Project X: A Falsification-First Approach to AI Trading Research

From failed candidates to a frozen XAUUSD shadow trial

Generated: 2026-06-23T11:19:30Z

Repository / project: AceFX-AI / Project X

Current phase: Project X Shadow Trial One

From failed candidates to a frozen XAUUSD shadow trial

Current phase: Project X Shadow Trial One

Status box

>

- Candidate 7: not created

- Demo/live: blocked

- Broker execution: blocked

- Official run state: prepared_not_started

- Official decisions: 0

- Official outcomes: 0

1. Executive Summary

Project X is not a live trading system. It is a falsification-first AI trading research project focused on XAUUSD and EURUSD. The project has spent most of its effort disproving ideas that looked attractive under weaker testing. That is the point of the process: weak ideas should die before they become dangerous.

Most branches failed. Candidate 4 looked promising internally and failed external validation. Price-derived indicators, support/resistance, multi-timeframe context, local structure, macro context, and economic surprise data all failed or were downgraded under stricter controls. A simple probabilistic sandbox also failed. The strongest surviving historical lead is a frozen XAUUSD H1 contextual-linear policy trained around risk-adjusted return and evaluated under conservative spread_p90 assumptions.

That lead has now moved into Project X Shadow Trial One. It will not trade. It will log future-only shadow decisions from local XAUUSD snapshots and resolve outcomes later. The trial requires at least 12 weeks and 200 resolved official outcomes before judgment. If the first real future snapshot were logged on the generation date of this paper, the earliest review date would be approximately September 15, 2026.

2. Why Project X Exists

Trading research has a basic credibility problem: many systems look good on historical data and fail when conditions change. Backtests can accidentally include future information, underestimate costs, tune too heavily to one feed, or reward chance patterns. The result is false confidence.

Project X exists to make that harder. It treats an attractive backtest as the beginning of an investigation, not the end. A strategy must survive simple baselines, leakage checks, realistic costs, external validation where possible, and future-only evidence collection before it deserves more serious consideration.

For a trader, the analogy is simple: a backtest is a practice exam. Shadow logging is a real exam without money. Demo or live trading comes only after future evidence improves, and even then only through a separate review.

3. Research Philosophy

- **Backtest:** a historical simulation. Useful, but never proof.
- **Walk-forward validation:** train on earlier data and test on later data, repeating through time.
- **Baselines:** simple comparison models that complex models must beat.
- **Shuffled controls:** tests that break the relationship between features and outcomes to check whether the apparent edge disappears.
- **Adversarial controls:** stronger challenges that ask whether a simpler or safer explanation beats the proposed model.

- **Leakage checks:** controls that make sure future information is not accidentally visible at decision time.
- **Cost realism:** accounting for spread, slippage, adverse entry/exit pricing, and broker/feed differences.
- **External validation:** testing whether an idea survives outside its original data source.
- **Shadow logging:** recording what a frozen model would have done on future data without placing trades.

4. Chronological Research Story

Candidate 4 internal promise

Goal: Test whether a locked early machine-learning candidate could produce a repeatable XAUUSD process.

What we tested: A session-sensitive classifier with internal historical process metrics.

What looked promising: Internal results looked attractive: roughly 803 trades, AUC near 0.5276, profit factor near 1.4111, and positive historical net result.

What failed: Internal promise did not prove portability.

Decision: REVIEW before external test

Lesson learned: A strong internal result is only a reason to test harder.

Candidate 4 external validation

Goal: Check whether Candidate 4 survived outside its original feed and cost environment.

What we tested: The locked candidate behavior on alternate external data.

What looked promising: The design was fixed before the external test, which made the validation meaningful.

What failed: External validation produced negative net result, and the behavior looked feed-specific or execution-sensitive.

Decision: FAIL

Lesson learned: A candidate that fails external validation should not be promoted, even if the internal backtest was attractive.

Project X v2 price/state framework

Goal: Build stricter walk-forward tooling before more model search.

What we tested: Random-policy checks, walk-forward folds, promotion gates, leakage controls, and simple evaluation plumbing.

What looked promising: The framework made later research more auditable and harder to game.

What failed: The framework itself did not create a predictive edge.

Decision: PASS infrastructure

Lesson learned: Better process is valuable even when it rejects more ideas.

Volatility and trend/range audits

Goal: Test whether market state transitions had predictive information.

What we tested: Trend, range, volatility-regime, unfavorable-condition, and spread/execution labels.

What looked promising: Some transition-style targets produced near-miss historical results.

What failed: Simple current-state and persistence baselines explained too much of the result.

Decision: FAIL / closed

Lesson learned: A complex model must beat the obvious current-state explanation.

Higher-timeframe near-miss audit

Goal: Check whether H1/H4 context added value beyond M15-only features.

What we tested: Higher-timeframe context and near-miss ensemble features.

What looked promising: The strongest XAUUSD H1 low-to-high volatility result reached AUC around 0.7923 versus FX-only around 0.5035.

What failed: Later disproof showed the result was baseline-explained, with trendiness_8 becoming the mandatory adversarial baseline.

Decision: FAIL after disproof

Lesson learned: A large AUC can still be the wrong story if a simpler baseline explains it.

Baseline registry and trendiness_8 lesson

Goal: Formalize clean simple baselines before further feature expansion.

What we tested: Whether trendiness_8 was past-only and whether transition targets were hybrid current-state/future labels.

What looked promising: The baseline registry clarified what every future model had to beat.

What failed: Some target families were not pure future labels and had to be interpreted more carefully.

Decision: REGISTER / REVIEW

Lesson learned: Strong baselines are not optional; they are the first adversary.

Multi-timeframe indicator and support/resistance audit

Goal: Test common trader-style features under strict no-leak rules.

What we tested: Technical indicators, H1/H4 context, objective support/resistance, and news context.

What looked promising: Some pre-correction results looked interesting, especially around transition-style targets.

What failed: After stronger baselines and target hygiene checks, the branch did not show robust incremental value.

Decision: FAIL

Lesson learned: Popular trader concepts still need the same validation standards as any model feature.

Local structure disproof and S/R leakage fixes

Goal: Disprove the narrow local-structure lead after support/resistance corrections.

What we tested: Delayed-confirmed swing levels and session-to-date high/low features.

What looked promising: The branch fixed a real leakage issue in prior support/resistance features.

What failed: After correction, XAUUSD H1 direction_next_4 reached AUC around 0.5258 versus a simple baseline around 0.5264.

Decision: FAIL

Lesson learned: Leakage fixes can turn apparent edge into noise; that is a success for the research process.

Project X v3 price-derived closeout

Goal: Close the exhausted price-derived research universe.

What we tested: The accumulated evidence from indicators, support/resistance, higher-timeframe context, and local structure.

What looked promising: The closeout preserved useful fixes and baseline rules.

What failed: No price-derived branch qualified for promotion.

Decision: CLOSED / FAIL

Lesson learned: When a feature universe is exhausted, reopening it requires materially new information.

FRED macro context

Goal: Test whether daily macro context added information beyond FX-only features.

What we tested: Treasury, equity, volatility, and macro proxy features aligned conservatively.

What looked promising: The data was clean and leakage-controlled.

What failed: The strongest lift was only about +0.002 AUC, too small to matter.

Decision: FAIL

Lesson learned: Clean data can still be too coarse for the target horizon.

Intraday cross-asset and paid-data feasibility

Goal: Find practical data sources after the daily macro branch failed.

What we tested: ETF proxies, broker symbols, Dukascopy-style data, Alpha Vantage, and Polygon/Massive-style paid feeds.

What looked promising: Polygon/Massive-style paid data looked structurally practical.

What failed: Alpha Vantage acquisition was blocked by key/entitlement constraints and budget remained paused.

Decision: PASS planning / blocked acquisition

Lesson learned: Data quality and access are part of model feasibility.

Project X v4 new-information feasibility

Goal: Find a new information source after price-derived research closed.

What we tested: Economic surprise, cross-asset data, news text, options, order-book, and other low-cost paths.

What looked promising: Economic calendar surprise data was the top practical candidate.

What failed: The branch did not authorize modeling by itself.

Decision: REVIEW

Lesson learned: New information must first be acquired and aligned before it can be tested.

Economic surprise data feasibility

Goal: Test whether historical economic surprise data could be normalized and aligned safely.

What we tested: Local MT5-normalized calendar data and credential-gated TradingEconomics access.

What looked promising: Local coverage included 2,259 events and 1,482 calculable surprises.

What failed: Timestamp caveats and credential limits prevented a clean pass.

Decision: REVIEW

Lesson learned: Useful event data still needs timestamp trust.

Economic surprise incremental-value audit

Goal: Test whether surprise features added value beyond price baselines.

What we tested: Schedule features, surprise features, combined features, shuffled controls, and timestamp shifts.

What looked promising: EURUSD M15 future_realized_range for USD events reached AUC around 0.6453.

What failed: Only one usable fold carried the best result.

Decision: REVIEW

Lesson learned: One good fold is not a conclusion.

Economic surprise focused disproof

Goal: Kill or confirm the one meaningful economic-surprise review result.

What we tested: More folds, timestamp shifts, surprise shuffles, event decomposition, and controls.

What looked promising: The branch was narrow and properly targeted.

What failed: The best result collapsed below baselines.

Decision: FAIL

Lesson learned: Event data may be contextually useful, but this branch did not survive as a predictive edge.

Candidate 7 Sandbox v1 probabilistic model

Goal: Create a quarantined research framework without creating Candidate 7.

What we tested: Logistic-style probabilistic models, rolling folds, prediction logs, and calibration checks.

What looked promising: Infrastructure worked.

What failed: The best model had AUC around 0.5000 and no meaningful lift over simple baselines.

Decision: FAIL model / PASS infrastructure

Lesson learned: Good infrastructure is useful even when the first model fails.

Candidate 7 RL Sandbox v2

Goal: Test a controlled policy-learning approach under walk-forward validation.

What we tested: Contextual-linear policy, flat/long/short action space, fixed holding assumptions, and cost-aware rewards.

What looked promising: Initial net result was around +0.115875 after costs with 6 positive folds.

What failed: The result had to be treated as suspicious until simulator accounting was checked.

Decision: Historical PASS, later superseded

Lesson learned: A strong result is a reason to audit the simulator.

RL disproof and execution accounting issue

Goal: Reproduce and try to kill the first strong RL result.

What we tested: Execution timing, reward accounting, costs, feature causality, fold boundaries, and controls.

What looked promising: The original result reproduced exactly.

What failed: The simulator undercharged full round-trip execution costs.

Decision: FAIL original result

Lesson learned: Exact reproduction is not enough if the simulator is optimistic.

Corrected execution rerun

Goal: Rerun the same frozen setup after correcting execution accounting.

What we tested: Adverse entry/exit pricing and full round-trip cost accounting.

What looked promising: Corrected base net stayed positive at about +0.099180.

What failed: The result weakened and failed 3x cost stress.

Decision: REVIEW

Lesson learned: The lead survived correction, but cost sensitivity mattered.

Cost realism audit

Goal: Compare fixed cost assumptions to observed spread behavior.

What we tested: Dynamic historical cost variants, spread/slippage stress, external/alternate source checks, and stronger baselines.

What looked promising: The policy survived base and 2x cost conditions.

What failed: 3x cost removed the edge and base cost was classified as optimistic.

Decision: REVIEW

Lesson learned: A gold policy must carry enough spread margin.

H1 spread aggregation correction

Goal: Replace H1 last-spread cost proxy with conservative within-hour M15 spread aggregation.

What we tested: spread_last, spread_mean, spread_p75, spread_p90, spread_p95, spread_max, and break-even cost margin.

What looked promising: spread_p90, spread_p95, and spread_max all remained positive at about +0.086681.

What failed: The result was still historical only.

Decision: PASS narrow cost issue

Lesson learned: The lead survived a major cost realism correction.

Frozen readiness gate

Goal: Check whether the frozen spread_p90 lead was ready for shadow logger design.

What we tested: Frozen reproduction, pseudo-sealed holdout, alternate replay, and M15 micro-replay.

What looked promising: The frozen result reproduced at +0.086681 with positive pseudo-holdout and micro-replay checks.

What failed: True future-sealed evidence was still unavailable.

Decision: REVIEW_MORE_VALIDATION_REQUIRED

Lesson learned: Historical cleanliness can justify shadow logging, not trading.

Final adversarial battery

Goal: Rerun shuffled and adversarial controls under the final frozen setup.

What we tested: Shuffled features, shuffled reward, block shuffle, random actions, sign flip, feature dropout, and timing variants.

What looked promising: Shuffled/randomized controls passed cleanly.

What failed: A lagged-feature variant outperformed the frozen policy.

Decision: REVIEW_MORE_VALIDATION_REQUIRED

Lesson learned: An outperforming variant must be classified before it can be interpreted.

Adversarial failure forensics

Goal: Classify the outperforming lagged-feature variant.

What we tested: Variant validity, timing matrix, fold/trade diagnostics, and decision taxonomy.

What looked promising: The lagged variant was decision-time safe and reached about +0.130377.

What failed: It was not part of the frozen policy and could not rescue the lead.

Decision: ALTERNATE_VALID_POLICY_DISCOVERED

Lesson learned: A new hypothesis must get its own frozen validation path.

Corrected final battery taxonomy

Goal: Separate valid controls, invalid diagnostics, and alternate hypotheses.

What we tested: The original frozen policy under corrected taxonomy.

What looked promising: Valid controls stayed below the frozen policy.

What failed: True future evidence was still missing.

Decision: SHADOW_LOGGER_ALLOWED

Lesson learned: The correct next step was future-only logging, not promotion.

Shadow logger design

Goal: Build future-only logging infrastructure for the frozen policy.

What we tested: Design-only mode, historical dry run, single-snapshot dry run, schemas, manifests, and safety checks.

What looked promising: The logger wrote shadow decision and outcome records without broker or order functionality.

What failed: Dry-run records needed quarantine before real future logging.

Decision: PASS infrastructure

Lesson learned: Logging infrastructure must separate test data from future evidence.

Fixture quarantine and official run reset

Goal: Prevent fixture or validation records from contaminating future evidence.

What we tested: Fixture detection, quarantine, clean official run state, and real snapshot intake rules.

What looked promising: Official run state was reset to zero decisions and zero outcomes.

What failed: No performance evidence was created; this was an integrity gate only.

Decision: PASS infrastructure

Lesson learned: Future evaluation is only meaningful if the starting state is clean.

On-demand catch-up runner

Goal: Make manual future shadow logging practical without scheduling or broker access.

What we tested: Local snapshot validation, chronological catch-up, outcome resolution, status-only mode, dry-run mode, and evaluation lock.

What looked promising: The runner can process missed completed H1 bars and resolve eligible outcomes.

What failed: It does not judge performance before the sealed window completes.

Decision: PASS infrastructure

Lesson learned: The current phase is auditable evidence collection, not trading.

Project X Shadow Trial One

Goal: Collect future-only evidence for the frozen policy.

What we tested: Not yet judged; future local snapshots will be processed under the frozen logger protocol.

What looked promising: This is the first branch strong enough to justify future-only shadow logging.

What failed: The trial has not accumulated future evidence yet.

Decision: PREPARED

Lesson learned: The next question must be answered by future data.

5. Full Experiment Summary Tables

5.1 Failed or rejected branches

Branch	Instrument/timeframe	Method	Best result	Decision	Why it failed	Current status
Candidate 4 external validation	XAUUSD M15	The locked candidate behavior on alternate external data.	The design was fixed before the external test, which made the validation meaningful.	FAIL	External validation produced negative net result, and the behavior looked feed-specific or execution-sensitive.	Closed or downgraded
Volatility and trend/range audits	EURUSD and XAUUSD, M15/H1/H4	Trend, range, volatility-regime, unfavorable-condition, and spread/execution labels.	Some transition-style targets produced near-miss historical results.	FAIL / closed	Simple current-state and persistence baselines explained too much of the result.	Closed or downgraded
Higher-timeframe near-miss audit	XAUUSD H1 lead	Higher-timeframe context and near-miss ensemble features.	The strongest XAUUSD H1 low-to-high volatility result reached AUC around 0.7923 versus FX-only around 0.5035.	FAIL after disproof	Later disproof showed the result was baseline-explained, with trendiness_8 becoming the mandatory adversarial baseline.	Closed or downgraded
Multi-timeframe indicator and sup port/resistance audit	EURUSD and XAUUSD	Technical indicators, H1/H4 context, objective support/resistance, and news context.	Some pre-correction results looked interesting, especially around transition-style targets.	FAIL	After stronger baselines and target hygiene checks, the branch did not show robust incremental value.	Closed or downgraded
Local structure disproof and S/R leakage fixes	XAUUSD H1	Delayed-confirmed swing levels and session-to-date high/low features.	The branch fixed a real leakage issue in prior support/resistance features.	FAIL	After correction, XAUUSD H1 direction_next_4 reached AUC around 0.5258 versus a simple baseline around 0.5264.	Closed or downgraded
Project X v3 price-derived closeout	EURUSD and XAUUSD	The accumulated evidence from indicators, support/resistance, higher-timeframe context, and local structure.	The closeout preserved useful fixes and baseline rules.	CLOSED / FAIL	No price-derived branch qualified for promotion.	Closed or downgraded
FRED macro context	EURUSD and XAUUSD M15	Treasury, equity, volatility, and macro proxy features aligned conservatively.	The data was clean and leakage-controlled.	FAIL	The strongest lift was only about +0.002 AUC, too small to matter.	Closed or downgraded
Economic surprise focused disproof	EURUSD M15	More folds, timestamp shifts, surprise shuffles, event decomposition, and controls.	The branch was narrow and properly targeted.	FAIL	The best result collapsed below baselines.	Closed or downgraded
RL disproof and execution accounting issue	XAUUSD H1	Execution timing, reward accounting, costs, feature causality, fold boundaries, and controls.	The original result reproduced exactly.	FAIL original result	The simulator undercharged full round-trip execution costs.	Closed or downgraded

5.2 Infrastructure and process branches

Branch	Instrument/timeframe	Method	Best result	Decision	Why it mattered	Current status
Project X v2 price/state framework	EURUSD and XAUUSD	Random-policy checks, walk-forward folds, promotion gates, leakage controls, and simple evaluation plumbing.	The framework made later research more auditable and harder to game.	PASS infrastructure	Better process is valuable even when it rejects more ideas.	Infrastructure or process retained
Baseline registry and trendiness_8 lesson	EURUSD and XAUUSD	Whether trendiness_8 was past-only and whether transition targets were hybrid current-state/future labels.	The baseline registry clarified what every future model had to beat.	REGISTER / REVIEW	Strong baselines are not optional; they are the first adversary.	Infrastructure or process retained
Intraday cross-asset and paid-data feasibility	EURUSD and XAUUSD	ETF proxies, broker symbols, Dukascopy-style data, Alpha Vantage, and Polygon/Massive-style paid feeds.	Polygon/Massive-style paid data looked structurally practical.	PASS planning / blocked acquisition	Data quality and access are part of model feasibility.	Infrastructure or process retained
Candidate 7 Sandbox v1 probabilistic model	XAUUSD and EURUSD H1	Logistic-style probabilistic models, rolling folds, prediction logs, and calibration checks.	Infrastructure worked.	FAIL model / PASS infrastructure	Good infrastructure is useful even when the first model fails.	Infrastructure or process retained
Candidate 7 RL Sandbox v2	XAUUSD H1	Contextual-linear policy, flat/long/short action space, fixed holding assumptions, and cost-aware rewards.	Initial net result was around +0.115875 after costs with 6 positive folds.	Historical PASS, later superseded	A strong result is a reason to audit the simulator.	Infrastructure or process retained
H1 spread aggregation correction	XAUUSD H1	spread_last, spread_mean, spread_p75, spread_p90, spread_p95, spread_max, and break-even cost margin.	spread_p90, spread_p95, and spread_max all remained positive at about +0.086681.	PASS narrow cost issue	The lead survived a major cost realism correction.	Infrastructure or process retained
Adversarial failure forensics	XAUUSD H1	Variant validity, timing matrix, fold/trade diagnostics, and decision taxonomy.	The lagged variant was decision-time safe and reached about +0.130377.	ALTERNATE_VALID_POLICY_DISCOVERED	A new hypothesis must get its own frozen validation path.	Infrastructure or process retained
Shadow logger design	XAUUSD H1	Design-only mode, historical dry run, single-snapshot dry run, schemas, manifests, and safety checks.	The logger wrote shadow decision and outcome records without broker or order functionality.	PASS infrastructure	Logging infrastructure must separate test data from future evidence.	Infrastructure or process retained
Fixture quarantine and official run reset	XAUUSD H1	Fixture detection, quarantine, clean official run state, and real snapshot intake rules.	Official run state was reset to zero decisions and zero outcomes.	PASS infrastructure	Future evaluation is only meaningful if the starting state is clean.	Infrastructure or process retained

Branch	Instrument/timeframe	Method	Best result	Decision	Why it mattered	Current status
On-demand catch-up runner	XAUUSD H1	Local snapshot validation, chronological catch-up, outcome resolution, status-only mode, dry-run mode, and evaluation lock.	The runner can process missed completed H1 bars and resolve eligible outcomes.	PASS infrastructure	The current phase is auditable evidence collection, not trading.	Infrastructure or process retained

5.3 Surviving, review, and current branches

Branch	Instrument/timeframe	Method	Best result	Decision	Why it matters	Current status
Candidate 4 internal promise	XAUUSD M15	A session-sensitive classifier with internal historical process metrics.	Internal results looked attractive: roughly 803 trades, AUC near 0.5276, profit factor near 1.4111, and positive historical net result.	REVIEW before external test	A strong internal result is only a reason to test harder.	Current or future-testable
Baseline registry and trendiness_8 lesson	EURUSD and XAUUSD	Whether trendiness_8 was past-only and whether transition targets were hybrid current-state/future labels.	The baseline registry clarified what every future model had to beat.	REGISTER / REVIEW	Strong baselines are not optional; they are the first adversary.	Current or future-testable
Project X v4 new-information feasibility	EURUSD and XAUUSD	Economic surprise, cross-asset data, news text, options, order-book, and other low-cost paths.	Economic calendar surprise data was the top practical candidate.	REVIEW	New information must first be acquired and aligned before it can be tested.	Current or future-testable
Economic surprise data feasibility	EURUSD and XAUUSD M15/H1/H4	Local MT5-normalized calendar data and credential-gated TradingEconomics access.	Local coverage included 2,259 events and 1,482 calculable surprises.	REVIEW	Useful event data still needs timestamp trust.	Current or future-testable
Economic surprise incremental-value audit	EURUSD M15 lead	Schedule features, surprise features, combined features, shuffled controls, and timestamp shifts.	EURUSD M15 future_realized_range for USD events reached AUC around 0.6453.	REVIEW	One good fold is not a conclusion.	Current or future-testable
Corrected execution rerun	XAUUSD H1	Adverse entry/exit pricing and full round-trip cost accounting.	Corrected base net stayed positive at about +0.099180.	REVIEW	The lead survived correction, but cost sensitivity mattered.	Current or future-testable
Cost realism audit	XAUUSD H1	Dynamic historical cost variants, spread/slippage stress, external/alternate source checks, and stronger baselines.	The policy survived base and 2x cost conditions.	REVIEW	A gold policy must carry enough spread margin.	Current or future-testable
Frozen readiness gate	XAUUSD H1	Frozen reproduction, pseudo-sealed holdout, alternate replay, and M15 micro-replay.	The frozen result reproduced at +0.086681 with positive pseudo-holdout and micro-replay checks.	REVIEW_MORE_VALIDATION_REQUIRED	Historical cleanliness can justify shadow logging, not trading.	Current or future-testable

Branch	Instrument/timeframe	Method	Best result	Decision	Why it matters	Current status
Final adversarial battery	XAUUSD H1	Shuffled features, shuffled reward, block shuffle, random actions, sign flip, feature dropout, and timing variants.	Shuffled/randomized controls passed cleanly.	REVIEW_MORE_VALIDATION_REQUIRED	An outperforming variant must be classified before it can be interpreted.	Current or future-testable
Corrected final battery taxonomy	XAUUSD H1	The original frozen policy under corrected taxonomy.	Valid controls stayed below the frozen policy.	SHADOW_LOGGER_ALLOWED	The correct next step was future-only logging, not promotion.	Current or future-testable
Project X Shadow Trial One	XAUUSD H1	Not yet judged; future local snapshots will be processed under the frozen logger protocol.	This is the first branch strong enough to justify future-only shadow logging.	PREPARED	The next question must be answered by future data.	Current or future-testable

6. Why So Many Branches Failed

The repeated failures were not accidental. The research process was designed to expose fragile ideas. Price-derived features were usually explained by simple state or trend baselines. Support/resistance collapsed after leakage fixes. FRED macro context was clean but too low-frequency. Economic surprise data had useful coverage but failed the focused disproof branch. The first probabilistic sandbox model did not beat baselines. The first RL result looked strong, then required an execution-accounting correction before it could be considered again.

Every failure reduced false confidence. That matters more than preserving a flattering historical story.

7. Current Lead: Project X Shadow Trial One

The current lead is a frozen XAUUSD H1 contextual-linear policy. It uses a risk_adjusted_return reward, corrected adverse entry/exit execution accounting, and a conservative spread_p90 cost model. It is frozen: no retraining, no tuning, no feature changes, no filters, and no broker execution.

Confirmed historical audit numbers:

- Frozen spread_p90 net: `+0.086681`
- Lift vs always-flat: `+0.086681`
- Lift vs matched random: `+0.126016`
- Lift vs best simple baseline: `+0.103866`
- always_flat: `0.000000`
- random_matched_frequency: `-0.039335`
- simple_mean_reversion: `-0.017186`
- delayed_execution_one_bar: `+0.072991`
- delayed_execution_two_bar: `+0.037578`
- spread_last: `+0.099180`
- spread_mean: `+0.099594`
- spread_p75: `+0.092836`

- spread_p90: `+0.086681`
- spread_p95: `+0.086681`
- spread_max: `+0.086681`
- break-even cost multiplier: `2.786`
- break-even spread: `1.3032`
- observed p95 spread: `0.8000`

These are historical audit numbers, not performance promises. The lead is promising because it survived unusually strict historical audits, not because historical performance guarantees anything.

8. Why This Is Still Not Proven

Historical evidence can still be selected. Project X has tested many branches, so false discovery risk remains. A pseudo-holdout is not the same as future data. The only credible next question is whether the frozen model still behaves well on future data collected after the rules were locked.

9. What We Are Doing Now

Project X Shadow Trial One uses an on-demand catch-up runner. The user supplies a local XAUUSD snapshot file. The runner validates that file, logs missed completed H1 bars, records `shadow_action`, resolves outcomes when enough future bars exist, and reports progress. It refuses broker, live, order, scheduled, and auto-trading modes.

The evaluation remains locked until both gates are met: 12 calendar weeks and 200 resolved official outcomes. Current official count is 0 decisions and 0 outcomes.

Official catch-up:

```
python -m acefx_ai.cli project-x-candidate7-shadow-logger --official-catch-up --input-file
<REAL_LOCAL_SNAPSHOT> --confirm-real-future-snapshot --json
```

Status only:

```
python -m acefx_ai.cli project-x-candidate7-shadow-logger --status-only --json
```

Dry-run catch-up:

```
python -m acefx_ai.cli project-x-candidate7-shadow-logger --official-catch-up --input-file
<REAL_LOCAL_SNAPSHOT> --confirm-real-future-snapshot --dry-run --json
```

10. What Happens If Shadow Trial One Succeeds

Success does not automatically authorize trading. A strong future-only result would justify a sealed shadow evaluation review. If that review passes, the next possible step would be a Candidate 7 review branch, followed by a separate demo mirror design, demo-only execution validation, and execution-quality comparison. Live micro-risk discussion would come much later, and only after another explicit review.

11. What Happens If It Fails

Failure means the frozen policy is downgraded or closed. The logs should be preserved, the failure mode should be documented, and the team should avoid emotional rescue tuning. The lag_features_one_bar variant can remain a separate future hypothesis, but it cannot rescue Shadow Trial One retroactively.

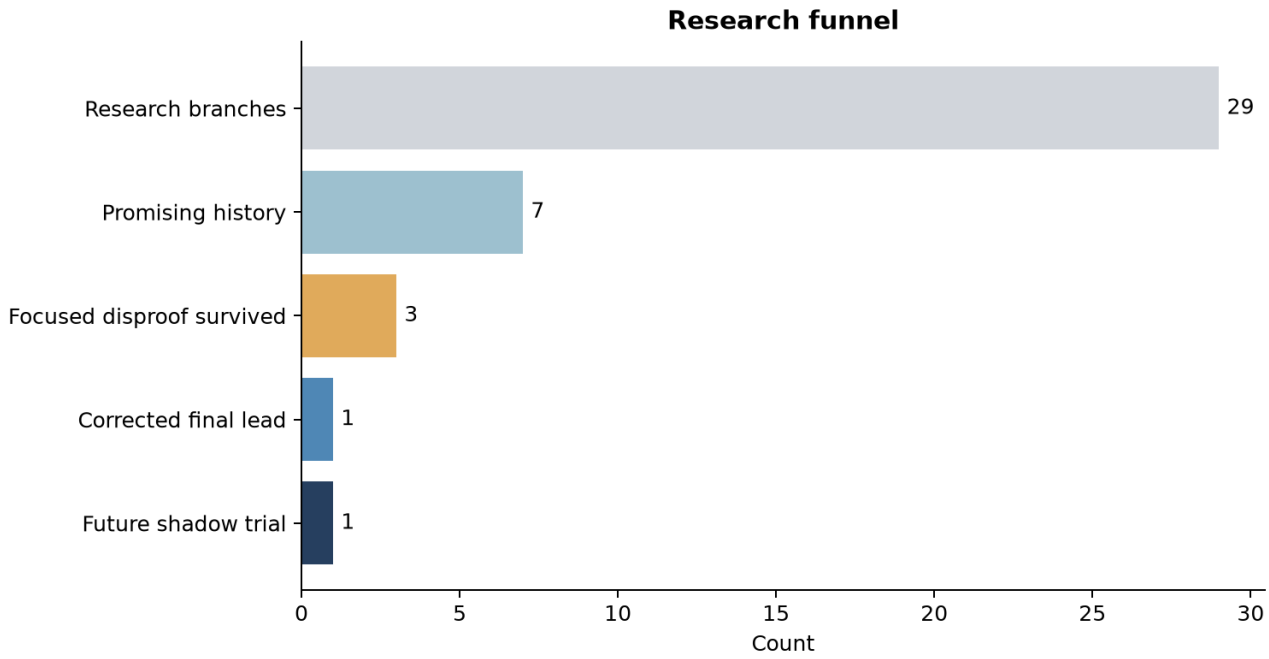
12. Investor-Relevant Takeaways

- The value is not just one model; it is the validation pipeline.
- Weak ideas were killed instead of marketed.

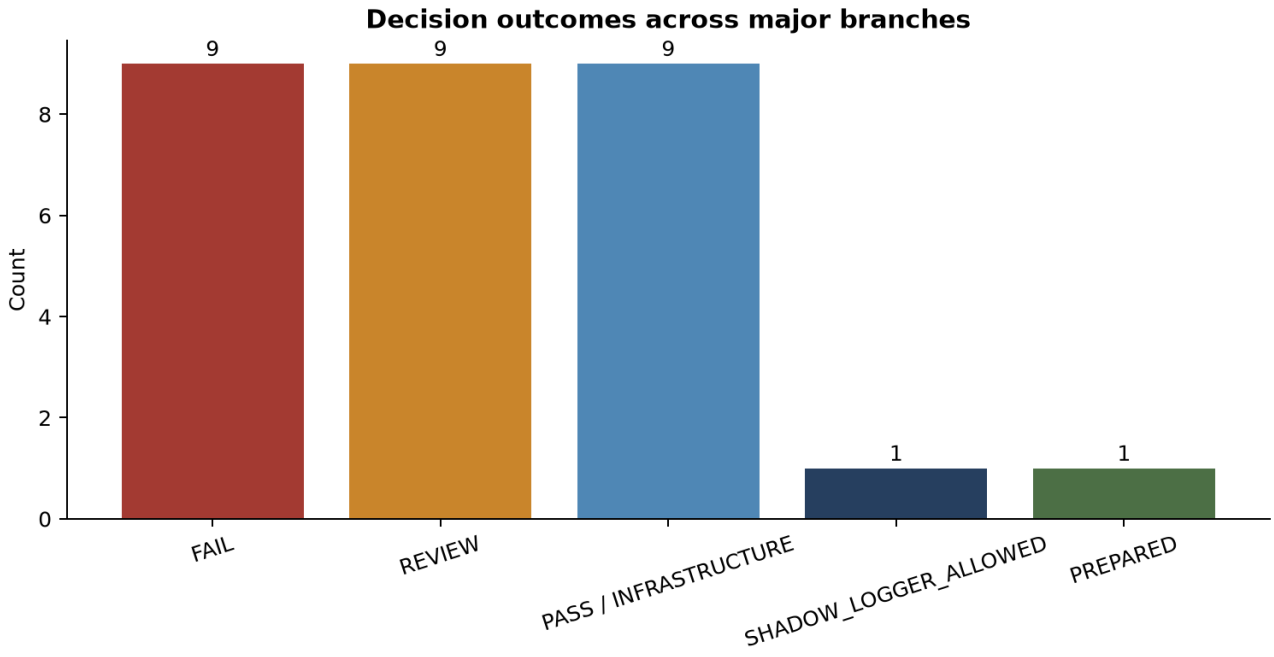
- Execution accounting and spread realism were treated as first-class risks.
- Future evidence collection is auditable and separated from broker execution.
- The project is not claiming a proven edge.
- The current phase is evidence collection, not monetization.

13. Visuals

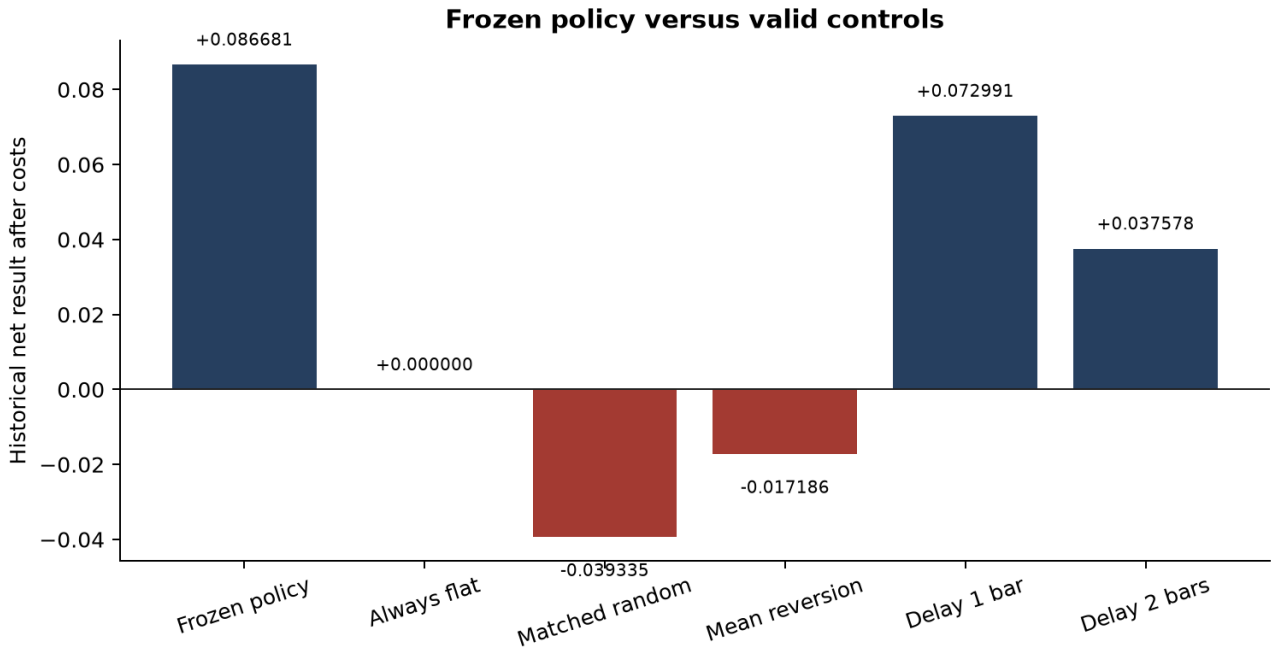
Research funnel



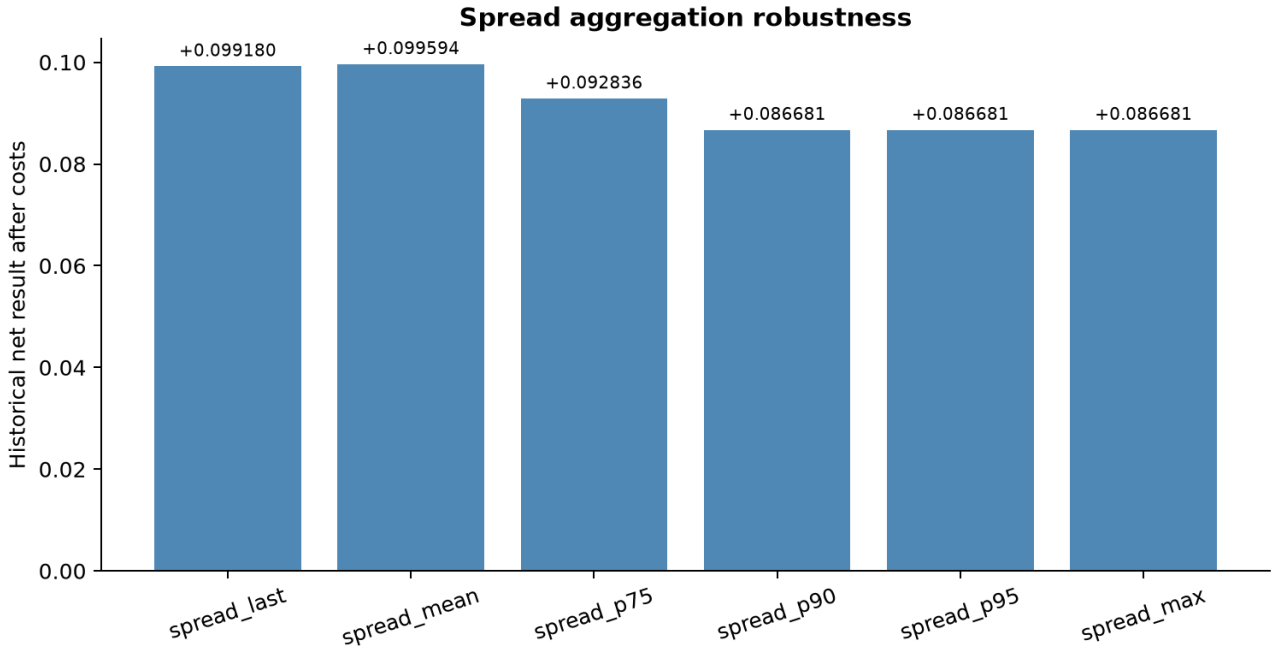
Decision count chart



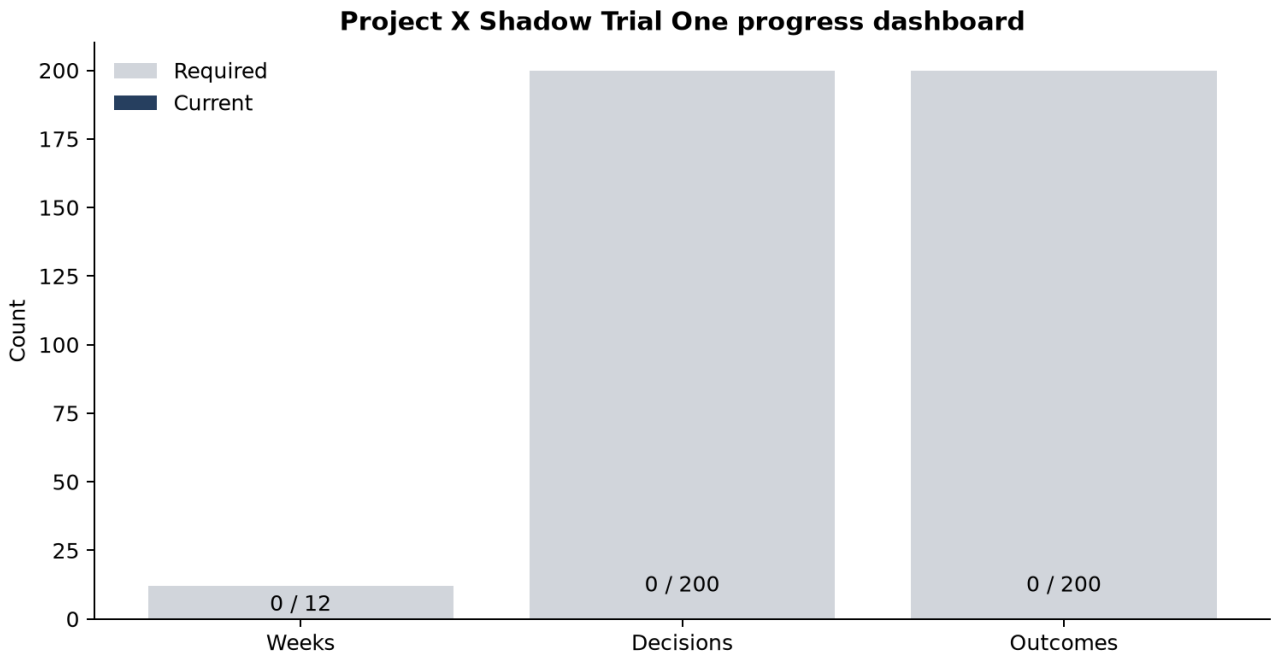
Latest lead versus valid controls



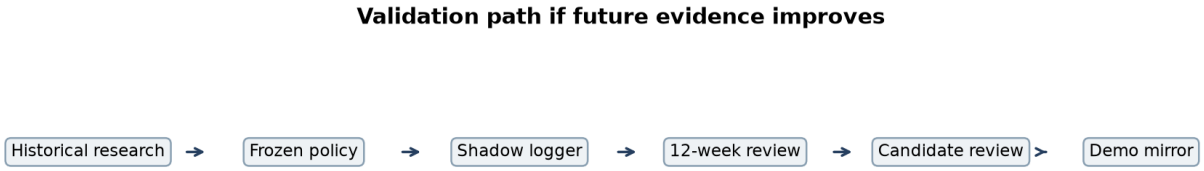
Spread robustness chart



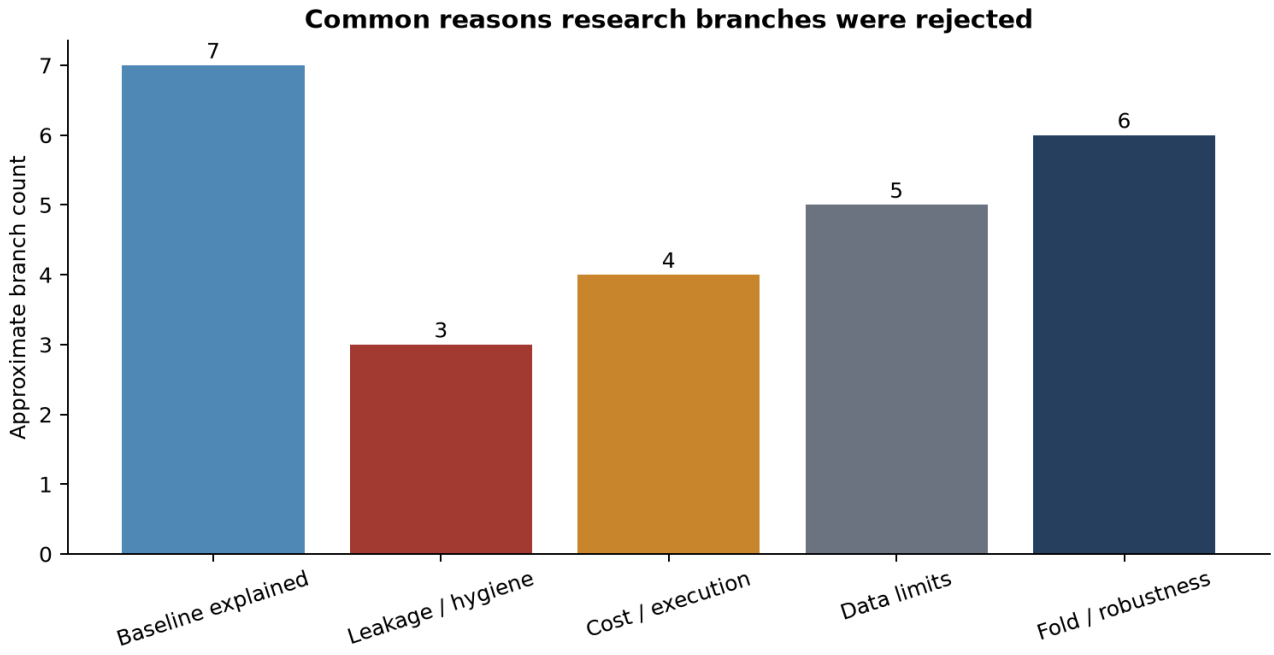
Shadow Trial One progress dashboard



Process diagram



Why branches failed



14. Glossary

- **XAUUSD:** spot gold priced in U.S. dollars.
- **EURUSD:** the euro priced in U.S. dollars.
- **H1 candle:** a one-hour market bar.
- **M15 candle:** a fifteen-minute market bar.
- **Backtest:** a historical simulation.
- **Walk-forward validation:** testing later periods after training on earlier periods.
- **Baseline:** a benchmark used to decide whether a model adds value.
- **Simple baseline:** a deliberately simple benchmark such as persistence, session, or volatility state.
- **trendiness_8:** a clean past-only trendiness baseline that became mandatory after near-miss audits.
- **Shuffled control:** a randomized test that should break a real relationship.
- **Adversarial control:** a stronger comparison designed to challenge the proposed model.
- **Leakage:** accidental use of information not available at decision time.
- **Overfitting:** fitting historical noise instead of repeatable structure.
- **False discovery risk:** the chance that a result looks good because many ideas were tried.
- **Spread:** the bid-ask cost of trading.
- **Slippage:** worse execution than the ideal quoted price.
- **p90 spread:** a conservative spread assumption around the 90th percentile of observed spreads.
- **Spread aggregation:** converting lower-timeframe spreads into an H1 cost estimate.
- **Contextual-linear policy:** a simple linear policy that maps market context to flat, long, or short actions.
- **Risk-adjusted return:** return measured relative to volatility or range.
- **Reward:** the feedback a policy learner uses during training.
- **Shadow logging:** logging what a model would have done without trading.
- **shadow_action:** the logged would-have action in the shadow logger.
- **Frozen policy:** a locked model and rule set with no tuning during evaluation.
- **Resolved outcome:** a shadow decision whose future scoring window has elapsed.
- **Demo mirror:** a possible later demo-only execution comparison system.
- **Candidate 7:** a possible future formal review status; it does not currently exist.
- **Execution accounting:** the simulator rules for entry, exit, costs, and timing.
- **External validation:** testing on a different source or feed.
- **Pseudo-holdout:** a historical holdout that is useful but not truly future-sealed.
- **Sealed future data:** data collected after the model and rules are frozen.

15. Risk Note / Disclaimer

This is research documentation, not financial advice. No live trading is authorized. No demo trading is authorized. Historical results do not guarantee future performance. The current phase is future-only shadow logging.

Appendix A - Source Artifacts

Artifact	Found	Used for	Caveat
JOURNAL.md	yes	timeline, current status	
reports/external_research/candidate_budget_ledger.md	yes	timeline, decision ledger	
reports/project_x_paper/project_x_research_paper.md	yes	v1 paper baseline	
reports/20260617T091133Z_candidate4-audit.json	yes	Candidate 4 internal audit	
reports/20260617T122351Z_candidate4-external-validation_memo.md	yes	Candidate 4 external validation	
reports/project_x_v2/project_x_v2_information_closeout_memo.md	yes	Project X v2 framework	
reports/project_x_v3/fred_macro_incremental_value_audit_memo.md	yes	FRED macro audit	
reports/project_x_v3/intraday_cross_asset_feasibility_audit_memo.md	yes	cross-asset feasibility	
reports/project_x_v3/near_miss_higher_timeframe_audit_memo.md	yes	higher-timeframe near miss	
reports/project_x_v3/narrow_review_disproof_audit_memo.md	yes	near-miss disproof	
reports/project_x_v3/baseline_registry_target_hygiene_audit_memo.md	yes	baseline registry	
reports/project_x_v3/multitimeframe_indicator_sr_audit_memo.md	yes	indicator and support/resistance audit	
reports/project_x_v3/local_structure_disproof_audit_memo.md	yes	local-structure disproof	
reports/project_x_v3/project_x_v3_price_derived_closeout_memo.md	yes	price-derived closeout	
reports/project_x_v4/new_information_source_feasibility_memo.md	yes	new information feasibility	
reports/project_x_v4/economic_surprise_data_feasibility_memo.md	yes	economic surprise feasibility	
reports/project_x_v4/economic_surprise_incremental_value_audit_memo.md	yes	economic surprise incremental audit	
reports/project_x_v4/economic_surprise_focused_disproof_audit_memo.md	yes	economic surprise focused disproof	
reports/project_x_sandbox/candidate7_sandbox_memo.md	yes	probabilistic sandbox	
reports/project_x_sandbox/candidate7_rl_sandbox_memo.md	yes	RL sandbox	Superseded by corrected execution accounting.
reports/project_x_sandbox/candidate7_rl_disproof_audit_memo.md	yes	RL disproof	
reports/project_x_sandbox/candidate7_rl_corrected_execution_rerun_memo.md	yes	corrected execution rerun	
reports/project_x_sandbox/candidate7_rl_cost_reality_audit_memo.md	yes	cost realism audit	
reports/project_x_sandbox/candidate7_rl_spread_aggregation_audit_memo.md	yes	spread aggregation audit	
reports/project_x_sandbox/candidate7_shadow_readiness_gate_memo.md	yes	shadow readiness gate	

Artifact	Found	Used for	Caveat
reports/project_x_sandbox/candidate7_final_adversarial_battery_memo.md	yes	final adversarial battery	
reports/project_x_sandbox/candidate7_adversarial_failure_forensics_memo.md	yes	adversarial forensics	
reports/project_x_sandbox/candidate7_corrected_final_battery_rerun_memo.md	yes	corrected final battery	
reports/project_x_shadow_logger/candidate7_shadow_logger_design_memo.md	yes	shadow logger design	
reports/project_x_shadow_logger/candidate7_shadow_catch_up_runner_memo.md	yes	catch-up runner	
reports/project_x_shadow_logger/candidate7_real_shadow_start_checklist.md	yes	start checklist	
reports/project_x_shadow_logger/candidate7_shadow_evaluation_plan.json	yes	evaluation plan	
data/project_x_shadow_logger/candidate7_frozen_xauusd_h1/run_state/official_shadow_run_state.json	yes	official run state	

Appendix B - Current Operating Commands

Official catch-up:

```
python -m acefx_ai.cli project-x-candidate7-shadow-logger --official-catch-up --input-file
<REAL_LOCAL_SNAPSHOT> --confirm-real-future-snapshot --json
```

Status only:

```
python -m acefx_ai.cli project-x-candidate7-shadow-logger --status-only --json
```

Dry-run catch-up:

```
python -m acefx_ai.cli project-x-candidate7-shadow-logger --official-catch-up --input-file
<REAL_LOCAL_SNAPSHOT> --confirm-real-future-snapshot --dry-run --json
```

Appendix C - Current Status

- Candidate 7 created: false
- Demo/live blocked: true
- Broker execution blocked: true
- Official run state: prepared_not_started
- Official decisions: 0
- Official outcomes: 0
- Current phase: Project X Shadow Trial One

Appendix D - Notes on the Lagged Feature Hypothesis

The lag_features_one_bar variant did better historically than the frozen policy, reaching about +0.130377 versus the frozen policy's +0.086681. The forensics branch classified it as decision-time safe but separate from the frozen policy. It cannot be used to rescue Project X Shadow Trial One. If pursued later, it needs its own frozen manifest, corrected execution audit, cost audit, adversarial battery, readiness gate, and future-only trial.