WHEN SHARING HURTS: A PRE-REGISTERED REFUTATION OF FIXED HIERARCHICAL CREDIT SHARING IN NON-STATIONARY STRUCTURED BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

When an online decision system detects that one option has degraded, should it update only that option, or also related options that share structure with it? We answer this question for one fixed hierarchical routing mechanism using a pre-registered, frozen experiment rather than a post-hoc benchmark. In $n = 97$ paired synthetic scenarios at shared-shock fraction $\rho = 0.50$, the hierarchical mechanism's mean restricted recovery time was 42.1% *higher* than that of edge-only adaptation (bootstrap 95% interval $+11.9\%$ to $+79.1\%$; one-sided upper bound $+72.1\%$). The pre-specified analysis therefore classifies the primary hypothesis as **refuted**. At $\rho = 0$, the safety test also failed: its $+16.3\%$ upper bound exceeded the $+10\%$ non-inferiority margin, so harmful transfer could not be excluded. A post-confirmatory audit found that identical nominal temperature and exploration parameters did not equalize the methods' realized action distributions (triage class EQ-B), preventing attribution of the loss to hierarchical representation alone. Our contribution is consequently not a claim that hierarchy is universally harmful. It is a bounded negative result, an auditable separation of confirmatory evidence from later diagnostics, and a concrete warning: sharing structure is itself an intervention whose scale must be calibrated and tested for negative transfer.

1 THE QUESTION, WITHOUT PREREQUISITES

Modern AI systems rarely make one isolated choice. A request can be routed through a prompt, retriever, model, tool, parser, guardrail, and fallback. These choices form a graph: a complete route is a path, and each edge is one local decision. After a disruption, a router must learn which routes remain useful.

There are two natural update strategies. An *edge-only* learner changes only the connections it observed. A *hierarchical* learner also updates the nodes incident to those connections, thereby sharing evidence across several routes. The potential benefit is speed: if many edges fail for a shared reason, one observation can inform many choices. The potential cost is negative transfer: if the failure is local, the same shared update can incorrectly penalize healthy choices.

The experiment in one sentence. We created identical disruptions for both learners, froze the hypotheses, parameters, seeds, outcome, and decision rule, and then asked whether the hierarchical learner recovered faster when half of the disruption was shared.

The answer in one sentence. It did not: the tested hierarchical mechanism recovered 42.1% more slowly on the frozen primary estimand, and it did not pass the local-disruption safety test.

Why experts should care. Partial pooling is a powerful statistical idea, but a routing policy combines representation, update geometry, score scale, and action selection. Equal hyperparameters do not guarantee equal effective policies. Our result shows why structured online learners need both a benefit test in a favorable regime and a negative-transfer gate in an unfavorable regime.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 CONTRIBUTIONS AND CLAIM BOUNDARY

This work makes four contributions.

1. We give an auditable, pre-registered test of fixed hierarchical credit sharing in a non-stationary combinatorial semi-bandit with immutable paired potential outcomes.
2. We report the terminal negative result without changing the estimand or relabeling post-hoc analyses as confirmation.
3. We introduce an explicit negative-transfer gate and demonstrate that a favorable-regime efficacy test is insufficient by itself.
4. We diagnose a material mismatch in effective policy scale and turn it into a design requirement for future comparisons: equalize behavior, not merely parameter names.

The strongest supported statement is deliberately narrow:

> Under the frozen graph, observation contract, shock construction, parameters, and estimand, the implemented fixed hierarchical routing update did not improve disruption recovery relative to edge-only adaptation.

We do *not* claim universal inferiority of hierarchical pooling, a causal explanation for the loss, a crossover $\rho^\star$, real-provider generalization, or production readiness.

3 RELATED WORK

Non-stationary bandits study decisions under changing rewards, including switching and variation-budget formulations (Garivier & Moulines, 2011; Besbes et al., 2014; Russac et al., 2019). Classical stochastic bandits isolate exploration–exploitation (Auer et al., 2002); combinatorial and semi-bandit models extend the action to a path or subset (Cesa-Bianchi & Lugosi, 2012; Chen et al., 2013; Kveton et al., 2015). Hierarchical Bayes and shrinkage share strength across related quantities (James & Stein, 1961; Gelman et al., 2013), and hierarchical bandit methods carry that principle online (Russo et al., 2018; Hong et al., 2022). When the correct representation is unknown, master algorithms and regret balancing instead treat algorithm selection as the learning problem (Agarwal et al., 2017; Pacchiano et al., 2020). Misspecification is a central risk of structured bandits (Lattimore & Szepesvári, 2020).

Our work is not a new regret-optimal algorithm and does not compare against these methods as if they were treatments. It contributes a falsification-focused empirical study of one implemented sharing rule. The closest methodological precedent is the broader reproducibility movement's emphasis on traceable decisions and artifact availability (Pineau et al., 2021); here, the freeze matters because it keeps a disappointing result from being rewritten after inspection.

4 EXPERIMENTAL DESIGN

4.1 ENVIRONMENT AND DISRUPTION

The environment is a layered directed acyclic graph with three internal layers and three alternatives per layer, yielding 27 source-to-sink paths. At each step, a method selects a path and observes rewards only on traversed edges (semi-bandit feedback). After 200 warm-up steps, a single disruption changes edge means.

The shock has constant ℓ_2 magnitude $m = 0.55$ and a controlled shared fraction ρ:

$$d(\rho) = -m\left(\sqrt{\rho}\, n + \sqrt{1-\rho}\, i\right), \tag{1}$$

where n is a unit node-incidence pattern and i is a disjoint unit edge-interaction pattern. Thus $\rho = 0$ is purely edge-specific and $\rho = 1$ is fully shared by construction. The grid is $\{0, 0.25, 0.50, 0.75, 1\}$, but only $\rho = 0.50$ (primary) and $\rho = 0$ (safety) support confirmatory statements.

Scenarios are immutable and paired: every method receives the same potential outcomes for a given scenario. Environment and agent random-number streams are namespace-separated. Unique pre- and post-shock optima are certified before a scenario is admitted.

Table 1: Pre-specified confirmatory decisions. Positive δ_{RRT} means slower hierarchical recovery. CI is a paired bootstrap 95% interval; UCB is the one-sided 95% upper bound.

Test	Edge	Hier.	δ_{RRT}	UCB	Decision
Primary, $\rho = .50$	36.01	51.18	$+42.1\%$	$+72.1\%$	Refuted
Safety, $\rho = 0$	17.65	17.92	$+1.5\%$	$+16.3\%$	Failed

Primary two-sided CI: $[+11.9\%, +79.1\%]$. Safety margin: $+10\%$.

4.2 Compared mechanisms

The edge-only comparator maintains one clipped conductance per edge, applies temporal decay and local reward updates, and selects paths with softmax temperature $T = 0.20$ plus basal exploration probability $\epsilon = 0.08$.

The hierarchical mechanism scores an edge $e = (u, v)$ as

$$\mathrm{score}(e, t) = \mathrm{base} + a_{u,t} + b_{v,t} + c_{e,t}, \tag{2}$$

with node and edge effects, shrinkage, and sum-to-zero projection. It uses the same nominal T and ϵ, but its scores, learning-rate decomposition, and projection differ. Node-only and structured sliding-window UCB conditions are secondary and cannot decide the primary hypothesis.

4.3 Outcome and decision rules

Recovery occurs when a ten-step trailing mean reaches 90% of the post-shock oracle utility for twenty consecutive steps. The restricted recovery time is

$$\mathrm{RRT} = \min(\text{recovery time}, \tau), \qquad \tau = 300. \tag{3}$$

Lower is better. The primary relative effect is

$$\delta_{\mathrm{RRT}} = \frac{\overline{\mathrm{RRT}}_{\mathrm{hier}} - \overline{\mathrm{RRT}}_{\mathrm{edge}}}{\overline{\mathrm{RRT}}_{\mathrm{edge}}} \quad \text{at } \rho = 0.50. \tag{4}$$

The frozen superiority test is $H_0 : \delta_{\mathrm{RRT}} \geq 0$ against $H_1 : \delta_{\mathrm{RRT}} < 0$. The safety test at $\rho = 0$ uses the non-inferiority null $H_0 : \delta_{\mathrm{RRT}} \geq 0.10$. Both use paired scenario bootstrap resampling with 10,000 draws and a one-sided 95% upper bound.

The decision taxonomy was frozen before confirmatory outcomes. Refuted applies when the upper bound is not below zero and the two-sided interval excludes the pre-specified practically relevant benefit -0.20. The safety gate passes only when its upper bound is below $+0.10$.

4.4 Freeze, pilot, and sample size

The protocol, analysis plan, configuration hash, and seed file were bound before confirmatory execution. The confirmatory population is the unfiltered prefix of a precommitted seed pool and is disjoint from development and pilot populations.

A 20-pair pilot estimated variance for a normal-approximation calculation with one-sided $\alpha = 0.05$, power 0.80, and target $\delta_{\mathrm{RRT}} = -0.20$, producing $N = 97$. The pilot point estimate was -10.2% with upper bound $+21.9\%$; the safety bound was $+18.0\%$. The pilot was allowed to size the study but not to change hypotheses, methods, parameters, or thresholds.

5 Confirmatory results

Table 1 reports the two pre-specified decisions. All $97/97$ primary pairs completed, with no method failures in frozen contrasts and no non-administrative censoring.

At $\rho = 0.50$, the hierarchical mean RRT was 51.18, versus 36.01 for edge-only adaptation. The estimate $\widehat{\delta_{\mathrm{RRT}}} = +42.1\%$ points in the wrong direction, and its interval excludes the target benefit -20%. The frozen rule therefore returns Refuted, not "inconclusive."

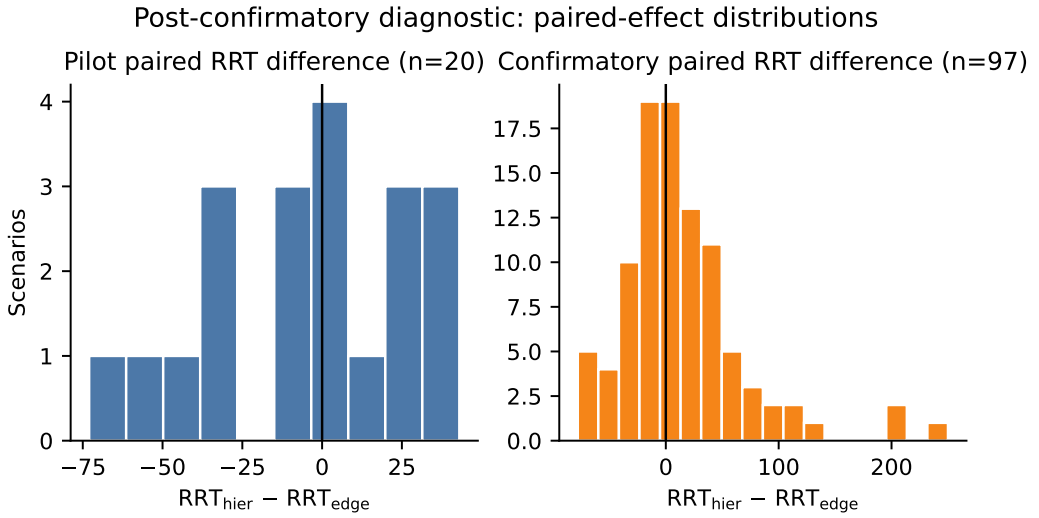


Figure 1: Paired RRT differences at the primary cell. Positive values mean slower hierarchical recovery. The pilot (left) sizes the study; only the confirmatory sample (right) decides the hypothesis.

Table 2: Warm-up action-distribution audit at $\rho = .50$ (post-confirmatory).

Diagnostic	Edge-only	Hierarchical
Path entropy / log 27	0.629	0.969
Maximum path probability	0.468	0.076
Exploration-step rate	0.224	0.221
Optimal pre-shock path rate	0.136	0.043

At $\rho = 0$, $\widehat{\delta_{\mathrm{RRT}}} = +1.5\%$, but the $+16.3\%$ upper bound crosses the $+10\%$ non-inferiority margin. This does not prove harmful transfer in every setting; it means the experiment failed to rule out practically relevant negative transfer in the setting designed to expose it.

6 What the Negative Result Teaches

All analyses in this section are Post-confirmatory diagnostics. They help design the next experiment but cannot change the frozen result.

6.1 A small pilot understated dispersion

At $\rho = 0.50$, paired RRT differences had SD 32.52 in the pilot and 55.95 in confirmation, a $1.72\times$ increase. The confirmatory distribution was right-skewed (skewness 1.63): its median difference was $+6$, mean $+15.16$, and range $[-78, +250]$. Hierarchical recovery was slower in 54.6% of pairs and faster in 44.3%. Thus the ratio-of-means result is not a uniform per-scenario delay. Future studies should support small-pilot normal approximations with simulation- or bootstrap-based power analyses.

6.2 Parameter parity was not policy parity

The methods shared the same nominal temperature, exploration probability, scenarios, and feedback. Yet their warm-up behavior differed materially (Table 2). Pre-declared tolerances assign class EQ-B: the comparison remains interpretable as a test of two complete mechanisms, but it is not an isolated causal test of hierarchical representation.

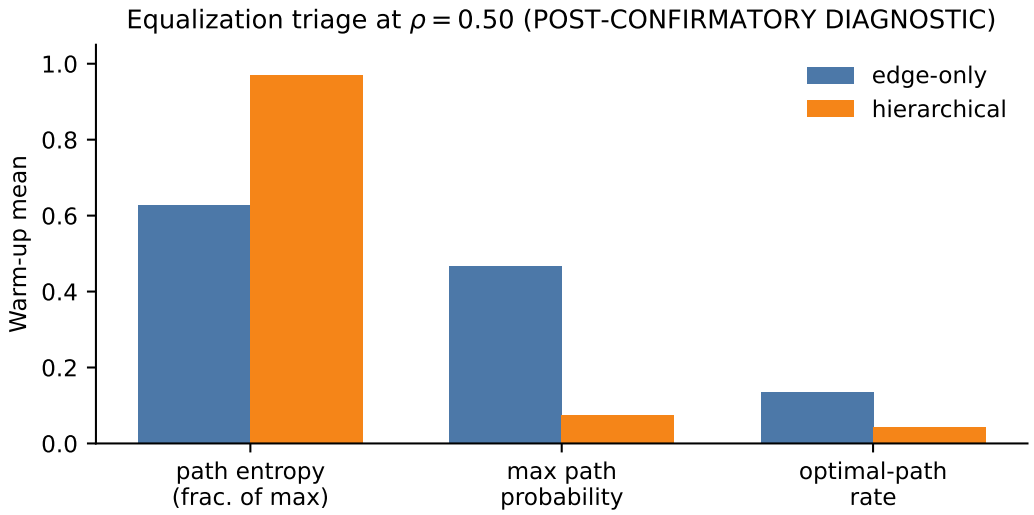


Figure 2: Realized warm-up behavior. Nominally shared T and ϵ did not produce comparable action distributions. Stored traces contain selected scores rather than every candidate score.

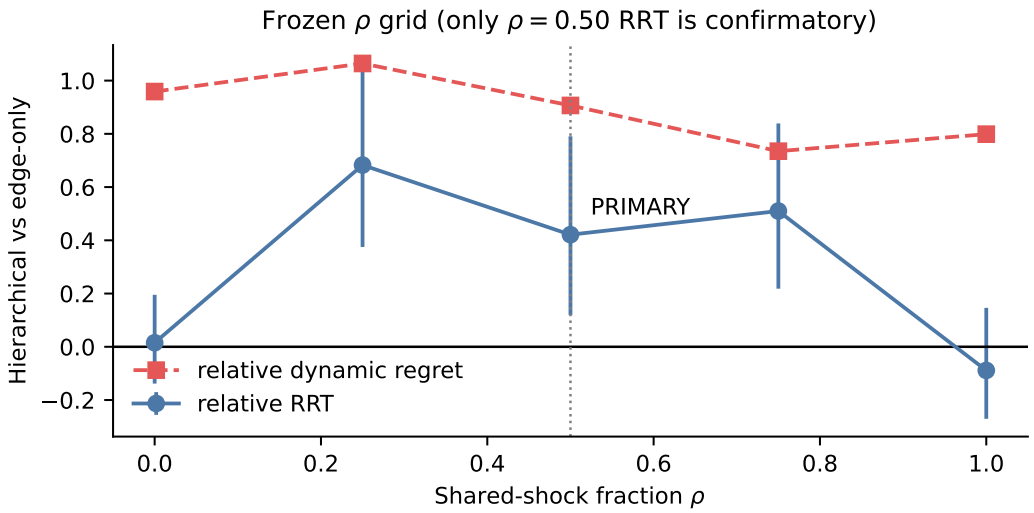


Figure 3: Hierarchical-versus-edge-only effects across the frozen grid. Only RRT at $\rho = .50$ is the primary confirmatory contrast; the remaining points are diagnostic or exploratory.

The selected-score standard-deviation ratio was 12.08, even though exploration-step rates were nearly identical. The hierarchical policy was near-diffuse rather than collapsed. This observation supports a design diagnosis, not a causal verdict: score geometry, update geometry, and representation changed together.

6.3 More shared shock did not rescue the frozen mechanism

Across the diagnostic ρ grid, hierarchical RRT was better only at $\rho = 1$: -8.9% with interval $[-27.0\%, +14.6\%]$ and one-sided upper bound $+10.2\%$. That cell does not provide a detectable benefit under the frozen one-sided rule. Moreover, relative dynamic regret at $\rho = 1$ was $+79.9\%$ against hierarchy, so RRT and regret disagreed in sign. This divergence suggests that threshold crossing and cumulative adaptation cost answer different questions; it does not license replacing RRT after observing the result.

7 THREATS TO VALIDITY AND IMPLICATIONS

Synthetic environment. No real-provider traces were used; external validity is unsupported. The graph family, feedback contract, noise, shock magnitude, and hyperparameters are narrow.

Entangled treatment (EQ-B). Because effective action distributions were not equalized, the observed loss cannot be attributed to representation alone. A cleaner factorial study should separate representation, update scale, projection, and action temperature.

Outcome choice. RRT is thresholded and administratively restricted at $\tau = 300$. At the primary cell, $0/97$ edge-only and $1/97$ hierarchical trials reached that restriction, so censoring frequency cannot explain the $+42.1\%$ effect; nevertheless, regret or survival summaries may capture distinct behavior and must be pre-specified in future work.

Power design. The 20-pair pilot underestimated confirmatory dispersion. This does not invalidate the observed paired comparison, but it weakens confidence in the prospective power calculation.

No independent reproduction or theorem. The artifact was internally replayed and mechanically checked; no external team has reproduced it. No regret or sample-complexity theorem is claimed.

Practical implication. Structured sharing should be treated as a calibrated intervention, not a free prior. A deployment-oriented successor should (i) measure realized action-distribution equivalence before the efficacy test, (ii) gate pooling using estimated sharedness, (iii) include a pure-local negative-transfer test, and (iv) pre-specify both recovery and cumulative-cost outcomes.

8 REPRODUCIBILITY AND AUDIT TRAIL

The artifact separates development, pilot, and confirmatory seed namespaces; stores the frozen protocol, analysis plan, configuration and seed hashes; validates paired-scenario identity and optima; records method failures and censoring; and seals derived summaries. The canonical full replay executes 97 seeds $\times$ five ρ values $\times$ four methods $= 1{,}940$ trials, then recomputes the frozen analysis and verifies equality with the sealed result.

The original sealed execution records 437.59375 decision CPU-seconds (0.122 CPU-hours), summed over trials. A later full replay on four workers took 403.8 seconds wall-clock and recorded 0.269 decision CPU-hours; these are different machine-dependent metrics. The result classification is deterministic with respect to the frozen scientific payload.

An anonymized code-and-artifact package accompanies the submission. Repository identity and author metadata are intentionally omitted during double-blind review.

9 CONCLUSION

A plausible structural idea failed its frozen test. At $\rho = 0.50$, the implemented hierarchical routing update recovered more slowly than edge-only adaptation, and at $\rho = 0$ it failed the negative-transfer safety gate. The confirmatory conclusion is bounded but unambiguous for the complete tested mechanism. The post-confirmatory audit adds a second lesson: nominal parameter equality can conceal radically different effective policies, so the experiment cannot isolate hierarchical representation as the cause. Publishing both facts is more informative than rescuing the original narrative.

AI USE STATEMENT

Generative AI tools were used for software-development assistance, repository maintenance, and drafting and editing portions of the manuscript, including structure and readability suggestions. All AI-assisted text and code were reviewed by the human author; numerical claims in this paper were checked against frozen, sealed artifacts and executable validation rules. Generative AI was not used to

generate the experimental outcomes, select confirmatory observations, or alter the frozen hypotheses, estimand, seed population, methods, or decision thresholds after outcome inspection. The author takes responsibility for the final content, claims, and artifacts.

ETHICS STATEMENT

The study uses synthetic rewards and no human subjects, personal data, or learned foundation models. Its main foreseeable risk is overgeneralization from a synthetic router to production systems; the paper mitigates that risk by bounding every central claim and explicitly withholding production and real-provider conclusions.

REPRODUCIBILITY STATEMENT

Section 8 describes the audit trail and canonical replay; Sections 4–5 define the estimand and decision rules; Appendix A gives commands, dependencies, and artifact boundaries. The supplementary package contains frozen configuration and seed hashes, sealed analyses, diagnostic outputs, figures, and source code.

REFERENCES

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E. Schapire. Corralling a band of bandit algorithms. In *Proceedings of the 30th Conference on Learning Theory*, pp. 12–38, 2017.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 151–159, 2013.

Aurélien Garivier and Éric Moulines. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory*, pp. 174–188. Springer, 2011.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 3 edition, 2013.

Joey Hong, Branislav Kveton, Manzil Zaheer, and Craig Boutilier. Hierarchical Bayesian bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pp. 7724–7741, 2022.

W. James and Charles Stein. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:361–379, 1961.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 535–543, 2015.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Aldo Pacchiano, My Phan, Yasin Abbasi-Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvári. Model selection in contextual stochastic bandit problems. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10328–10337, 2020.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

A Artifact and verification details

The tested runtime uses Python 3.13.3, with Python 3.11 also covered in continuous integration, and pins `numpy==2.2.6`, `scipy==1.15.3`, `PyYAML==6.0.2`, and `matplotlib==3.10.1`. After installing the locked requirements and the editable package, the canonical command is:

```
python reproduce_confirmatory.py --full
```

The default invocation without `--full` verifies bindings and sealed hashes and, when local raw trials exist, reanalyzes them; it does not regenerate all $1{,}940$ trials.

Historically frozen hashes refer to stored bytes. New text hashes use LF-canonicalized content so that CRLF checkout cannot silently redefine a declared hash. Raw confirmatory trial records are regenerated by the full replay rather than stored in git history; sealed summaries and their hashes are committed.

B Claim–evidence boundary

Table 3: Status of central claims.

Claim	Status	Stronger wording not supported
Tested mechanism slower at $\rho = .50$	Frozen confirmatory	Hierarchical pooling is universally inferior
Safety UCB exceeds $+10\%$	Frozen confirmatory	Negative transfer is always harmful
Warm-up policies differ materially	Post-confirmatory	Score scale caused the refutation
Representation alone caused the loss	Unsupported	Any representation-only causal claim
Crossover $\rho^\star$ exists	Unsupported	Location or shape of a phase transition
Real-system benefit or readiness	Unsupported	Provider or production superiority
Reproduction	Internal	Independent reproduction

C Frozen and diagnostic analysis roles

The sole primary family contains hierarchical versus edge-only RRT at $\rho = .50$. The $\rho = 0$ non-inferiority gate is a pre-specified secondary safety test. Other ρ values, node-only and sliding-window UCB comparisons, dynamic regret, distributional inspection, variance comparison, and action-distribution equalization are exploratory or post-confirmatory diagnostics. No uncorrected exploratory comparison is presented as confirmation.