

FORESIGHT-9: Prospective and Process-Aware Evaluation of Adaptive Trading Agents

Xiangxin Luo¹, Chengtian Hong¹, Haohua Li², Yongyi Xie³

¹University of Science and Technology of China

²University of Hong Kong ³Shanghai Jiao Tong University

Abstract

Retrospective backtests provide a limited test of adaptive trading agents: they cannot rule out historical contamination, expose sensitivity to a single realized market path, or reveal internal degeneration during long-horizon adaptation. We introduce FORESIGHT-9, a prospective and process-aware benchmark built from nine auditable counterfactual stress worldlines branching from a common July 2026 information boundary. Each worldline specifies staged macro-financial events and joint multi-asset anchors; a deterministic generator realizes the trajectories, while observations are disclosed according to in-world time. A common contract standardizes observations and execution while preserving each agent’s native adaptation loop. We evaluate two adaptive trading-agent frameworks with two foundation-model backbones across 36 long-horizon runs. Agent rankings vary substantially across worldlines and backbones, and a fixed equal-weight policy outperforms 31 of 36 runs. Process telemetry exposes failures that terminal returns conceal: in one high-return run, the live factor library collapsed while executed holdings converged to the equal-weight fallback, even though decision records continued to report an active factor ensemble. FORESIGHT-9 therefore evaluates not only portfolio outcomes, but whether adaptive agent state and execution remain coherent across alternative futures. We release the worldlines, trajectories, audit traces, and regeneration scripts.

1 Introduction

Foundation-model agents are increasingly used for financial research, factor discovery, and portfolio construction, yet they are still evaluated primarily through retrospective backtests on realized market history. This habit has three failure modes. First, it cannot separate *transferable decision-making* from *historical memorization*: a strong backtest is compatible with an agent that has effectively overfit one realized path [Glasserman and Lin, 2023, Didisheim et al., 2025, Zhu et al., 2026]. Second, a single realized path cannot reveal whether an agent’s ranking is robust to different macro-financial regimes [White, 2000, Bailey et al., 2017]. Third, it looks only at *outcomes*, while the failures of long-horizon agents—stalled loops, degenerate factor libraries, silently frozen strategies, decision records that no longer match executed portfolios—live in the process layer that terminal scores compress away.

We present **FORESIGHT-9**, a prospective benchmark that evaluates foundation-model trading agents across nine auditable counterfactual futures with staged macro-financial events, multi-asset anchors, and time-gated observations (§3). Agents observe only what each in-world date discloses and must search for, validate, and admit factors, rank signals, construct portfolios, and execute under a fixed observation–action–execution contract while evolving their factor libraries and research memories (§4). The contract standardizes universe, cadence, constraints, costs, and failure handling, but deliberately preserves each agent’s native search and self-evolution mechanisms—a benchmark should not erase the behavior it intends to measure. Figure 1 summarizes the evaluation pipeline.

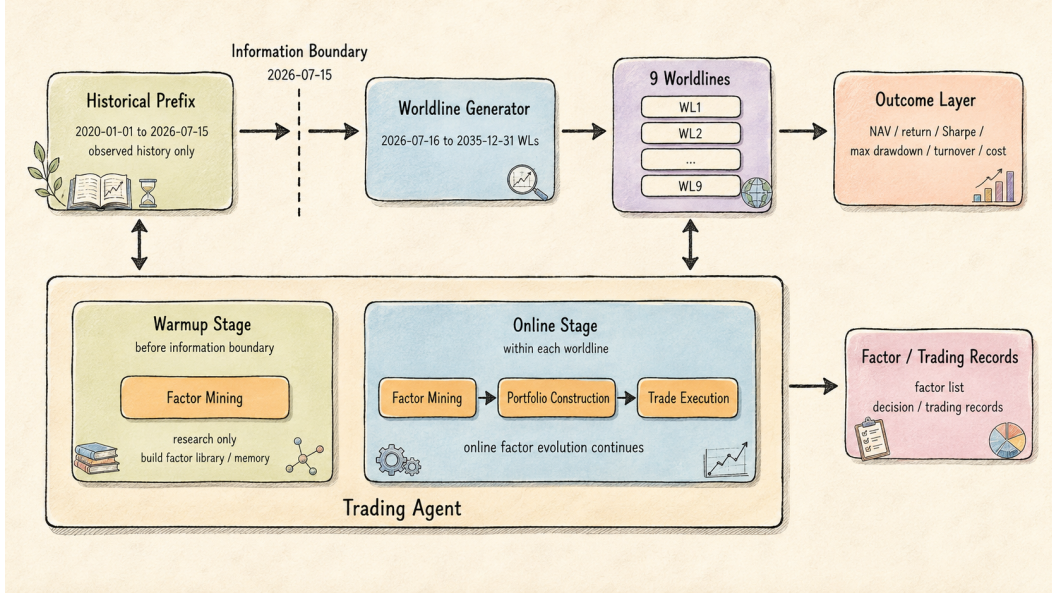


Figure 1: **FORESIGHT-9 overview.** The trading agent first mines factors on the shared pre-boundary history (warmup stage: mining only); from the shared information boundary it then mines factors, constructs portfolios, and executes trades online across the nine worldlines. Online operation feeds two audit surfaces: the outcome layer (NAV, drawdown, Sharpe) and the process records (factor lists and trade records).

FORESIGHT-9 makes three contributions.

- **Prospective multi-worldline evaluation.** We replace the single realized test path with nine controlled counterfactual stress futures branching from a shared information boundary. Each worldline combines staged macro-financial events, joint multi-asset anchors, deterministic regeneration, and time-gated disclosure of the shared 20-series panel.
- **Process-aware evaluation.** The benchmark records persistent research state, portfolio proposals, executed holdings, factor lifecycle events, and failure states, allowing evaluation to distinguish outcome quality from internal degeneration or execution inconsistency.
- **Agent-preserving evaluation contract.** FORESIGHT-9 standardizes observations, portfolio constraints, cadence, transaction costs, temporal isolation, and execution while leaving each agent’s native search and adaptation loop intact.

We instantiate the benchmark with two agent frameworks—**AlphaCrafter** (AC), a multi-agent miner–scraper–trader system, and **FactorMiner** (FM), an adaptive factor-research agent—each under two foundation-model backbones: gpt-5.6-terra and deepseek-v4-flash (“DS”). All $2 \times 2 \times 9 = 36$ runs completed to the 2035-12-31 horizon from identical initial capital.

Across these 36 runs, we find that performance varies substantially across both worldlines and model backbones. A fixed equal-weight policy outperforms 31 of 36 agent runs, showing that positive absolute returns do not by themselves establish value from adaptation. Process traces further reveal cases in which recorded research state remains active while executed portfolios have already converged to a fallback policy (§5).

FORESIGHT-9 releases the nine worldline panels and narratives, all 36 trajectories (NAV, holdings, per-decision factor usage, factor-library audit trails with admission/eviction reasons, ensemble histories), a staged factor inventory—warmup libraries, online admissions, and mining-stage candidates with formulas (Appendix E)—equal-weight baselines, and the processing scripts that regenerate every number and figure in this paper from archived artifacts.

2 Related Work

Retrospective evaluation of financial agents is not enough. LLM-agent systems now mine alpha factors, screen signals, and manage portfolios [Yu et al., 2024, Zhang et al., 2024, Xiao et al., 2024, Yuan et al., 2026, Wang et al., 2026], but they are typically evaluated by a single backtest on one realized path—exposed to look-ahead leakage and path overfitting [Glasserman and Lin, 2023, Didisheim et al., 2025, Zhu et al., 2026]. KTD-Fin controls memorized knowledge when evaluating agents on realized financial history; FORESIGHT-9 instead moves the evaluation environment itself beyond realized history and additionally audits persistent adaptation and execution state. A single-history leaderboard cannot tell whether a ranking reflects decision quality or the difficulty of one draw [White, 2000, Bailey et al., 2017]. Existing financial-agent evaluations therefore provide rich agent mechanisms but limited variation in the evaluation environment.

Synthetic paths and predefined stresses fall short as evaluation environments. Stress testing evaluates policies across synthetic adverse scenarios [Glasserman et al., 2015]; generative time-series models synthesize alternative trajectories from learned historical dynamics [Yoon et al., 2019, Wiese et al., 2020]; counterfactual evaluation estimates off-policy performance from logged data [Dudík et al., 2011, Thomas and Brunskill, 2016]. These methods generate alternative paths or predefined stresses, but are not designed as persistent, time-gated environments for adaptive foundation-model agents: the agent under test lives inside the environment for years of in-world time, must maintain state across stages, and must never observe beyond the current in-world date. FORESIGHT-9’s worldlines are documented counterfactual continuations of a shared history designed precisely for that use.

Process-aware evaluation of long-running agents. Agent benchmarks moved evaluation from static datasets to interactive environments [Liu et al., 2024, Zhou et al., 2024, Xie et al., 2024], with follow-up work pushing trajectory-level scoring—intermediate states, tool-use quality, latency, cost [Ma et al., 2024, Yao et al., 2025, Lu et al., 2025]—and process supervision distinguishing intermediate-step feedback from outcome-only supervision [Uesato et al., 2022, Lightman et al., 2024]. FORESIGHT-9 combines forward scenario variation with persistent process telemetry in a financial decision environment: admission and eviction audits are step-level records of research decisions, and the run-evidence layer captures operational failure and recovery as first-class evaluation objects.

3 The FORESIGHT-9 Benchmark

The evaluation unit of FORESIGHT-9 is a worldline (WL): a complete forward market path with a documented anchor table, staged regime transitions, a volatility profile, and event/news timing. The nine WLs are controlled stress scenarios, not forecasts: the 10-year probabilities elicited by the scenario debate (§3.2) are provenance metadata and never enter evaluation as sampling frequencies or likelihood weights.

3.1 Shared information boundary and panel contract

All nine WLs fork from a shared information boundary $t_0 = 2026-07-15$. Before t_0 , every run sees the same daily panel of 20 series: 15 tradable market exposures—equity indices {SPX, NDX, SOX, SX5E, N225, HSI, 000300.SH, 000688.SH}, commodities {XAU, COPPER, WTI}, crypto {BTC, ETH}, and sovereign yields {US10Y, CN10Y}—plus 5 observation-only signals (DXY, USDCNY, USDJPY, EURUSD, VIX). Sovereign yields are quoted in level points and traded directly as synthetic yield-level exposures—a long position profits when the yield level rises—directionally analogous to a short-duration exposure, though position P&L is the percentage change of the level rather than duration-scaled (Appendix C). Returns on all exposures are combined in local-currency percentage terms without FX conversion, so a portfolio is a normalized cross-market allocation rather than a currency-denominated account. After t_0 , each WL unfolds its own path to 2035-12-31 (~2,468 trading days). An agent at in-world date t sees exactly the rows with date $\leq t$; any exception is a failed data audit, not a modeling choice. Panel construction details and data-quality boundaries are documented in Appendix C.

WL	Worldline crisis	Main transmission
WL1	Geopolitical	conflict → sanctions → trade fragmentation
WL2	Technology fragmentation	internet splintering → bipolar digital economy
WL3	Nuclear escalation	nuclear test → conventional conflict → new nuclear order
WL4	Sovereign debt / liquidity	Treasury stress → Minsky moment → stagflation
WL5	Currency disorder	JPY disorderly devaluation → Asian currency fragmentation
WL6	Inflation regime	inflation re-ignition → credibility collapse → stagflation
WL7	Financial-system fragility	fragility build-up → algorithmic bank run → regulatory rebuild
WL8	Energy	Hormuz blockade → energy-system collapse → restructuring
WL9	Pandemic / automation	labor shock → recession → automation supercycle

Table 1: **The nine worldlines at a glance.** Each worldline is a sequence of 4–5 stages; the full stage end dates and narratives are in Table 7 (Appendix C). All WLs end 2035-12-31.

3.2 Scenario origination: a specialist agent team and debate

The scenario set was produced by a three-role financial-research agent team covering strategy (policy), macro-finance, and chief analyst (synthesis), run on a commercial financial-data and agent platform. Each role proposed three 2026–2035 scenarios, followed by three rounds of cross-review (independent derivation, bidirectional cross-review, defended responses), scoring every scenario on logical rigor, 10-year cumulative probability, and shock severity. The resulting scenario specifications—staged event narratives and joint multi-asset endpoints—are what the trajectory generator consumes (§3.3). The full score matrix is released as a provenance artifact (Appendix G), and the anchor-level provenance is documented in Appendix C.

3.3 Generator: scenario anchors with bridge noise

Each WL is specified as a sequence of 4–5 *stages*, each with a named narrative endpoint and a terminal date (Table 1; full stage narratives in Appendix C). Prices are interpolated in log space between stage endpoints; yield and volatility paths follow the stage specifications; foreign-exchange series without an explicit trajectory are derived from DXY via declared betas. Within each stage, local variation is generated by a seeded Brownian bridge [Karatzas and Shreve, 1991] while preserving the scenario-defined endpoints,

$$\ell_u = \frac{T-u}{T} \ell_0 + \frac{u}{T} \ell_T + \sigma \left(B_u - \frac{u}{T} B_T \right), \quad u = 0, \dots, T, \quad (1)$$

whose volatility σ is calibrated to the realized volatility of the warmup window and whose endpoints are pinned to zero, so the perturbation hits each designated stage endpoint exactly: scenario anchors determine the macro path, and the bridge supplies only local variation. At stage boundaries the path is *re-anchored* to the next stage’s endpoint table, preventing accumulated noise from drifting a WL away from its macro scenario. News and event narratives are timestamped inside a stage only after the corresponding price information, with in-world publication at a controlled offset from the stage start; the generator records the seed and parameters of every stage, making each WL deterministic conditional on its manifest.

3.4 Auditability

Three properties make a WL auditable rather than merely synthetic: (i) *endpoint anchoring*—the stage endpoint table is part of the released manifest, so any realized path can be checked against its declared scenario; (ii) *leakage control*—the information boundary is enforced by the panel contract of §3.1, and news timing follows the price-leads-news rule; (iii) *determinism*—conditional on the manifest and seed, regeneration is reproducible, and the worldline inputs consumed by all 36 runs

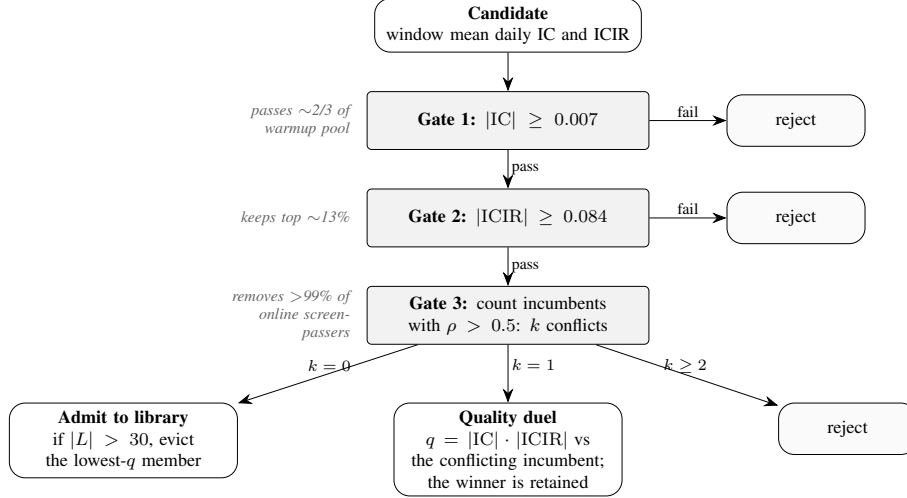


Figure 2: **The admission funnel.** Gate values are stipulated and deliberately permissive; selectivity comes from the staged tightening. Gate 3 counts the candidate’s pairwise Spearman conflicts ($\rho > 0.5$) with incumbents: $k \geq 2$ rejects the candidate; $k = 1$ triggers a quality duel on $q = |\text{IC}| \cdot |\text{ICIR}|$ with the conflicting incumbent, the winner retained; $k = 0$ admits it, with the capacity-30 trim evicting the lowest- q member. Gray rates are measured on the warmup candidate pool and on online screen-passers.

are byte-identical. We also record what the WLs are *not*: they are not sampled from a calibrated real-world probability law, and robustness summaries across WLs (medians, quantiles, worst-WL) describe behavior under documented scenario spread, not probabilistic expectations.

4 Evaluated Agents and Protocol

FORESIGHT-9 evaluates agents as test subjects: they observe the panel, maintain persistent state, and emit portfolios through a fixed execution interface, while their internal search and self-evolution mechanisms remain intact. We evaluate two frameworks—FactorMiner (FM) and AlphaCrafter (AC)—under two foundation-model backbones: gpt-5.6-terra and deepseek-v4-flash (“DS”).

4.1 FactorMiner (FM): adaptive factor research

FM follows the FactorMiner framework [Wang et al., 2026] and runs a mining–admission–combination loop with a strict separation of concerns: only *mining* consumes the LLM; admission, combination, and forward execution are deterministic. In the shared warmup (2020-01-01–2026-07-15) the loop mines a large candidate pool; online, each 10-trading-day window appends one further iteration. Admission is a deterministic contract built on one principle: at a fifteen-asset cross-section, no single statistic can be trusted to certify significance, so the gate values— $|\text{IC}| \geq 0.007$ and $|\text{ICIR}| \geq 0.084$ on the window mean daily series—are fixed by stipulation and deliberately permissive, and selectivity is engineered downstream where it can be audited stage by stage (Figure 2): Gate 1 removes candidates without measurable signal, Gate 2 removes unstable ones, and Gate 3 enforces diversity by routing each survivor on its conflict count k against the incumbents, so that a correlated but strictly better candidate can displace a weaker incumbent instead of being silently blocked. Capacity then keeps the library at its thirty best slots by $q_i = |\text{IC}_i| \cdot |\text{ICIR}_i|$ —a greedy, diversity-constrained library update rule, repeatedly re-optimized, pairwise near-orthogonal ($\rho < 0.5$), and trim-protected against stale members. At each decision the combination step takes the IC-weighted top-10 of the current library, preserving $\text{sign}(\text{IC})$; absent a valid ensemble the contract’s built-in fallback is the equal-weight $1/15$ vector. Every lifecycle event—proposed, parsed, screened, rejected, admitted, evicted—is appended with the candidate’s formula to an auditable factor-lifecycle log.

4.2 AlphaCrafter (AC): multi-agent miner–screener–trader

AC follows the AlphaCrafter framework [Yuan et al., 2026] and chains specialist agents inside a per-WL sandbox workspace, each research cycle comprising three parallel Miners, one Screener, and one Trader. The shared warmup runs 40 such cycles on the pre-boundary history, and each WL run is seeded from its own experiment’s (strictly disjoint) warmup workspace. Miner agents each propose a research idea, implement and run a metrics stage, and persist a factor artifact; admission recomputes pairwise correlations on the artifacts’ actual signals rather than trusting the miner’s self-report, then applies the same stipulated IC/ICIR gates and the same conflict-count routing as FM ($k \geq 2$ reject; $k = 1$ quality duel on $q = |\text{IC}| \cdot |\text{ICIR}|$; $k = 0$ admit), with the capacity-30 trim evicting the lowest- q member and every eviction carrying a machine-readable reason. The screener selects the active ensemble (≤ 10 factors, quality/IC-tilted weights); the trader consumes the ensemble, computes asset preferences, and emits target weights. AC’s native path has no DSL parser—factor generation is proposal-and-screen rather than expression search—and we make no claim that AC and FM perform identical factor search. Account state, the per-cycle factor-library audit trail, research memory, and the evolving workspace are all released artifacts.

4.3 Benchmark contract and evaluation setup

Both frameworks were wired to the benchmark through a thin, logged adaptation layer; nothing inside either agent’s search or evolution loop was rewritten.

- **Unified panel and universe.** Both frameworks consume the same 15+5 series with the same naming and calendar; worldline inputs are byte-identical across runs.
- **Decision cadence and the proposal/gate contract.** Both frameworks decide on the same 10-trading-day grid. A decision may only produce a *proposal*; a deterministic execution layer is the sole writer to accounts. With w the current and w^* the proposed target weights and $\hat{r}$ the agent’s forecast returns over the 10-day horizon,

$$\tau = \frac{1}{2} \sum_k |w_k^* - w_k|, \quad e = 10^4 \sum_k (w_k^* - w_k) \hat{r}_k,$$

and the rebalance executes only if the run is the initial full allocation or $e > 3\tau$; the realized cost is $3 \text{ bps} \times \tau \times V$. A gated-out decision persists the proposal and forecasts but leaves holdings untouched—a logged no-trade, not a failure. Timing is pinned to the observation boundary: at decision date t every observation runs through the close of t ; FM fills at the next trading mark—the last observed close, since forward opens equal prior closes (Appendix C)—while AC’s native trader fills at the decision day’s open; the implied same-day overlap for AC is quantified in Appendix B and does not alter any headline comparison. The simulator executes emitted weights verbatim, and direct order placement is disabled in the benchmark account, so any divergence between a decision record and the executed portfolio is an auditable event, not a silent correction.

- **Portfolio constraints and costs.** Long-only, fully invested, non-negative weights summing to 1 with zero cash; first allocation free; 3 bps one-way on migrated notional at every rebalance.
- **Per-WL isolation and persistence.** Each run lives in an isolated sandbox with persisted state and supports checkpoint/resume without changing in-world time.
- **Failure handling and guaranteed window advance.** A rejected or skipped proposal is logged with its reason and the portfolio state is retained; silently replacing a failed action with a successful one is prohibited. The calendar never stalls: whatever a module does or fails to do, each cycle advances the in-world date cursor by exactly 10 trading days—if the trader cannot act, the existing strategy (or a deterministic hold) advances the window under mark-to-market. Module failures therefore degrade a run; they cannot freeze it.
- **Trajectory logging and process telemetry.** For every decision block we record the observation payload, tool calls, candidate factors and validation outcomes, the portfolio proposal and execution result, and skip reasons—the process layer on which §5.3 relies. Reflection is treated as an intervention, not free post-hoc information: a reflection may consume the current result and write a bounded memory item that affects later prompts only after its write timestamp. We distinguish logged executable actions from model reasoning text; hidden chain-of-thought is never an evaluation target.

Worldline	AC-terra	AC-DS	FM-terra	FM-DS	EW-15
WL1	+34.9%	+57.5%	+53.4%	+34.9%	+64.5%
WL2	+43.4%	+49.9%	+87.8%	+41.4%	+82.2%
WL3	+38.6%	+62.4%	+96.8%	+13.8%	+54.6%
WL4	+5.8%	+25.1%	+8.8%	+20.3%	+26.6%
WL5	+39.4%	+59.4%	+138.7%	+8.4%	+71.8%
WL6	+72.2%	+67.0%	+77.1%	+85.7%	+103.2%
WL7	+24.7%	+21.2%	+53.5%	+39.8%	+59.7%
WL8	+58.4%	+86.3%	+75.0%	+29.6%	+98.2%
WL9	+48.0%	+15.7%	+109.4%	-6.7%	+80.3%
Mean	+40.6%	+49.4%	+77.8%	+29.7%	+71.2%
Median	+39.4%	+57.5%	+77.1%	+29.6%	+71.8%
Range	[+5.8, +72.2]	[+15.7, +86.3]	[+8.8, +138.7]	[-6.7, +85.7]	[+26.6, +103.2]

Table 2: **Cumulative total return per worldline** for the four agent–backbone combinations and the equal-weight baseline. Every agent run consumed the same byte-identical worldline inputs, constraints, and fee rule as its EW-15 column.

	AC-terra	AC-DS	FM-terra	FM-DS	EW-15
Mean cumulative return	+40.6%	+49.4%	+77.8%	+29.7%	+71.2%
Median cumulative return	+39.4%	+57.5%	+77.1%	+29.6%	+71.8%
Ann. return (median)	+3.5%	+4.8%	+6.0%	+2.7%	+5.7%
Ann. volatility (median)	7.4%	8.8%	10.9%	9.9%	7.9%
Sharpe (median, $r_f=0$)	0.46	0.50	0.54	0.27	0.70
Max drawdown (median)	-16.9%	-19.1%	-26.6%	-28.0%	-22.6%

Table 3: **Standard portfolio metrics from the daily NAV series** (annualization over 252 trading days; $r_f = 0$; median = across-worldline median of the per-run statistic; per-run Sharpe and drawdown tables are in Appendix B).

We evaluate AC and FM under the two backbones on all nine worldlines, yielding 36 completed runs to 2035-12-31 from identical initial capital. All runs use the same 10-day decision cadence, long-only fully invested portfolios, and 3 bps one-way costs. We report cumulative return, annualized volatility, Sharpe ratio, maximum drawdown, turnover, and transaction costs; per-run metrics, run-status bookkeeping, and NAV-reconstruction conventions are documented in Appendix B, and the model, reasoning-effort, and sampling configuration of every LLM call in Appendix H.

5 Results

We report the completed $2 \times 2 \times 9$ matrix—36 runs, each reaching 2035-12-31 from initial capital 1,000,000; returns are cumulative (terminal net assets over initial capital). We additionally compute a *15-asset equal-weight baseline* (EW-15) per WL from the same panel: equal weights on the 15 market exposures, the agents’ 10-day cadence, first allocation free, 3 bps one-way on migrated notional. All numbers and figures regenerate from archived artifacts via the released scripts.

5.1 Outcome Robustness Across Worldlines

Performance varies substantially across both worldlines and model backbones. FM has the highest mean return under terra (+77.8% versus +40.6% for AC), whereas AC leads under DS (+49.4% versus +29.7%). The ordering also changes within individual worldlines; for example, WL5 is the strongest outcome for FM-terra (+138.7%) but among the weakest for FM-DS (+8.4%). Framework rankings inferred from a single environment and a single backbone are therefore not stable across the benchmark.

Worldline difficulty itself is shared: WL4 (Minsky/stagflation) suppresses everything, agents and baseline alike, while WL6 and WL8 yield the highest returns for EW-15 (Figures 8–16, Appendix A, decompose each worldline into its five runs). Yet the per-worldline ordering of agents still reshuffles from one worldline to the next (Table 2).

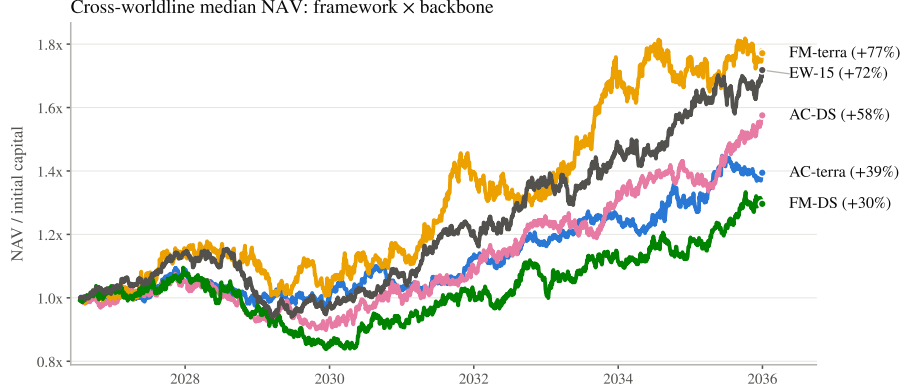


Figure 3: **Cross-worldline median NAV of the four agent-backbone combinations, with the EW-15 median (dark gray).** A single cross-worldline median per combination is what the matrix is designed to expose: FM-terra leads, FM-DS trails, and each framework’s ranking inverts when the backbone changes. Per-version overlays across the nine worldlines are in Figures 4–7 (Appendix A).

5.2 Comparison with the Equal-Weight Reference

The equal-weight reference is difficult to beat. EW-15 achieves a mean cumulative return of +71.2% and outperforms 31 of the 36 agent runs (Table 2): AC-terra 0/9, AC-DS 1/9, FM-DS 0/9, and FM-terra 4/9—the only configuration that exceeds the reference on multiple worldlines. On risk-adjusted performance, EW-15 also attains the highest median Sharpe ratio (0.70); FM-terra combines the highest mean return with a larger median drawdown than the reference (−26.6% vs. −22.6%; Table 3). These comparisons show that positive absolute return alone does not establish value from online adaptation; much of the portfolio gain is also available to a fixed allocation under the same market path and cost model. Window-level comparisons and worldline-clustered significance tests, which reach the same conclusion, are reported in Appendix D. The best DS run (+86.3%, AC-DS on WL8) illustrates the point from the process side in §5.3 below.

5.3 Process Consistency

Because the benchmark’s promise is *self-evolution*, we evaluate not only outcomes but whether the three states a run maintains stay coherent: the *research state* (the live factor library), the *declared decision state* (the ensemble and portfolio the agent records), and the *executed state* (the holdings actually held). All three are recorded persistently as trajectory series and terminal registers (Appendix F), so a run is auditable as a decision system rather than as a terminal score.

Process traces reveal failures that terminal portfolio metrics cannot. In AC-DS on WL8—one of the best agent runs in the matrix (+86.3%)—the live factor library eventually fell to zero while decision records continued to reference a non-empty factor ensemble, and executed holdings converged to the equal-weight fallback for most of the remaining horizon. The run still produced a comparatively high absolute return, illustrating that outcome performance alone does not establish that the intended adaptive mechanism remained active. Appendix F dissects the lifecycle and execution traces and tabulates the process registers for all runs.

More generally, most runs end with a populated factor library at or near capacity, yet the audit trails behind similar terminal snapshots are heterogeneous—from stable libraries with sparse, cleanly reasoned evictions to continuous churn until little survives. Terminal snapshots, the only thing an outcome leaderboard sees, cannot distinguish a stable library from one that churned to zero. Transaction costs are negligible throughout (at most 126 bps of initial capital over 9.5 years), so the observed return dispersion cannot be explained by transaction-cost differences alone.

6 Limitations

FORESIGHT-9 evaluates each configuration with one seed per run; multi-seed replication per worldline would separate agent-level variance from worldline-level variance but multiplies an already

large compute budget. The worldlines are controlled scenarios, not sampled forecasts: we make no probabilistic claim that any WL is more likely than another, and cross-WL summaries (medians, quantiles, worst-WL) describe behavior under documented scenario spread, not expectations under a generative law. The anchors enforce scenario-consistent multi-asset endpoints, but we do not claim the daily trajectories are samples from a calibrated joint return distribution, and scenario plausibility was produced and cross-reviewed by a single agent team rather than independently validated by human domain experts. Wall-clock events (service interruptions, retries) do not enter in-world outcomes—the in-world calendar advances deterministically—so operational reliability is reported but not scored. Finally, every claim in this paper re-derives from the released artifacts via the released scripts, but the benchmark evaluates two agent frameworks and two backbones; coverage of further mechanisms (e.g., memory-augmented or multi-agent debate traders) remains future work. The long-only, fully invested contract evaluates relative cross-market allocation rather than cash timing or short selling.

7 Conclusion

FORESIGHT-9 evaluates trading agents along two dimensions that retrospective backtests largely collapse: robustness across alternative market futures and consistency between adaptive state and executed behavior. Across nine worldlines, agent rankings vary substantially with both environment and backbone, while a simple equal-weight reference remains competitive. Process traces further show that favorable portfolio outcomes can coexist with failures of the intended adaptation mechanism. These results motivate evaluating long-horizon financial agents as persistent decision systems rather than only by terminal portfolio scores. We release the worldline panels and manifests, all 36 trajectories, the staged factor inventories, baselines, and regeneration scripts.

Ethics. All evaluation runs on synthetic counterfactual worldlines; no live-market execution or client funds are involved, and the benchmark is a research instrument, not investment advice.

References

- David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. The Probability of Backtest Overfitting. *The Journal of Computational Finance*, 20(4):39–69, 2017. doi: 10.21314/JCF.2016.322. URL <https://doi.org/10.21314/JCF.2016.322>.
- Antoine Didisheim, Martina Fracchini, and Luciano Somoza. AI’s Predictable Memory in Financial Analysis. *Economics Letters*, 256:112602, 2025. doi: 10.1016/j.econlet.2025.112602. URL <https://doi.org/10.1016/j.econlet.2025.112602>.
- Miroslav Dudík, John Langford, and Lihong Li. Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104, 2011. URL https://icml.cc/2011/papers/554_icmlpaper.pdf.
- Paul Glasserman and Caden Lin. Assessing Look-Ahead Bias in Stock Return Predictions Generated By GPT Sentiment Analysis, 2023. URL <https://arxiv.org/abs/2309.17322>.
- Paul Glasserman, Chulmin Kang, and Wanmo Kang. Stress Scenario Selection by Empirical Likelihood. *Quantitative Finance*, 15(1):25–41, 2015. doi: 10.1080/14697688.2014.926019. URL <https://doi.org/10.1080/14697688.2014.926019>.
- Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. Springer, 2 edition, 1991. doi: 10.1007/978-1-4612-0949-2. URL <https://doi.org/10.1007/978-1-4612-0949-2>.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s Verify Step by Step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as Agents. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e9df36b21ff4ee211a8b71ee8b7e9f57-Abstract-Conference.html.
- Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. ToolSandbox: A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1160–1183. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.65. URL <https://aclanthology.org/2025.findings-naacl.65/>.
- Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2365. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/877b40688e330a0e2a3fc24084208dfa-Abstract-Datasets_and_Benchmarks_Track.html.
- Philip Thomas and Emma Brunskill. Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2139–2148. PMLR, 2016. URL <https://proceedings.mlr.press/v48/thomasa16.html>.
- Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving Math Word Problems with Process- and Outcome-Based Feedback, 2022. URL <https://arxiv.org/abs/2211.14275>.
- Yanlong Wang, Jian Xu, Hongkang Zhang, Shao-Lun Huang, Danny Dongning Sun, and Xiao-Ping Zhang. FactorMiner: A Self-Evolving Agent with Skills and Experience Memory for Financial Alpha Discovery, 2026. URL <https://arxiv.org/abs/2602.14670>.
- Halbert White. A Reality Check for Data Snooping. *Econometrica*, 68(5):1097–1126, 2000. doi: 10.1111/1468-0262.00152. URL <https://doi.org/10.1111/1468-0262.00152>.
- Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. Quant GANs: Deep Generation of Financial Time Series. *Quantitative Finance*, 20(9):1419–1440, 2020. doi: 10.1080/14697688.2020.1730426. URL <https://doi.org/10.1080/14697688.2020.1730426>.
- Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. TradingAgents: Multi-Agents LLM Financial Trading Framework, 2024. URL <https://arxiv.org/abs/2412.20138>.
- Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1650. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.
- Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R. Narasimhan. τ -Bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b126cc38b8638e07bef37e7b2bb72bf-Abstract-Conference.html.
- Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-Series Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, volume 32, pages 5508–5518, 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/c9efe5f26cd17ba6216bbe2a7d26d490-Abstract.html>.

- Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W. Suchow, and Khaldoun Khashanah. FinMem: A Performance-Enhanced LLM Trading Agent with Layered Memory and Character Design. *Proceedings of the AAAI Symposium Series*, 3(1): 595–597, 2024. doi: 10.1609/aaais.v3i1.31290. URL <https://ojs.aaai.org/index.php/AAAI-SS/article/view/31290>.
- Yishuo Yuan, Jiayi Sheng, Sirui Zeng, Jiaqi Wang, and Jiaheng Liu. AlphaCrafter: A Full-Stack Multi-Agent Framework for Cross-Sectional Quantitative Trading, 2026. URL <https://arxiv.org/abs/2605.05580>.
- Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. A Multimodal Foundation Agent for Financial Trading: Tool-Augmented, Diversified, and Generalist. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2024. doi: 10.1145/3637528.3671801. URL <https://arxiv.org/abs/2402.18485>.
- Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4410c0711e9154a7a2d26f9b3816d1ef-Abstract-Conference.html.
- Taojie Zhu, Wentao Zhao, Rui Sun, Beidi Luan, Jiacheng Lu, Sinuo Wang, Jing Li, Daxin Jiang, Yonghong He, and Zuo Bai. From Knowing to Doing: A Memory-Controlled Benchmark for LLM Trading Agents on Stock Markets, 2026. URL <https://arxiv.org/abs/2605.28359>.

A Additional NAV Views

Per-version overlays across the nine worldlines (Figures 4–7), followed by one panel per worldline: the four completed runs (AC–terra, AC–DS, FM–terra, FM–DS) against the 15-asset equal-weight baseline (Appendix B tabulates the corresponding risk-adjusted metrics). All curves are solid in a validated five-color scheme; end-of-horizon labels give terminal returns. AC curves are daily reconstructions from the sandbox `rebalance_history` marked on the panel closes (held target weights between rebalances, terminal segment anchored to the account’s closing net assets; reconstruction error against the next mark’s pre-trade NAV is $< 0.03\%$ mean for 17 of 18 runs, $< 0.5\%$ max for all).

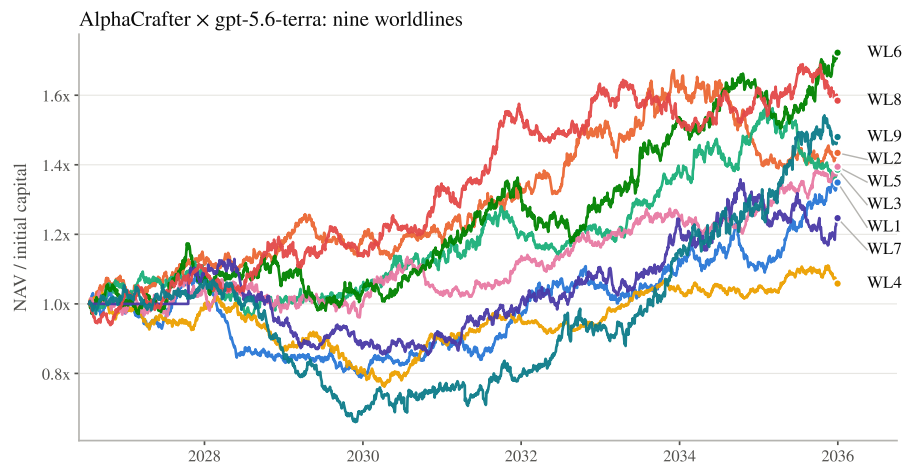


Figure 4: **AC–terra across the nine worldlines** (nine solid runs; colorblind-validated palette; right-edge labels mark the terminal worldline). Conventions are identical in Figures 5–16.

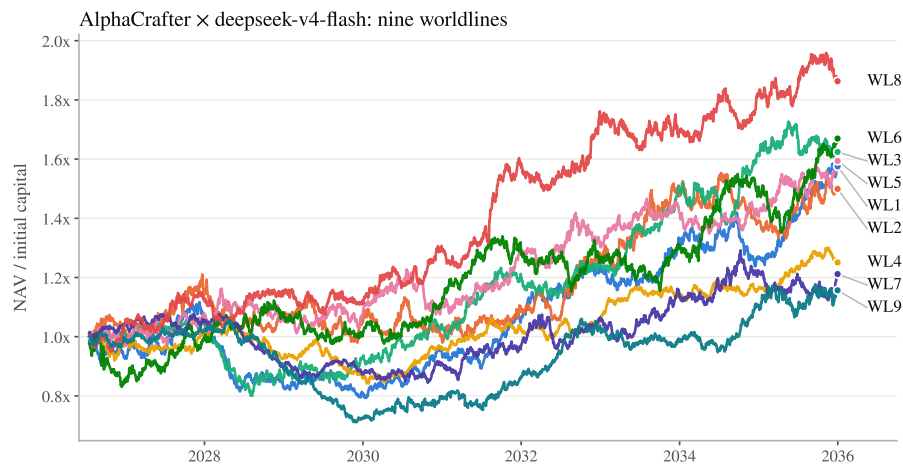


Figure 5: **AC–DS across the nine worldlines.**

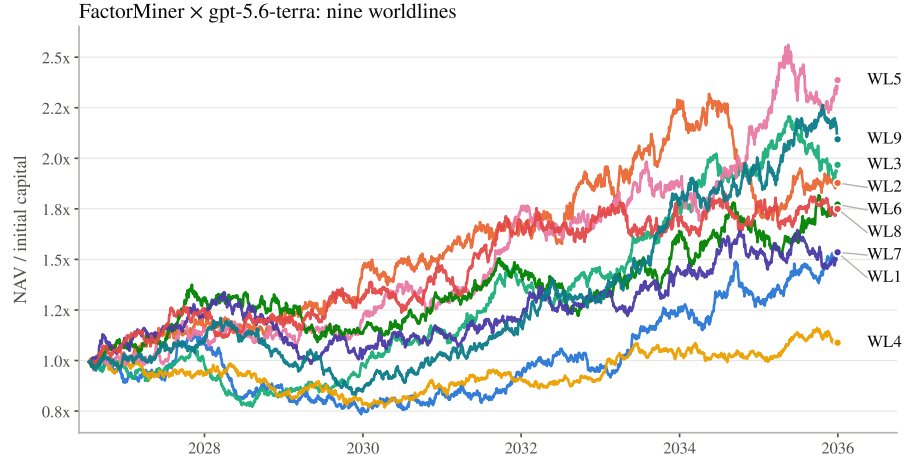


Figure 6: FM-terra across the nine worldlines.

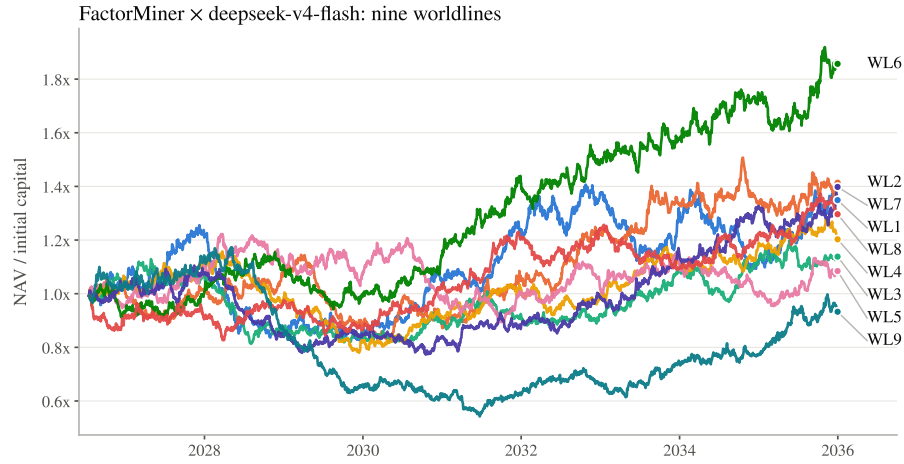


Figure 7: FM-DS across the nine worldlines. The framework that leads under terra (Figure 6) trails under DS: the per-worldline ranking is substantially backbone-dependent.

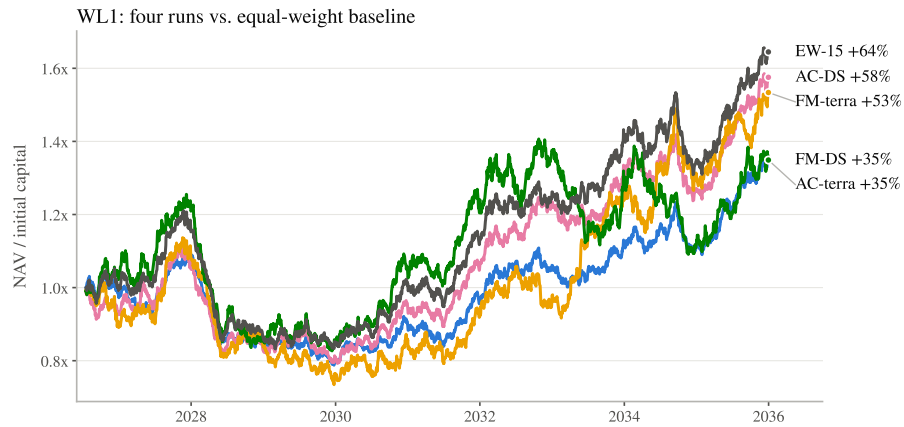


Figure 8: WL1: four runs vs. the equal-weight baseline. AC-terra (blue), AC-DS (pink), FM-terra (amber), FM-DS (green), EW-15 (gray); labels give terminal returns.

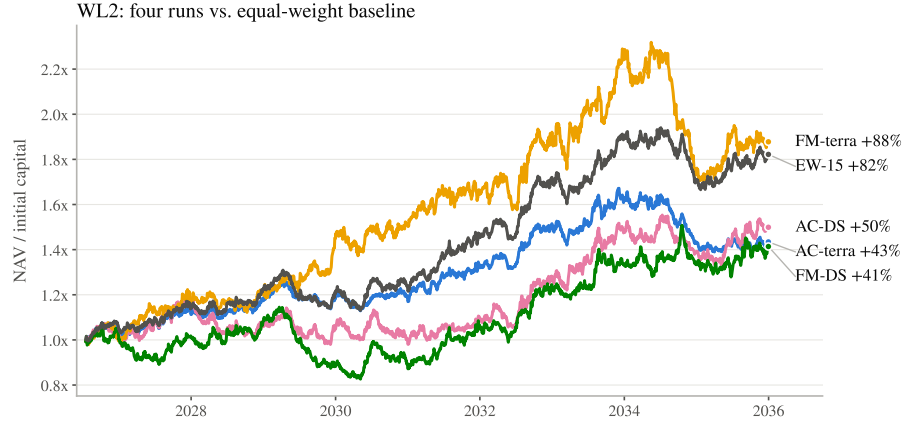


Figure 9: **WL2: four runs vs. the equal-weight baseline.** AC-terra (blue), AC-DS (pink), FM-terra (amber), FM-DS (green), EW-15 (gray); labels give terminal returns.

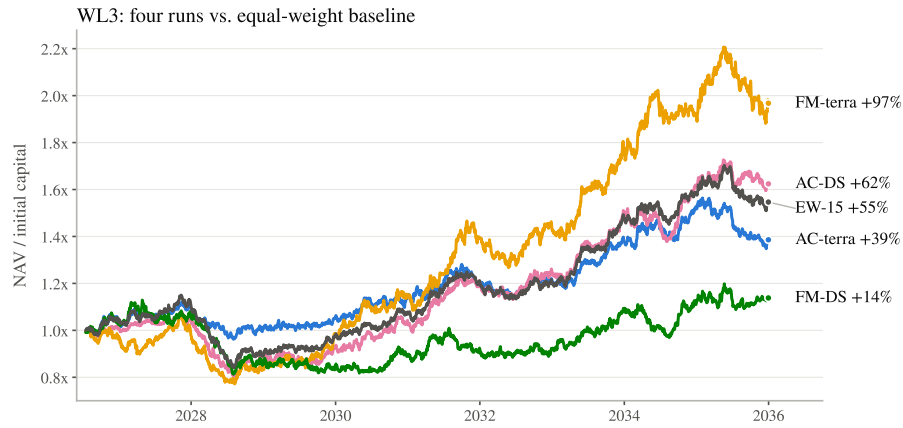


Figure 10: **WL3: four runs vs. the equal-weight baseline.** AC-terra (blue), AC-DS (pink), FM-terra (amber), FM-DS (green), EW-15 (gray); labels give terminal returns.

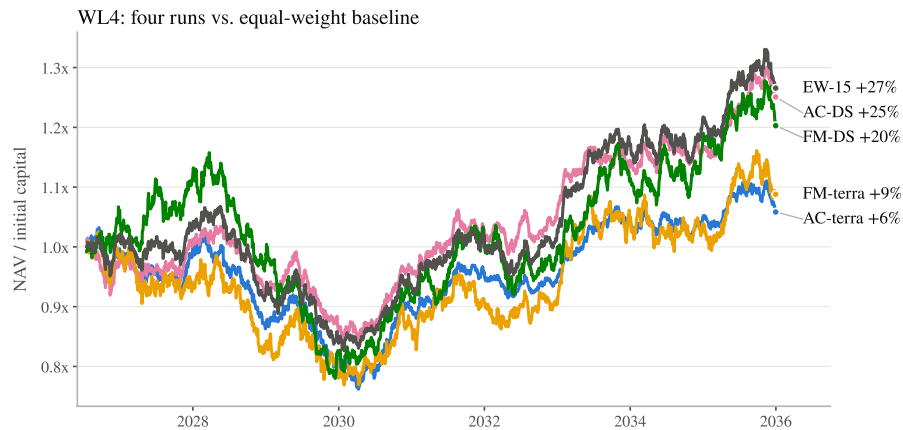


Figure 11: **WL4: four runs vs. the equal-weight baseline.** AC-terra (blue), AC-DS (pink), FM-terra (amber), FM-DS (green), EW-15 (gray); labels give terminal returns.

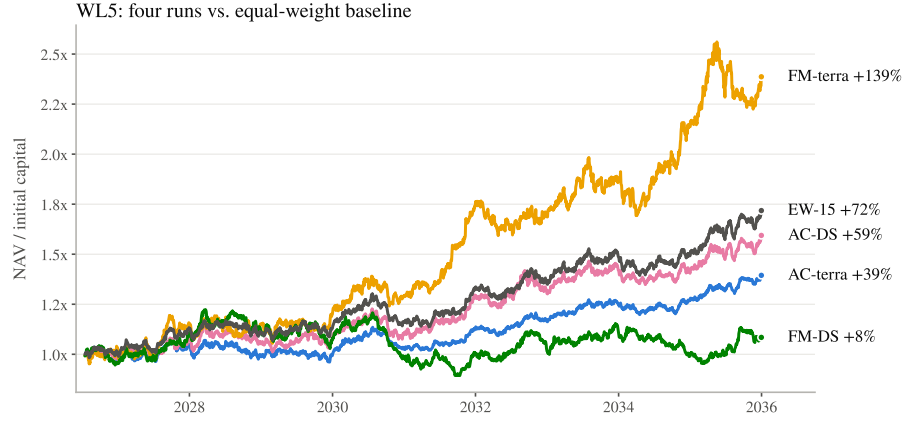


Figure 12: **WL5: four runs vs. the equal-weight baseline.** AC-terra (blue), AC-DS (pink), FM-terra (amber), FM-DS (green), EW-15 (gray); labels give terminal returns.

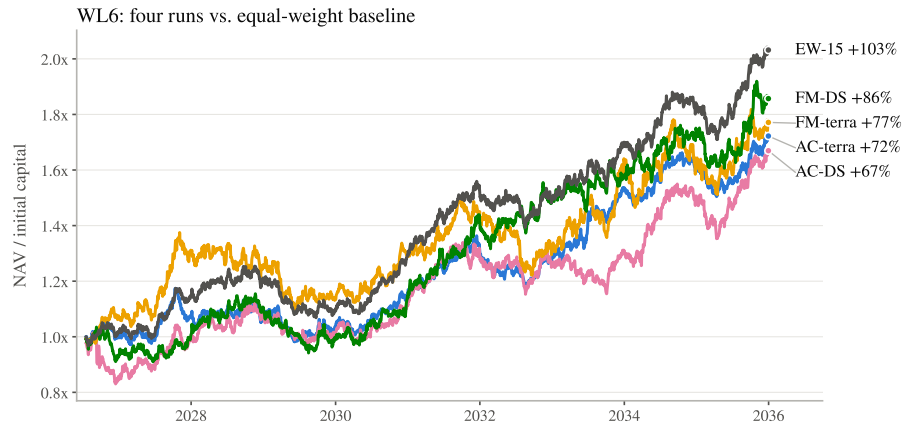


Figure 13: **WL6: four runs vs. the equal-weight baseline.** AC-terra (blue), AC-DS (pink), FM-terra (amber), FM-DS (green), EW-15 (gray); labels give terminal returns.

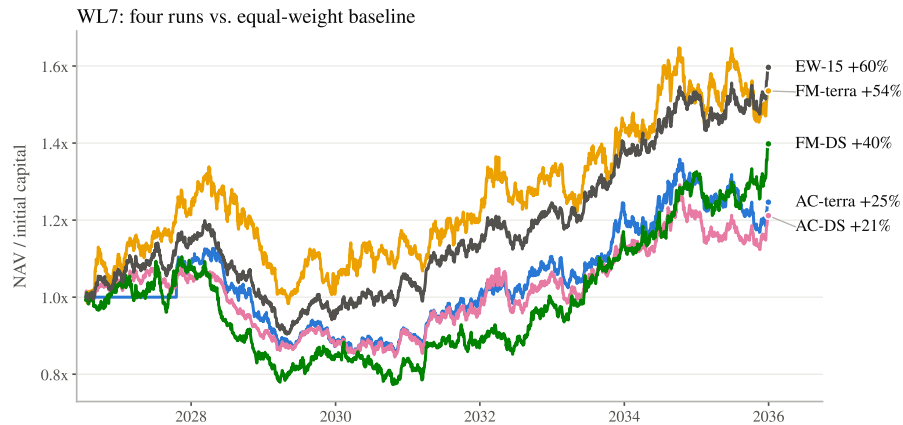


Figure 14: **WL7: four runs vs. the equal-weight baseline.** AC-terra (blue), AC-DS (pink), FM-terra (amber), FM-DS (green), EW-15 (gray); labels give terminal returns.

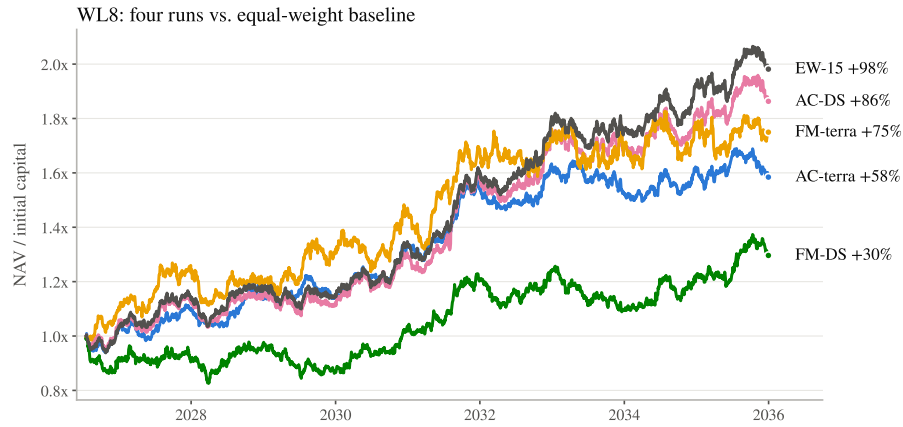


Figure 15: **WL8: four runs vs. the equal-weight baseline.** AC-terra (blue), AC-DS (pink), FM-terra (amber), FM-DS (green), EW-15 (gray); labels give terminal returns.

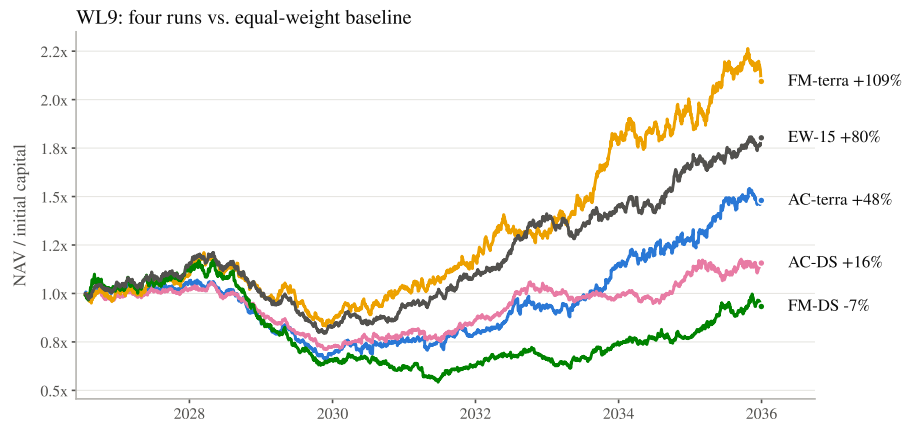


Figure 16: **WL9: four runs vs. the equal-weight baseline.** AC-terra (blue), AC-DS (pink), FM-terra (amber), FM-DS (green), EW-15 (gray); labels give terminal returns.

B Per-Run Risk-Adjusted Metrics

Daily-series metrics per run ($r_f = 0$; AC series are the daily reconstructions of Appendix A). Median refers to the across-worldline median of each column.

	AC-terra	AC-DS	FM-terra	FM-DS	EW-15
WL1	0.41	0.52	0.37	0.27	0.58
WL2	0.53	0.41	0.61	0.35	0.82
WL3	0.46	0.58	0.67	0.14	0.60
WL4	0.10	0.39	0.10	0.23	0.37
WL5	0.50	0.55	0.81	0.09	0.70
WL6	0.67	0.50	0.54	0.64	0.92
WL7	0.28	0.26	0.41	0.35	0.63
WL8	0.66	0.81	0.54	0.28	0.92
WL9	0.36	0.18	0.70	-0.06	0.72

Table 4: **Sharpe ratio per worldline** (daily NAV, annualized, $r_f = 0$).

	AC-terra	AC-DS	FM-terra	FM-DS	EW-15
WL1	-28.4%	-29.0%	-35.4%	-34.1%	-31.2%
WL2	-16.9%	-19.1%	-27.0%	-27.8%	-14.2%
WL3	-14.2%	-25.7%	-27.5%	-28.0%	-26.8%
WL4	-26.2%	-18.6%	-25.0%	-32.7%	-22.6%
WL5	-10.3%	-13.0%	-14.7%	-26.9%	-12.7%
WL6	-16.5%	-18.1%	-20.7%	-18.5%	-14.8%
WL7	-24.5%	-21.6%	-26.6%	-30.1%	-24.6%
WL8	-9.1%	-10.3%	-12.3%	-16.9%	-11.3%
WL9	-38.3%	-32.8%	-31.4%	-53.8%	-34.5%

Table 5: **Maximum drawdown per worldline** (daily NAV).

B.1 Run-status bookkeeping and NAV conventions

The primary unit of analysis is a completed run identified by agent, model, worldline, and seed. Every reported series—agent NAV, holdings, the equal-weight baseline, and every figure in this paper—shares one horizon: each run terminates at 2035-12-31, with AC NAV reconstructed *daily* from each executed rebalance’s post-trade account value and target weights marked on the worldline panel closes and anchored at the horizon’s terminal account value, and FM and baseline series marked daily. Every run is assigned a status of *complete*, *partial*, *running*, *pending*, or *failed*; partial and running artifacts are excluded from headline comparisons. Released tables carry a row-level manifest so that missing combinations are visible rather than silently dropped.

For drawdown, $DD_t = V_t / \max_{s \leq t} V_s - 1$ and $MDD = \min_t DD_t$. Cumulative transaction cost is audited per run as $C_{bps} = 10^4 \sum_t c \Delta_t / V_0$ over one-way migrated notional Δ_t at rate $c=3$ bps, exactly the cost leg of the gate contract of \$4.3. Robustness is summarized across the nine worldlines with the median, interquartile range, minimum, and maximum of each metric; statistical intervals and paired tests are computed from completed run units only and never treat decision blocks from one worldline as independent replications. A configuration field is marked *observed* only when present in the recorded request or runner—never substituted from an official provider default—so comparisons are not silently based on undocumented defaults.

Execution and fill conventions. At a decision date t , every observation is masked to dates $\leq t$, i.e., through the close of t . FM emits its target after the observation finalizes and fills at the next trading mark; since the forward panels carry no overnight gap (each forward open equals the previous close, Appendix C), FM’s fill price is the last close it observed. AC’s native trader fills at the open of the decision day itself—the close of $t - 1$ —so the change in weights it executes is exposed to the $t-1 \rightarrow t$ close-to-close span, whose endpoint the agent had already observed; positions are then marked at the close of t . We quantify this same-day overlap for every AC run by applying each executed rebalance’s weight change to its decision day’s close-to-close return and summing over the horizon: the cumulative effect averages -0.8 percentage points of initial capital (per-run range $[-4.8, +4.6]$, positive in only 5 of 18 runs)—sign-inconsistent and small relative to the reported

return spread, so no reported ranking or comparison changes. The EW-15 baseline rebalances at the close of each grid date under static weights, so its fills involve no information.

C Worldline Panels, Generator Details, and Data Quality

This appendix records the artifact-level details behind the benchmark of §3: panel construction, the full stage table, and the input-quality boundaries released with the panels.

Panel construction. Each of the nine final panels carries 83,347 daily rows over 2020-01-01–2035-12-31 with no duplicated (asset, date) keys; the forward window opens on 2026-07-16 and closes on 2035-12-31 (~2,468 trading days per WL). Panels are byte-identical across all 36 runs; WL1–WL8 regenerate from the released warmup and waypoint files to floating-point tolerance, while WL9’s third stage carries no parseable endpoint in its source and is therefore distributed as the panel itself with an explicit extended anchor.

Input-quality boundaries. Released alongside the panels, rather than silently patched, are two input-quality boundaries: seven online-synthesized series (CN10Y, DXY, EURUSD, SOX, US10Y, USDCNY, USDJPY) carry empty high/low fields where the warmup window ends in zero-amplitude bars, and weekend stage endpoints (e.g., WL7’s 2029-06-30) are anchors, not tradable dates. Open/close consumers are unaffected; full-OHLC consumers must treat those fields as absent.

Forward observation schema. Forward (post-boundary) rows are fully determined by the generated close path and by pre-boundary statistics; Table 6 states the rule per field. The forward panel therefore contains no information channel beyond the close path: open, high, low, volume, amount, and vwap are deterministic functions of the closes and of warmup statistics, so factor expressions built on them degenerate to close-derived signals in the forward period (vwap equals the close exactly). The yield series are synthetic yield-level instruments: their level path preserves the worldline’s basis-point moves anchored at the boundary level, a unit long position tracks the level (directionally analogous to a short-duration position; P&L is the percentage change of the level), volume is zero, and high/low are undefined as noted above.

Field	Warmup (real data)	Forward (synthetic)
close	observed	generator output (Eq. 1, re-anchored)
open	observed	previous close
high / low	observed	$\max / \min(\text{open}, \text{close}) \times (1 \pm \text{rng}/2)$
volume	observed	per-asset warmup median (constant)
amount	observed	$\text{close} \times \text{volume}$
vwap	amount / volume	equals close
PE, PS, PB, DYR	observed	undefined forward

Table 6: **Observation schema, warmup vs. forward.** rng is the per-asset median of (high – low)/close over the last 60 pre-boundary days; high/low are undefined for the seven zero-amplitude series listed above, and volume is zero (hence vwap undefined) for the yield, FX, and VIX series.

Scenario provenance artifacts. The original two debate reports are released with SHA-256 pins. Of the unified 2026-07 baseline anchors, three instruments (000300.SH, SPX, HSI) are explicitly designated platform (Suntime) values and the remainder are market-checked estimates. The scenario layer attributes anchors and event narratives to an agent-team debate workflow for which no executable platform export has been located; we therefore release the anchors as declared provenance artifacts rather than as a verified third-party implementation, and any future executable stage must add its input, seed, and output hash to the manifest. The full post-debate score matrix is in Appendix G.

Full stage table. Table 7 lists, for each worldline, the stage terminal dates and narratives parsed from the machine-readable waypoint file that accompanies each panel; Table 1 in the main text is its compact form.

WL	Stage end dates	Stage narratives (in order)
WL1	27-12, 28-05, 29-12, 31-12, 35-12	geopolitical escalation: red-line warnings → 72-hour blitz → sanctions & global restructuring → new equilibrium → new normal
WL2	27-12, 29-12, 31-12, 35-12	internet splintering → fintech-system splintering → bipolar consolidation → bipolar digital economy
WL3	27-12, 28-06, 29-12, 31-12, 35-12	nuclear test escalation → conventional conflict → deterrence failure & restructuring → non-proliferation rebuild → new nuclear order
WL4	27-12, 29-12, 31-12, 35-12	cracks spreading → Minsky moment → stagflation & de-dollarization → early institutional rebuild
WL5	27-12, 29-12, 31-12, 35-12	cracks emerge → JPY disorderly-devaluation spiral → Asian currency fragmentation → Japan's "new normal"
WL6	27-12, 29-12, 31-12, 35-12	inflation re-ignition → central-bank credibility collapse → entrenched stagflation → new monetary order
WL7	27-12, 29-06, 30-12, 32-12, 35-12	fragility build-up & first resonance → homogenization culminating in an algorithmic run on banks → regulatory rebuild → AI-regulation paradigm → new equilibrium
WL8	27-12, 29-12, 31-12, 33-12, 35-12	Middle-East escalation → Hormuz blockade & energy-system collapse → violent energy restructuring → new-energy order → end of the fossil era
WL9	27-12, 29-12, —, 33-12, 35-12	pre-pandemic growth → XPV-27 outbreak & labor collapse → deep recession, automation sprouts [†] → automation/AI supercycle → digital-economy normal

Table 7: **The nine worldlines: full stage table.** Stage terminal dates (YY-MM) and narratives, parsed from the machine-readable waypoint file that accompanies each panel. WL9's third stage carries no explicit date in the waypoint parser ([†]; boundary inferred between 2029-12 and 2033-12). All WLS end 2035-12-31.

D Window-Level Persistence Analysis

Section 5.2 compares agents against the EW-15 baseline on terminal returns, which yields only nine paired comparisons per agent column. This appendix repeats the comparison at decision granularity as a *temporal persistence diagnostic*: it tests whether the baseline’s advantage arises only from a few terminal episodes or accumulates steadily across the horizon. The shared ten-trading-day decision grid partitions each run into 245 non-overlapping windows; for each window we compare the agent’s realized window return (net of costs, from the daily NAV series of Appendix B) with EW-15’s window return over the same dates, a *win* being a strictly higher agent return. This yields 2,205 paired comparisons per agent column and 8,820 in total. Windows within a worldline are serially dependent—strategy state and market regimes persist across windows—so inference clusters at the worldline level: all results below use a cluster bootstrap that resamples the nine worldlines with replacement (10^4 resamples, fixed seed); because the nine worldlines are fixed scenarios rather than draws from a population, the resulting intervals and one-sided p -values (null of no EW-15 advantage) are descriptive scenario-resampling summaries.

On terminal returns themselves, exact one-sided sign tests across the nine worldlines, treating worldlines as the unit of aggregation, give $p = 0.002$ for AC-terra and FM-DS, $p = 0.020$ for AC-DS, and $p = 0.50$ for FM-terra (n.s., consistent with its 4/9 record); pooling all 36 paired comparisons with the same worldline-clustered bootstrap yields a mean EW-agent spread of +21.8 percentage points (95% CI [+14.1, +29.1], $p < 10^{-3}$).

Agent column	Window hit rate (%, 95% CI)	Mean excess / window (bps, 95% CI)	p
AC-terra	45.7 [44.0, 47.4]	−8.0 [−8.9, −7.0]	$< 10^{-4}$
AC-DS	47.8 [45.5, 49.9]	−5.6 [−9.9, −2.0]	$< 10^{-3}$
FM-terra	50.3 [48.9, 51.6]	+2.4 [−1.9, +7.3]	0.86 (n.s.)
FM-DS	47.0 [45.4, 48.7]	−10.3 [−14.9, −5.9]	$< 10^{-4}$
Pooled	47.7 [47.1, 48.3]	−5.4 [−7.0, −3.8]	$< 10^{-4}$

Table 8: **Per-window comparison against EW-15** (245 non-overlapping ten-trading-day windows per run; 8,820 paired comparisons). Hit rate is the share of windows the agent beats EW-15; excess is the agent’s window return minus EW-15’s, in basis points. CIs and p -values are worldline-clustered bootstrap (one-sided, null of no EW-15 advantage).

Three columns lose the window-level comparison decisively: hit rates seven-to-three points below one half and strictly negative mean excess, with confidence intervals excluding zero by a wide margin (Table 8). FM-terra is the exception—its hit rate is statistically indistinguishable from a coin flip and its positive mean excess per window (+2.4 bps) is not significant, consistent with its 4/9 endpoint record (the only column that beats EW-15 overall). The per-window gaps compound to the endpoint spread: AC-terra’s −8.0 bps per window over 245 windows reproduces its terminal gap almost exactly, so the window-level and endpoint views tell one story. The window-level view additionally shows *when* the baseline wins: no agent column shows statistically reliable short-horizon outperformance over EW-15 (FM-terra’s point estimate is slightly positive but not significant), indicating that the endpoint gaps of \$5.2 are not artifacts of a few large terminal drawdowns but accumulate steadily across the horizon.

E Staged Factor Inventory

The factor lists are staged by lifecycle phase: warmup libraries are shared across the nine worldlines of an experiment and would otherwise be repeated ninefold, so we separate (i) each experiment’s warmup qualified library (2020-01-01–2026-07-15), (ii) factors admitted online into a terminal live library (at most capacity 30), and (iii) fast-screen-passed mining candidates that never passed the in-library correlation gate. Table 9 summarizes the counts; Tables 10–11 give representative factors with their exact formulas; the complete machine-readable inventory is released as described below.

	AC–terra	AC–DS	FM–terra	FM–DS
Warmup qualified library (enters every WL)	6	4	21	8
Online-admitted terminal members listed per WL*	270 (30 per WL; all distinct)	99 (3–20 per WL; 94 distinct)	11 (on 7 WLs)	0
Terminal live-library size across WLs	30 (at capacity)	0–22	21–24	8 (warmup only)
Mining-screen candidates ($\Sigma \geq 3$), never admitted	—	—	—	109 (Σ 824)

Table 9: **Factor lifecycle in numbers.** *AC–DS’s WL8 is the citation-listed exception: its library was evicted to zero before the horizon (§5.3). FM–DS admitted nothing online; its 109 unadmitted candidates accumulated $\Sigma = 824$ screenings across the nine mining streams. Every number regenerates from the released factor-lifecycle and factor-library audit logs.

Experiment	Factor	IC / ICIR	Formula
AC–terra	peer_median_leadlag_5d	+0.056 / 0.17	median({return_5d[j] for j != i})
AC–terra	miner_3_clv_1d	+0.068 / 0.19	-(2 * (close - low) / (high - low) - 1)
AC–DS	mom_10d_skip5	+0.041 / 0.12	close.shift(5) / close.shift(15) - 1.0
AC–DS	vix_beta_cond_60x20	-0.038 / -0.09	-beta(asset_ret, VIX_ret, 60) * (VIX/VIX.shift(20) - 1.0)
FM–terra	return_rank_mean_reversion	+0.031 / 0.09	Neg(CsRank(TsRank(Sum(\$returns, 5), 30)))
FM–terra	intraday_vwap_pressure	+0.028 / 0.09	CsZScore(Mul(Div(Sub(\$close, \$vwap), Add(Abs(\$vwap), 0.001)), Div(Sub(\$vwap, \$open), Add(Abs(\$open), 0.001))))
FM–DS	decayed_return_signal	+0.033 / 0.09	CsRank(Decay(\$returns, 10))
FM–DS	cumulative_amount_per_share	+0.028 / 0.11	CsZScore(Div(CumSum(\$amt), Add(CumSum(\$volume), 1)))

Table 10: **Representative warmup factors** (IC/ICIR from the warmup evaluation; each library seeds all nine worldlines of its experiment). The complete lists of 6/4/21/8 factors are in the released manifest.

E.1 Complete machine-readable inventory

The tables above are an excerpt. The released manifest carries every factor admitted by any run—id, exact formula, category, warmup IC/ICIR, per-decision citation spans, and admission/eviction timestamps with machine-readable reasons—in the per-run factor-lifecycle logs and library audit trails shipped with the trajectories. One gap is exposed, not hidden: ten AC–terra ids cited by decisions have no retained artifact, and the manifest marks them as missing. The full tables (all 6/4/21/8 warmup factors, the 270, 99, and 11 online memberships per WL, and all 109 supplement candidates with formulas) and every count of Table 9 regenerate from the archived data with the released scripts.

Stage / stream	Factor	Evidence	Formula
AC-terra, WL1	miner_1_20281116_defensive_relative_lead_20d	cited 29-04-19~35-12-21 (144 dec.)	(\$asset_return_20d - mean(XAU, US10Y, CN10Y 20d)) * (0.25 + 0.75*I(breadth_20d < 0.40)), lagged
AC-terra, WL1	macro_stress_resilience_20d	cited 29-12-13~35-12-21 (100 dec.)	shift1((r20 - median_cs(r20)) / (downside_vol20 + 0.02) * (1 + 0.35*clip((VIX-median60(VIX))/std60(VIX), -1, 2)*tanh(5*(r20-median_cs(r20)))))
AC-DS, WL2	spx_corr60	cited 26-08-13~35-12-06 (198 dec.)	close.pct_change().rolling(60).corr(close['SPX'].pct_change()))
AC-DS, WL3	cn10y_beta_60	cited 26-07-30~35-12-20 (227 dec.)	rolling_beta(asset_daily_ret, cn10y_yield_daily_change, 60)
FM-terra, WL4	range_adjusted_return_rank	admitted 2026-08-01	CsRank(Div(TsRank(\$returns, 13), Add(TsRank(Div(Sub(\$high, \$low), Add(Abs(\$close), 0.001)), 13), 0.001)))
FM-terra, WL5	volume_acceleration_pressure	admitted 2026-08-03	CsZScore(Delta(EMA(Log(Add(\$volume, 1))), 6), 14))
FM-DS (mining)	return_residual_on_volume	$\Sigma = 43$	CsRank(Resid(\$returns, \$volume, 20))
FM-DS (mining)	dema_ema_momentum	$\Sigma = 33$	CsRank(Sub(DEMA(\$close, 10), EMA(\$close, 20)))
FM-DS (mining)	kama_cross_signal	$\Sigma = 17$	CsRank(Sub(KAMA(\$close, 20), KAMA(\$close, 5)))

Table 11: **Representative online admissions and mining candidates.** AC spans show first–last decision citing the factor (n decisions); FM-terra gives the admitting WL and timestamp; mining rows give the total screen count Σ across the nine FM-DS mining streams (all failed the correlation gate; none was ever admitted).

F Process Registers and the WL8 Forensic Note

§5.3 evaluates each run’s coherence across three states—research state, declared decision state, executed state. This appendix records the observables and registers behind that evaluation and reconstructs the WL8 degeneration at mechanism level. It is a forensic record rather than a new result: every fact regenerates from the released decision, ensemble, and execution traces.

F.1 Process observables and registers

Each run is summarized by five trajectory-level observables:

$$\begin{aligned} L_t &= |\text{factor library}_t|, & C_t &= \text{library churn rate}, & A_t &= \text{new-factor admission rate}, \\ D_t &= \|w_t^{\text{declared}} - w_t^{\text{executed}}\|_1, & F_t &= \|w_t - w_t^{\text{fallback}}\|_1, \end{aligned} \quad (2)$$

where L_t is the live factor-library size, C_t and A_t the per-window velocity of its composition (admissions plus evictions), D_t the distance between the portfolio weights the agent *records* and those it *executes*, and F_t the distance of executed weights from the equal-weight fallback policy. All five are released as trajectory series. Table 12 complements them with the terminal values of the raw registers from which they derive: the two event ledgers (*dec* and *reb*, below), cumulative transaction cost, and terminal library size versus ensemble selection.

WL	AC-terra				AC-DS			
	dec	reb	cost	lib→ens	dec	reb	cost	lib→ens
1	43	205	30.9	30→8	35	212	100.3	14→6
2	121	159	33.8	30→9	47	200	102.8	22→5
3	294	134	31.1	30→7	19	228	71.8	22→10
4	98	45	16.1	30→5	64	187	45.2	3→2
5	76	159	37.8	30→7	235	12	3.7	12→0
6	157	90	43.1	30→6	42	244	125.8	15→8
7	174	30	4.2	30→6	255	16	3.8	8→7
8	185	57	18.0	30→5	202	47	7.9	0→3
9	120	121	26.8	30→10	100	135	30.2	20→10

Table 12: **AC process registers.** *dec* = declined proposals: each cycle’s proposed portfolio shift is ruled on by comparing its expected gross edge against friction (migration) cost, and every *dec* entry is a rejection by that ruling (tagged `gross_edge_not_above_migration_cost`, unchanged target, or missing forecast); each entry carries the proposed weights, forecasts, and factor ids. *reb* = executed rebalances: the initial build plus every proposal the gate accepted, each with its transferred notional and realized cost. The two registers partition the proposal stream under the shared gate (see below); *cost* = cumulative transaction cost in bps of initial capital over 9.5 years (3 bps on migrated notional, 0 on the initial build); *lib→ens* = terminal live-library size → ensemble-selected factors. FM terminal libraries hold 21–24 factors (terra) and exactly the 8 warmup factors (DS) in every WL; the staged inventory—warmup libraries, online admissions, and the DS mining supplement—is in Appendix E.

One gate, two ledgers. The migration gate of \$4.3 is the single ruling instance in both frameworks, and the two registers record its two outcomes. A rejected proposal is appended to *dec* with its skip reason and holdings are untouched: all 43 proposals in AC-terra WL1, all 174 in WL7, and all 294 in WL3 were declined, which is why *dec* can exceed *reb* without any trade having occurred. A proposal the gate accepts executes and is appended to *reb* with its transferred notional and realized cost. The only executions outside the gate are the initial build (once per run) and the executor’s self-healing full-investment repair, which is tagged as such and never occurred in the released runs; *reb* $\gg$ *dec* therefore marks cycles that proposed and were approved, never trades that bypassed a ruling. Recording both outcomes separately keeps such divergences visible; the WL8 forensic below dissects the extreme case of a run that mostly declined to trade.

F.2 Forensic note: anatomy of the silent degeneration

§5.3 reports the AC–DS WL8 degeneration in outline; this note reconstructs the mechanism-level chain that the row $0 \rightarrow 3$ in Table 12 summarizes.

Research state: the library evicted to zero ($L_t \rightarrow 0$). The live factor library was progressively emptied by the admission contract’s own quality arbitration: each eviction removed the lower- q member of a conflicting pair, so the library contracted as the agent shed factors that no longer cleared the diversity/quality bar. In the agent’s own decision logic this is deliberate pruning—actively deleting factors whose quality no longer meets the bar, to cut losses from weak signals early rather than let them drag the portfolio; a majority of audit cycles already recorded $\text{kept} = 0$ well before the horizon, so the terminal empty library was not an end-state accident but the endpoint of a sustained trajectory. The pruning nonetheless left the agent without live signals.

Declared state: a frozen, unreachable ensemble ($D_t > 0$). Decision records nonetheless kept citing a three-factor ensemble. The cited file matched a hardcoded fallback ensemble last written in 2031 and was never updated; a second ensemble, written to a different directory, never reached the trader. The writer/reader path divergence means the declared portfolio—built from the three factor ids—corresponded to no executed configuration: $w_t^{\text{declared}} \neq w_t^{\text{executed}}$ on the remaining horizon.

Executed state: exact equal weight ($F_t \rightarrow 0$). With the preference mapper’s 22% single-asset cap binding on every one of the fifteen assets, all weights clip and renormalize to exactly $1/15$. For nearly the entire horizon the run therefore held a basket numerically identical to the EW-15 policy while its records showed three factor ids—execution had behaviorally converged to the fallback and $F_t = 0$ on those intervals. On WL8 that policy returned +98.2% against the run’s +86.3%.

Reading the row. Table 12 lists the WL8 AC–DS row as $0 \rightarrow 3$ (terminal live-library size $\rightarrow$ ensemble-selected factors) across 202 recorded decision entries and 47 executed rebalances. The three registers this note dissects—research state, declared state, and realized holdings—are precisely the ones the evaluation contract records separately, which is what makes the dissociation visible at all.

G Scenario Provenance: Post-Debate Review Scores

The scenario set of §3 was produced by the debate workflow of §3.4. Each scenario was independently derived by one of the team’s three members—strategy (policy) / macro-finance / chief analyst (synthesis)—and then cross-reviewed by the full team through three debate rounds, scoring every scenario on logical rigor, 10-year cumulative probability, and shock severity. Table 13 releases the full score matrix. The scores document how the scenario set was produced; the benchmark itself assigns no likelihood weights.

WL	Theme	Rigor (/5)	10-yr prob. (%)	Severity (/10)
1	geopolitical escalation	3	10–15	9
2	tech “silicon curtain”	4	40–50	7.5
3	nuclear-proliferation chain	3	8–12	8.5
4	Treasury “Minsky moment”	3.5	18–25	10
5	JPY devaluation spiral	3	15–22	7
6	stagflation & CB collapse	3	10–15	8
7	algorithmic run on banks	3.5	18–25	8
8	energy-system collapse	4	12–18	9
9	pandemic supply-chain rebuild	4	10–18	9.5

Table 13: **Post-debate review scores per worldline**, from the released debate matrix of the scenario workflow: logical rigor (5-point scale), elicited 10-year cumulative probability, and shock severity (10-point scale).

H Model Configuration

All 36 runs share the configuration of Table 14: both models operate at reasoning effort medium, with sampling parameters left at provider defaults.

Model	Reasoning effort	Temperature	top_p
gpt-5.6-terra	medium	default	default
deepseek-v4-flash	medium	default	default

Table 14: **LLM configuration for all runs.**