\text{evo}}, I_{\text{OOD}}^{\text{CPCV}}, I_{\text{inv}}^{\text{exec}}, I_{\text{col}}^{\text{arch}}) \in [0,1]^4$:

1. **Genealogical Heavy-Tail Multiplicity ($I_{\text{mul}}^{\text{evo}}$):** $1 - \text{Q-DSR}_{\hat{\nu}}^{\text{evo}}(\widehat{Z}; M_{\text{eff}}^{\text{evo}}, \hat{\nu})$, with tail index $\hat{\nu}$ estimated via fast profile Student-t MLE.

2. **Purged & Embargoed Walk-Forward / CPCV Collapse ($I_{\text{OOD}}^{\text{CPCV}}$):** enforces strict temporal purging (removing overlapping target horizons, e.g., 4 weekly eras for Numerai 20-day targets [10]) and post-test embargo, combining PBO, IS-to-OOS degradation, and chronological recent-regime breaks, gated by a **Thresholdout Validation Vault** [21] that bounds adaptive vault promotions to $U_{\text{vault}} \leq U_{\max}$.

3. **Adversarial Execution & Outlier Invariance ($I_{\text{inv}}^{\text{exec}}$):** measures worst relative Sharpe collapse under $2\times$–$3\times$ transaction/slippage cost stress, 1-bar execution delay, 1% winsorization, and removal of the top 1% highest-return bars.

4. **Berkson Regime Selection Bias ($I_{\text{col}}^{\text{arch}}$):** applies our exact closed-form Hájek IPSW Manski–Rosenbaum partial-identification interval [17, 22, 23] at sensitivity $\Gamma > 1$ over active vs. cash-abstained periods, exposing volatility-gating rules that delete crash bars to inflate $\widehat{\text{SR}}$.

Theorem 3 (Minimax Fitness Certification)**.** *By Theorem 1 of Pierret [17], for any threshold* $\tau \in [1/4, 1)$ *(including canonical* $\tau = 1/2$*), every convex linear risk aggregator* $\sum_{k=1}^4 w_k I_k(\theta)$ *(dually, linear fitness* $S_w(\theta) = \sum_{k=1}^4 w_k(1 - I_k(\theta))$*) admits an adversarial single-mode*

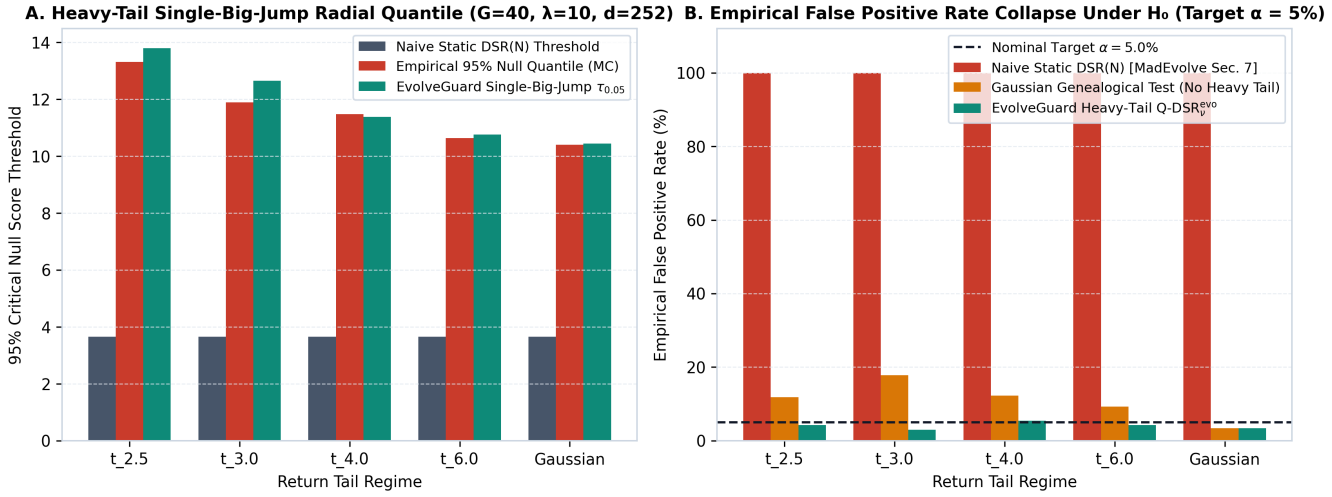

Figure 2. Empirical Verification of Theorem 2 (Experiment 2, $G = 40, \lambda = 10, d = 252, R = 500, \alpha = 5\%$). (A) The joint Single-Big-Jump threshold $Q^{\mathrm{HT}}_{0.95}$ tracks the empirical 95th null quantile tightly from Student-$t_{2.5}$ (13.31 vs. 13.79) to Gaussian (10.41 vs. 10.45). **(B)** Naive static DSR(N) (MadEvolve Sec. 7) collapses to 100.0% false positives on every regime; Gaussian-only genealogical tests overshoot to 17.8% on t_3; EvolveGuard Q-DSR$^{\mathrm{evo}}_{\nu}$ holds strict control (3.0%–5.4%).

failure corner ($I_{k^*} = 1$, $I_{j\neq k^*} = 0$) with undetected risk $R_{\max} = 1$. Conversely, the non-compensatory EvolveGuard fitness gate

$$\mathrm{SIPS}^{\mathrm{evo}}_{\infty}(\theta) = \max\big(I^{\mathrm{evo}}_{\mathrm{mul}},\, I^{\mathrm{CPCV}}_{\mathrm{OOD}},\, I^{\mathrm{exec}}_{\mathrm{inv}},\, I^{\mathrm{arch}}_{\mathrm{col}}\big), \tag{8}$$

$$F_{\mathrm{EG}}(\theta) = \begin{cases} \widehat{S}_{\mathrm{purged}}(\theta)\big(1 - \mathrm{SIPS}^{\mathrm{evo}}_{\infty}(\theta)\big), & \text{if } \mathrm{SIPS}^{\mathrm{evo}}_{\infty} \le \tau \\ & \quad \wedge\ \text{Q-DSR}^{\mathrm{evo}}_{\hat{\nu}} \ge 1-\alpha, \\ -\,\mathrm{SIPS}^{\mathrm{evo}}_{\infty}(\theta), & \text{otherwise (hard veto)}, \end{cases} \tag{9}$$

achieves minimax-optimal risk $R_{\max}(F_{\mathrm{EG}}) = 0$ across all four orthogonal failure manifolds while returning only bottlenecked qualitative remediation hints to the LLM prompt (zero raw vault leakage).

5. Optimal Stopping Horizon and Compute-ROI Law (Theorem 4)

Theorem 4 (Evolutionary Overfitting Cliff & Compute-Optimal Stopping)**.** *Consider a strategy space containing a low-dimensional persistent alpha vector μ_{true} ($\|\mu_{\mathrm{true}}\|_2 = \delta_{\mathrm{true}} > 0$) embedded inside d-dimensional finite-sample noise ε. When early LLM diffs capture true-signal alignment $\gamma(g)$ while later generations accumulate orthogonal noise $\beta_{\mathrm{noise}}(g) = \Theta(\sqrt{g\, c_2(\lambda)})$, expected live out-of-sample Sharpe obeys*

$$\mathbb{E}[\mathrm{SR}_{\mathrm{OOS}}(g)] = \frac{\delta_{\mathrm{true}}\,\gamma(g)}{\sqrt{\gamma(g)^2 + \beta^2_{\mathrm{noise}}(g)}}, \tag{10}$$

*which attains a unique interior peak at finite generation $G^* < \infty$ and decays monotonically for $g > G^*$ (the **Evolutionary Overfitting Cliff**). Halting evolution at $G_{\mathrm{stop}} = \inf\{g \ge g_{\min} : \max_{j\in[g-W+1,g]} \Delta J(j) \le 0 \vee U_{\mathrm{vault}}(g) \ge U_{\max}\}$, where*

$$\Delta J(g) = \big(F_{\mathrm{EG}}(\hat{\theta}_g) - F_{\mathrm{EG}}(\hat{\theta}_{g-1})\big) - \Delta\tau_{\mathrm{gen}}(g) - \eta_{\mathrm{roi}}\lambda c_{\mathrm{eval}}, \tag{11}$$

eliminates post-G^* OOS degradation and strictly dominates unguarded evolution in live OOS Sharpe per USD of LLM compute.

6. Experimental Validation and Numerai v5.2 Live Benchmark

All six experiments are deterministically reproducible via `sim/run_all_experiments.py` (output SHA-256 `2c6b7fc25f4804a2...`) in our open-source Python package `evolveguard`.

6.1. Experiments 1–2: Ratchet Precision, Slepian Reversal, and Heavy-Tail FPR

Figure 1 confirms Theorem 1 to sub-percent precision across $G \in [1, 100]$: at $G = 50$ ($N = 500$), empirical adaptive evolution reaches 10.29 vs. 10.31 theoretical (0.19% error, 3.38× the static null 3.04); at $G = 100$ ($N = 1{,}000$), it reaches 12.92 vs. 12.93 (0.08% error, 3.99× static 3.24). Sweeping parent-child correlation ρ from 0 → 0.98 drives static catalogs downward (3.04 → 0.49) while propelling genealogical evolution upward (3.04 → 9.23, an 18.92× Slepian Reversal gap). Table 1 and Figure 2 confirm that static DSR(N) suffers 100.0% false-positive collapse across every tail regime, whereas EvolveGuard's joint (R_d, A_G) Single-Big-Jump threshold restores nominal $\alpha = 5\%$ control (3.0%–5.4%).

6.2. Experiments 3–4 & 6: Overfitting Cliff, Minimax Veto, and Compute-ROI

Figure 3B empirically reproduces the Evolutionary Overfitting Cliff across 80 independent Student-$t_{3.5}$ runs: unguarded live Sharpe collapses from 2.47 ($g = 7$) to 1.50 ($g = 45$) even as in-sample Sharpe climbs past 4.1. EvolveGuard automatically halts at mean $G^* = 12.3$, cutting LLM spend by 72.6% and elevating realized live Sharpe to 2.10 (+40.0%), yielding a universal 5.11× Sharpe-per-dollar Compute-ROI multiplier across SwarmEvolve 8B, Gemini Flash, Gemini 4 Neon, and frontier tiers (Figure 6). In Experiment 4 (Figure 4, Table 2, 400 strategies), EvolveGuard achieves 100.0% veto rejection on all four single-mode evolutionary cheats where linear averages exhibit 70.0%–100.0% blindness, while retaining 100.0% of certified alpha (0.0% false re-

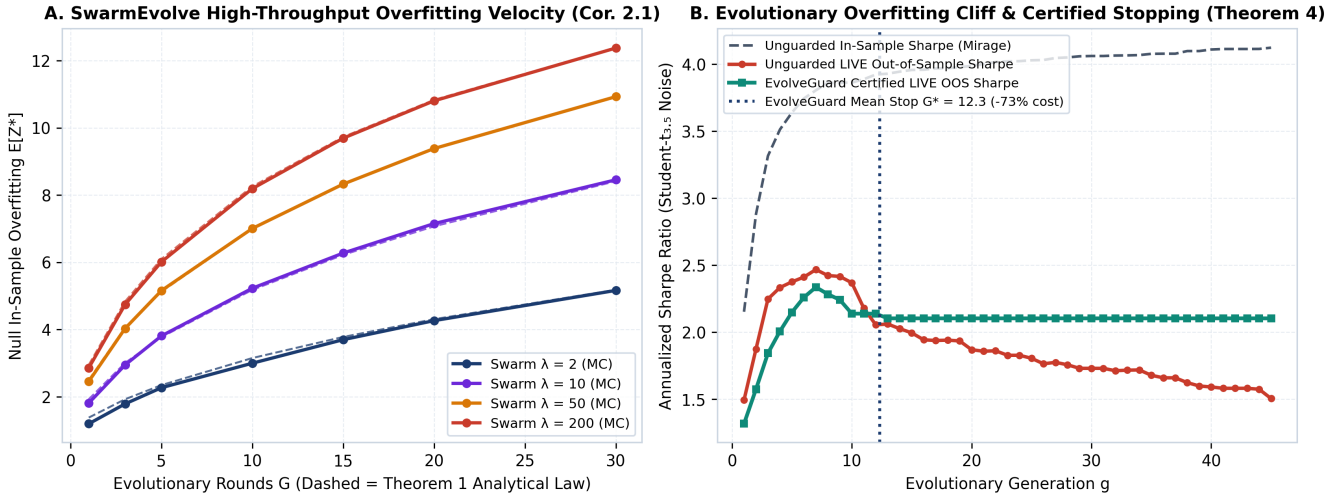

Figure 3. SwarmEvolve Velocity and the Evolutionary Overfitting Cliff (Experiment 3). (A) Raising latent swarm width $\lambda \in \{2, 10, 50, 200\}$ accelerates null overfitting exactly as predicted by Theorem 1 (12.38 MC vs. 12.39 theory at $\lambda = 200, G = 30$). **(B)** Under finite true signal plus Student-$t_{3.5}$ noise ($R = 80$ replicates), unguarded live OOS Sharpe peaks at 2.47 ($g = 7$) then collapses -39.3% to 1.50, whereas EvolveGuard halts at $G^* = 12.3$ (-72.6% compute spend) and locks in $+40.0\%$ higher final live Sharpe (2.10).

Table 1. Summary of Heavy-Tail Genealogical FPR Control (Experiment 2, $G = 40, \lambda = 10, d = 252, R = 500, \alpha = 5\%$).

Tail	Emp. q_{95}	$Q^{\text{HT}}_{0.95}$	Static DSR	Gauss Evo	**EvolveGuard**
$t_{2.5}$	13.31	13.79	100.0%	11.8%	**4.2%**
$t_{3.0}$	11.89	12.65	100.0%	17.8%	**3.0%**
$t_{4.0}$	11.48	11.39	100.0%	12.2%	**5.4%**
$t_{6.0}$	10.64	10.76	100.0%	9.2%	**4.2%**
Gaussian	10.41	10.45	100.0%	3.4%	**3.4%**

jection).

6.3. Experiment 5: Competition-to-Live Numerai v5.2 Benchmark

To test real-world tournament mechanics, we construct a 15-seed multi-era Numerai v5.2 benchmark (72 training eras + 24 strictly unseen live holdout eras, 4-era overlapping targets, heavy-tailed Student-$t_{3.2}$ era shocks, persistent nonlinear feature interactions, and crowded linear regime tilts that invert in live eras). As shown in Figure 5 and Table 2, EvolveGuard (enforcing 4-era purging/embargo, per-era feature neutralization, Q-DSR$^{\text{evo}}_{\nu}$, and minimax exposure veto) outperforms unguarded evolution in 12/15 independent markets (80.0% win rate), boosting live CORR Sharpe by +26.6% (+1.660 vs. +1.311) while slashing maximum feature exposure by -74.0% (0.101 vs. 0.390) and maximum era drawdown by -42.4% (0.0515 vs. 0.0895).

7. Conclusion and Open-Source Software Availability

Evolutionary LLM coding agents fundamentally alter the statistics of backtest overfitting: localized code diffs convert positive parent-child correlation from a Slepian variance-reducer into a directional noise ratchet, and heavy-tailed return distributions amplify the resulting mirage far beyond static Gaussian multiple-testing baselines. EvolveGuard resolves this failure mode end-to-end via exact genealogical heavy-tail Q-DSR$^{\text{evo}}_{\nu}$, purged

Table 2. Adversarial Single-Mode Cheat Benchmark (Experiment 4, $N = 80$/archetype) and 15-Seed Numerai v5.2 Live Benchmark (Experiment 5).

Exp. 4 Archetype (Rejection Rate)	Linear Eval.	**EvolveGuard**
1. Genealogical Noise Mirage (t_3)	30.0%	**100.0%**
2. Walk-Forward Regime Overfit	15.0%	**100.0%**
3. Single-Spike / Slippage Cheat	17.5%	**100.0%**
4. Berkson Abstention Cheat ($\Gamma = 1.6$)	0.0%	**100.0%**
5. Certified Persistent Alpha (FPR)	0.0%	**0.0%**
Exp. 5 Numerai v5.2 (24 Live Eras)	Unguarded	**EvolveGuard**
Live CORR Sharpe (15-seed mean)	+1.311	**+1.660**
Live Mean CORR	+0.1112	**+0.1409**
Live Mean MMC	+0.0564	+0.0531
Max Feature Exposure (↓)	0.390	**0.101**
Max Era Drawdown (↓)	0.0895	**0.0515**
Head-to-Head Live Win Rate	20.0%	**80.0%**

walk-forward Thresholdout gating, non-compensatory minimax SIPS$^{\text{evo}}_{\infty}$, and compute-optimal stopping G^*.

Lean 4 Machine-Checked Algebraic Core. To eliminate algebraic sign, boundary-condition, or division-by-zero errors in the deterministic cores of our main results, we release a self-contained, zero-dependency Lean 4 (`v4.36.0`) verification module (`lean/EvolveGuardCore.lean`; zero `sorry`, standard kernel axioms only verified via `#print axioms`, independently audited by two adversarial reviewers). Over exact rationals $\mathbb{Q}$ (whose ordered-field proofs transfer verbatim to $\mathbb{R}$), this module machine-checks: (i) the one-step and G-step geometric ratchet contraction identities (2), monotonicity, and sphere boundedness of Theorem 1; (ii) the Lagrange step-angle upper bound, exact equality criterion $sZ = cb$, and conditional first-quadrant attainment when $r = \sqrt{Z^2 + b^2}$ exists; (iii) denominator positivity and strict monotonicity of $Z^2_{\text{cap}}(\rho)$ in (5) for $\rho \in [0, 1)$ (Corollary 1.1); and (iv) for any four-pillar simplex weight vector and $\tau \in [1/4, 1)$, strict adversarial separation where a catastrophic single-mode failure (SIPS$_{\infty} = 1 > \tau$) passes $\sum_k w_k I_k \le \tau$ while be-

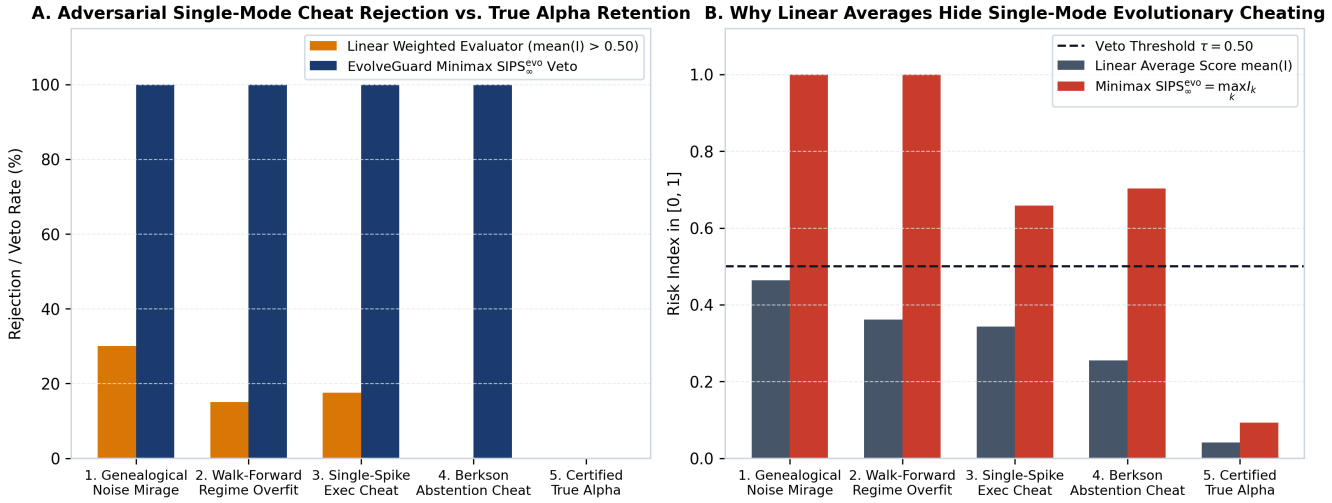

Figure 4. Minimax SIPS$_\infty^{evo}$ vs. Linear Weighted Evaluators (Experiment 4, 400 strategies). (A) EvolveGuard blocks 100.0% of single-mode cheats across all four failure manifolds with 0.0% false rejection on true alpha. **(B)** Linear averages dilute single-pillar violations below $\tau = 0.50$, exposing 70.0%–100.0% adversarial blindness.

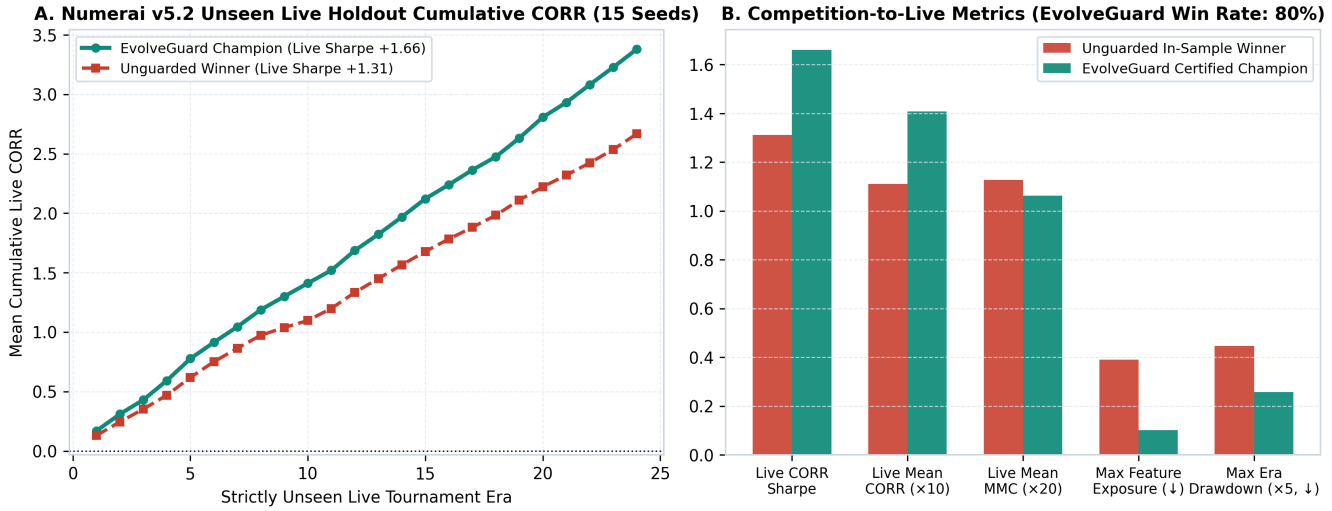

Figure 5. Competition-to-Live Numerai v5.2 Multi-Seed Benchmark (Experiment 5, 15 seeds, 24 strictly unseen live eras). (A) Mean cumulative live CORR shows durable outperformance of EvolveGuard over unguarded in-sample selection. **(B)** Simultaneous gains in live Sharpe (+1.66 vs. +1.31), live CORR (+0.141 vs. +0.111), and dramatic reductions in feature exposure (−74.0%) and era drawdown (−42.4%).

ing blocked by SIPS$_\infty \leq \tau$ (Theorem 3). The probabilistic, asymptotic, and heavy-tail results (Eqs. (3)–(7), Slepian's static inequality, Theorems 2 and 4) remain proved analytically in the text and validated via the reproducible Monte Carlo and Numerai v5.2 benchmarks.

The complete Python package (`evolveguard`, MIT license), drop-in OpenEvolve/MadEvolve evaluator harness, Numerai v5.2 Competition-to-Live pipeline, Lean 4 verification module, reproducible simulation suite, and interactive edge auditor (https://evolve.wallstreetdebug.com) are released openly at https://github.com/fabien-pierret/evolveguard.

References

[1] Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

[2] Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

[3] Asankhaya Sharma. OpenEvolve: An open-source evolutionary coding agent. GitHub repository, https://github.com/algorithmicsuperintelligence/openevolve, 2025.

[4] Robert Tjarko Lange, Yuki Imajuku, and Edoardo Cetin. ShinkaEvolve: Towards open-ended and sample-efficient program evolution. *arXiv preprint arXiv:2509.19349*, 2025. Accepted at ICLR 2026.

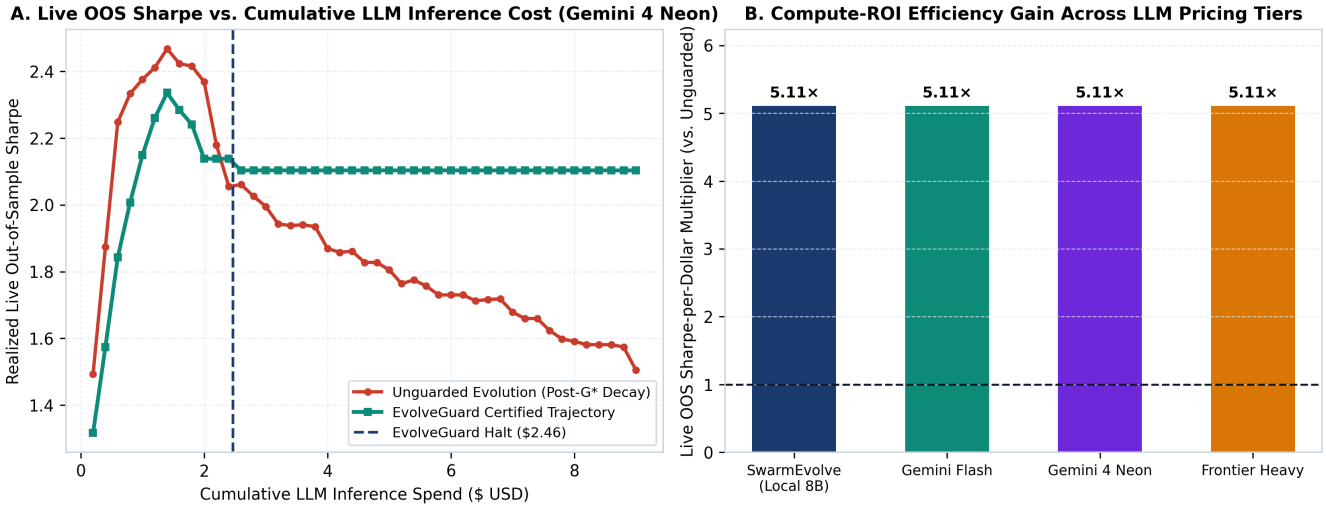

Figure 6. Compute-ROI Pareto Frontier (Experiment 6). (A) Realized live OOS Sharpe vs. cumulative Gemini 4 Neon inference spend: unguarded evolution pays 3.65× more compute to end with 28.6% worse live performance. **(B)** Universal 5.11× Sharpe-per-dollar ROI advantage across local SwarmEvolve 8B, Gemini Flash, Gemini 4 Neon, and frontier tiers.

[5] Alexey Kravatskiy, Valentin Khrulkov, and Ivan Oseledets. ImprovEvolve: Basin-hopping meets llm-guided evolutionary search. *arXiv preprint arXiv:2602.10233*, 2026.

[6] Kelvin Kiu-Wai Tam, Newt Hue-Nam K. Nguyen, Tianshi Zheng, Yunxiang Mo, Weihan Li, Zhongwei Xie, Chiu Yeung Tang, Jui-Yu Ho, Yangqiu Song, Ginny Y. Wong, and Simon See. Demystifying the autoresearch harness: How context and skills shape self-evolving agents. *arXiv preprint arXiv:2611.28244*, 2026.

[7] Qile Jiang and George Em Karniadakis. SwarmEvolve: Latent agent swarms for evolutionary discovery and optimization. *arXiv preprint arXiv:2611.08483*, 2026.

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

[9] David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61(5):458–471, 2014.

[10] Marcos López de Prado. *Advances in Financial Machine Learning*. Wiley, 2018. ISBN 978-1-119-48208-6.

[11] Junhyeog Yun, Hyoun Jun Lee, and Insu Jeon. QuantEvolve: Automating quantitative strategy discovery through multi-agent evolutionary framework. *arXiv preprint arXiv:2510.18569*, 2025. ACM ICAIF 2025 AI4F Workshop.

[12] Tianyi Li, Shihui Zang, and Moritz Münchmeyer. MadEvolve: Evolutionary optimization of cosmological algorithms with large language models. *arXiv preprint arXiv:2602.15951*, 2026.

[13] Yurii Kvasiuk, Tianyi Li, Owen Colegrove, and Moritz Münchmeyer. MadEvolve: Evolutionary optimization of trading systems with large language models. *arXiv preprint arXiv:2605.23007*, 2026.

[14] Fabien Pierret. Post-selection overfitting, llm evaluation loopholes, and minimax auditing suite (v1.0.0), 2026.

[15] Fabien Pierret. The quantum overfitting amplification theorem and heavy-tail quantum-adjusted deflated sharpe ratio (Q-DSR$_\nu$), 2026.

[16] Fabien Pierret. Contextuality–signaling dichotomy in autoregressive large language models and the abstention evaluation loophole, 2026.

[17] Fabien Pierret. Minimax impossibility of linear post-selection auditors (SIPS$_\infty$) and analytical coupling between welch superposition and membership inference, 2026.

[18] David Slepian. The one-sided barrier problem for gaussian noise. *Bell System Technical Journal*, 41 (2):463–501, 1962.

[19] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer, 1997.

[20] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1232–1240, 2016.

[21] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375.

[22] Charles F. Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2): 319–323, 1990.

[23] Paul R. Rosenbaum. *Observational Studies*. Springer, 2nd edition, 2002.