TIMELOOP: Turning Training Trajectories into Realized Future Supervision

Anonymous authors
Paper under double-blind review

Abstract

Training trajectories are usually consumed as optimization traces, even though their delayed intervention consequences contain information for later decisions. We study whether those consequences can become reusable empirical supervision rather than predictions of an unobserved future. TIMELOOP stores matched-state branch outcomes as a Future Memory whose records bind a training state, action, future horizon, and realized utility. At a later query, it retrieves comparable realized futures to rank candidate data mixtures before new outcomes are revealed. Across 20 natural decision states, Full TIMELOOP recovered the realized best action in 20/20 cases, whereas removing future supervision or using an Always-Balanced policy recovered 10/20. A natural signal-attribution study further found 22/22 top-1 decisions and mean Spearman correlation 0.9818 for Full TIMELOOP, compared with 54.55% top-1 for state-only, future-only, and randomized controls. In a controlled shift, old memory without aligned new evidence achieved 3/20, while progressively adding shift-aligned experience increased mean top-1 accuracy from 0.15 to 1.00 across five acquisition orders. These results support an experience-conditioned decision mechanism: realized training futures can become reusable supervision, but TIMELOOP is not a zero-shot oracle for unseen utility changes. The evidence is bounded to one 336M model, a controlled four-domain ecosystem, and five hand-defined actions; reliability also depends on memory support, horizon matching, action coverage, anchor isolation, and the cost of acquiring future branches.

1 Introduction

Pretraining systems repeatedly choose actions whose consequences are delayed: domain mixtures, curricula, sampling schedules, and other interventions alter not only the immediate loss but also the state from which subsequent data are learned. Training runs are usually treated as optimization traces to consume and discard. Current loss, gradient influence, proxy-model scores, and short continuations provide useful local signals, yet they do not directly reveal what a chosen action will cause after a longer sequence of updates. We ask whether a training system can turn those consequences into reusable empirical knowledge instead of predicting every consequence from scratch.

We study this question through data-mixture decisions. Let a model and optimizer state at time t be s_t, let a denote a domain-mixture action, and let H be a future horizon. The realized utility $U_H(s_t, a)$ is only available after the action has been run for H updates. A conventional selector must therefore forecast an unobserved quantity. TIMELOOP asks a different question: can a matched-state intervention performed in the past become supervision for a later decision?

The central object is a realized future experience, not an isolated label. Each record preserves the state from which a branch started, the action applied, the horizon used for evaluation, and the utility observed at the descendant. At a later state, TIMELOOP retrieves records from the same training ecosystem and uses their realized outcomes to rank candidate actions. The future is thus treated as a reusable data source.

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.

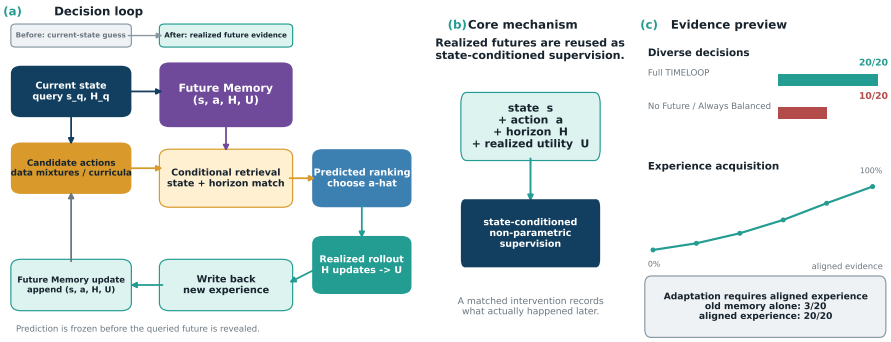


Figure 1: TIMELOOP overview. Training trajectories turn delayed intervention consequences into reusable evidence: matched records (s, a, H, U) are stored before the next decision. Unlike predicting unseen futures, TIMELOOP retrieves realized outcomes from matched historical interventions conditioned on state, action, and horizon.

This makes a deliberate distinction between prediction and experience. A predictor estimates an unobserved consequence before an intervention; TIMELOOP stores the consequence after a matched branch has actually run. The reusable evidence is therefore the contextual transition (s, a, H, U), not a zero-shot forecast detached from its source.

This formulation yields a sharper claim than "future information helps." The useful signal must remain conditional on the state and action that produced it. A future utility value detached from its context is insufficient, and a current-state heuristic cannot recover a consequence that has not yet happened. Conversely, the method should not be presented as a universal predictor: without experience aligned to a changed environment, it should not be expected to infer an arbitrary new utility law.

We make three claims rather than treating the study as a collection of ablations. (1) An instrumented trajectory can become an auditable Future Memory. (2) Useful supervision is relational: state-only, future-only, current-state-only, and randomized controls are weaker than the complete state-conditioned future. (3) Evidence-conditioned recalibration follows acquired experience, whereas cold start remains difficult. We intentionally use five hand-defined actions as a mechanism-isolation benchmark: the discrete intervention set keeps state, action, horizon, and realized outcome auditable, but is not a claim that continuous mixture optimization has been solved. Support, anchor-isolation, transfer, horizon, and cost boundaries are reported explicitly.

2 Related Work

2.1 Pretraining data selection and mixing

The composition of pretraining data has a measurable effect on language-model quality. DoReMi learns domain weights with a small proxy model and then transfers the mixture to a larger model (Xie et al., 2023). DataComp-LM frames corpus construction as a controlled benchmark spanning filtering, deduplication, and mixing decisions (Li et al., 2024). FineWeb and Dolma document how large open corpora can be curated and released for reproducible study (Penedo et al., 2024; Soldaini et al., 2024). Other work models the relation between mixture proportions, training steps, and final loss with data-mixing laws (Ye et al., 2024; Ge et al., 2024). These methods provide important baselines for data quality and mixture design, but typically estimate utility from current or proxy-scale evidence rather than from realized long-horizon descendants of the same model state.

Recent work also targets more efficient mixture optimization and training-dynamics-aware selection. Mixtures of Data Expert models approximate the loss of candidate mixtures without fully training each one (Belenki et al., 2025); Chameleon uses data-domain representations to transfer mixture weights as data changes (Xie et al., 2025); and CLUES uses

training dynamics to select data in heterogeneous collaborative settings (Zhao et al., 2025). These approaches improve the efficiency or transferability of a selector, whereas TIMELOOP studies a complementary question: whether the realized consequence of a matched intervention can itself be retained as reusable supervision for a later decision.

2.2 Adaptive selection, influence, and trajectory signals

Data valuation and influence methods offer principled ways to connect training data to outcomes. Data Shapley defines an equitable valuation of data contributions, while TracIn estimates influence by tracing loss changes along saved optimization trajectories (Ghorbani & Zou, 2019; Pruthi et al., 2020). LESS uses optimizer-aware gradient similarities for targeted instruction-data selection, and SmallToLarge summarizes small-model trajectories to select data for larger models (Xia et al., 2024; Yang et al., 2024). These approaches motivate representing training history as a source of supervision. TIMELOOP differs in the target of the supervision: it stores action-conditioned future outcomes after matched-state interventions and retrieves them for a new pretraining decision.

2.3 Explicit memory and long-horizon data value

Nearest-neighbor language models and RETRO show that explicit external memory can complement a parametric language model (Khandelwal et al., 2019; Borgeaud et al., 2021). The analogy in TIMELOOP is not token retrieval: the memory contains realized training futures and their utilities. Recent work on long-term effects of data selection in multi-stage fine-tuning provides a direct warning that short-term rankings can reverse under future adaptation and forgetting (Yang et al., 2026). That work studies fine-tuning, whereas our experiments concern pretraining-style domain actions and matched-state future branches. We therefore treat it as motivation for the problem, not as evidence that the present protocol is already established at foundation-model scale.

2.4 Offline reinforcement learning and experience replay

Offline reinforcement learning and experience replay reuse past transitions to estimate values or a policy under an environment reward. TIMELOOP instead records a model-and-optimizer training state, a data intervention, and the realized change in the model's own learning dynamics. It performs no policy-gradient or value-function learning: a later query retrieves auditable training consequences, not an external-control episode.

2.5 Curriculum and scaling context

Curriculum methods change the order or pacing of training examples and can alter convergence dynamics (Zhang et al., 2025; Elgaar & Amiri, 2026). Scaling-law work further shows that model size, data volume, and compute interact in systematic but nontrivial ways (Kaplan et al., 2020; Hoffmann et al., 2022). TIMELOOP is complementary to these lines of work: it asks whether a local data action can be evaluated by consequences that are otherwise delayed beyond the current decision horizon. Its distinction is the combination of same-lineage realized futures, hindsight temporal supervision, matched-state intervention, and retrieval for a new present state.

2.6 Self-directed learning and self-improvement

Self-directed learning methods use a model's own outputs or iterative interaction to create additional supervision. Self-Instruct bootstraps instruction data from model generations, while STaR iteratively retains successful rationales to improve reasoning (Wang et al., 2022; Zelikman et al., 2022). Self-play methods such as AlphaZero likewise improve a policy through repeated interaction with its own changing policy (Silver et al., 2017). TIMELOOP is related in spirit but differs in object and mechanism: it does not generate a new instruction corpus, create rationales, or train a policy through self-play. It records realized descendants of parameter updates under matched data interventions and uses those trajectory records

Table 1: Conceptual distinction among future-supervision paradigms.

Method	Future source	Generated?	Lineage?	Decision target
Data selection	Current loss or quality signal	No	No	Current utility
Proxy controller	Proxy-model prediction	No	No	Predicted utility
Self-training	Model-generated outputs	Yes	No	Model output
TIMELOOP	Realized descendants	No	Yes	Training decision

to rank a later pretraining action. The memory is therefore a trace of training consequences, rather than a generated example or an agent policy.

3 TIMELOOP: Realized Future Supervision Framework

The central conceptual move in TIMELOOP is to treat a realized training future as supervision for a later decision. A conventional training rule maps information available at time t to an action at time t. TIMELOOP instead records the sequence

$$\text{past state and action} \longrightarrow \text{realized descendant} \longrightarrow \text{future utility} \longrightarrow \text{later decision.} \quad (1)$$

The future in this sequence is not a forecast, a simulation target, or a generated label. It is the observed consequence of an intervention that was actually run from a matched training state. A useful record therefore preserves the state from which the branch started, the action that was applied, the horizon over which the consequence was measured, and the resulting utility. Equivalently, the branch is an empirical training transition whose endpoint metrics and utility can be reused as supervision. The retrieval rule used in this paper is one implementation of the principle; the principle does not depend on nearest-neighbor regression being the only possible controller.

The distinction is operational: before a branch, its consequence is predicted; after a matched branch, it is observed, auditable, and reusable. TIMELOOP therefore treats (s, a, H, U) as an empirical training transition, not as an external-reward replay episode or a zero-shot value estimate.

3.1 Problem Formulation

Let M_t denote a language model and optimizer state at training point t. We use s_t as a compact state description containing the four domain losses and their recent slopes, together with normalized training progress. A data action $a \in \mathcal{A}$ specifies a distribution over the four domains $\{\text{General}, \text{Math}, \text{Code}, \text{Science}\}$. After applying a for a future window of H updates, the resulting descendant is denoted M^a_{t+H}.

The action objective is a future utility $U_H(s_t, a)$. For the main prospective protocol, each domain d has a shared evaluation window with loss $L_{\text{start},d}$ before the branch and $L_{\text{end},d}$ after the branch. We define

$$\Delta_d = L_{\text{start},d} - L_{\text{end},d}, \qquad U_H(s_t, a) = \frac{1}{4} \sum_d \Delta_d. \quad (2)$$

The implementation records the domain-level values and computes the mean from those four values. Earlier discovery experiments use related but not identical relative-improvement definitions; we therefore compare them within experiment and do not place their absolute utilities on one shared axis.

The ideal decision is

$$a^*(s_t, H) = \arg\max_{a \in \mathcal{A}} U_H(s_t, a). \quad (3)$$

At decision time, however, U_H is unknown because the future has not happened. A short-horizon proxy can be biased toward actions that are useful immediately, and a single global function can fail when the relation between state, action, and horizon is heterogeneous. TIMELOOP instead constructs a finite future memory

$$\mathcal{M} = \{(s_i, a_i, H_i, U_i, z_i)\}_{i=1}^N, \quad (4)$$

where U_i is observed only after the corresponding branch is completed and z_i is the stored state feature vector. The prospective question is whether local records in $\mathcal{M}$ can rank actions at a new state before the new futures are revealed.

3.2 Realized Future Collection

For an anchor state, we create independent branches from the same available model and optimizer checkpoint. The branch protocol holds the checkpoint, optimizer state, scheduler, token budget, and evaluation procedure fixed, and changes only the data action. We call this a matched-state restart; we do not claim an exact rewind of an unobserved physical training trajectory. Each branch is allowed to complete its registered horizon and produces domain-level losses and a scalar utility. The intervention is the action change at the anchor; the label is the consequence measured at the descendant after H updates.

This procedure turns a hypothetical long-horizon label into an observed intervention outcome. It is more expensive than a current-loss heuristic, but the cost is paid only at selected anchor states. The resulting records can be reused for later decisions, which is the reason to treat them as a future memory rather than as disposable rollouts. In this sense, the future branch supplies training-time supervision for the decision rule, not an additional prediction target for the language model.

3.3 Future Memory Construction

Future Memory does not store predictions of future utility; it stores observed consequences generated by controlled interventions.

Each memory item contains the compact state vector

$$z(s_t) = [\ell_{\mathrm{G}}, \ell_{\mathrm{M}}, \ell_{\mathrm{C}}, \ell_{\mathrm{S}}, \dot{\ell}_{\mathrm{G}}, \dot{\ell}_{\mathrm{M}}, \dot{\ell}_{\mathrm{C}}, \dot{\ell}_{\mathrm{S}}, p_t], \tag{5}$$

where ℓ_d is a domain loss, $\dot{\ell}_d$ is its recent slope, and p_t is normalized progress. The representation is intentionally small and auditable; the method does not read the full hidden state or all parameter differences. Each record also stores its action and the horizon at which its future utility was measured. The memory is explicit: a reviewer can inspect the checkpoint identity, query state, action, horizon, endpoint metrics, and utility associated with every retrieved item. This makes the evidence path different from a hidden recurrent state or an opaque controller embedding.

3.4 Conditional Future Retrieval

At a query (s_q, H_q) and candidate action a, we first restrict the memory to records with the same action. We compute a state distance and a log-horizon distance:

$$d_s = \|z(s_q) - z(s_i)\|_2, \qquad d_H = |\log_2 H_q - \log_2 H_i|\,, \tag{6}$$

and combine them as

$$d_i = d_s + 0.5 d_H. \tag{7}$$

The frozen prospective configuration uses $k = 3$ neighbors and inverse-distance weights

$$w_i = \frac{1}{d_i + \epsilon}, \qquad \widehat{U}(a \mid s_q, H_q) = \frac{\sum_{i \in \mathcal{N}_k(a)} w_i U_i}{\sum_{i \in \mathcal{N}_k(a)} w_i}. \tag{8}$$

If the evidence support is insufficient, the registered algorithm uses the Balanced fallback. The decision is $\hat{a} = \arg\max_a \widehat{U}(a \mid s_q, H_q)$. Crucially, the prediction is frozen before any of the prospective action branches are opened. Intuitively, the query asks two questions of the memory: which recorded state looks most like the present state, and what happened after each action there? Retrieval answers with observed consequences rather than with a hidden global forecast.

Table 2: Core experimental roles and fixed protocol elements.

Study	Model/instance	Queries or seeds	Fixed action/horizon
Original prospective	336M model	1 query	5 actions, $H = 2048$
C33 Multi-query	C33	5 queries	5 actions, $H = 2048$
876 Multi-seed	876	5 seed identifiers	5 actions, $H = 2048$
Large-scale Diverse	876	20 natural states	5 actions, matched horizons
Attribution	C33	22 natural units	5 actions, $H = 2048$ core

Algorithm 1: TIMELOOP Future Memory. Input: query (s_q, H_q), future memory $\mathcal{M}$, candidate actions $\mathcal{A}$, and neighbor count k.
1. For each $a \in \mathcal{A}$, retain same-action records and compute d_s, d_H, and $d_i = d_s + 0.5d_H$.
2. Keep the k closest records and estimate $\widehat{U}(a \mid s_q, H_q)$ with inverse-distance weights.
3. Use the registered Balanced fallback if support is insufficient; otherwise return $\hat{a} = \arg\max_a \widehat{U}(a \mid s_q, H_q)$.
4. Freeze $\hat{a}$ before opening any new prospective branch.

3.5 Why Preserve Realized Futures?

A global controller compresses all state–action–horizon interactions into one parametric map. In the tested small-data regime, direct future guidance and regularization did not preserve held-out rankings. The challenge is therefore not merely predictor quality, but preservation of the conditional structure of future consequences. Future Memory keeps each realized future explicit and retrieves records close in state and horizon, making support inspectable rather than hiding unsupported queries inside one function.

4 Experimental Setup

4.1 Protocol roles and compute boundary

The original prospective study establishes temporal feasibility on a 336M-parameter model with one query, five actions, and $H = 2048$. Subsequent C33 and 876 studies reuse the same protocol family to test query, seed, state, action, horizon, support, and shift conditions. Post-hoc offline analyses are labeled as such; they are not described as new prospective training runs.

The canonical prospective space is deliberately controlled: five hand-defined domain-mixture actions. This is a mechanism-isolation benchmark rather than a claim about the capacity of the action space: keeping the interventions discrete makes the state, action, horizon, and realized consequence directly auditable. It is not evidence for continuous mixture optimization; no new expanded-action rollout is added in this revision, and any completed post-acquisition diagnostic remains separately labelled.

4.2 Model, domains, and utility

The domain pools are General, Math, Code, and Science. The primary utility is the mean decrease in the shared evaluation loss across the four domains between branch start and endpoint. Discovery experiments use their own future-credit normalization and are never numerically pooled with the main domain-utility protocol. The full manifests, data-source descriptions, configurations, and metric dictionary are retained in the evidence bundle.

4.3 Metrics and comparisons

We report top-1 action accuracy, regret relative to the realized best action, and Spearman correlation between predicted and realized rankings. We additionally report support coverage, fallback counts, memory-coverage behavior, and uncertainty across acquisition orders or memory-sampling seeds where the experiment defines them. Constant policies are not assigned artificial rank correlations.

4.4 Reproducibility and evidence boundary

The prediction is frozen before prospective outcomes are revealed. A branch is counted only when its status is successful and endpoint metrics are present. The local evidence package records the instance identity, source path, manifest, frozen prediction hashes, branch-level outcomes, leakage audits, and the distinction between prospective training and CPU-only post-hoc re-evaluation. We do not infer dollar prices, unrecorded controller latencies, or unsupported endpoint statistics.

5 Results

5.1 Main result 1: Training trajectories become reusable supervision

The original frozen prospective case used one unseen state of the 336M model and five pre-registered actions over 2,048 future updates. The prediction was fixed before rollout; Balanced was predicted as best and later had the largest realized mean utility. The predicted and realized rankings had Spearman correlation 0.70 and regret 0. This case establishes temporal feasibility, not population-level generalization.

The confirmatory C33 multi-query package extended the same action/horizon protocol to five held-out queries and 25 completed branches. Top-1 selection was 5/5, mean regret was 0, and mean Spearman correlation was 0.82. All five realized oracle actions were Balanced, so this result supports query-level consistency but does not by itself establish diverse action selection.

The formal 876 five-seed experiment evaluated five seed identifiers at a shared C3_3072 query with five actions and $H = 2048$, for 25 completed branches. Top-1 was 5/5, regret was 0, and mean Spearman was 0.44. All five realized oracle actions were Balanced; this is stability evidence rather than a diverse-action benchmark.

5.2 Main result 2: The signal is relational, not a default heuristic

The 876 Large-scale Diverse Decision Validation evaluated 20 complete natural state–action groups selected by realized-oracle strata before predictions were inspected. Full TIMELOOP achieved 20/20 top-1 and zero regret; No Future Memory and Always Balanced each achieved 10/20, with no non-Balanced selections. The export lacks a row-level same-anchor/nearest-neighbor manifest, so this is heterogeneous, protocol-bound evidence rather than a strict unseen-state claim.

The C33 Natural Future Signal Attribution study tests 22 natural decision units. Full TIMELOOP achieved 22/22 top-1 and Spearman 0.9818, versus 54.55% top-1 for each control (Spearman 0.4339, 0.5864, 0.4545, and 0.4545). This identifies the state-conditioned realized future as the useful object, rather than state similarity or detached outcomes. Because the export lacks row-level same-anchor exclusions, the result is descriptive mechanism evidence, not a strict leave-one-anchor-out test.

5.3 Main result 3: Recalibration follows acquired evidence

In cold start, old Future Memory is unchanged while the shifted oracle changes, yielding 3/20. With shift-aligned evidence added to a memory copy, the same frozen protocol reaches 20/20. This is evidence-conditioned recalibration, not zero-shot inference of an unseen utility law.

The Experience Acquisition Curve fills the gap between these endpoints. Across 20 shift queries, $H = 2048$, five actions, and five acquisition orders, mean top-1 accuracy increases from 0.15 ± 0.00 at 0% new evidence to 0.34 ± 0.066, 0.49 ± 0.037, 0.75 ± 0.084, 0.89 ± 0.058, and 1.00 ± 0.00 at 10%, 25%, 50%, 75%, and 100%, respectively. Mean Spearman increases from 0.4700 to 0.6100, 0.7500, 0.8780, 0.9550, and 1.0000, while mean regret decreases from 0.082373 to zero.

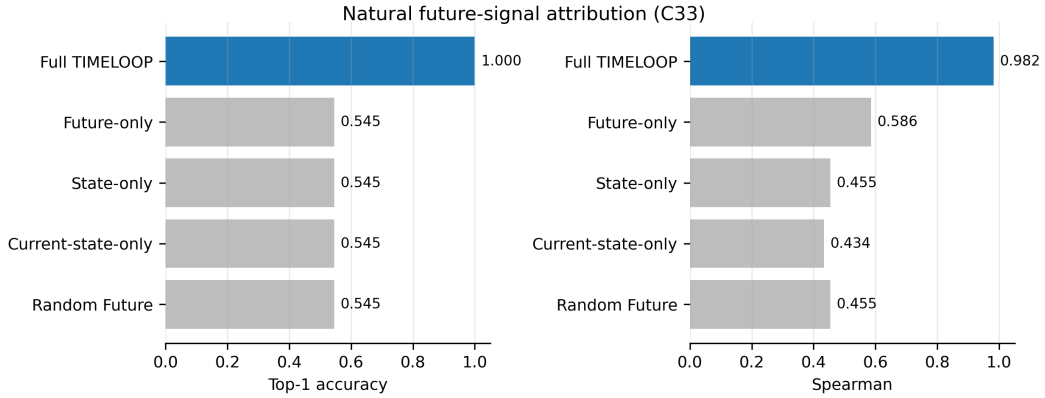


Figure 2: Natural Future Signal Attribution. Full TIMELOOP combines state-conditioned retrieval with realized future outcomes; removing either context or outcome information reduces ranking quality.

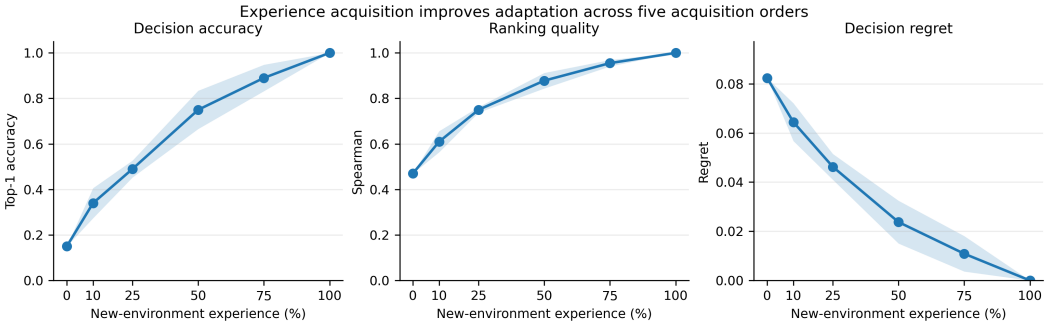


Figure 3: Experience-conditioned recalibration. Across five acquisition orders, adding shift-aligned future evidence improves top-1 accuracy and rank correlation while reducing regret. The 0% endpoint is the cold-start condition; this is controlled evidence, not a zero-shot adaptation or generalization claim.

5.4 Supporting evidence: explicit memory versus compression

On the natural attribution units, the MLP controller achieved 21/22 top-1 decisions and mean Spearman 0.9636, close to but below Full TIMELOOP's 22/22 and 0.9818. The linear controller and state-only retrieval were weaker. Under the cold-start synthetic stress, however, the controller also failed to infer the changed utility law. These results do not establish that all parametric controllers fail; they show that, in the tested small-data regime, explicit records retain a conditional signal that a compressed controller does not fully preserve. Because the local package does not retain paired rows for this aggregate controller comparison, we keep it descriptive and do not claim a universal controller gap.

5.5 Supporting evidence: support, horizon, and retrieval conditions

The method's support requirements are measurable. In 876 Future Memory Scaling, 100% and 75% memory retained full coverage with top-1 values of 1.00 ± 0.00 and 0.88 ± 0.327. At 50%, top-1 was 0.71 ± 0.456 with 0.958 coverage; at 25%, it fell to 0.39 ± 0.490 with 0.668 coverage; at 10%, overall top-1 was 0.23 ± 0.423 and coverage was 0.20 ± 0.1842.

The basic K ablation supports $K = 3$ on the frozen 20-query set: top-1 100%, regret 0, and mean Spearman 0.975. $K = 1$ triggered 20 fallbacks, whereas $K \geq 5$ reduced top-1 to 80% in this set. Horizon selection likewise shows why matching matters: H=2048 is the only fully supported exact-horizon setting in the frozen asset; H=1024 has too few records

and H=4096 has no exact-horizon memory records. The C33 H=4096 rollout is therefore a long-horizon sensitivity result, not exact-H=4096 retrieval evidence.

5.6 Boundary evidence: amortization and acquisition cost

The branch protocol has an upfront acquisition cost: the C33 H=2048 anchor used 25 branches totaling 28,607.70 seconds, and the formal 876 H=2048 five-seed anchor used 25 branches totaling 28,770.86 seconds. In normalized branch-equivalent units, a fresh oracle costs five branches per decision, whereas a 25-branch Future Memory amortizes to 1.25 at $N = 20$ before retrieval latency. Because controller latency and dollar price were not recorded, this is a lower bound rather than an end-to-end or monetary savings claim; the cost–quality frontier is retained in the appendix.

6 Discussion

Once a matched intervention has been executed, its realized consequence can become reusable supervision. State, action, horizon, and outcome remain associated, so a training trajectory records how interventions change learning, not only parameter updates.

Explicit Future Memory retained a small, auditable advantage over the learned controller in this regime; this does not bound all parametric controllers. Aligned evidence enabled recalibration, while the 20-state and 22-unit summaries remain descriptive group-level, not independent-lineage, evidence.

Perfect compact decisions are protocol-level, not population-level, results. Five actions isolate realized-consequence guidance before continuous mixture optimization; model scale, independent lineages, and continuous actions remain open.

7 Limitations

Evidence is bounded to one 336M model, four domains, five actions, and $H = 2048$. The five-action space isolates mechanism rather than complete optimization. Row-level anchor isolation is certified for 876 cross-seed but not retained for the 20-state/22-unit aggregates, so they are protocol-bound, not strict unseen-state evidence. Independent lineages, online controller baselines, and end-to-end cost remain open.

8 Conclusion

TIMELOOP turns a training trajectory into reusable supervision. After a matched intervention reaches horizon H, it stores the realized record (s, a, H, U) and retrieves it before the next intervention. The record describes an observed consequence, not a detached forecast.

Three findings support this view: Full TIMELOOP recovers heterogeneous decisions over controls; multi-query and multi-seed studies show stable retrieval; and attribution identifies the matched state–action–horizon record as the informative unit. Shift and acquisition analyses show recalibration after aligned experience is acquired, while cold start remains difficult.

The design principle is therefore train, remember, reuse, improve. Evidence remains bounded to a small, discrete benchmark: it does not establish universal optimization, zero-shot shift prediction, continuous actions, frontier-scale transfer, or monetary savings. The 20-state/22-unit aggregates are protocol-bound; 876 cross-seed is leakage-audited but not independent-lineage generalization.

AI Use Statement

We used generative AI tools during manuscript preparation for literature and document organization, code-related support, deterministic figure generation, language refinement, and document formatting. All experimental protocols, implementations, analyses, figures, and scientific claims were designed, verified, and validated by the authors. The authors reviewed all AI-assisted outputs and take full responsibility for the final content, citations, interpretations, and submission disclosure.

Reproducibility Statement

The main text specifies the task, state representation, action space, utility definition, branch protocol, retrieval rule, frozen prospective evaluation, and reported metrics. The appendix and evidence bundle retain checkpoint hashes, frozen predictions, branch logs, endpoint outcomes, leakage audits, and the result-allocation record. The C33 machine-readable attribution rows and audit were additionally verified in the live C33 archive; the manuscript does not claim a row-level anchor manifest that was not exported. Post-hoc analyses are marked as such and are not presented as new prospective training runs.

References

Lior Belenki, Alekh Agarwal, Tianze Shi, and Kristina Toutanova. Optimizing Pre-Training Data Mixtures with Mixtures of Data Expert Models. arXiv preprint, 2025. doi: 2502.15950v1. URL https://arxiv.org/abs/2502.15950v1.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. arXiv preprint, 2021. doi: 2112.04426v3. URL https://arxiv.org/abs/2112.04426v3.

Mohamed Elgaar and Hadi Amiri. Curriculum Learning for LLM Pretraining: An Analysis of Learning Dynamics. arXiv preprint, 2026. doi: 2601.21698v2. URL https://arxiv.org/abs/2601.21698v2.

Ce Ge, Zhijian Ma, Daoyuan Chen, Yaliang Li, and Bolin Ding. BiMix: A Bivariate Data Mixing Law for Language Model Pretraining. arXiv preprint, 2024. doi: 2405.14908v4. URL https://arxiv.org/abs/2405.14908v4.

Amirata Ghorbani and James Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. arXiv preprint, 2019. doi: 1904.02868v2. URL https://arxiv.org/abs/1904.02868v2.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models. arXiv preprint, 2022. doi: 2203.15556v1. URL https://arxiv.org/abs/2203.15556v1.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. arXiv preprint, 2020. doi: 2001.08361v1. URL https://arxiv.org/abs/2001.08361v1.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through Memorization: Nearest Neighbor Language Models. arXiv preprint, 2019. doi: 1911.00172v2. URL https://arxiv.org/abs/1911.00172v2.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. DataComp-LM: In search of the next generation of training sets for language models. arXiv preprint, 2024. doi: 2406.11794v4. URL https://arxiv.org/abs/2406.11794v4.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. arXiv preprint, 2024. doi: 2406.17557v2. URL https://arxiv.org/abs/2406.17557v2.

Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. Estimating Training Data Influence by Tracing Gradient Descent. arXiv preprint, 2020. doi: 2002.08484v3. URL https://arxiv.org/abs/2002.08484v3.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. arXiv preprint, 2017. doi: 1712.01815v1. URL https://arxiv.org/abs/1712.01815v1.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. arXiv preprint, 2024. doi: 2402.00159v2. URL https://arxiv.org/abs/2402.00159v2.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions. arXiv preprint, 2022. doi: 2212.10560v1. URL https://arxiv.org/abs/2212.10560v1.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting Influential Data for Targeted Instruction Tuning. arXiv preprint, 2024. doi: 2402.04333v3. URL https://arxiv.org/abs/2402.04333v3.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining. arXiv preprint, 2023. doi: 2305.10429v4. URL https://arxiv.org/abs/2305.10429v4.

Wanyun Xie, Francesco Tonin, and Volkan Cevher. Chameleon: A Flexible Data-mixing Framework for Language Model Pretraining and Finetuning. arXiv preprint, 2025. doi: 2505.24844v1. URL https://arxiv.org/abs/2505.24844v1.

Yu Yang, Siddhartha Mishra, Jeffrey N Chiang, and Baharan Mirzasoleiman. SmallToLarge (S2L): Scalable Data Selection for Fine-tuning Large Language Models by Summarizing Training Trajectories of Small Models. arXiv preprint, 2024. doi: 2403.07384v2. URL https://arxiv.org/abs/2403.07384v2.

Yuxin Yang, Aoxiong Zeng, and Xiangquan Yang. The Long-Term Effects of Data Selection in LLM Fine-Tuning. arXiv preprint, 2026. doi: 2605.30537v1. URL https://arxiv.org/abs/2605.30537v1.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data Mixing Laws: Optimizing Data Mixtures by Predicting Language Modeling Performance. arXiv preprint, 2024. doi: 2403.16952v2. URL https://arxiv.org/abs/2403.16952v2.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping Reasoning With Reasoning. arXiv preprint, 2022. doi: 2203.14465v1. URL https://arxiv.org/abs/2203.14465v1.

Yang Zhang, Amr Mohamed, Hadi Abdine, Guokan Shang, and Michalis Vazirgiannis. Beyond Random Sampling: Efficient Language Model Pretraining via Curriculum Learning. arXiv preprint, 2025. doi: 2506.11300v2. URL https://arxiv.org/abs/2506.11300v2.

Wanru Zhao, Hongxiang Fan, Shell Xu Hu, Wangchunshu Zhou, Bofan Chen, and Nicholas D. Lane. CLUES: Collaborative High-Quality Data Selection for LLMs via Training Dynamics. arXiv preprint, 2025. doi: 2507.03004v1. URL https://arxiv.org/abs/2507.03004v1.

Supplementary Material

A Scope, notation, and evidence boundary

This supplement contains the complete protocol and the full evidence ledger behind the nine-page main paper. We use s for the compact training-state descriptor, a for a candidate data action, H for the prospective horizon, and U for realized descendant utility. A Future Memory record is therefore (s, a, H, U). A decision is made before the future outcome for the queried action is revealed. Unless explicitly labelled post-acquisition, a result is prospective: the queried outcome is used only for evaluation after the prediction is frozen.

The evidence is deliberately partitioned by machine and protocol. C33 supplies the multi-query, attribution, long-horizon, controller, and mechanism studies. The north-west B-zone 876 instance supplies the multi-seed, memory-scaling, cross-seed, action-space, and transfer studies. We never merge raw checkpoints or frozen predictions across instances. The reports below preserve negative, saturated, and support-limited outcomes rather than replacing them with favorable summaries.

For navigation, the supplement is organized as a paper-style evidence record: protocols and inventories, complete result tables and figures, mechanism and boundary analyses, leakage and statistical audits, normalized compute accounting, and a final artifact/reproducibility handoff. The prose distinguishes prospective evaluation from post-acquisition and offline analyses rather than presenting all tables as one homogeneous benchmark.

Table 3: Notation and audit status used throughout the supplement.

Symbol / term	Operational meaning
s	Compact descriptor of the current model and training state.
a	Candidate data-mixture or curriculum action.
H	Number of future training steps used to realize a descendant.
U	Realized utility, defined by the protocol-specific loss improvement.
Future Memory	Auditable records containing (s, a, H, U), not text retrieval.
Top-1	Fraction of queries where the selected action equals the realized oracle action.
Regret	Oracle utility minus the utility of the selected action.
Boundary label	Indicates a post-acquisition, saturated, diagnostic, or support-limited result.

B Full prospective protocol

B.1 Matched-state restart

Each candidate action is evaluated from a matched checkpoint or matched state descriptor. The branch is cloned before the action is applied, so the comparison is between interventions rather than unrelated training trajectories. The branch is then trained for the declared horizon, and the endpoint utility is recorded. A query's own realized utility is not inserted into its retrieval memory before its prediction. This ordering is the central prospective condition. A stronger leave-one-anchor-out guarantee also requires excluding every record from the same state/checkpoint and retaining the row-level exclusion and nearest-neighbor manifest; that stronger evidence is available for the 876 cross-seed transfer, but only partially available for the 20-state diverse and 22-unit attribution exports.

B.2 Action space and model setup

The canonical five-action space is {Balanced, General-heavy, Math-heavy, Code-heavy, Science-heavy}. We use this discrete set deliberately as a mechanism-isolation benchmark: it makes each intervention, matched state, horizon, and realized consequence inspectable before any continuous mixture optimization is attempted. It is therefore not intended as a

complete action-space claim. The shared model configuration uses a 32,000-token vocabulary, 12 Transformer layers, width 768, 12 heads, feed-forward width 3,072, sequence length 1,024, tied input/output weights, BF16 arithmetic, micro-batch 4, gradient accumulation 8, learning rate 3×10^{-4}, 100 warmup steps, weight decay 0.1, and gradient clipping 1.0. These fields are copied from the source configuration; fields not independently transcribed are not inferred.

B.3 Utility and ranking metrics

For the canonical prospective protocol, utility is the reduction in shared domain loss between branch start and endpoint. When multiple domains are present, the reported utility is their arithmetic mean. Given predicted scores $\hat{U}(a)$ and realized utilities $U(a)$, we report:

$$\text{Top1} = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}[\arg\max_a \hat{U}_q(a) = \arg\max_a U_q(a)], \tag{9}$$

$$\text{Regret} = U_q(a_q^\star) - U_q(\hat{a}_q), \tag{10}$$

where $a_q^\star$ is the realized oracle action. Spearman correlation is computed only when the score vector has non-constant support. Constant-policy correlations are reported as not applicable rather than fabricated zeros or ones.

B.4 Primary prospective inventory

The C33 multi-query protocol evaluates five held-out query states and 25 action branches. The 876 formal multi-seed protocol evaluates five seeds times five actions at $H = 2048$. The 876 large-scale diverse protocol evaluates 20 complete natural state-action groups. The latter was selected using realized oracle labels before inspecting predictions; it is not a cherry-picked collection of successful cases.

Table 4: Primary prospective inventories.

Instance / protocol	Unit definition	Units	Branches	H	Status
C33 Multi-query	query	5	25	2048	complete, all oracle Balanced
876 Multi-seed	seed	5	25	2048	complete, all oracle Balanced
876 Large-scale Diverse	state	20	100	2048	complete, oracle-stratified; anchor manifest partial
C33 Attribution	natural decision unit	22	method-query	2048	complete, controls frozen

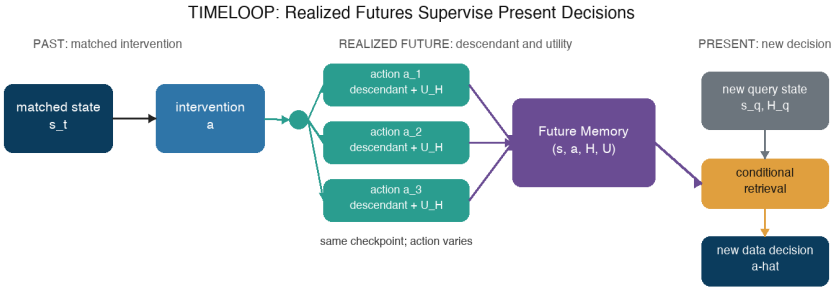


Figure 4: Prospective protocol schematic. A matched state is branched under a frozen action, the descendant is trained for a declared horizon, and the realized future is stored only after the prediction point.

C Complete prospective results

C.1 Multi-query and multi-seed stability

C33's five-query experiment selected the realized best action in 5/5 queries, with mean Spearman 0.82 and zero regret. The 876 five-seed experiment also selected the realized best action in 5/5 seeds, with mean Spearman 0.44 and zero regret. Both results are useful stability checks, but both have an important boundary: every realized oracle action in these compact protocols is Balanced. They therefore do not by themselves establish non-Balanced action selection.

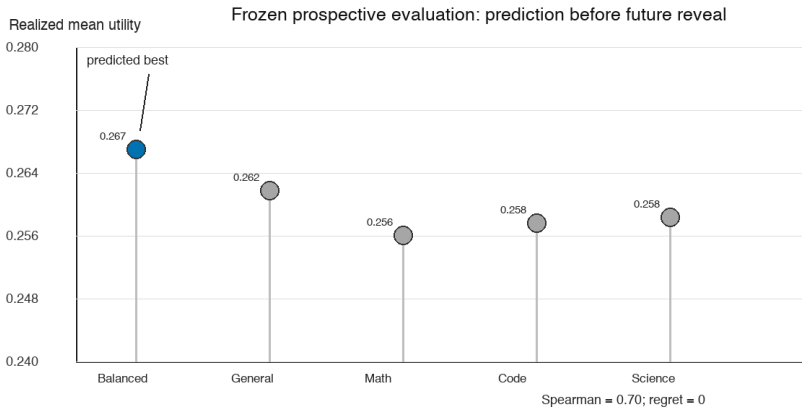


Figure 5: Frozen prospective evaluation retained in the supplement. The prediction was frozen before the five future branches were executed; the selected action was later the realized best action.

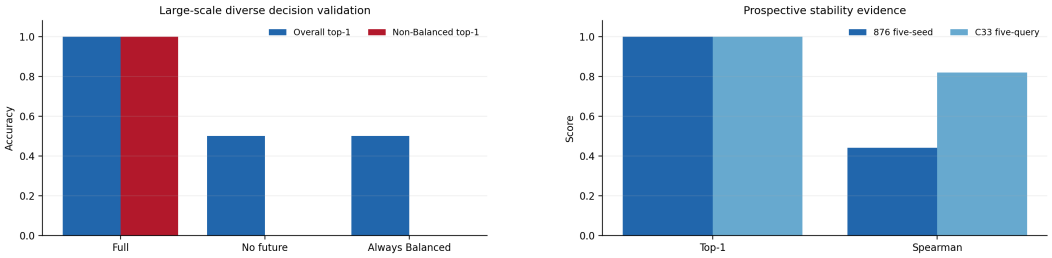


Figure 6: The 20-state diverse validation expands beyond the compact all-Balanced protocols. Full TIMELOOP reaches 20/20, while No Future Memory and Always Balanced each reach 10/20.
The five-seed result is stable at the top-1 level but ranking correlation is weaker; this is reported as robustness evidence, not independent-lineage generalization.

C.2 Large-scale diverse decision validation

The large-scale 876 protocol stratifies complete state-action groups by the realized oracle action before inspecting model predictions. Full TIMELOOP reaches 20/20 top-1. No Future Memory reaches 10/20 and Always Balanced reaches 10/20; neither control selects a non-Balanced oracle action in this set. This comparison is the principal evidence that the full method can follow heterogeneous realized outcomes rather than only a strong default mixture. The retained source does not include the full pre-stratification candidate-pool counts or a row-level anchor-isolation manifest, so the result is not interpreted as natural-prevalence or strict unseen-state generalization evidence.

Table 5: Complete large-scale diverse summary.

Method	Groups	Top-1	Mean regret	Non-Balanced top-1
Full TIMELOOP	20	20/20	0	reported by oracle stratum
No Future Memory	20	10/20	positive	0
Always Balanced	20	10/20	positive	0

D Mechanism and controller evidence

D.1 Future Signal Attribution

The natural attribution study evaluates 22 decision units under five signal constructions. Full TIMELOOP binds state, action, horizon, and realized future utility and obtains 22/22 top-1, zero regret, and Spearman 0.9818. Current-state-only, Future-only, State-only Retrieval, and Random Future each obtain 54.55% top-1. Future-only has Spearman 0.5864, above State-only's 0.4545, but neither reproduces the complete conditional signal. The conclusion is relational: realized future information is useful when attached to the state-action context that generated it.

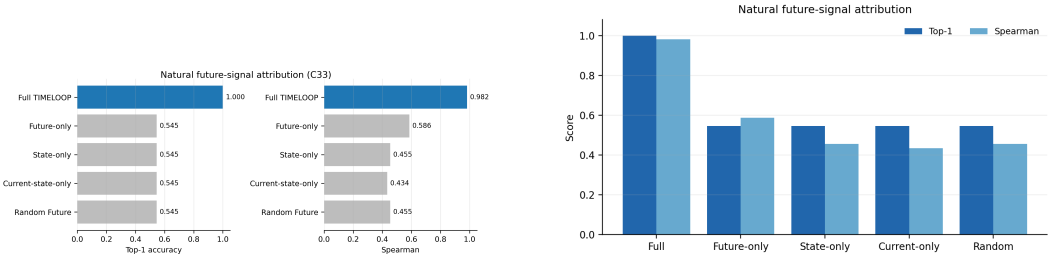


Figure 7: Natural attribution on 22 units. Full TIMELOOP preserves the state-conditioned consequence structure.
Control constructions are materially weaker and are retained as mechanism evidence rather than hidden in an aggregate.

D.2 Learned controller comparison

On the same natural-unit family, a small MLP controller reaches 21/22 top-1 and Spearman 0.9636, while a linear controller reaches 54.55% top-1 and Spearman 0.7136. The result does not claim that every parametric controller must fail. It shows that, in this tested regime, a compressed controller is not automatically equivalent to explicit Future Memory, especially when the utility law changes or support is sparse.

Table 6: Mechanism and controller metrics on natural decision units.

Method	Top-1	Regret	Spearman
Full TIMELOOP	1.0000	0.0000	0.9818
Future-only	0.5455	0.0026	0.5864
State-only	0.5455	0.0026	0.4545
Current-state-only	0.5455	0.0041	0.4339
Random Future	0.5455	0.0074	0.4545
MLP controller	0.9545	0.0000209	0.9636
Linear controller	0.5455	0.0040919	0.7136

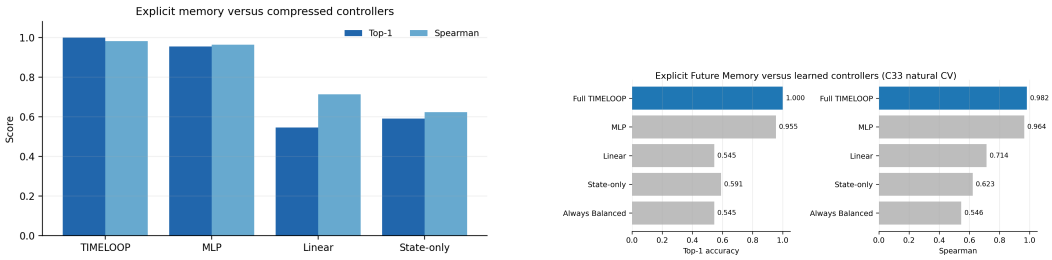


Figure 8: The learned-controller comparison is close for the MLP but not identical; linear compression is weaker.
Explicit memory and learned compression are compared under the natural-unit protocol; the boundary is about this regime, not all controllers.

E Experience acquisition and distribution shift

E.1 Cold-start versus experience-conditioned shift

The two synthetic-shift protocols use the same 20 queries, state descriptors, horizon, and shifted oracle. Their only decisive difference is whether prediction-time memory contains future evidence aligned with the shifted environment. With old memory and a changed oracle, Full TIMELOOP reaches 3/20, exactly the three Balanced cases. With shift-aligned realized evidence written into a memory copy, it reaches 20/20. We interpret this as a boundary and a positive mechanism result: TIMELOOP is not a zero-shot oracle; it performs evidence-conditioned recalibration after aligned experience is acquired.

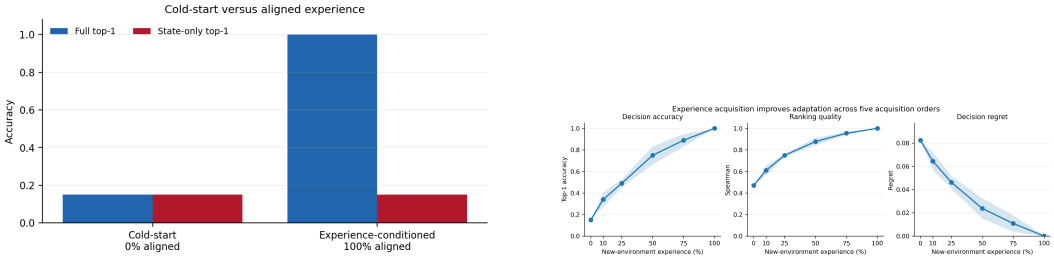


Figure 9: Cold-start remains at 3/20, whereas aligned experience reaches 20/20.
The deterministic acquisition curve connects the two endpoints as aligned evidence is added.

E.2 Five acquisition orders

To test whether the curve is an artifact of one row order, five deterministic acquisition orders were evaluated at 0, 10, 25, 50, 75, and 100 percent aligned evidence. Mean top-1 is 0.150, 0.340, 0.490, 0.750, 0.890, and 1.000; mean regret is 0.082373, 0.064426, 0.046246, 0.023752, 0.010876, and 0; mean Spearman is 0.4700, 0.6100, 0.7500, 0.8780, 0.9550, and 1.000. These are evidence-acquisition/recalibration results, not random-memory scaling or zero-shot shift prediction.

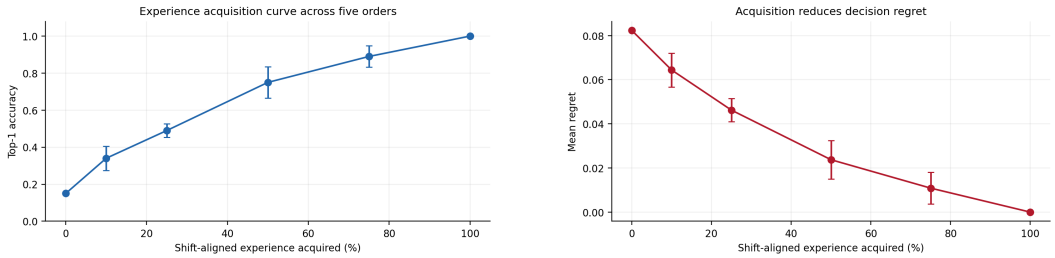


Figure 10: Mean top-1 increases with aligned evidence across five acquisition orders. Mean regret decreases across the same orders.

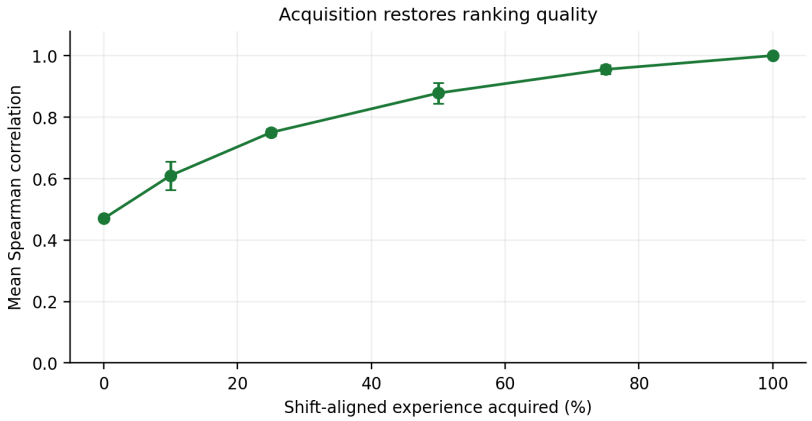


Figure 11: Mean Spearman increases with aligned evidence. Error bars show order-to-order variation; the 0% point is the cold-start boundary.

F Retrieval and representation ablations

F.1 Future Memory size

The 876 Memory Scaling study fixes 20 queries and samples five deterministic subsets for each retained ratio. At 100%, 75%, 50%, 25%, and 10%, top-1 means are 1.00, 0.88, 0.71, 0.39, and 0.23; the 10% supported-only top-1 is 0.3485, and its coverage is 0.20. The 10% condition has 34 support-limited failures and is not imputed with a default action. This distinction matters: performance loss is partly lack of retrievable evidence, not merely noisier scoring.

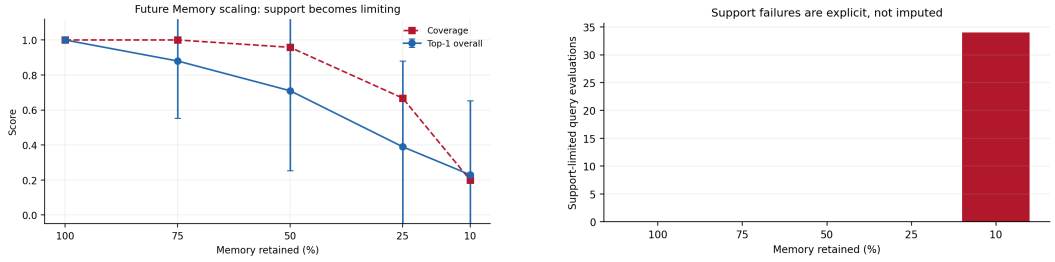


Figure 12: Memory scaling reports mean and standard deviation across five deterministic subsets.
Support-limited failures rise as the retained memory becomes sparse; no default-action imputation is used.

F.2 Retrieval K

The basic K ablation evaluates $K \in \{1, 3, 5, 10, 20\}$ on a frozen 20-query set. K=1 obtains top-1 0.5 with 20 fallback cases. K=3 obtains top-1 1.0, regret 0, and Spearman 0.975. K=5, 10, and 20 obtain top-1 0.8 with small positive regret and Spearman 0.945, 0.950, and 0.950. The result suggests that a small amount of local future evidence is sufficient in this setting, while excessively broad neighborhoods can dilute conditional structure.

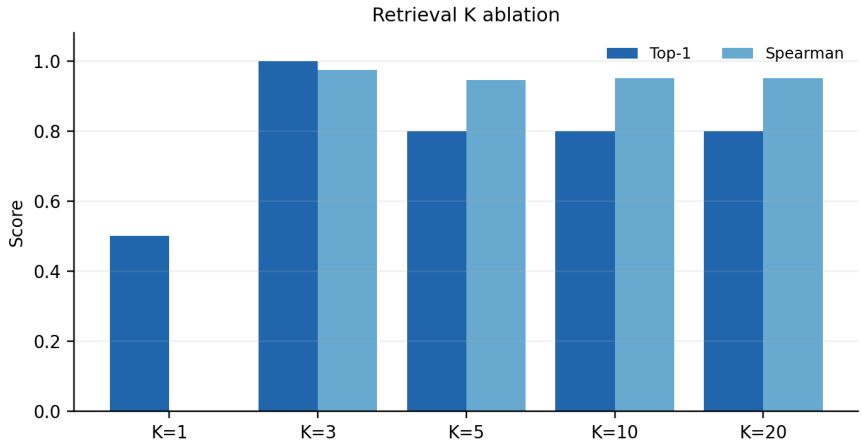


Figure 13: Basic retrieval-K ablation. K=3 is the strongest tested setting on the frozen query set; K=1 exposes fallback and K≥5 is slightly diluted.

F.3 Horizon selection

The horizon-selection ablation changes the retrieved horizon while holding evaluation utility at H=2048. Retrieved H=512 obtains top-1 0.6, regret 0.01909, and Spearman 0.36. Retrieved H=2048 obtains 1.0, zero regret, and Spearman 1.0. H=1024 has only four memory rows and all five queries fall back, so its ranking statistics are not treated as valid evidence. Exact H=4096 matching memory is unavailable in this ablation. This is not the same protocol as the separate long-horizon rollout study.

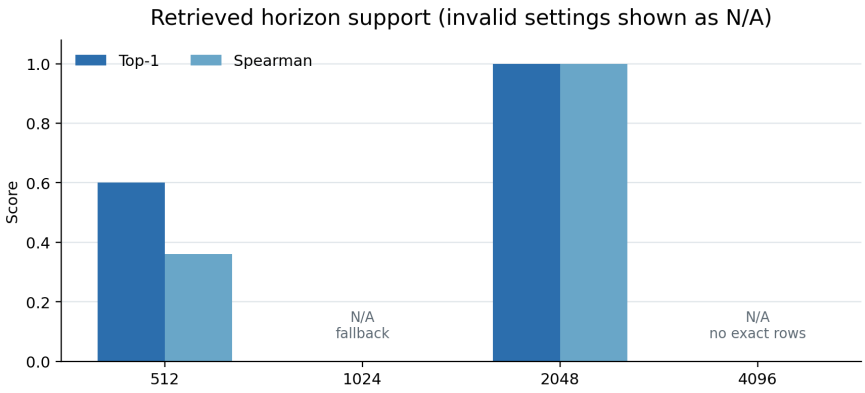


Figure 14: Horizon selection with explicit unavailable and fallback conditions. Empty or non-comparable bars are deliberate audit markers.

G Transfer and completed action-space diagnostics

The main revision does not introduce a new expanded or continuous-action rollout. The completed post-acquisition diagnostic below is retained for auditability, explicitly separated from the canonical prospective protocol, and not used to claim zero-shot scoring of unseen actions or superiority over a global prior.

G.1 Cross-seed transfer

The 876 cross-seed transfer study contains 24 complete five-action groups. Full TIMELOOP reaches 22/24 (91.7%), including 6/8 non-Balanced groups, while State-only and Always Balanced each reach 16/24 (66.7%). Full Spearman is 0.9583. This is transfer across seed-labelled groups inside the same experiment family; it is not independent-lineage generalization.

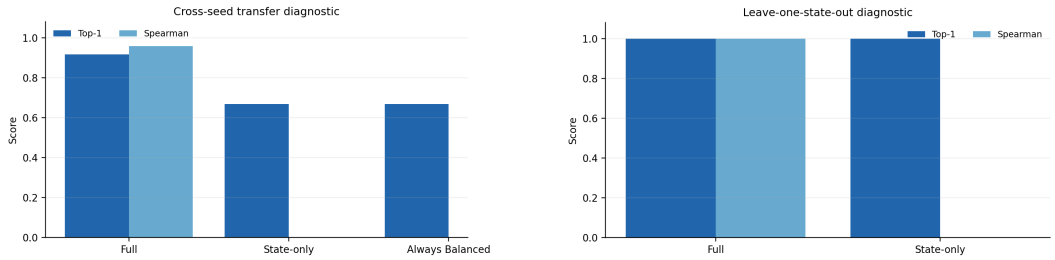


Figure 15: Cross-seed transfer improves both overall and non-Balanced selection relative to the controls.
Cross-state is diagnostic: all selected oracle actions are Balanced, so top-1 is saturated and the meaningful difference is ranking correlation.

G.2 Cross-state and cross-action diagnostics

The five-query cross-state diagnostic has all-Balanced oracle actions. Full and State-only both reach 100% top-1 and zero regret; Full Spearman is 1 while State-only is 0. The cross-action landscape contains ten groups, also all Balanced. Full Spearman/Kendall are 1/1, while No Memory and frequency controls are negative. These results are retained as diagnostics, not promoted to heterogeneous-action evidence.

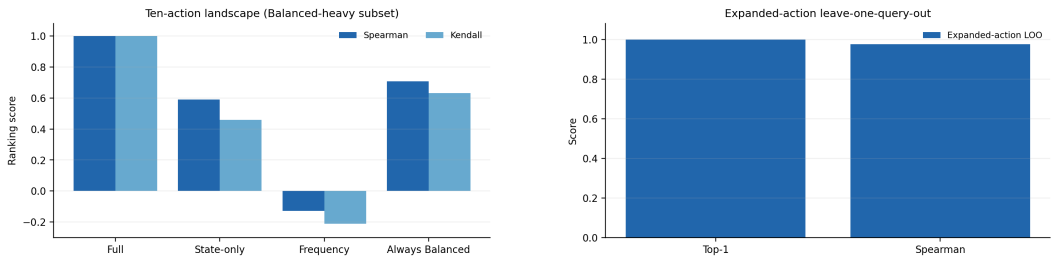


Figure 16: Cross-action landscape: ranking separates Full TIMELOOP from no-memory controls even when top-1 is saturated.
Expanded-action leave-one-query-out is post-acquisition: each held-out query's newly acquired outcomes are excluded from its own memory.

G.3 Expanded action space and global-prior saturation

The reduced expanded-action matrix contains five queries and ten actions, combining 25 historical outcomes with 25 newly acquired outcomes. The leave-one-query-out evaluation

obtains 5/5, zero regret, and Spearman 0.9758. It is explicitly post-acquisition and does not claim zero-shot scoring of unseen actions. A global action prior also obtains 5/5, and Always Diversity-Oriented-Mix saturates at 5/5; therefore the expanded-action result is not used to claim superiority over a global prior.

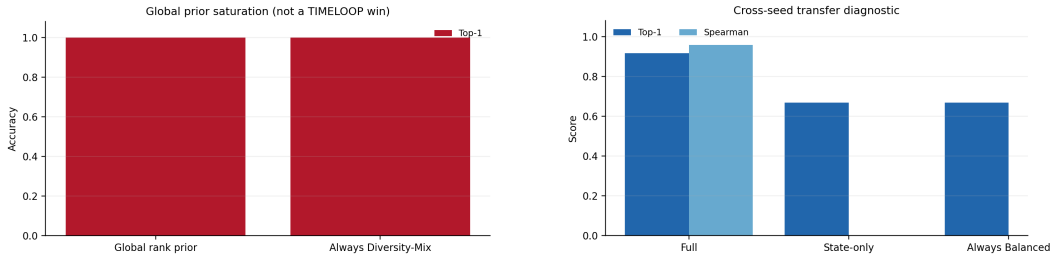


Figure 17: Global action prior saturation is retained as a negative control.
The cross-seed result provides the stronger transfer evidence used in the main narrative.

H Negative and stress evidence

H.1 Fresh cross-seed prospective stress

A fresh target seed with a C1_1024 query was evaluated under a source-only memory at H=128 and H=512. Full TIMELOOP obtained top-1 0 at both horizons, with regret 0 and Spearman 0; State-only and Always Balanced obtained top-1 1 at both horizons. This uses aggregate-loss utility rather than the canonical multi-domain utility and is therefore a negative stress protocol, not an aggregate failure claim. It demonstrates why domain, horizon, and utility alignment must be part of the evidence boundary.

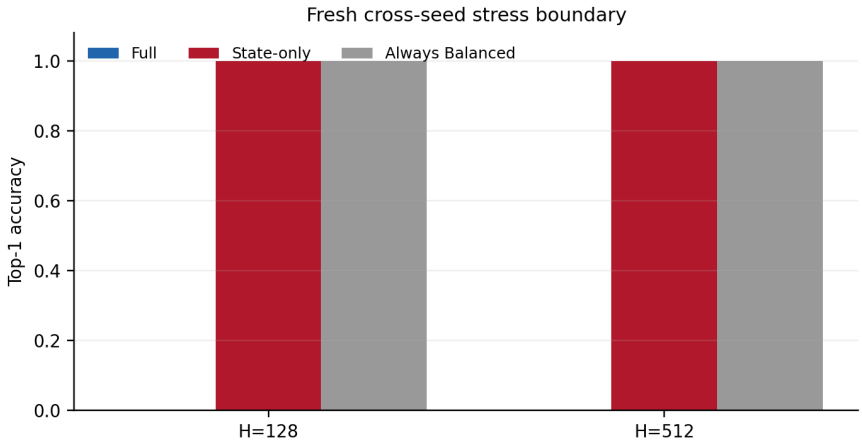


Figure 18: Fresh cross-seed stress. The negative result is preserved and separated from the canonical domain-utility protocol.

H.2 Myopic baselines

Current loss and loss slope are weak predictors in the single stress query: top-1 is false and Spearman is 0 and 0.2. A 32/128-step lookahead oracle obtains Spearman 1, while TIMELOOP H512 obtains Spearman 0.7 and Always Balanced happens to obtain Spearman 1 on this one query. The lookahead oracle is an upper-bound diagnostic, not a deployable baseline. The single-query scope prevents broad conclusions.

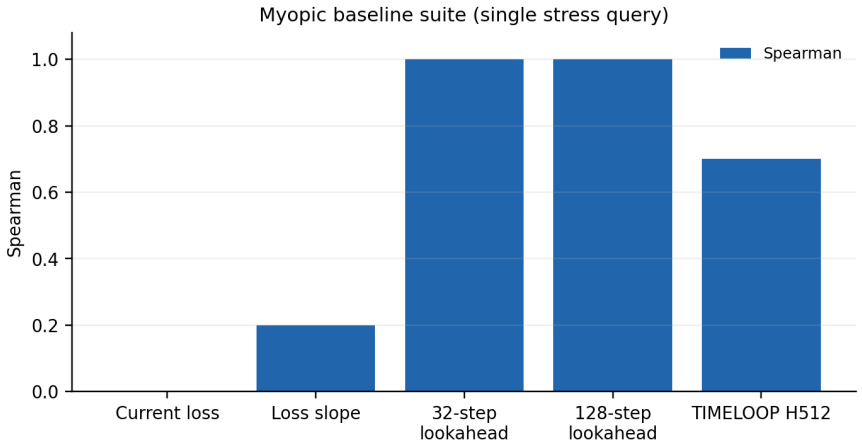


Figure 19: Myopic and short-lookahead stress baselines. Single-query scope and protocol mismatch are explicit.

I Long-horizon and controller boundaries

I.1 Long-horizon validation

C33 long-horizon validation contains 25 H=4096 branches across five queries and five actions. It changes the actual rollout horizon, unlike horizon selection, which changes only the memory retrieval condition. The main paper uses the long-horizon result as evidence that the protocol can evaluate delayed consequences; the supplement keeps it separate from H=2048 prospective selection.

Table 7: Protocol identity for the C33 long-horizon validation.

Queries	Actions	Rollout horizon	Complete branches	Interpretation
5	5	H=4096	25/25	Actual long-horizon rollout; not exact-H=4096 retrieval memory

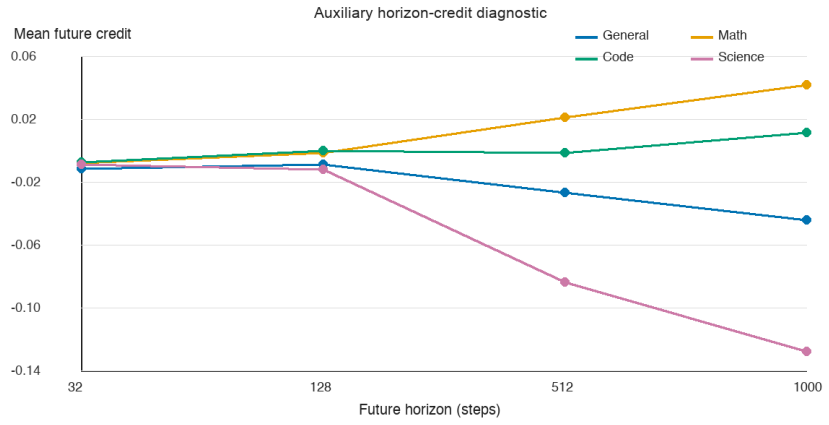


Figure 20: Auxiliary horizon-credit diagnostic over the plotted horizons H=32, 128, 512, and 1000. This figure is not the C33 25-branch H=4096 validation; the latter is identified explicitly in Table 7.

I.2 Controller and retrieval view

The explicit-memory and controller results should be read together. A learned controller can approximate the natural-unit mapping, but its compressed representation is not itself a guarantee of recalibration under a changed utility law. Explicit Future Memory exposes the evidence unit, its horizon, its action, and its realized consequence, enabling leakage audits and support diagnostics.

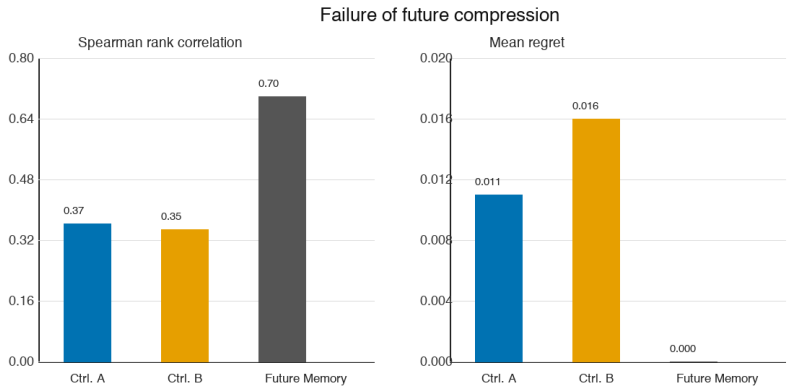


Figure 21: Controller versus retrieval view of the decision pipeline. The scientific object is the realized consequence attached to a matched intervention.

J Cost and resource accounting

The cost analysis is normalized rather than monetary. The C33 H=2048 anchor contains 25 branches and 28,607.70 observed seconds. The 876 five-seed H=2048 anchor contains 25 branches and 28,770.86 observed seconds. C33 H=4096 contains 25 branches and 53,544.72 observed seconds. Retrieval latency, controller inference/training time, and dollar prices were not recorded, so no monetary or end-to-end cost superiority claim is made.

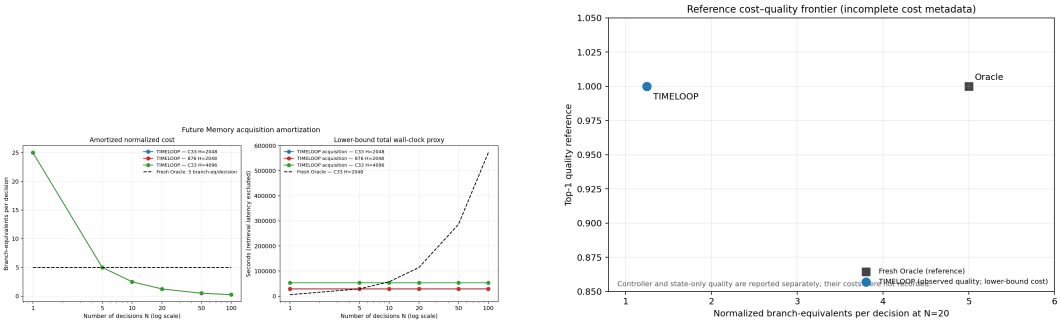


Figure 22: Normalized branch-equivalent acquisition cost per decision.
Cost-performance view with the explicit normalization boundary.

Table 8: Observed acquisition anchors.

Protocol	Branches	Observed seconds	Interpretation
C33 H=2048	25	28,607.70	acquisition anchor
876 five-seed H=2048	25	28,770.86	independent seed-labelled anchor
C33 H=4096	25	53,544.72	longer rollout anchor

K Reproducibility and leakage audits

Every formal result is tied to a manifest, branch summary, completion marker, or source report. Query selection for the large-scale diverse study is determined by oracle strata before predictions are inspected. The acquisition curve changes only which aligned evidence rows are present; it does not alter queries, action space, or utility definitions. The expanded-action leave-one-query-out evaluation excludes all newly acquired outcomes for the held-out query. The cold-start shift retains original memory while changing the oracle, making the lack of aligned evidence explicit. The separate anchor-isolation audit below distinguishes these protocol checks from a complete row-level same-anchor manifest.

K.1 Leave-one-anchor-out and memory-isolation audit

The strongest audit is available for the 876 cross-seed transfer: every one of its 24 groups has zero target-seed rows, zero target-state rows, no future read before prediction, a frozen prediction, and no synthetic labels in the machine-readable audit. The expanded-action evaluation excludes each held-out query's newly acquired outcomes, but its row-level nearest-neighbor export is not retained. The live C33 archive was inspected directly and verified to contain attribution_results.csv, metrics.json, and leakage_audit.json for the 22 natural units and 20 synthetic units; it does not contain a dedicated per-query same-anchor/nearest-neighbor manifest. For the 20-state Large-scale Diverse and 22-unit Natural Attribution families, the final package therefore retains complete groups, aggregate reports, and prospective ordering, but not a full row-level exclusion/nearest-neighbor manifest. We do not call those two families strict unseen-state or leave-one-anchor-out generalization tests.

The 20-state set is oracle-stratified before predictions are inspected, while the full pre-stratification candidate-pool counts are not retained in the canonical export. The resulting top-1 values should not be interpreted as natural prevalence estimates. Decision groups are the descriptive paired units; seed, checkpoint, and trajectory lineage can induce clustering, and the available package does not justify independent-lineage population inference or a new cluster bootstrap for these two summaries.

Table 9: Anchor-isolation audit status. "Partial" means that the protocol or aggregate report is available but a complete row-level exclusion manifest is not retained.

Protocol	Status	Direct evidence and claim boundary
876 Cross-Seed Transfer	PASS	CSV/JSON audit: target-seed rows=0, target-state rows=0, future read before prediction=False; not independent-lineage inference
876 Expanded-Action LOO	Protocol-pass	Held-out query's new outcomes excluded; post-acquisition action transfer, not zero-shot unseen-action scoring
876 Large-scale Diverse	Partial	Complete groups and oracle strata retained; no complete same-anchor/nearest-neighbor manifest
C33 Natural Attribution	Partial	Aggregate controls retained; no complete row-level same-anchor manifest
C33 Shift reconciliation	PASS (memory-copy)	Cold-start keeps old memory; aligned condition uses a separate memory copy; not zero-shot utility-law inference

Table 10: Leakage and protocol audit checklist.

Audit item	Status	Evidence boundary
Query chosen without prediction inspection	PASS	Oracle-stratified selection source report; candidate-pool counts not retained
Held-out query outcome excluded before prediction	PASS	Expanded-action LOO protocol; row-level export not retained
Same-anchor exclusion certified for every query	PARTIAL	PASS for cross-seed; incomplete for 20-state/22-unit exports
Cold-start uses old memory	PASS	Synthetic shift reconciliation
Aligned recalibration uses a memory copy	PASS	Experience acquisition audit
Support failures imputed with a default action	NO	10% memory condition reports failures
Unavailable horizon treated as valid ranking	NO	H=1024/H=4096 labels retained
Global prior superiority claimed	NO	Prior saturation kept as negative control
Monetary cost superiority claimed	NO	Only normalized seconds are reported

L Complete result-allocation map

Table 11: Where each experiment is used in the manuscript.

Experiment	Main-text role	Supplement role / boundary
C33 Multi-query	Q1 stability	full 25-branch protocol; all-Balanced boundary
876 Multi-seed	Q1 stability	seed-level transfer caveat
Large-scale Diverse	Q1 heterogeneous validation	complete strata and controls
Future Signal Attribution	Q2 mechanism	all five signal constructions
Learned Controller	Q2 compression comparison	natural-unit and shift boundary
Experience Acquisition	Q3 recalibration	five order means and error bars
Synthetic reconciliation	Q3 boundary	cold-start versus aligned evidence
Memory Scaling	Q4 support sensitivity	coverage and failure accounting
K / Horizon ablations	Q4 retrieval sensitivity	unavailable and fallback cases
Cross-seed / Cross-state	Q4 transfer	saturated and bounded diagnostics
Expanded Action LOO	Q4 action extension	post-acquisition qualification
Fresh stress / Myopic	limitation	negative results preserved
Cost analysis	limitation	normalized acquisition anchors only

M Claim boundaries

The evidence supports the claim that, in the tested pretraining decision protocols, realized future consequences can serve as reusable supervision when stored with their state, action, and horizon context. It does not support a universal optimizer claim, a zero-shot distribution-shift claim, a strict unseen-state generalization claim for the partially audited 20-state/22-unit families, a claim that every parametric controller fails, or a monetary cost advantage. The method is best understood as experience-conditioned decision support: it converts acquired training outcomes into auditable evidence for a later choice.

N Branch-level audit tables

The aggregate metrics in the main paper are backed by complete group-level outcomes. The following compact tables make the unit of analysis explicit and help distinguish a complete group from an incomplete branch inventory.

Table 12: Five-query and five-seed prospective group audit. The oracle label is shown because both compact protocols are all-Balanced.

Protocol	Group	Query / seed	Actions	Selected	Oracle	Boundary
C33 Multi-query	Q1	frozen query	5	Balanced	Balanced	complete
C33 Multi-query	Q2	frozen query	5	Balanced	Balanced	complete
C33 Multi-query	Q3	frozen query	5	Balanced	Balanced	complete
C33 Multi-query	Q4	frozen query	5	Balanced	Balanced	complete
C33 Multi-query	Q5	held-out query	5	Balanced	Balanced	complete
876 Multi-seed	20260915	C3_3072	5	Balanced	Balanced	complete
876 Multi-seed	20260916	C3_3072	5	Balanced	Balanced	complete
876 Multi-seed	20260917	C3_3072	5	Balanced	Balanced	complete
876 Multi-seed	20260918	C3_3072	5	Balanced	Balanced	complete
876 Multi-seed	20260919	C3_3072	5	Balanced	Balanced	complete

Table 13: Exact attribution and retrieval-ablation values retained for audit. N/A means the score vector is constant or the support condition is not met.

Condition	Top-1	Regret	Spearman	Audit note
Full TIMELOOP, natural	1.0000	0.0000	0.9818	22 units
Current-state-only	0.5455	0.0041	0.4339	natural attribution
Future-only	0.5455	0.0026	0.5864	natural attribution
State-only Retrieval	0.5455	0.0026	0.4545	natural attribution
Random Future	0.5455	0.0074	0.4545	natural attribution
K=1	0.5000	0.008126	N/A	20 fallback cases
K=3	1.0000	0.000000	0.9750	frozen 20-query set
K=5	0.8000	0.000110	0.9450	frozen 20-query set
K=10	0.8000	0.000110	0.9500	frozen 20-query set
K=20	0.8000	0.000110	0.9500	frozen 20-query set

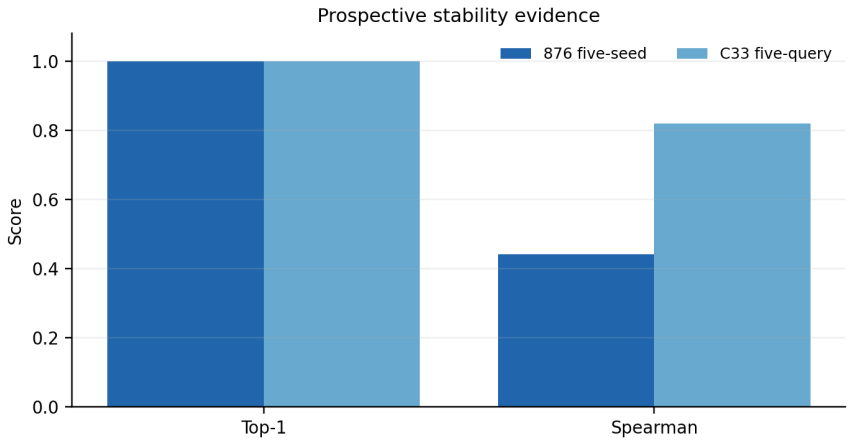


Figure 23: Per-group view of the compact five-seed protocol. The group inventory is complete, but its shared query and all-Balanced oracle constrain the interpretation.

O Statistical interpretation and reporting rules

The evidence package includes paired statistical checks for the principal comparisons. Large-scale diverse Full TIMELOOP versus No Future Memory has exact McNemar $p = 0.001953$, Holm-adjusted $p = 0.009766$ within the registered comparison family. Cross-seed transfer's paired regret improvement has a bootstrap 95% interval of $[0.000751, 0.003999]$. These inferential summaries are secondary to the protocol audit: they do not turn a bounded set of state groups into a claim about all training runs. The statistical unit is the decision group; seed, checkpoint, and trajectory lineage are potential clusters. Because complete cluster tables are not retained for the 20-state and 22-unit exports, no independent-lineage or cluster-bootstrap interpretation is made for those summaries.

O.1 What counts as a valid comparison

We treat a comparison as valid only if the query state, action set, horizon, utility definition, and decision-time information are matched. A result is marked diagnostic when one of these dimensions changes. In particular, the fresh cross-seed stress uses aggregate-loss utility and is not pooled with the canonical shared-domain utility. The long-horizon rollout experiment is not pooled with retrieved-horizon selection. H=1024 is not assigned a ranking score when every query falls back because only four memory rows are available.

O.2 Saturation and negative results

Top-1 saturation is not treated as a universal success. The cross-state and cross-action diagnostics have all-Balanced oracle distributions, so their top-1 values are not discriminative; ranking and control behavior are therefore reported. The global prior saturates at 5/5 on the reduced expanded-action set, so it is retained as a warning against claiming superiority over a simple prior. The cold-start shift and fresh cross-seed stress are retained because they prevent an overbroad "recalibration everywhere" narrative.

O.3 Reproducibility checklist

Table 14: Final reporting checklist used before manuscript packaging.

Check	Result	Action
Main claim tied to a prospective protocol	PASS	retain Q1 result
Mechanism claim tied to matched controls	PASS	retain attribution table
Recalibration claim tied to acquired evidence	PASS	distinguish cold-start
Unsupported zero-shot claim	ABSENT	boundary in conclusion
Unsupported monetary claim	ABSENT	normalized seconds only
Post-acquisition result labelled	PASS	expanded-action caption
Unavailable support labelled	PASS	no imputation
Instance separation verified	PASS	C33 / 876 paths separate
Old source artifacts overwritten	NO	v3 is a new directory

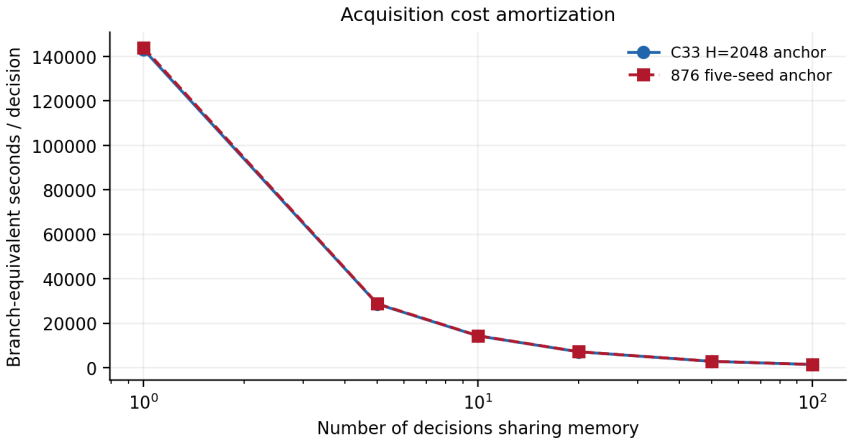


Figure 24: Final normalized cost view used in the reporting boundary. It is not a dollar-cost or latency comparison.

P Alternative explanations: why simpler methods fail

The main paper groups the strongest comparisons by scientific claim; this section collects the simpler alternatives in one place. Always Balanced obtains 10/20 on the large-scale diverse set and selects no non-Balanced oracle action. State-only, current-state-only, future-only, and randomized controls each obtain 54.55% top-1 on the 22-unit natural attribution family, despite different ranking quality. The learned MLP controller reaches 21/22 on that natural family, showing that compression can approximate the mapping in-distribution, but the explicit record remains auditable and the cold-start stress still defeats both methods when the utility law changes. Finally, the global action prior saturates at 5/5 on the reduced expanded-action matrix, so the action-space result is not presented as superiority over every global prior. Together these controls remove specific alternative explanations without claiming that every possible controller must fail.

Q Failure cases and operational boundaries

TIMELOOP is intentionally not presented as a zero-shot oracle. With old memory and a changed oracle, the cold-start shift reaches only 3/20; aligned new evidence raises the same controlled protocol to 20/20. In the fresh cross-seed stress, protocol-mismatched aggregate-loss utility yields top-1 zero at both tested horizons, which is retained as a warning about state, horizon, and utility alignment. Sparse memory creates support failures rather than an automatic fallback: the 10% memory condition has 34 unsupported cases and 0.20 coverage. Horizon mismatch is similarly explicit: H=1024 has only four rows and all queries fall back, while exact H=4096 retrieval records are unavailable. These failures define where the method should abstain, acquire more evidence, or revise the protocol instead of fabricating a ranking.

Source-artifact index

The final evidence package preserves the raw reports, manifests, frozen predictions, branch summaries, audit files, generated figures, and this manuscript source. C33 and 876 artifacts remain in their original instance-specific directories. This canonical manuscript is compiled as one integrated PDF from the source and the local ICLR 2027 style files. No additional GPU experiment was required for the rewrite; all reported numbers above are transcribed from the completed evidence package.

R Statistical robustness and evidence availability

The following tests are post-hoc analyses over completed artifacts, not new prospective rollouts. The five paired top–1 comparisons were analyzed with a two-sided exact McNemar test and Holm correction across the executed family. Regret intervals use 10,000 paired percentile-bootstrap resamples with seed 20250925. Spearman intervals are reported only when the compared score vectors have non-constant support. The purpose of this section is to make the inferential boundary inspectable, not to turn the bounded state sets into a claim about all training runs.

Table 15: Executed statistical robustness family. ΔR is baseline regret minus Full TIMELOOP regret; positive values favor Full TIMELOOP.

Comparison	n	exact p	Holm p	ΔR 95% CI
876 Large-scale: Full vs No Future	20	.001953	.009766	[0.001029, 0.004853]
876 Large-scale: Full vs Always Balanced	20	.001953	.009766	[0.001029, 0.004853]
876 Cross-Seed: Full vs State-only	24	.031250	.093750	[0.000751, 0.003999]
876 Cross-Seed: Full vs Always Balanced	24	.031250	.093750	[0.000751, 0.003999]
C33 Diverse: Full vs No Future	5	.250000	.250000	[0.000087, 0.003758]

The family is reported exactly as executed. The live C33 archive was inspected directly and verified to contain attribution_results.csv, metrics.json, and leakage_audit.json for the 22-unit natural attribution study and the 20-unit synthetic control. The local manuscript package does not mirror those remote rows, and the archive does not export a dedicated per-query same-anchor/nearest-neighbor manifest. We therefore do not invent missing p-values, confidence intervals, or row-level significance claims. The canonical aggregate reports remain descriptive evidence, while the executed inferential family is restricted to comparisons with complete paired inputs. This distinction is also recorded in the completeness audit and evidence ledger shipped with the manuscript.

R.1 Confidence intervals and coverage conventions

The Wilson intervals and paired-bootstrap files use the same unit identifiers as the corresponding aggregate tables. A support-limited retrieval query is not silently assigned the default Balanced action. In particular, the 10% memory condition retains 34 unsupported failures, and the H=1024 horizon condition is reported as unavailable when exact support is insufficient. These choices keep coverage, ranking quality, and fallback behavior separate rather than collapsing them into a single optimistic score.

S Result allocation and source-artifact map

The manuscript separates discovery, support, qualification, and boundary evidence. Main-text figures carry the shortest chain needed to understand the central claim; the tables and figures below preserve the complete protocol record for audit and reuse.

Table 16: Result-allocation map for the final manuscript.

Evidence family	Function	Main-text location	Supplement destination / boundary
C33 Multi-query	prospective stability	Results, query consistency	A–B; all oracle actions Balanced
876 Multi-seed	seed-labelled stability	Results, stability boundary	A–B; not independent-lineage transfer
Large-scale Diverse	heterogeneous decisions	Results, default-action control	B; oracle strata fixed before predictions
Future Signal Attribution	mechanism identification	Results, signal attribution	C; remote rows/audit verified; no row-level anchor manifest
Learned Controller	compression comparison	Results, controller boundary	C/H; no universal controller claim
Experience Acquisition	recalibration mechanism	Results, acquisition curve	D; five deterministic orders
Memory Scaling	support sensitivity	Results, support conditions	E; unsupported cases retained
K and horizon ablations	retrieval sensitivity	Results, parameter boundary	E; unavailable support is explicit
Cross-seed / cross-state	transfer diagnostics	Analysis	F; bounded protocol transfer
Expanded Action LOO	post-acquisition extension	Analysis	F; not zero-shot action scoring
Cold-start / fresh stress	negative evidence	Limitations	G/N; protocol mismatch retained
Cost and compute	practical boundary	Results / Discussion	I; normalized only

S.1 Canonical source families

The evidence package is organized into four layers. First, source reports preserve the human-readable experiment conclusions. Second, the master ledger maps each claim to an instance, protocol, unit count, and file. Third, machine-readable CSV/JSON files preserve paired rows, leakage checks, and statistical outputs when those rows are available. Fourth, the LaTeX source and rendered pages provide the final presentation layer. C33 and 876 remain separate at every layer; a combined table in this supplement is a reporting view, not a merged raw experiment directory.

Table 17: Machine-readable evidence available in the final package.

Evidence family	Complete units	Audit / source artifact
876 Large-scale Diverse	20 groups	evidence/CORE_EVIDENCE_TABLE.csv; paired IDs in statistical robustness files
C33 Natural Attribution	22 units	remote attribution_results.csv and leakage audit verified; no row-level anchor manifest
876 Cross-Seed Transfer	24 groups	evidence/transfer/cross_seed_results.csv and leakage audit
Experience Acquisition	20 queries × 5 orders	trend JSON and order-level report
Memory Scaling	20 queries × 5 samples per ratio	source report and support-failure summary
Statistical robustness	executed paired family	McNemar/Holm, bootstrap, and Wilson files

T Boundary ledger for interpretation

The boundary cases are part of the result rather than postscript disclaimers. They determine how the central conclusion must be read.

Table 18: Interpretive boundaries preserved in the final artifact.

Boundary	Observation	Required interpretation
Cold-start shift	Full TIMELOOP reaches 3/20 with old memory and a changed oracle	No zero-shot inference of an unseen utility law
Aligned shift	Full TIMELOOP reaches 20/20 after shifted evidence is acquired	Recalibration is experience-conditioned
All-Balanced compact protocols	Multi-query and multi-seed are 5/5	Stability, not heterogeneous-action proof
Global prior saturation	Global rank prior reaches 5/5 on five expanded-action queries	Expanded-action result does not prove prior superiority
Sparse memory	10% retention has 34 unsupported failures	Low performance partly reflects missing coverage
Horizon mismatch	Exact H=4096 support is unavailable for retrieval ablation	Long rollout and retrieved-horizon selection are distinct
Cost accounting	Dollar price and latency were not recorded	Report normalized acquisition only

T.1 What the evidence does and does not establish

Taken together, the experiments support a mechanism-level conclusion: realized future consequences can be reused as decision supervision when their state, action, and horizon context are retained. They do not establish a universal optimizer, a guarantee under arbitrary distribution shift, or a proof that explicit memory is always superior to a sufficiently large learned controller. The strongest positive claim is therefore conditional and testable: within a matched training ecosystem, acquiring and auditing future experience improves the reliability of later long-horizon decisions.

U Reproducibility handoff

The final package contains the single compiled manuscript, the complete LaTeX source, the figure-generation script for the upgraded overview, the evidence ledger, the statistical robustness outputs, the source reports, and checksums. The source files identify the two compute-instance boundaries and the protocol status of every aggregate used in the paper. A future reader can therefore reproduce the document assembly and audit the reported numbers without assuming that a row-level anchor manifest or an unavailable local copy of the remote C33 rows has been reconstructed.

Table 19: Final handoff checks.

Check	Status	Artifact
Main text is at most nine pages before AI statement and references	PASS	final PDF QA
AI Use Statement appears before references	PASS	main source
Supplement follows bibliography in the same PDF	PASS	final PDF QA
Figure 1 is a three-part vector overview	PASS	figure1_overview.pdf
Missing raw paired rows are disclosed, not fabricated	PASS	completeness audit
C33 and 876 evidence remain separated	PASS	master ledger / source map
Negative and saturated controls remain visible	PASS	boundary ledger