EPISODESCOPE: BUDGET-AWARE EVALUATION OF DIVERSITY CORRECTIONS IN MOLECULAR AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards can reduce output diversity, and several corrections restore it within a rollout group or a selected batch. We ask whether local diversity improves discovery per oracle call when a multi-turn agent also controls its oracle spend. We introduce EpisodeScope, an evaluation protocol for multi-turn molecular agents under an unenforced oracle budget. It counts distinct molecules above a baseline-calibrated score threshold per charged call within a validation window and reports realised spend and coverage at common call counts. In four seed-matched Qwen3-4B pairs trained separately under the same per-step scoring cap, farthest-first selection raises within-episode fingerprint spread but lowers qualified discovery per call by 0.170–0.309 in every pair. Repeated evaluations concentrate on its first-position anchor. A registered, non-concurrent random-anchor control still trails uniform selection in all three same-machine pairs, and determinantal selection gives mixed results. On replayed candidate pools, avoiding already-scored molecules reduces repeats most consistently. Enabling the complete memory configuration raises discovery per call in all five prospectively specified pairs (median difference $+0.720$, one-sided sign test $p = 0.03125$), although one memory run spends 15.67 calls per episode. At this operating point, higher local diversity need not improve qualified discovery per call.

1 INTRODUCTION

Reinforcement learning for molecular design must preserve structural variety as the policy learns to score well. Reduced output diversity after reinforcement learning with verifiable rewards is well documented (GX-Chen et al., 2025). Corrections act at different horizons. Reward reweighting, conditional uniformity, cooperative credit and set-level objectives act on a rollout group (Sinha et al., 2026; Lochab et al., 2026; Chen et al., 2026; Park et al., 2026; Cao et al., 2026; Bahlous-Boldi et al., 2026; Cai et al., 2026; Yuan et al., 2026; Ye et al., 2026), diverse mini-batch selection acts on the molecules chosen for scoring (Svensson et al., 2025), and persistent molecular memory penalises molecules found in earlier episodes (Blaschke et al., 2020; Guo & Schwaller, 2024). The distinction between within-batch diversity and cross-iteration recycling is established in self-play reasoning (Li et al., 2026). Multi-turn agents also control cost by choosing how many proposals to send to the oracle. Higher discovery counts may therefore reflect greater expenditure rather than more efficient exploration (Gao et al., 2022).

However, existing evaluations of local corrections fix the batch size or the oracle budget, whereas a multi-turn agent under a soft budget sets its own spend. The molecular evaluation of inverse-probability scaling (IPS) reports thresholded unique-molecule yield and the calls needed to reach a yield target at fixed budgets, with memory active in both arms (Sinha et al., 2026). Diverse mini-batch selection has been compared with generating only the target batch size (Svensson et al., 2025), which is equivalent in distribution to a uniform subset when candidates are sampled independently. A multi-turn agent instead proposes a joint list at each turn and controls its charged calls. Neither design tests selector choice within an agent-generated proposal list when realised spend and later proposals can both change.

To address this gap, we introduce EpisodeScope, an evaluation protocol that combines the established qualified-yield numerator with charged-call accounting and same-cap online selector compar-

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.

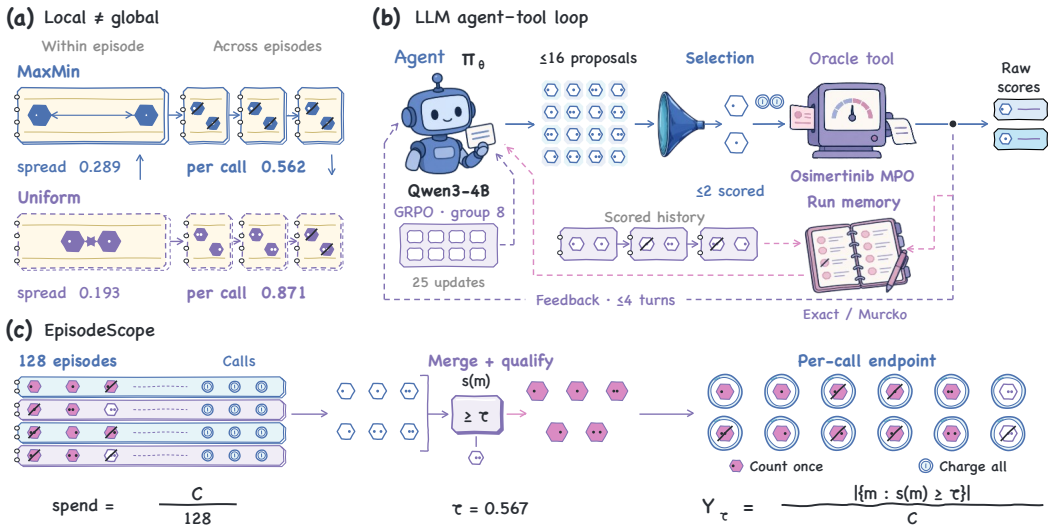


Figure 1: **EpisodeScope separates local diversity from discovery per charged call.** (a) In matched pair 82, farthest-first (MaxMin) selection raises within-episode spread and lowers qualified discovery per call against uniform selection. The four values are measured. The trajectories are schematic. (b) The agent loop with the selection and memory sites. Figure E.1 shows all three intervention sites. (c) Every charged call counts in the denominator C and each qualifying molecule counts once. Glyph counts are schematic.

isons. The rate counts distinct canonical molecules above a baseline-calibrated threshold per charged call, repeats included. We report realised spend and coverage at common call counts alongside it (Figure 1). In four seed-matched Qwen3-4B pairs with different assigned selectors, farthest-first selection increases the within-episode fingerprint spread it targets but lowers qualified discovery per call by 0.188, 0.292, 0.309 and 0.170 (Figure 2b). Separate training yields a total online contrast, not a fixed-pool comparison. Farthest-first spends up to 2.8 times as many calls per episode and leads per episode where it spends more, yet at 224 common charged calls uniform selection has found 44, 49, 43 and 59 more distinct molecules. Repeated evaluations concentrate on the selector's starting molecule, its anchor. Our implementation takes it from the first valid proposal of a turn, whereas published MaxMin-A draws it at random. A registered, non-concurrent random-anchor control still trails uniform selection in all three same-machine pairs. Avoiding previously scored molecules reduces repeats most consistently on logged candidate pools (Section 4.1).

The determinantal selector and IPS added to memory give mixed results (Table 1). Within-group outcome multiplicity tracks raw discovery across runs ($\rho = +0.80$) without determining it (Figure C.4). Enabling the complete memory configuration raises discovery per call in all five prospectively specified pairs at commanded budget two (median paired difference $+0.720$, exact one-sided sign test $p = 0.03125$, Figure 4). The contrast neither isolates the scaffold clause nor separates training from validation effects. A historical cohort of twenty-nine completed runs provides descriptive corroboration (Appendix A.1). A pre-registered Qwen3-1.7B replication reproduces the ranking without full separation (Section 6).

We make three contributions.

1. We define EpisodeScope for multi-turn agents under an unenforced call budget (Section 3). It combines thresholded yield with charged-call rates, realised spend, a same-cap selector control and coverage at common calls. Repeated evaluations count in the denominator, and distinctness is defined within the validation window (Sinha et al., 2026; Henderson et al., 2018).
2. We evaluate diverse selection in separately trained, same-cap pairs. Four seed-matched pairs compare first-start farthest-first with uniform selection, while a registered, non-concurrent random-anchor control and a replay on logged candidate pools test the reversal. Three pairs evaluate the determinantal rule (Section 4.1).

3. We evaluate complete memory configurations in five prospectively specified pairs at the same commanded budget and service (Section 4.2). Realised spend remains an outcome, not a matched condition.

These findings show that the measured local diversity statistics do not determine within-window discovery efficiency under this soft-budget protocol. The established calculations in Section 5 provide context, not an explanation of these effects.

2 Related work

The nine group-level corrections cited above differ in what they reweight, from outcome frequency to rarity and update geometry. Our evaluation covers IPS and a top-k reward only. The limits in Section 5 apply to a subset of the nine. Niu et al. (2026) and Gai et al. (2025) examine related group-reward behaviour.

IPS averages three seeds in its molecular evaluation and keeps Murcko memory active in both arms (Sinha et al., 2026). Our rate uses the same qualified-yield numerator to compare multi-turn runs with unequal realised calls, alongside common-call coverage. Supergroup GRPO compares with memory-assisted GRPO and reports gains in its single-turn setting (Ye et al., 2026), which rules out any field-wide claim that an established filter dominates newer learning rules. Tool-guided molecular RL, group degeneracy and the loss of diversity after RL have all been studied before (Yang et al., 2025b; Zhong et al., 2026; GX-Chen et al., 2025). Li et al. (2026) separate within-batch diversity from cross-iteration recycling in self-play reasoning and address the latter with persistent memory. We take that separation as a premise. The molecular selection study (Svensson et al., 2025) counts distance-separated active molecules and scaffolds across ten runs at a fixed evaluated batch size, whereas our qualified count has no distance radius and our online control chooses from each agent-generated proposal list at the same per-step cap. Multi-turn tool-use agents have used call budgets for rollout allocation (Hu et al., 2026), and molecular scaffold memory has been combined with score caching on the practical molecular optimisation (PMO) benchmark (Guo & Schwaller, 2024). We contribute the measured selector reversal, its anchor analysis and variable-spend accounting in this setting. Qualified yield, the memory rule and the local-versus-global distinction all come from prior work.

Thomas et al. (2025) find diversity filters the only exploration strategy for chemical language models that avoids severe penalties, and attribute this to short budgets that leave memory buffers too empty to penalise rewards. Our treated runs have a non-zero filter footprint (Table A.4), confirming reward penalties at these budgets.

3 Setting and evaluation

Task and training. An episode allows the policy up to four turns on Osimertinib_MPO from the Therapeutics Data Commons (Huang et al., 2021). Each turn may propose up to sixteen candidate SMILES strings. The environment parses and scores valid proposals and returns their scores. An unparseable action incurs a penalty of 0.1. Qwen3-4B (Yang et al., 2025a) is trained with group relative policy optimisation (GRPO) (Shao et al., 2024) through verl-agent (Feng et al., 2025) for 25 updates, using groups of eight rollouts and vLLM generation. Replicates vary the environment seed, not independent model initialisations. Appendix F gives the remaining settings, the configuration differences between machines (Appendix F.2) and the choices required for reimplementation (Appendix F.1).

Soft budgets and commanded values. The prompt states the episode's oracle budget, but nothing enforces it. The call-penalty coefficient of the training objective is zero, and no rule stops an episode early. The experiments use two distinct commanded budgets. The historical cohort, the prospective memory pairs, the memory-plus-IPS pairs, the budget-two repeat-only pairs and the second-size replication command two calls per episode. The selection screens of Section 4.1, the budget-eight memory-clause ablation and the bucket screen of Appendix D.6 command eight, shared by both arms of each pair. Final mean spend in the historical cohort ranges from 1.59 to 11.98 calls per episode, exceeding the command in 23 of 29 runs. We report every rate alongside realised spend and cumulative coverage at common call counts. This compares soft-budget efficiency without

guaranteeing hard-budget performance. Appendix A.4 reports training calls, validation passes and over-budget episodes.

Endpoint and evaluation window. Each completed run contains 3968 episode records. The final validation pass contains 128 environment restarts at temperature 1.0. Persistent memory changes across those episodes, so the training run remains the unit of replication. Let C denote charged calls in that pass, one per molecule the oracle scores, repeats included, and $s(m)$ the best raw oracle score for canonical molecule m within that pass. For a qualification threshold τ we report

$$Y_\tau = \frac{|\{m : s(m) \geq \tau\}|}{C}, \qquad \text{spend} = \frac{C}{128}. \tag{1}$$

Raw novelty counts all valid distinct molecules. The qualification threshold on the primary target is $\tau = 0.567$, the ninetieth percentile of the untrained policy's pooled-budget scores. It was set after the original raw-novelty analysis and differs from the environment's success threshold of 0.5, so we report both readings and a threshold sweep (Figure D.1, Appendix D.1). Novelty is relative to the evaluation window, not a claim that a molecule is new to chemistry or absent from the policy's training history. Qualification uses raw scores. Resubmitting a molecule already counted in the pass adds a charged call without increasing the numerator, so it cannot inflate the endpoint (Figure 3a). If the numerator is positive and other calls are fixed, that repeat strictly lowers the rate. A molecule seen before this pass still counts if scored in it, so Equation 1 measures within-window, not campaign-new, discovery.

Interventions. Plain GRPO uses the episode's best score. Our top-k distinct reward averages the three highest scores among distinct canonical molecules ($k = 3$), with missing entries contributing zero. IPS reweights each outcome's reward by the inverse of its empirical frequency in the group, $w(o) = n/k_o$ for an outcome seen k_o times in $n = 8$ rollouts (Sinha et al., 2026), applied before group standardisation (Appendix F.1). The memory arm uses persistent exact-repeat suppression followed by a Bemis–Murcko (Bemis & Murcko, 1996) scaffold bucket. It zeroes the reward of repeated molecules, or those clearing the filter's minimum score of 0.4 (Loeffler et al., 2024) whose scaffold bucket of 25 is full. Every molecule scored updates memory during both training and validation passes, and the agent's observation marks zeroed molecules. The filter changes policy observations, while the endpoint uses raw scores (Appendix D.2).

Populations and comparisons. The historical cohort is fixed at 29 completed runs from four arms sharing one training recipe (Appendix A). Three requested memory interventions did not execute because an older scoring service ignored the filter parameter. They remain in the unfiltered group. Their requested labels are retained in the per-cell table (Appendix A.1), and the reclassification is independently checked against each run's recorded rewards (Appendix A.5). The same three seeds were re-run on a verified service and enter the memory group (Appendix A.2). The resulting memory group has seven runs. Selection pairs are matched within a machine. The second-target extension uses complete machine–seed blocks and reports each machine separately. Selector process randomisation was not recorded, limiting bitwise reproduction across service restarts (Appendix B.9). Repeated validation episodes and checkpoints do not increase the number of independent pairs.

4 Discovery, local diversity, and persistent memory

4.1 Selection at the same scoring cap

Within each oracle step, every selector receives that step's valid proposal list and may score at most two molecules. The arms train separately, producing different proposal lists and later policy states. Each contrast measures the total online effect. Both arms use persistent memory under a commanded budget of eight oracle calls per episode. Our farthest-first rule (MaxMin) uses Svensson et al.'s MaxMin-A distance, the average Morgan and scaffold atom-pair dissimilarity. It starts from the first valid proposal, whereas MaxMin-A draws the first element at random. The control samples its own list uniformly without replacement, seeding the generator from the hash of the ordered candidate list. Identical lists therefore receive identical draws within one service process. Contrasts are paired by training seed within machine, and the scoring service was checked against a reference score before training. Appendix B details the selection analyses.

Table 1: Every paired contrast, as qualified discovery per charged call ($\tau = 0.567$, final 128-episode pass), first arm minus second. The arms of a pair share seed and machine and train separately. Budget is the commanded calls per episode. Only the prospective memory contrast has a pre-specified test (one-sided sign test $p = 0.03125$). The other rows, with three or four pairs, are descriptive.

Contrast	Budget	Positive	Median	Range
Farthest-first (first start) − uniform	8	0/4	−0.240	−0.309 to −0.170
Farthest-first (random start) − uniform	8	0/3	−0.162	−0.289 to −0.156
Determinantal − uniform	8	1/3	−0.088	−0.175 to +0.092
Memory + IPS − memory	2	1/3	−0.135	−0.213 to +0.050
Memory on − off, prospective	2	5/5	+0.720	+0.115 to +0.893
Memory − repeat-only	8	2/4	+0.055	−0.113 to +0.130
Memory − repeat-only	2	2/3	+0.031	−0.039 to +0.106

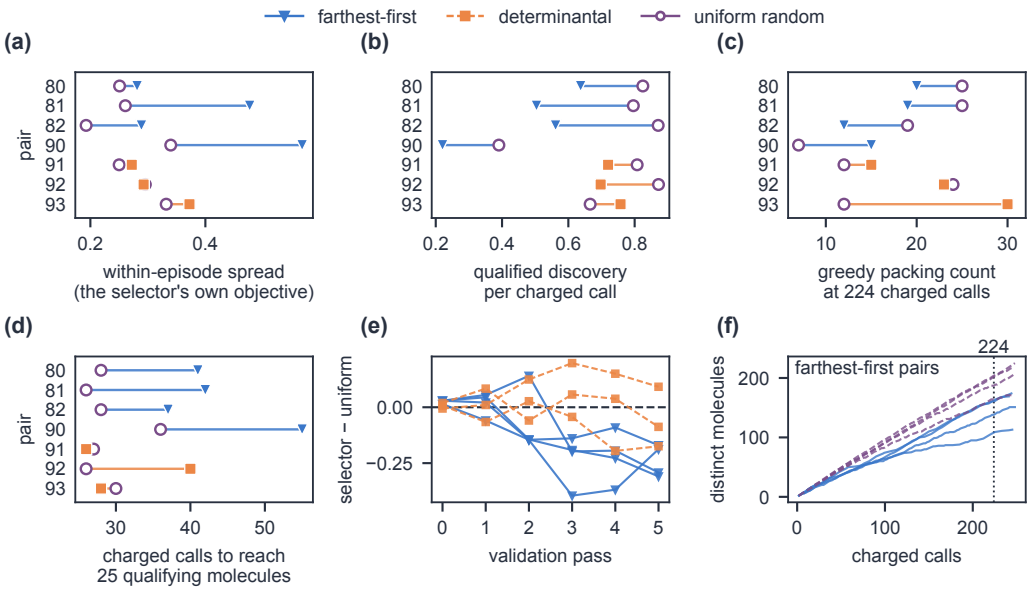


Figure 2: **Farthest-first raises within-episode spread but lowers final-pass discovery per call in all four pairs.** In (a–d), each row is a final-pass matched pair. Open circles mark uniform-random controls and filled markers mark selectors. (a) Within-episode fingerprint spread, the quantity farthest-first optimises. (b) Qualified discovery per charged call. (c) At a common 224 charged calls its greedy packing count, a lower bound on the packing number, is also lower in three of the four, pair 90 being the exception. (d) It needs more charged calls to reach twenty-five qualifying molecules, in four of four. (e) The paired difference (selector minus uniform) is positive at the first passes after training starts in three pairs, so a comparison stopped early would have reported the opposite. (f) Cumulative distinct molecules against charged calls, dashed for the control. The determinantal results in (a–e) are mixed throughout.

Farthest-first increases within-episode spread in all four pairs on its own distance metric (Figure 2a), yet its qualified discovery per call trails uniform selection's by 0.188, 0.292, 0.309 and 0.170 at $\tau = 0.567$, and trails in every pair at all nine thresholds tested from 0 to 0.7. Uniform selection leads cumulative coverage in all 28 pair–checkpoint comparisons, with mean margins growing from +7.0 at 32 calls to +48.8 at 224 (Appendix B.5). Crediting every budget-straddling episode to farthest-first leaves uniform ahead in 27 of 28 (+37, +49, +43, +55 at 224 calls). In all four pairs, farthest-first also spends a larger share of its calls buying molecules it has already scored. Repeats thus contribute to the deficit alongside distinct molecules that fail to qualify (Figure 3).

Repeated evaluations concentrate on the anchor. The first molecule a farthest-first episode scores is the anchor of its first scoring turn, chosen by position, not distance. In the final pass

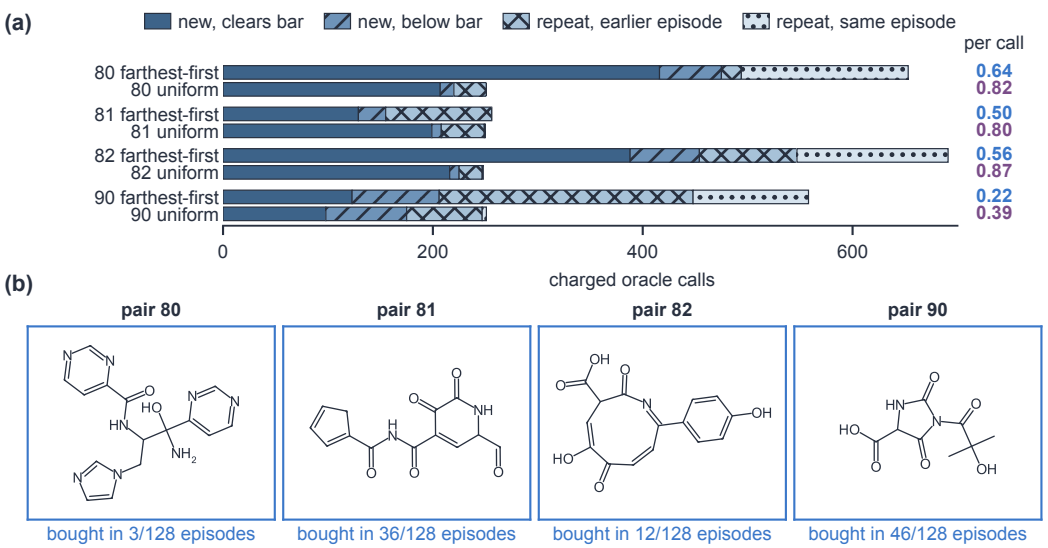


Figure 3: **Farthest-first spends a greater share of its charged calls on repeated molecules.** (a) All charged oracle calls in the four matched pairs' final 128-episode windows, split into four categories. Bar length is total cost. The leading segment divided by the full bar is qualified discovery per call, printed at the right. Farthest-first finds more qualifying molecules in three pairs but fewer per charged call in all four. (b) The compound each farthest-first run paid for in the most episodes, among those the endpoint counts, drawn from its recorded output and counted once per episode. Farthest-first pays again for it in 12, 36 and 46 episodes in three pairs against 4 to 9 for uniform selection's own most re-bought compound, and it was the episode's first-scored anchor in 11, 36 and 46 of them. In pair 80 its repeats fall inside single episodes instead. Figure C.3 shows both arms' compounds.

that anchor had already been bought in an earlier episode in 7, 62, 48 and 85% of episodes, against 3, 17, 8 and 42% for the other molecules of the same episodes, and it carries 45–78% of the arm's earlier-episode repeats (Table B.1). The compound each farthest-first run re-bought most often was that anchor in 36 of 36, 11 of 12 and 46 of 46 of its episodes in pairs 81, 82 and 90. Within-episode spread cannot register such a repeat. A registered control that draws the anchor at random, run after its uniform controls, still trails uniform selection in all three same-machine pairs, by 0.289, 0.162 and 0.156, and narrows the gap to first-start selection in two of them (Table 1, Appendix B.2). The anchor contributes to repeated evaluations but does not fully explain the reversal.

Fixed-pool comparisons. The random-start runs logged every candidate pool, so each rule can be replayed on one fixed stream, keeping its own purchase history from the start of training (Appendix B.3). On all four streams first-start farthest-first buys the largest share of repeats, and a random start stays within 0.010 of a uniform draw in new qualifying molecules per call. Avoiding previously scored molecules gives the largest consistent gain over a fresh uniform draw, by 0.026 to 0.041. Because the replayed pools cannot respond to the purchases, these contrasts do not identify how much of the online deficit arises through learning.

A second diversity measure. We apply the packing number of Svensson et al. (2025), the largest set of molecules separated by a Morgan-fingerprint distance threshold of 0.7, to our records at 224 charged calls, without their activity gate. Uniform selection's greedy count exceeds farthest-first's by 5, 6 and 7 in three pairs, alongside 44, 49 and 43 more distinct molecules. In the fourth, farthest-first scores 15 against uniform's 7 on the greedy count, while uniform still finds 59 more distinct and 26 more qualified molecules. Greedy counts are lower bounds and do not order the exact maxima (Appendix B.6).

Checkpoint and cost sensitivity. At each arm's best validation pass, the difference remains negative in all four pairs, at −0.241, −0.227, −0.305 and −0.091, although farthest-first leads early

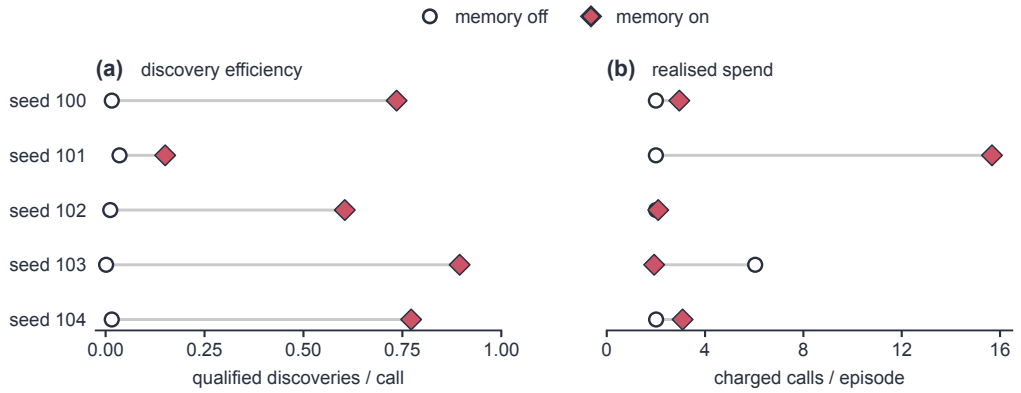


Figure 4: **Five prospectively specified memory pairs at commanded budget two.** Each row is one matched training seed. Open circles show memory off and filled diamonds memory on. (a) Final-pass qualified discoveries per charged call. (b) Realised calls per episode. Seed 101 spends 15.67 versus 2.00 calls per episode. Memory on changes the complete memory configuration, not only its scaffold clause.

in training in three pairs (Figure 2e, Appendix B.7). To reach twenty-five qualifying molecules, farthest-first needs 1.32 to 1.62 times as many charged calls as its control, and 1.41 to 1.82 times as many to reach fifty, in all four pairs. It also spent 1.14 to 2.35 times the control's training calls (Appendix B.8). Per episode, farthest-first leads only in the three pairs where it spends 2.2 to 2.8 times as much, and lags where spend is nearly equal.

Mixed results for determinantal selection. The determinantal rule samples an exact k-determinantal point process (k-DPP) over the same kernel and has no positional anchor (Appendix F.2). Against a uniform control in three matched seeds its qualified differences are mixed in sign, its raw-novelty differences are negative in all three, and two qualified differences turn positive at $\tau = 0.7$ (Table 1, Appendix B.9). Selection can raise local spread without reliably improving discovery. These results do not establish that diverse selection fails generally.

Scope relative to published selection work. Svensson et al. (2025) evaluate a packing-style measure among active molecules and recommend the determinantal rule, whereas our endpoint counts distinct qualifying molecules without a distance radius. Our agent and call accounting differ from their training setup. Applying their metric tests whether evaluation objectives explain the reversal, not whether their results reproduce. Proposal drift remains untested. The selection ratio, recorded only in the random-start runs, is about two in four or five against their one in ten (Appendix B.4).

4.2 Persistent memory in matched pairs and descriptive analyses

Prospective matched intervention. Memory on improves qualified discovery per charged call in all five prospectively specified Qwen3-4B pairs on Osimertinib MPO, with paired differences $+0.720$, $+0.115$, $+0.593$, $+0.893$ and $+0.757$ (median $+0.720$, exact one-sided sign test $p = 0.03125$ in the pre-specified direction, Figure 4). The complete memory configuration changes, so this contrast does not isolate scaffold suppression from other training and evaluation effects. The filter footprint is non-zero in every treated run and zero in every control (Table A.4). Seed 101 spends 15.67 calls per episode against its control's 2.00. In a post-hoc common-call analysis, memory leads at 224 calls in all five pairs, by 21 to 201 qualifying molecules under conservative episode-boundary bounds. Seed 101 trails at 32 and 64 calls, then leads at every reported checkpoint from 96 through 224 calls. Of each treated pass's 162–305 qualifying molecules, 20–51 were scored earlier in the same run. Without them memory still leads in all five pairs, by 0.109 to 0.781 per call. None of this is a hard-budget guarantee (Appendix A.3).

Historical cohort. At $\tau = 0.567$, the seven memory runs reach 0.684–0.865 qualified discoveries per call, compared with 0.003–0.234 in the 22 unfiltered runs. Their median spend is 1.87 against 2.55 calls per episode. The complete rank separation is $U = 154/154$. Here U counts the memory–unfiltered run pairs in which the memory run is higher. The cohort was not randomised, and three of its memory runs are re-runs, on a verified service, of runs whose first execution never applied the filter (Appendix A.2), so the separation is descriptive corroboration (Table A.1). The best pass of any non-memory run, 0.357, still falls short of the lowest best pass among the memory runs, 0.704, with complete rank separation, $U = 154/154$ (Appendix D.2). The same ordering holds without a rate denominator. At 192 calls, the memory runs have discovered 144–178 distinct molecules, while the other runs have found 1–62, with separated ranges at all six checkpoints up to common support (Appendix A.4).

Spend and efficiency in the historical cohort. Under the soft budget, realised spend inside the unfiltered arm ranges from 2.00 to 11.98 calls per episode. Across those 22 runs, the endpoint has no positive association with spend (Spearman $\rho = -0.19$, two-sided permutation $p = 0.39$ over 200,000 shuffles). The five highest-spending unfiltered runs reach at most 0.112 against the memory arm's floor of 0.684, and the highest-spending run, at 11.98 calls an episode against the memory arm's median of 1.87, returns 0.012. Spend is an outcome, so this association does not predict the effect of raising the commanded budget. Against the untrained policy the comparison depends on the threshold, while the arm ordering holds at all ten swept thresholds (Appendix D.1).

High group multiplicity is insufficient. IPS preserves within-group outcome multiplicity in all six completed runs, ending at or above 0.769, compared with plain GRPO's median 0.19 and top-k's 0.453. Across the 26 runs carrying both readings, multiplicity correlates with raw novelty at $\rho = +0.800$ (two-sided permutation $p < 10^{-4}$, 200,000 shuffles, Figure C.4), a pooled association across arms, not an estimate of a causal effect (Appendix D.3). The four memory runs with a multiplicity reading lie at 0.8688–0.9172 and the six IPS runs at 0.7686–0.9174, yet their raw discovery rates do not overlap (0.720–0.856 against 0.068–0.289). The highest-multiplicity run of either arm is an IPS run, at 0.9174, delivering 0.289, while the lowest-multiplicity memory run, at 0.8688, delivers 0.856, so multiplicity does not determine the endpoint across these arms. Memory may still gain partly through multiplicity, since no IPS run falls inside memory's range and memory's median multiplicity, 0.87, exceeds IPS's 0.84. Memory also yields greater fixed-size fingerprint spread and scaffold yield (Appendix C).

The verified re-run. Three runs requested as memory runs reached a service that did not apply the filter (Appendix A.5). On a verified service the bridge reached 0.655, inside $[0.604, 0.804]$, and the three seeds returned 0.827, 0.684 and 0.865, each independently above the unfiltered ceiling. Their endpoint-dependent inclusion criterion was dropped after the outcomes were read. They enter the memory arm by filter footprint, with the exclude-three assignment kept as a sensitivity analysis (Appendices A.2, A.6 and A.7).

Adding IPS to the same memory. As the nearest molecular IPS result keeps memory on in both arms (Sinha et al., 2026), we compare memory+IPS with memory alone in three matched seeds. The registered raw-novelty differences (combined minus memory) are -0.146, $+0.165$ and $+0.078$, and qualified discovery improves in one pair of three (Table 1). Realised spend also changes, including a combined run at 0.68 calls per episode. The result is mixed incremental value at this operating point, not proof that IPS is ineffective when combined with memory (Figure A.2).

Memory-clause contributions. Replaying the historical streams attributes 87.4–100.0% of final-window suppression to exact repeats (Appendix C.2), but the scaffold bucket is more active during training and replay cannot identify its effect on the learned policy. A separate four-pair ablation at commanded budget eight gives a median memory-minus-repeat-only difference of $+0.055$ (Table 1), below that experiment's registered 0.15 decision threshold, but it is not an equivalence test and has no unfiltered or IPS-only arm. Three further pairs at the cohort's own commanded budget, registered without a decision threshold, put repeat-only at 0.659–0.746, against 0.704–0.765 for the original four scaffold-memory runs and 0.003–0.234 without memory. The exact-repeat clause alone therefore reaches the memory arm's range, but the paired differences are small, mixed in sign and taken at unmatched spend, precluding an equivalence claim (Appendix C.3).

4.3 A SECOND ORACLE

On Perindopril_MPO the original registered analysis on raw novelty uses two complete machine–seed blocks, and a secondary extension uses five, with three arms in each, so all 15 completed runs enter the balanced medians (Appendix D.4, with the other sensitivity analyses in Appendix D).

The same pooled-baseline percentile rule gives a secondary qualification threshold of 0.142 for Perindopril. At this threshold the five-block median rates, divided by that target's untrained rate, are 0.47 for plain GRPO, 2.33 for IPS and 9.91 for memory. Memory ranks first within each machine. Its median rate is 4.26 times IPS's in this extension, compared with 4.3 for the original four Osimertinib memory runs. These small-population ratios are descriptive, not matched effect sizes. Target-specific percentile calibration does not equate task difficulty.

5 SCOPE OF IDEALISED GROUP CALCULATIONS

Two established results give context, not selector guarantees or an explanation of the five-pair effect (Figure E.1). Neither covers finite-step autoregressive training, clipping, entropy, KL regularisation or history-dependent rewards.

Without an additive stabiliser, standardisation of two-valued rewards is invariant to positive affine rescaling (Vojnovic & Yun, 2025; Liu et al., 2025b). Outcome weights can preserve, reverse or tie their ordering but cannot interpolate between standardised values. Bay & Yearick (2026) give the binary count-only form. Appendix E.3 covers the deployed stabiliser. Binary-reward Mirror-GRPO lacks an unregularised interior fixed point, but reference regularisation can admit one (Mroueh, 2025). The idealised mode-weight flow of Liu et al. (2025a) lacks an interior stationary mixture for unequal rewards. Appendix E gives the calculation and its exclusions.

6 LIMITATIONS

A pre-registered Qwen3-1.7B replication with nine runs on one machine reproduces the ordering without full separation. Memory ranks above every other run but one, $U = 17/18$, exact one-sided $p = 0.024$ (Appendix D.5). At that size the budget barely binds, and eight of nine runs spend at most one call an episode against the commanded two. The study otherwise uses one model family, short training runs and two surrogate molecular objectives. Soft constraints produce unequal spend. Common-call evaluation leaves training costs unequal, and the historical cohort includes non-executed interventions and unequal arm counts. The memory-clause ablations leave each clause's necessity at budget two unresolved. The prospective contrast changes the reward, the observations and the training and validation histories together, and realised spend differs between its arms. Selector arms follow different online proposal trajectories. Our farthest-first rule starts from the first valid proposal, where MaxMin-A draws at random. The random-start control ran after its uniform controls, not concurrently. The original uniform draw was seeded from the candidate list. Replayed pools come from random-start policies and cannot respond to replayed purchases. Cross-target calibration is secondary, and few paired runs leave mixed effects unresolved. The historical filtered group has a lower median per-run mean episode quality (0.438 versus 0.569) and does not lead the external quality benchmark, with a top-ten area under the curve of 0.750 against a cited best of 0.837 under a different protocol (Appendix D.6). We claim neither chemistry validation nor state-of-the-art molecular optimisation.

7 CONCLUSION

At this operating point, higher within-episode spread need not improve qualified discovery per charged call. Farthest-first has lower discovery per call than uniform selection in every same-machine pair with a fixed or random anchor. Avoiding previously scored molecules gives the largest consistent reduction in repeats on replayed pools. The complete memory configuration has higher discovery per call in five prospective pairs. Evaluations of diversity corrections for agents that set their own oracle spend should therefore report each rate beside realised spend and coverage at common call counts.

AI USE STATEMENT

We used large language models solely for code implementation assistance, manuscript copy-editing and polishing, generating conceptual schematic icons, and assistance with the derivations in Appendix E, which the authors checked. All experimental workflows, data analyses and scientific claims were designed, conducted and verified by the authors.

ETHICS STATEMENT

This work uses public molecular property oracles and involves no human subjects or personal data. The objectives are screening surrogates, and oracle scores are not evidence of clinical efficacy or safety. Improved molecular exploration could be applied to other objectives. We study a published memory mechanism, not a new experimentally validated capability.

REPRODUCIBILITY STATEMENT

The supplementary material provides environment, oracle-integration, training and analysis code, with tests and instructions for reproducing every reported table and figure. The per-cell tables in Appendix A list every run and every exclusion, and Appendix F.2 gives the training settings and the differences between machines. Proofs and their scope are in Appendix E.

REFERENCES

Ryan Bahlous-Boldi, Isha Puri, Idan Shenfeld, Akarsh Kumar, Mehul Damani, Sebastian Risi, Omar Khattab, Zhang-Wei Hong, and Pulkit Agrawal. Vector policy optimization: Training for diversity improves test-time search. *arXiv preprint arXiv:2605.22817*, 2026. URL https://arxiv.org/abs/2605.22817.

Yong Yi Bay and Kathleen A. Yearick. GRPO, Dr. GRPO, and DAPO are three operations on one number: The group-standard-deviation identity. *arXiv preprint arXiv:2607.00152*, 2026. URL https://arxiv.org/abs/2607.00152.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. URL https://pubs.acs.org/doi/10.1021/jm9602928.

Thomas Blaschke, Ola Engkvist, Jürgen Bajorath, and Hongming Chen. Memory-assisted reinforcement learning for diverse molecular de novo design. *Journal of Cheminformatics*, 12(1), 2020. doi: 10.1186/s13321-020-00473-0.

Zhicheng Cai, Xinyuan Guo, Hanlin Wu, Mingxuan Wang, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Beyond euclidean clipping: Overcoming exploration collapse in LLM RL via riemannian isometric policy optimization. *arXiv preprint arXiv:2607.10169*, 2026. URL https://arxiv.org/abs/2607.10169.

Zhe Cao, Miaowen Wen, and Fangjiong Chen. When correct solutions repeat: Rarity-aware credit redistribution for GRPO. *arXiv preprint arXiv:2608.03467*, 2026. URL https://arxiv.org/abs/2608.03467.

Haoxuan Chen, Tianming Liang, Wei-Shi Zheng, and Jian-Fang Hu. Breaking *Winner-Takes-All*: Cooperative policy optimization improves diverse LLM reasoning. *arXiv preprint arXiv:2605.11461*, 2026. URL https://arxiv.org/abs/2605.11461.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *NeurIPS*, 2025.

Jingchu Gai, Guanning Zeng, Huaqing Zhang, and Aditi Raghunathan. Differential smoothing mitigates sharpening and improves llm reasoning. *arXiv preprint arXiv:2511.19942*, 2025. URL https://arxiv.org/abs/2511.19942.

Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor Coley. Sample efficiency matters: A benchmark for practical molecular optimization. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022. URL https://arxiv.org/abs/2206.12411.

Jeff Guo and Philippe Schwaller. Augmented memory: Sample-efficient generative molecular design with reinforcement learning. *JACS Au*, 4(6), 2024. doi: 10.1021/jacsau.4c00066. URL https://doi.org/10.1021/jacsau.4c00066.

Anthony GX-Chen, Jatin Prakash, Jeff Guo, Rob Fergus, and Rajesh Ranganath. KL-regularized reinforcement learning is designed to mode collapse. *arXiv preprint arXiv:2510.20817*, 2025. URL https://arxiv.org/abs/2510.20817.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI*, 2018.

Yuelin Hu, Zhenbo Yu, Zhengxue Cheng, Wei Liu, and Li Song. Maximizing rollout informativeness under a fixed budget: A submodular view of tree search for tool-use agentic reinforcement learning. *arXiv preprint arXiv:2605.05262*, 2026. URL https://arxiv.org/abs/2605.05262.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2102.09548.

Gengsheng Li, Jinghan He, Shijie Wang, Dan Zhang, Ruiqi Liu, Renrui Zhang, Zijun Yao, Junfeng Fang, Haiyun Guo, and Jinqiao Wang. R-Diverse: Mitigating diversity illusion in self-play LLM training. *arXiv preprint arXiv:2602.13103*, 2026. URL https://arxiv.org/abs/2602.13103.

Henglin Liu, Huijuan Huang, Jing Wang, Chang Liu, Xiu Li, and Xiangyang Ji. DiverseGRPO: Mitigating mode collapse in image generation via diversity-aware GRPO. *arXiv preprint arXiv:2512.21514*, 2025a. URL https://arxiv.org/abs/2512.21514.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025b. URL https://arxiv.org/abs/2503.20783.

Anamika Lochab, Bolian Li, and Ruqi Zhang. Uniform-correct policy optimization: Breaking RLVR's indifference to diversity. *arXiv preprint arXiv:2605.00365*, 2026. URL https://arxiv.org/abs/2605.00365.

Hannes H. Loeffler, Jiazhen He, Alessandro Tibo, Jon Paul Janet, Alexey Voronov, Lewis H. Mervin, and Ola Engkvist. REINVENT 4: Modern AI-driven generative molecule design. *Journal of Cheminformatics*, 16(20), 2024. doi: 10.1186/s13321-024-00812-5.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *ICML*, pp. 6820–6829, 2020.

Youssef Mroueh. Reinforcement learning with verifiable rewards: GRPO's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025. URL https://arxiv.org/abs/2503.06639.

Wenzhe Niu, Wei He, Zongxia Xie, Jinpeng Ou, Huichuan Fan, Yuchen Ge, Yanru Sun, Ziyin Wang, Yizhao Sun, Chengshun Shi, Jiuchong Gao, Jinghua Hao, and Renqing He. From absolute to relative: Rethinking reward shaping in group-based reinforcement learning. *arXiv preprint arXiv:2601.23058*, 2026. URL https://arxiv.org/abs/2601.23058.

Junoh Park, Junseo Hwang, Wonguk Cho, and Taesup Kim. ReCo: Reweighting GRPO against distributional concentration. *arXiv preprint arXiv:2607.26862*, 2026. URL https://arxiv.org/abs/2607.26862.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Abhijeet Sinha, Sundari Elango, and Dianbo Liu. Expected return causes outcome-level mode collapse in reinforcement learning and how to fix it with inverse probability scaling. *arXiv preprint arXiv:2601.21669*, 2026. URL https://arxiv.org/abs/2601.21669.

Hampus Gummesson Svensson, Ola Engkvist, Jon Paul Janet, Christian Tyrchan, and Morteza Haghir Chehreghani. Diverse mini-batch selection in reinforcement learning for efficient chemical exploration in de novo drug design. *arXiv preprint arXiv:2506.21158*, 2025. URL https://arxiv.org/abs/2506.21158.

Morgan Thomas, Albert Bou, Jose Carlos Gómez-Tamayo, Gary Tresadern, Mazen Ahmad, and Gianni De Fabritiis. REINFORCE-ING chemical language models for drug discovery. *arXiv preprint arXiv:2501.15971*, 2025. URL https://arxiv.org/abs/2501.15971.

Milan Vojnovic and Se-Young Yun. What is the alignment objective of GRPO? *arXiv preprint arXiv:2502.18548*, 2025. URL https://arxiv.org/abs/2502.18548.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a. URL https://arxiv.org/abs/2505.09388.

Zhuo Yang, Yeyun Chen, Jiaqing Xie, Ben Gao, Shuaike Shen, Wanhao Liu, et al. MolAct: An agentic RL framework for molecular editing and property optimization. *arXiv preprint arXiv:2512.20135*, 2025b. URL https://arxiv.org/abs/2512.20135.

Xinwu Ye, He Cao, Hao Li, Bin Feng, Zijing Liu, Xiangru Tang, Yu Li, and Shenghua Gao. Pushing biomolecular utility-diversity frontiers with supergroup relative policy optimization. *arXiv preprint arXiv:2605.08659*, 2026. URL https://arxiv.org/abs/2605.08659.

Suqin Yuan, Jinkun Chen, Jiyang Zheng, Muyang Li, Lei Feng, Dadong Wang, Tao Xiang, Tongliang Liu, and Bo An. Understanding diversity collapse in RLVR via the lens of overtraining. *arXiv preprint arXiv:2606.15455*, 2026. URL https://arxiv.org/abs/2606.15455.

Haitian Zhong, Jixiu Zhai, Lei Song, Jiang Bian, Qiang Liu, and Tieniu Tan. RC-GRPO: Reward-conditioned group relative policy optimization for multi-turn tool calling agents. *arXiv preprint arXiv:2602.03025*, 2026. URL https://arxiv.org/abs/2602.03025.

APPENDIX

The appendix follows the order of the claims in the main text. Appendix A lists every run, the confirmation re-runs, the prospective memory pairs and the spend accounting, together with the execution checks and the departures from registered analyses that define the populations of Sections 3 and 4.2. Appendix B gives the selection analyses behind Section 4.1. Appendix C examines memory suppression and molecular outputs. Appendix D reports sensitivity to the threshold and the checkpoint, the second target, the second model size and baseline parity. Appendix E gives the calculations behind Section 5, and Appendix F the implementation and training settings.

CONTENTS

A Cohorts, records and registrations **14**

A.1 Per-cell records . . . 14

A.2 Confirmation on a verified service and the memory+IPS control . . . 14

A.3 Matched memory intervention at commanded budget two . . . 17

A.4 Training cost, validation cost and soft-budget compliance . . . 18

A.5 Provenance and execution checks . . . 19

A.6 Departures from registered analyses . . . 20

A.7 Pre-specified criteria . . . 21

B Selection experiments **21**

B.1 Repeat purchases by position . . . 21

B.2 Registered random-start control . . . 22

B.3 Fixed-pool replay . . . 23

B.4 Proposal drift and selection ratio . . . 23

B.5 Common-call coverage of the selection pairs . . . 24

B.6 Both diversity measures on the same records . . . 25

B.7 The selection pairs at every validation pass . . . 26

B.8 Cost of a discovery target and of training . . . 27

B.9 Determinantal selection against a uniform control . . . 27

C Memory and multiplicity **29**

C.1 Scaffold-level yield . . . 29

C.2 Suppression by the two memory clauses . . . 30

C.3 Training without the scaffold bucket . . . 30

C.4 Molecules emitted by each arm . . . 31

C.5 Variety at a fixed subsample size . . . 33

C.6 Multiplicity against yield, by arm . . . 33

D Robustness and scope **33**

D.1 Threshold sensitivity and within-arm variation . . . 33

D.2 Validation trajectories and untrained references . . . 36

D.3 Descriptive stratification of the historical association . . . 37

D.4 Registered analysis and balanced extension on the second target . . . 37

D.5 Pre-registered replication at a second model size 38
D.6 Baseline parity and screening 39

E Theory **41**
E.1 Proof of Proposition 1 41
E.2 Scope of Proposition 1 41
E.3 Enumeration of the two-value trichotomy 42

F Implementation and reproducibility **43**
F.1 Implementation specification 43
F.2 Hyperparameters and cross-machine configuration differences 44

A Cohorts, records and registrations

A.1 Per-cell records

The main text reports medians, ranges and counts. Each cell in these tables is one completed training run. Table A.1 lists the runs underlying the aggregate results, and Table A.2 lists both arms of every selection pair. Later confirmation and DPP runs appear in Appendices A.2 and B.9.

Table A.1 shows two details absent from the aggregate results. The three runs in which the filter never executed are assigned by the executed intervention, not the requested label. Their values lie within the interior of the plain-GRPO range. The reclassification is material because the initial groups overlap, as reported in Appendix A.6. The inverse-probability arm has six runs here and in the multiplicity comparison. A seventh run has a multiplicity value of 0.894 but is excluded from both because its molecule log ends at 2304 of the required 3968 episodes. Incomplete runs are excluded throughout.

A.2 Confirmation on a verified service and the memory+IPS control

Design. Three runs labelled as memory runs reached a scoring service that predated the filter, so the filter never ran in them (Appendix A.5). They were re-run on a verified service under a design fixed before the first re-run started and more than five hours before any re-run produced an endpoint. The three runs differ from an executed memory run only in the scoring-service configuration. Because the re-runs had to move to another machine, the executed seed was first re-run under its original configuration as a bridge. Both cohort assignments were specified in advance. With the three runs in the plain-GRPO arm, the arms are disjoint at $U = 88/88$. With those runs assigned to memory, the arms overlap at all ten thresholds and U falls to $96/133$. The design had three criteria.

1. Bridge. The re-run of the executed seed must fall within ± 0.10 of its original 0.7044, inside $[0.604, 0.804]$. The width is the frozen memory band (0.0605) plus the measured run-to-run spread of that arm. A failed bridge would have left the three re-runs confounded with the machine.
2. Confirmation. At least two of the three seeds exceed 0.4691, the midpoint of the gap between the arms. Otherwise the reported separation would revert to $U = 96/133$. A two-one split would be reported explicitly, with all four runs shown.
3. Execution. A cell that reaches the old service again is discarded. Before launch, the service was checked against a reference score and the filter's response to a six-molecule test batch at bucket capacity one. Each completed run's recorded filter, service and budget were then checked.

Outcome. The bridge reached 0.6553, inside its band, so the transfer between machines is supported. The bridge is separate from the three confirmation seeds. Seeds 0, 1 and 2 reached 0.8268, 0.6840 and 0.8647, all above 0.4691. No run reached the old service. The original historical cohort is unchanged.

Table A.1: **Every completed run at the operating point, one row per cell.** The arm is the *measured* arm. Three cells carrying the scaffold label reached a scoring service predating the filter, so the filter never executed in them and they are listed under plain GRPO with a dagger. Waste is the complementary share of charged calls that did not yield a new distinct canonical molecule in this window (within-episode and across-episode repeats), so distinct and waste sum to one. Best mol. is the highest raw score the run reached in any episode. Mean ep. is the mean of each episode's own best score. Cliff is the update at which within-group multiplicity first falls below sixty per cent of the run's own early maximum, and a dash means the run never fell.

arm	seed	spend	distinct	qualified	waste	best mol.	mean ep.	cliff
IPS-GRPO	0	2.05	0.153	0.145	0.847	0.768	0.567	–
IPS-GRPO	1	2.12	0.192	0.188	0.808	0.774	0.573	–
IPS-GRPO	2	2.70	0.162	0.151	0.838	0.783	0.573	–
IPS-GRPO	3	4.16	0.068	0.064	0.932	0.749	0.538	–
IPS-GRPO	5	2.05	0.229	0.202	0.771	0.768	0.551	–
IPS-GRPO	6	2.41	0.289	0.234	0.711	0.762	0.541	–
plain GRPO	0	2.88	0.089	0.079	0.911	0.759	0.559	12
plain GRPO†	0	2.13	0.110	0.099	0.890	0.789	0.596	22
plain GRPO	1	2.27	0.028	0.028	0.972	0.762	0.581	15
plain GRPO†	1	2.08	0.011	0.011	0.989	0.736	0.588	10
plain GRPO	2	2.40	0.033	0.020	0.967	0.723	0.562	13
plain GRPO†	2	2.02	0.066	0.058	0.934	0.775	0.570	14
plain GRPO	4	3.73	0.086	0.065	0.914	0.742	0.578	10
plain GRPO	5	2.21	0.155	0.155	0.845	0.789	0.595	–
plain GRPO	6	3.01	0.003	0.003	0.997	0.724	0.563	6
scaffold memory	0	1.80	0.844	0.827	0.156	0.787	0.457	–
scaffold memory	1	1.95	0.728	0.684	0.272	0.823	0.434	–
scaffold memory	2	2.37	0.891	0.865	0.109	0.802	0.489	–
scaffold memory	3	1.59	0.783	0.704	0.217	0.785	0.431	–
scaffold memory	4	1.87	0.720	0.707	0.280	0.788	0.460	–
scaffold memory	5	3.16	0.772	0.765	0.228	0.796	0.438	–
scaffold memory	6	1.84	0.856	0.737	0.144	0.766	0.433	–
top-k reward	0	6.33	0.072	0.072	0.928	0.745	0.564	15
top-k reward	1	2.00	0.020	0.020	0.980	0.756	0.568	15
top-k reward	2	11.98	0.012	0.012	0.988	0.761	0.586	15
top-k reward	3	8.70	0.093	0.064	0.907	0.803	0.604	–
top-k reward	4	7.93	0.051	0.051	0.949	0.761	0.561	10
top-k reward	5	6.70	0.055	0.055	0.945	0.743	0.583	8
top-k reward	6	7.61	0.139	0.112	0.861	0.752	0.554	17

Table A.2: **The selection-side pairs, both arms of each.** Within a pair the two cells share a seed and differ only in how the two scored molecules are chosen from the step's proposals. The rendered selector is read back from each run's own resolved configuration rather than assumed from its label.

pair	machine	arm	rendered	spend	distinct	qualified	waste	best mol.
80	A	farthest-first	farthest-first	5.10	0.727	0.637	0.273	0.795
80	A	uniform random	uniform	1.96	0.876	0.825	0.124	0.810
81	A	farthest-first	farthest-first	2.00	0.605	0.504	0.395	0.774
81	A	uniform random	uniform	1.95	0.832	0.796	0.168	0.781
82	A	farthest-first	farthest-first	5.40	0.657	0.562	0.343	0.812
82	A	uniform random	uniform	1.94	0.907	0.871	0.093	0.783
90	B	farthest-first	farthest-first	4.36	0.369	0.220	0.631	0.770
90	B	uniform random	uniform	1.96	0.697	0.390	0.303	0.791

Role of the confirmation criterion. Criterion 2 tests the treatment effect at these seeds, not data validity. Failure would restore the null interpretation of the reported separation. The three seeds therefore enter the memory arm solely on execution evidence from criterion 3 and the filter footprint in Appendix A.5. Before any endpoint is read, these checks distinguish the original unfiltered runs from the filtered re-runs. Their values are reported without further inclusion criteria. This moves the

Table A.3: **Whether the scaffold filter executed, measured per cell from retained output.** The reported best score is the maximum over raw oracle scores and the terminal return is the maximum over filtered scores, so their gap is the filter's footprint. A gap above zero proves the filter executed. A gap at zero does not prove it did not, since a filtered molecule that was not that episode's best leaves the maximum unchanged (Appendix A.5). Every cell whose own configuration record requests the filter is listed. The three with no footprint are the cells reassigned in Appendix A.1, identified here without reference to that reassignment.

run	screen	updates	raw best	return	footprint
memory, seed 0	primary	25	0.5997	0.5998	-0.0%
memory, seed 1	primary	25	0.5888	0.5888	0.0%
memory, seed 2	primary	25	0.5735	0.5735	0.0%
memory, seed 3	primary	25	0.4282	0.2978	30.4%
memory, seed 4	primary	25	0.4553	0.3011	33.9%
memory, seed 5	primary	25	0.4357	0.2897	33.5%
memory, seed 6	primary	25	0.4352	0.2940	32.4%
pair 80, farthest-first	selection	25	0.5093	0.3224	36.7%
pair 80, uniform	selection	25	0.4369	0.3458	20.8%
pair 81, farthest-first	selection	25	0.4591	0.3261	29.0%
pair 81, uniform	selection	25	0.4270	0.2990	30.0%
pair 82, farthest-first	selection	25	0.4710	0.2895	38.5%
pair 82, uniform	selection	25	0.4575	0.3243	29.1%
pair 90, farthest-first	selection	25	0.4226	0.2271	46.3%
pair 90, uniform	selection	25	0.4384	0.2512	42.7%

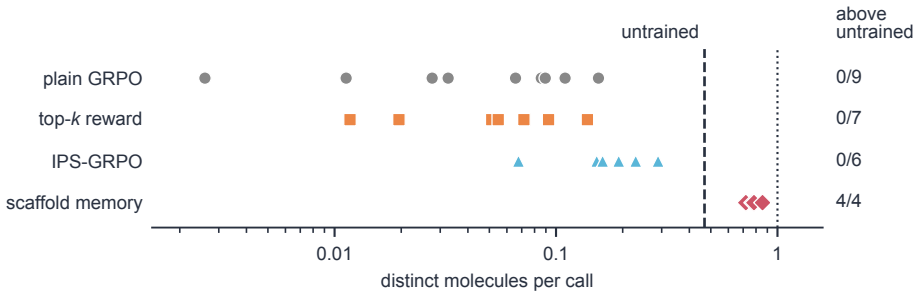


Figure A.1: **Persistent memory separates from the other arms at roughly two oracle calls per episode.** Each point is one completed training run on its final 128-episode pass, counting every distinct molecule regardless of score. The dashed line is the untrained policy at this spend, and the right-hand column counts each arm's runs at or above it. The axis is logarithmic, and the dotted rule at 1 is its ceiling, since a call returns at most one molecule. The four runs with executed persistent memory, exact-repeat suppression plus a 2020 Bemis–Murcko scaffold bucket, reach 0.720–0.856 distinct molecules per call. All twenty-two runs without it reach at most 0.289 and fall below their own untrained references, $22/22$. The groups do not overlap ($U = 88/88$, exact $p = 0.000067$), and memory runs spend less, at a median of 1.86 calls per episode against 2.55. The gap thus accompanies lower oracle spend. Three runs carrying the filter's label reached a scoring service predating it, so the filter never ran in them. They are plotted with the unfiltered arm and re-run on a verified service in Figure A.2. The main text and Appendix A.2 report all seven memory runs together, including that re-run. This panel plots the four runs with verified filter execution. The qualified endpoint and its threshold dependence are in the text and Appendix D.1.

memory population of Section 4 from four runs to seven, and Appendix A.6 lists the change with the other departures from the registered plans.

The memory+IPS comparison uses three paired L20 seeds with memory held on in both arms. Raw novelty is the registered endpoint. Qualified novelty is reported alongside it, and the two have opposite signs at seed 4. No equivalence is claimed. Calls per episode for memory versus

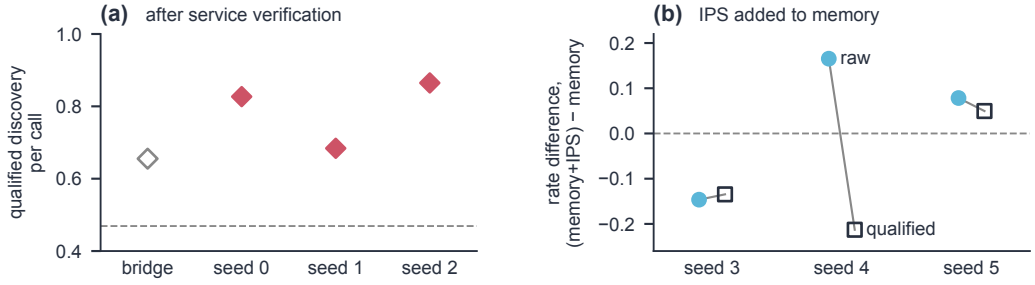


Figure A.2: **The three verified re-runs exceed the unfiltered ceiling, while adding IPS to memory gives mixed results.** (a) Qualified discovery after service verification. The open marker is the bridge. The three filled markers are confirmation seeds included in the main-text memory arm. The dashed line marks the disjoint gap's midpoint, for reference. (b) Paired differences from adding IPS to memory, with a line joining each seed's raw and qualified novelty differences. They have different signs for seed 4. Each marker is a training run or paired difference, not an evaluation episode.

memory+IPS are $1.59/2.80$, $1.87/0.68$ and $3.16/1.73$. Rate differences therefore compare the arms at their actual, unequal spend.

run	seed	qualified / calls	rate	calls / episode
hardware bridge	3	154 / 235	0.6553	1.84
confirmation	0	191 / 231	0.8268	1.80
confirmation	1	171 / 250	0.6840	1.95
confirmation	2	262 / 303	0.8647	2.37

	raw distinct / call			qualified distinct / call		
seed	filter	filter + IPS	difference	filter	filter + IPS	difference
3	0.7833	0.6369	-0.1464	0.7044	0.5698	-0.1346
4	0.7197	0.8851	+0.1654	0.7071	0.4943	-0.2129
5	0.7723	0.8507	+0.0784	0.7649	0.8145	+0.0496

A.3 Matched memory intervention at commanded budget two

A prospective experiment specified before any run tests whether the complete memory-on configuration raises qualified discovery at the same commanded budget, model and oracle service. Five new seeds (100–104) each pair memory-on and memory-off runs on Osimertinib MPO with Qwen3-4B at commanded budget two. All ten runs execute sequentially on one machine and oracle service. Each completes 3 968 episodes.

The pre-specified replication criterion requires all five treated–control differences to be positive and the median difference to exceed 0.20. Both conditions hold. Table A.4 reports per-seed qualified discovery rates, oracle spend, and filter footprint. Memory runs reach 0.151–0.895 (median 0.735) while controls collapse to 0.001–0.035 (median 0.016). The paired sign-test gives exact one-sided $p = 0.03125$. It tests the direction of five independent pair signs, not whether the population median exceeds the 0.20 threshold. The secondary unpaired statistic is $U = 25/25$. Its exact one-sided $p = 0.00397$ assumes exchangeability of the ten runs and ignores their pairing, so inference rests primarily on the paired test. Realised spend differs between arms, most sharply at seed 101 (15.67 versus 2.00 calls per episode).

The filter footprint confirms execution in all five treated runs (22.1%–50.9%) and is zero in every control. This contrast does not separate scaffold suppression from the rest of the memory-on training and evaluation configuration.

The cost and coverage analyses below were defined after the primary outcome was read. They retain all ten runs, the same final validation pass and the same threshold. Across their final passes, common

Table A.4: **Prospective matched memory intervention at commanded budget two.** Ten cells across five fresh seeds (100–104) run sequentially on one machine and oracle service, comparing complete memory-on and memory-off configurations. The paired median difference is $+0.7198$ (5/5 positive, exact one-sided sign-test $p = 0.03125$). Realised spend is reported for both arms, including the high-spend treated run at seed 101.

	memory on				memory off			
seed	qual / calls	rate	spend	foot. %	qual / calls	rate	spend	foot. %
100	278 / 378	0.7354	2.95	+22.1%	4 / 256	0.0156	2.00	0.0%
101	302 / 2006	0.1505	15.67	+50.9%	9 / 256	0.0352	2.00	0.0%
102	162 / 268	0.6045	2.09	+37.0%	3 / 257	0.0117	2.01	0.0%
103	221 / 247	0.8947	1.93	+26.8%	1 / 773	0.0013	6.04	0.0%
104	305 / 395	0.7722	3.09	+35.1%	4 / 257	0.0156	2.01	0.0%
median	—	0.7354	2.95	—	—	0.0156	2.01	—

Table A.5: **Post-hoc accounting for the five fresh memory pairs.** Training calls exclude all six validation passes. All-validation calls include the final pass, whose charged calls are shown separately. Q is the number of qualifying distinct molecules in the final pass. Pre-seen Q counts those scored in earlier recorded episodes of the same run. The 224-call column brackets within-window qualified coverage between the last complete episode and the episode straddling 224 calls. Common support across all ten cells ends at 247 calls.

seed	arm	train calls	all val. calls	final calls	Q	pre-seen Q	Q at 224
100	memory	7997	2019	378	278	46	171–173
100	off	7174	1588	256	4	4	4–4
101	memory	21076	6164	2006	302	51	30–35
101	off	7062	1559	256	9	5	9–9
102	memory	6674	1590	268	162	20	131–131
102	off	6831	1555	257	3	3	3–3
103	memory	5722	1482	247	221	28	202–204
103	off	16647	3761	773	1	1	1–1
104	memory	6704	1788	395	305	42	180–184
104	off	6601	1511	257	4	4	4–4

support ends at 247 charged calls. At 224 calls, the conservative memory-on-minus-off margins are $+167$, $+21$, $+128$, $+201$ and $+176$ qualifying molecules. The seed-101 margin is -4 at 32 calls and -2 at 64 calls, then positive at every reported checkpoint from 96 through 224 calls. The bounds account for unknown call order within the episode that straddles a checkpoint. They describe the observed policies without testing a hard stop.

Of the 162–305 qualifying molecules in each treated final pass, 20–51 were scored in earlier recorded episodes of that run. Earlier-window overlap in the controls is 1–5 qualifying molecules. Counting only qualifying molecules not scored earlier in the run, over the same final-pass calls, memory leads its control in all five pairs, by $+0.614$, $+0.109$, $+0.530$, $+0.781$ and $+0.666$ per call, because nearly every qualifying molecule of a control pass had been scored before. This is overlap with prior records, not an estimate of deliberate archive replay. Table A.5 reports each run's training, validation and final-pass calls beside that overlap. Table A.6 lists all evaluated common-call margins, including the early negative values.

A.4 Training cost, validation cost and soft-budget compliance

For each historical run, we separate the 3200 training episodes from six 128-episode validation passes. Validation costs already include the final pass. The last column counts the fraction of final-pass episodes with more than two charged calls. This differs from whether a run's *mean* spend exceeds two. The group medians of these episode fractions are 0.0391 for the seven memory runs and 0.4414 for the 22 unfiltered runs.

The coverage table uses completed-episode boundaries from the final pass. At a requested call count, it includes all episodes completed within that count. The within-episode proposal order is unavailable. The table covers the expanded twenty-nine-run cohort reported in the body and

Table A.6: **Post-hoc qualified coverage at common charged-call checkpoints.** Each entry is the memory-on count after the last complete episode minus the memory-off count after its episode straddling that checkpoint. It is a conservative lower bound on the paired margin under any within-episode call order. Negative values favor memory off. All five pairs favor memory on at the reported checkpoints from 96 through 224 calls. Seed 101 favors memory off at 32 and 64 calls. The final-window rate remains the pre-specified endpoint.

seed	32	64	96	128	160	192	224
100	+22	+52	+75	+100	+125	+148	+167
101	-4	-2	+7	+10	+18	+18	+21
102	+19	+37	+59	+82	+101	+116	+128
103	+30	+61	+88	+117	+144	+176	+201
104	+24	+52	+78	+100	+126	+157	+176

abstract. All 29 curves have support through 203 calls. A dagger marks the three cells whose filter never executed, listed under plain GRPO as in Table A.1. The six listed checkpoints stop at 192. This is a descriptive comparison of evaluation efficiency. Training costs remain unequal and are shown separately.

		calls				> 2			calls				> 2
arm	seed	train	all val.	final	calls/ep.	fraction	arm	seed	train	all val.	final	calls/ep.	fraction
scaffold memory	0	5967	1511	231	1.80	0.062	IPS-GRPO	2	8077	1931	345	2.70	0.625
scaffold memory	1	5620	1435	250	1.95	0.008	IPS-GRPO	3	8252	2098	532	4.16	0.641
scaffold memory	2	6894	1659	303	2.37	0.250	IPS-GRPO	5	8885	1841	262	2.05	0.125
scaffold memory	3	6454	1546	203	1.59	0.016	IPS-GRPO	6	9472	1951	308	2.41	0.477
scaffold memory	4	5960	1455	239	1.87	0.008	plain GRPO†	0	6726	1550	273	2.13	0.047
scaffold memory	5	7324	1724	404	3.16	0.523	plain GRPO†	1	6333	1541	266	2.08	0.062
scaffold memory	6	6222	1459	236	1.84	0.039	plain GRPO†	2	6506	1530	259	2.02	0.117
plain GRPO	0	7090	1664	369	2.88	0.406	top-k reward	0	15981	3703	810	6.33	1.000
plain GRPO	1	7254	1708	290	2.27	0.219	top-k reward	1	6293	1499	256	2.00	0.000
plain GRPO	2	7436	1757	307	2.40	0.242	top-k reward	2	21950	5458	1533	11.98	1.000
plain GRPO	4	8675	2288	478	3.73	0.664	top-k reward	3	12247	3392	1113	8.70	1.000
plain GRPO	5	7743	1811	283	2.21	0.203	top-k reward	4	12092	2826	1015	7.93	1.000
plain GRPO	6	9397	2167	385	3.01	1.000	top-k reward	5	7563	2222	857	6.70	1.000
IPS-GRPO	0	6854	1599	262	2.05	0.039	top-k reward	6	18875	4331	974	7.61	0.992
IPS-GRPO	1	7266	1712	271	2.12	0.133							

calls	filtered min	filtered max	other min	other max
32	26	32	1	17
64	53	63	1	30
96	78	91	1	38
128	99	121	1	45
160	125	147	1	57
192	144	178	1	62

A.5 Provenance and execution checks

Training and evaluation environments. Molecules are scored during training and counted in a separate analysis environment. The training and oracle environment runs Python 3.10.20 with RDKit 2023.09.6. All analyses run under Python 3.12.2 with RDKit 2025.03.3, NumPy 2.3.1 and Matplotlib 3.10.3. All counts depend on canonical SMILES, which can vary across RDKit versions. We therefore recomputed final-window distinct-molecule counts under both versions for one run of each principal arm.

run	RDKit 2023.09.6 (scored)	RDKit 2025.03.3 (analysed)
scaffold memory	202	202
IPS-GRPO	52	52
plain GRPO	8	8

All three counts match exactly.

Hardware and cross-machine differences. Runs used two machines. One has NVIDIA L20 accelerators with driver $580.82.07$ and tensor-parallel degree two. The other has an NVIDIA RTX PRO 6000 Blackwell with driver $580.126.09$ and degree one. The machine is recorded for every selection-side cell in Table A.2, and Appendix F.2 lists the ten differing configuration keys, five experimental and five infrastructural.

Filter execution checks. The scoring service is a long-running process. A service started before the filter was implemented accepts but ignores filter requests, as occurred in three historical runs. The selection experiments checked the service against a reference score before the first training step, but the historical runs checked only that the service responded, which cannot tell a current service from a stale one. Cohort assignment therefore relies on a retrospective check of each run's output, using quantities recorded before and after filtering. The best score of an episode is the maximum over raw oracle scores, and its return is the maximum over filtered scores, so a molecule whose scaffold bucket is full lowers the second and not the first. The gap between them is the filter's footprint. A gap above zero shows that the filter ran. A zero gap is inconclusive because filtering a molecule below the episode's best score leaves both maxima equal. Across all fifteen cells whose own configuration record requests the filter (Table A.3), twelve show a footprint between 20.8 and 46.3 per cent and three show none, with no overlap and a separation of 20.8 points. These are exactly the three runs reassigned in Appendix A.1, clearly separated from all others. The best-score series comes from episode records and the return series from training logs. They are aligned by position, not episode, so the comparison is distributional, not pointwise. This is why one cell shows a footprint of -0.0 per cent. Its per-update differences lie between -0.0005 and $+0.0005$, the scale of the alignment noise.

The same gap gives a prediction for every arm. Plain GRPO and IPS use the ordinary reward without a filter, so their gaps must be zero. IPS rescales the token-level reward after the environment has logged the episode return, so it cannot change the gap. The top-k reward averages the best distinct molecules, which falls below the maximum whenever more than one molecule scored, so its gap must be positive. All twenty-six completed runs follow this pattern. The two null arms span -0.0 to 0.0 per cent, the top-k arm 12.5 to 40.4, and the filtered arm 30.4 to 33.9. The three reassigned runs lie in the null band at exactly zero, so their reward records support the reclassification independently of the service history. The environment refuses to run if a filter is configured but the service returns no multiplier. Each run has a separate scoring namespace to prevent memory leakage through a shared service.

Model, data and seeds. The primary setting uses Qwen3-4B, and one replication uses Qwen3-1.7B. Each objective has its own frozen data split. The replicate index is the environment seed. Selector draws also depend on unrecorded service-process randomisation. The environment seed therefore specifies only part of the stochastic state.

A.6 Departures from registered analyses

Changes made before the affected outcome is read are calibration. Later changes make the affected analysis exploratory, irrespective of the result. Table A.7 lists every departure from a registered analysis that bears on a claim in this paper, with that distinction.

The reassignment and the change to the confirmation criterion were decided after the outcomes were known. Both rely on execution evidence computed without endpoints. Because the decisions followed outcome inspection, the historical separation provides descriptive corroboration, not confirmatory evidence.

Results under the original assignment. Retaining the three runs under their initial arm assignment changes the threshold sweep. At the calibrated threshold, the filter arm spans 0.0113 to 0.7649 over seven runs, against 0.0026 to 0.2338 over nineteen. The arms overlap at all ten thresholds, whereas under the executed assignment they are disjoint at all ten. The rank statistic falls from $88/88$ to $96/133$. The original assignment removes the separation. The three endpoints lie in the unfiltered range, lowering the arm's minimum from 0.7044 to 0.0113 without changing its maximum. The seven-run arm of the main text replaces the three unexecuted runs with their verified re-runs (Appendix A.2) and reads 0.6840 to 0.8647, disjoint at $U = 154/154$. The reassignment

Table A.7: Departures from registered analyses and whether the affected outcome had been read at the time of each decision.

departure	what changed	outcome already read?
Three runs reassigned out of the memory arm	Three runs labelled as memory runs reached a scoring service that predated the filter and were moved to plain GRPO. The historical memory arm went from seven runs to four.	Yes. The reassignment rests on the filter footprint (Appendix A.5), and the three seeds were re-run under criteria fixed in advance (Appendix A.2).
Role of the confirmation criterion	Criterion 2 of the confirmation design no longer decides whether the three re-run seeds count. They enter on execution evidence, the memory arm moves from four runs to seven, and U from $88/88$ to $154/154$.	Yes. The endpoints 0.8268, 0.6840 and 0.8647 had been read.
Control set of the bucket screen	Amended from sixteen control runs to eighteen, moving the promotion threshold to 0.7086.	No. Control endpoints had been read, no treated bucket result had.
Read on the second target	Two failed runs were not repeated, and the primary read became the balanced set of two runs an arm.	No. Completion status and spend were known, no endpoint had been computed.
Allocation across machines	A balanced allocation across two machines was dropped, and the second machine became a separately reported replication.	No. Decided before any run of the new allocation completed.

criterion is treatment execution, not outcome. The three runs have zero filter footprint, while the other twelve span 20.8 to 46.3 per cent, separated by 20.8 points. Their reward records also fall in the null band of unfiltered arms. Both checks can be computed without the endpoint.

A.7 Pre-specified criteria

Three analyses were fully specified before their runs produced any outcome. Table A.8 gives each criterion, the observed outcome and the alternative registered branch.

The recorded filter, service, budget and episode count of every run were checked before the run was counted.

B Selection experiments

B.1 Repeat purchases by position

Our farthest-first selector starts every selection from the first valid proposal of the turn and adds the proposal farthest from it, whereas MaxMin-A as published draws the first element at random. With two molecules scored per step, the anchor is half of every selection and is chosen by the policy's proposal order rather than by distance. An episode record lists its distinct molecules in the order they were first scored, and within a turn in proposal order, so the first listed molecule of a farthest-first episode is the anchor of its first scoring turn. Anchors of later turns occur among the other positions, making the contrast below conservative. For uniform selection the first listed molecule is the earlier-positioned of two random picks and is shown only as a reference.

Table B.1 splits the earlier-episode repeat purchases of the final 128-episode pass by position. A repeat is a canonicalised molecule already scored in an earlier episode of the same pass, as in the repeat category of Figure 3. The two positions sum to that category in every run. In farthest-first cells the anchor is a repeat in 7–85% of episodes against 3–42% elsewhere, and it carries 45–78% of the repeats. Uniform cells also repeat more often at their first recorded molecule, at 14–32% against 2–26%, which suggests that the policy tends to open its proposal list with familiar molecules. Farthest-first scores that opening molecule whenever selection is required, whereas uniform selection scores it only when drawn. Across their 128 episodes the farthest-first cells have 20 to 118 distinct anchors, against 88 to 111 distinct first molecules in the uniform cells, fewer in every pair

Table A.8: Three criteria specified before any outcomes were available, with observed outcomes and alternative registered branches.

experiment	criterion, fixed in advance	outcome, and what would have happened otherwise
Re-running the three reassigned cells	A bridge cell at the executed seed within ± 0.10 of its original 0.7044, *then* at least two of three seeds above 0.4691. A cell that reaches the old service is discarded.	Bridge 0.6553, inside $[0.604, 0.804]$. Seeds at 0.827, 0.684, 0.865, three of three. No cell was discarded. A missed bridge would have made the three cells box-confounded and settled nothing. Two or three at or below 0.4691 would have restated the headline at $U = 96/133$ with the arms overlapping.
Training without the scaffold bucket, at budget eight	Paired difference memory − repeat-only, median of four, against a threshold of 0.15 fixed before any cell ran.	Median $+0.055$, below the threshold, so at budget eight the registered rule did not support a bucket contribution of the specified size. This is not an equivalence result and does not transport to budget two (Appendix C.2). Above 0.15 and positive in four of four, the bucket would have carried it and the mechanism sentence in this paper would read the other way.
The same control at budget two	No threshold, and the budget-eight threshold is not reused. Registered as a measurement of the levels, not as a test.	Repeat-only at 0.659–0.746 against the original four scaffold-memory cells' 0.704–0.765. None was added afterwards and no verdict is drawn from the paired signs, which are mixed.

Table B.1: Earlier-episode repeat purchases by position in the final 128-episode pass. The first listed molecule of a farthest-first episode is the anchor of its first scoring turn, chosen by position. For uniform selection it is the earlier-positioned of two random picks. A repeat is a molecule already paid for in an earlier episode of the same pass, canonicalised. Anchor and other-position repeats sum to the repeat bucket of Figure 3. The last column counts how many of the episodes that bought the most re-bought qualifying compound had it in the first position (Top). Share is the fraction of all earlier-episode repeats that sit at the first position, and Distinct counts the different first molecules across the 128 episodes.

Pair	Arm	First (%)	Other (%)	Share (%)	Distinct	Top
80	farthest-first	7 (9/127)	3 (10/367)	47	118	3/3
80	uniform	14 (18/128)	11 (13/123)	58	111	0/4
81	farthest-first	62 (79/128)	17 (22/128)	78	49	36/36
81	uniform	21 (26/126)	12 (15/123)	63	103	8/8
82	farthest-first	48 (61/128)	8 (32/419)	66	68	11/12
82	uniform	16 (20/125)	2 (2/122)	91	108	4/4
90	farthest-first	85 (109/128)	42 (133/320)	45	20	46/46
90	uniform	32 (40/125)	26 (32/122)	56	88	9/9

but pair 80. Shuffling the order inside each episode lowers the anchor repeat rate of pair 81 from 0.617 to 0.393, so the split responds to position and not only to the episode's contents. This post-hoc descriptive analysis locates repeats but cannot estimate the effect of a random start.

B.2 Registered random-start control

The control changes only the farthest-first anchor, drawing it uniformly from the turn's valid proposals with a fresh per-run random stream instead of choosing the first proposal. Its design, decision branches and exclusion rules were registered before any cell ran. Each cell's resolved configuration matches its seed's original farthest-first cell key by key except the selector, run identity, service

Table B.2: The registered random-start control. Qualified discovery per charged call on the final 128-episode pass for the random-start farthest-first cell and the original first-start and uniform cells of the same seed, which ran earlier. D is random-start minus uniform and E random-start minus first-start. Spend is charged calls per episode. Seed 90's controls ran on the other machine, so its row is a cross-machine comparison and does not enter the registered reading.

Pair	random-start	first-start	uniform	D	E	spend (rs)	spend (fs)	spend (u)
80	0.536	0.637	0.825	-0.289	-0.101	5.98	5.10	1.96
81	0.634	0.504	0.796	-0.162	$+0.130$	5.04	2.00	1.95
82	0.715	0.562	0.871	-0.156	$+0.153$	2.27	5.40	1.94
90*	0.797	0.220	0.390	$+0.406$	$+0.576$	2.00	4.36	1.96

address and output paths, on the same machine and recipe. The pool log shows that the treatment was applied. The anchor coincides with the first valid proposal in 0.170, 0.211, 0.180 and 0.193 of selecting requests, against 0.173, 0.207, 0.187 and 0.196 expected from a uniform draw on the same pools. The controls are the original uniform and first-start runs, completed before the new runs rather than concurrently. The fourth pair's controls ran on the L20 machine while its random-start cell ran on the RTX PRO 6000 machine, because write access to L20 was unavailable. This cross-machine pair was excluded from the registered analysis before any endpoint was inspected.

The results meet the first registered branch. The random-start arm trails uniform selection in all three same-machine pairs (Table B.2), so the deficit does not require the positional anchor. Compared with the earlier first-start cells, the random-start cells raise the endpoint by 0.130 and 0.153 in two pairs and not in the third. Because the two arms ran at different times, this comparison does not isolate an anchor effect. At 224 common calls uniform selection has found 37, 10 and 30 more distinct molecules. The random start raises within-episode spread above uniform selection's in two of the three same-machine pairs. The cross-machine difference is $+0.406$, but the control machine consistently produced much lower values, making this pair uninformative about the arm.

B.3 Fixed-pool replay

The online comparisons train each arm separately and therefore measure selection together with the resulting policy adaptation. The random-start cells logged every selection request with its complete ordered proposal list, which makes the direct effect of a rule measurable. Each rule is replayed on one cell's logged stream in request order, keeping its own purchase history from the start of training, and every valid candidate is scored offline by the service's oracle. For the $44{,}240$ molecules scored online, the offline and service scores match exactly. The replay uses the service's chemistry toolkit, whose validity decisions define the pools, and a fixed hash seed, because the pool-hash uniform rule depends on it. The realised valid pool has a median of four or five proposals, so two of four or five are scored, far less selective than the published one in ten.

Table B.3 reports the results. First-start farthest-first has the highest repeat share on every stream and the fewest new qualifying molecules per call on three of four. Pool-hash uniform selection has the lowest yield on the fourth. A random start stays within 0.010 of a fresh uniform draw, and the determinantal rule within 0.012 in repeat share. Selection restricted to previously unscored molecules gives the largest consistent change relative to fresh uniform selection, lowering the repeat share by 5.5 to 7.7 points and raising new qualifying molecules per call by 0.026 to 0.041. The replay keeps each rule's history from the start of training, whereas the endpoint deduplicates within the final pass, so these rates are not endpoint values. First-start farthest-first trails the fresh uniform draw by 0.006 to 0.028 per call on the same pools, an order of magnitude below the online gaps. These pools come from random-start policies and remain fixed during replay. The analysis therefore measures the direct effect on a fixed stream, not the effect of training.

B.4 Proposal drift and selection ratio

Two further explanations concern proposal drift, which remains unmeasured, and the selection ratio, measured in the random-start runs.

Table B.3: Fixed-pool replay. Every rule is applied to the same logged sequence of candidate pools of one random-start cell, keeping its own purchase history from the start of training. Repeat is the share of charged calls spent on a molecule the rule had already bought, and new qualified is new molecules at or above $\tau = 0.567$ per charged call. Distance is the selector's own distance between the two molecules of a selecting request. The pools do not respond to the replayed purchases, so the table measures the direct effect of the rule on a fixed stream, not a training outcome.

Pools	Rule	Repeat	New qualified	Distance
80	farthest-first, first start	0.478	0.270	0.225
80	farthest-first, random start	0.450	0.280	0.222
80	uniform, pool-hash draw	0.472	0.270	0.135
80	uniform, fresh draw	0.454	0.276	0.137
80	determinantal	0.461	0.276	0.183
80	uniform over unbought	0.394	0.303	0.131
81	farthest-first, first start	0.413	0.322	0.266
81	farthest-first, random start	0.379	0.340	0.261
81	uniform, pool-hash draw	0.388	0.340	0.166
81	uniform, fresh draw	0.371	0.350	0.168
81	determinantal	0.383	0.341	0.216
81	uniform over unbought	0.294	0.391	0.155
82	farthest-first, first start	0.347	0.387	0.208
82	farthest-first, random start	0.316	0.404	0.204
82	uniform, pool-hash draw	0.337	0.397	0.128
82	uniform, fresh draw	0.323	0.401	0.131
82	determinantal	0.322	0.403	0.169
82	uniform over unbought	0.268	0.434	0.127
90	farthest-first, first start	0.340	0.373	0.279
90	farthest-first, random start	0.307	0.396	0.273
90	uniform, pool-hash draw	0.314	0.394	0.174
90	uniform, fresh draw	0.306	0.397	0.175
90	determinantal	0.305	0.398	0.225
90	uniform over unbought	0.241	0.429	0.171

Proposal drift. The generating policy in Svensson et al. (2025) is fine-tuned sequentially with a persistent memory rather than reset at each step, so a drifting proposal distribution is not peculiar to our setting. We did not measure whether the two distributions drift at comparable rates or whether drift interacts differently with selection, leaving this explanation unresolved.

The selection ratio. That work selects one molecule in ten from a pool of 640. We cap the agent at two per oracle step against a ceiling of sixteen proposals a turn, so the ceiling ratio, one in eight, is of the same order. The original runs did not record realised proposal counts. The random-start runs recorded a median valid pool of four or five proposals, only occasionally reaching sixteen (Appendix B.3). Selection in this setting is therefore two of four or five, far less selective than one in ten, assuming the first-start policies proposed pools of similar size.

B.5 Common-call coverage of the selection pairs

The main text reports uniform random selection's advantage at every budget in every pair and the mean margin. Table B.4 gives both arms' values for each margin.

All twenty-eight margins are positive. We stop at 224 calls because the smallest total spend across all arms and pairs is 248. Beyond that point, completed curves are compared with continuing ones. This artefact alone reverses two of four pairs at 320 calls.

Bounds at the episode boundary. Coverage curves advance at episode boundaries because the logs record calls and molecules by episode, without the within-episode proposal order. A point at B calls therefore counts the episodes that finished within B, and an arm spending five calls an episode leaves a larger unread tail there than one spending two. The uncertainty this leaves is bounded on both sides from the same logs. The reported value is the smallest coverage the arm can have at B,

Table B.4: Distinct molecules discovered by each arm of the four farthest-first pairs at each charged-call budget within the same 128-episode window. Margins are uniform random selection minus farthest-first.

pair	arm	32	64	96	128	160	192	224
80	farthest-first	26	50	59	86	114	139	157
80	uniform random	30	59	87	117	146	176	201
	margin	+4	+9	+28	+31	+32	+37	+44
81	farthest-first	22	45	60	81	100	118	138
81	uniform random	32	62	88	115	141	166	187
	margin	+10	+17	+28	+34	+41	+48	+49
82	farthest-first	24	45	68	86	106	140	161
82	uniform random	31	63	92	122	150	178	204
	margin	+7	+18	+24	+36	+44	+38	+43
90	farthest-first	23	45	62	74	85	92	105
90	uniform random	30	57	82	104	124	141	164
	margin	+7	+12	+20	+30	+39	+49	+59
	mean margin	+7.0	+14.0	+25.0	+32.8	+39.0	+43.0	+48.8

and the value after the episode straddling B is the largest, since the budget remaining at B cannot buy more than that episode's own new molecules. Crediting every straddling episode to farthest-first and none to uniform random, which is the worst case for the reported margin, leaves random ahead in 27 of the 28 cells and at every checkpoint from 64 calls onward. The exception is pair 82 at 32 calls, where $+7$ becomes -1. In the 224-call column the worst case is $+37$, $+49$, $+43$ and $+55$ against the reported $+44$, $+49$, $+43$ and $+59$. The determinantal screen's margins are small enough that the bracket covers the sign in 8 of its 21 cells, five of them at 64 calls or fewer and three in the pair whose control spent 5.96 calls an episode against its treated arm's 2.09. At 224 calls its reported $+13$, $+21$ and -4 stand, with a worst case of $+13$, $+21$ and -5. Recording calls in order would resolve this uncertainty, alongside the candidate-pool logging discussed above.

B.6 Both diversity measures on the same records

Our endpoint counts distinct qualifying molecules per charged call. The selection literature uses a packing number, defined as the largest subset of molecules that are all separated by at least a fixed distance, approximated greedily. We compute both our endpoint and a ported greedy packing statistic on each arm's recorded outputs at 224 charged calls. This checks whether the observed ordering is specific to our endpoint under that approximation.

Only the diversity definition is ported, a Tanimoto distance of 0.7 on Morgan fingerprints at radius 2 and 2048 bits. Their task-specific activity gate requires a predicted-activity model and a drug-likeness score and is not imported. The qualified column of Table B.5 instead uses our calibrated threshold. This is a comparison on a ported measure, not a reproduction of their benchmark. The packing number is computed greedily, so every value is a lower bound on its arm's exact maximum. We report both greedy input orderings, by discovery and by score, but their agreement does not establish the ordering of exact maxima.

The decision rule was fixed in the analysis specification before any value was computed. Specifically, if uniform random holds at least as many mutually dissimilar molecules in at least three of the four farthest-first pairs under both greedy orderings and both episode-boundary bounds, the reversal also appears under this ported statistic. If farthest-first is higher in at least three, the result is a trade-off between the two measured objectives. Anything else is reported per pair. The first branch holds in three of four pairs. Candidate-pool and online-policy differences remain possible explanations, and the ordering of exact maximum packing numbers remains unresolved.

In pairs 80, 81 and 82 uniform random holds more mutually dissimilar molecules than farthest-first, by 5, 6 and 7 in discovery order and 4, 9 and 8 in score order, while also finding 44, 49 and 43 more distinct molecules. Pair 90 is the exception and runs in the opposite direction. Farthest-first packs 15 against 7, demonstrating that the selector does achieve the structural spread it optimises in that instance, while still returning 59 fewer distinct and 26 fewer qualified molecules for a spend

Table B.5: Distinct molecules, qualified molecules and greedy packing numbers (Tanimoto distance at least 0.7) for each selection pair at 224 charged calls, at the lower episode-boundary bound. Packing is computed in discovery order and in score order, and over the qualified molecules in discovery order.

pair	arm	distinct	qualified	packing (discovery)	packing (score)	packing (qualified)
80	MaxMin	157	135	20	21	16
	random	201	191	25	25	22
81	MaxMin	138	114	19	19	15
	random	187	178	25	28	20
82	MaxMin	161	134	12	14	10
	random	204	197	19	22	16
90	MaxMin	105	68	15	16	7
	random	164	94	7	10	5
91	DPP	178	161	15	13	13
	random	191	183	12	14	10
92	DPP	187	149	23	25	21
	random	208	200	24	31	23
93	DPP	184	176	30	38	29
	random	180	156	12	10	8

Table B.6: Qualified discovery per charged call ($\tau = 0.567$) of each selection pair at every validation pass, from pass 0 before training to pass 5 after the last training cycle, with each arm's best pass.

pair	arm	pass 0	1	2	3	4	5	best
80	MaxMin	0.068	0.084	0.289	0.321	0.509	0.637	0.637
	random	0.038	0.063	0.435	0.716	0.878	0.825	0.878
81	MaxMin	0.068	0.168	0.452	0.458	0.569	0.504	0.569
	random	0.038	0.113	0.312	0.656	0.763	0.796	0.796
82	MaxMin	0.068	0.135	0.284	0.500	0.566	0.562	0.566
	random	0.038	0.091	0.428	0.691	0.793	0.871	0.871
90	MaxMin	0.040	0.120	0.106	0.262	0.346	0.220	0.346
	random	0.026	0.179	0.252	0.401	0.437	0.390	0.437
91	DPP	0.068	0.180	0.269	0.498	0.688	0.720	0.720
	random	0.049	0.096	0.328	0.441	0.650	0.808	0.808
92	DPP	0.057	0.086	0.425	0.595	0.562	0.697	0.697
	random	0.049	0.152	0.398	0.639	0.757	0.872	0.872
93	DPP	0.057	0.090	0.310	0.533	0.655	0.757	0.757
	random	0.062	0.080	0.186	0.336	0.505	0.666	0.666

of 4.36 calls an episode against 1.96. The determinantal screen is mixed on this measure as it is on ours. Random selection is ahead in pair 92 under both orders, the two orders disagree in pair 91, and pair 93, the pair whose control spent nearly three times what its treated arm did, runs strongly in the determinantal rule's favour on every column. The bracket of Appendix B.5 moves none of these signs.

B.7 The selection pairs at every validation pass

The selection endpoint uses the sixth and final validation pass. A run that peaked earlier is scored after its peak, and the two arms need not peak together. Table B.6 reports both final-pass and best-pass results. The same extraction function identifies pass boundaries in the selection and historical analyses.

Farthest-first leads early and loses by the end. In three of the four pairs its difference against uniform random is positive at the first passes after training starts, by up to $+0.141$, and only one pair is negative at every pass after training starts. An early-stopped comparison would therefore have

Table B.7: Charged calls each arm needed to reach 10, 25 and 50 qualifying molecules in the final validation pass, counted at episode boundaries, with its training and validation calls. n/r marks a target the pass did not reach.

pair	arm	charged calls to reach 10	25	50	training calls	validation calls
80	MaxMin	31	41	98	14577	3553
	random	10	28	60	6206	1541
81	MaxMin	16	42	102	6787	1673
	random	10	26	56	5964	1539
82	MaxMin	12	37	76	11079	2599
	random	10	28	54	7218	1691
90	MaxMin	15	55	145	14825	3054
	random	16	36	96	7545	1857
91	DPP	10	26	58	7728	1896
	random	11	27	53	8344	1925
92	DPP	21	40	75	10327	2679
	random	10	26	52	6096	1531
93	DPP	12	28	67	8342	2055
	random	12	30	60	16124	3960

reported the opposite result. The reversal persists when each arm is scored at its own best pass. This hindsight-optimal checkpoint, unavailable in deployment, leaves the difference negative in all four pairs, at -0.241, -0.227, -0.305 and -0.091. The determinantal screen stays mixed under both readings, negative in two of three at the final pass and at the best pass.

B.8 Cost of a discovery target and of training

A rate compares arms but does not quantify the calls needed to reach a specified molecular yield. Table B.7 therefore gives the charged calls needed to reach three targets, fixed before the table was made, in the final validation pass. Calls are counted at episode boundaries, making each entry an upper bound within one episode. Targets not reached during a pass are marked accordingly, without extrapolation.

Target-based comparisons preserve the rate ordering and express its magnitude in calls. Farthest-first needs 1.32 to 1.62 times as many charged calls as its control to reach twenty-five qualifying molecules, and 1.41 to 1.82 times as many to reach fifty, in four pairs of four. At ten molecules the comparison is mixed. Three pairs still favour uniform random by 1.20 to 3.10 times, while the fourth is level at 15 calls against 16. The determinantal screen is mixed at every target.

Training cost is not equalised. The common-call analysis equalises evaluation spend only, which the paper states as a limitation. Across the four pairs the farthest-first arm spent 1.14 to 2.35 times its control's training calls, so its lower evaluation yield also followed more costly training. In the determinantal screen the direction varies with the pair, and in the pair whose control spent nearly twice the treated arm's training calls the determinantal rule is the one that leads.

Wall-clock time is accounted separately in Table B.8, as the sum of the trainer's per-update step times, which include the five periodic validation passes. The farthest-first training loop took 1.00 to 1.71 times as long as its uniform control on the RTX PRO 6000 machine and 1.22 times on the L20 machine, and the pair with near-equal spend is the one at 1.00. The memory pairs range from 0.59 to 1.61. Each comparison used one shared machine. Wall-clock time therefore also reflects concurrent load and is reported descriptively, not as an arm effect.

B.9 Determinantal selection against a uniform control

The three complete DPP pairs (Figure B.1) use the same final-window endpoint and cumulative-coverage estimator as MaxMin. Every run's resolved configuration confirms its assigned selector. All runs use the memory arm's persistent memory, as in the farthest-first pairs. Differences in the text

Table B.8: Wall-clock hours of the training loop per cell, the sum of the 25 per-update step times, which include the five periodic validation passes (in parentheses) but not the pass before training. Hours depend on the machine, so only the two arms of one pair are comparable. The ratio is treated over control.

Comparison	Seed	Machine	Treated (h)	Control (h)	Ratio
farthest-first vs uniform	80	RTX PRO 6000	8.2 (0.2)	4.8 (0.1)	1.71
farthest-first vs uniform	81	RTX PRO 6000	4.8 (0.1)	4.8 (0.1)	1.00
farthest-first vs uniform	82	RTX PRO 6000	6.9 (0.1)	5.2 (0.1)	1.32
farthest-first vs uniform	90	L20	8.4 (0.4)	6.9 (0.3)	1.22
determinantal vs uniform	91	RTX PRO 6000	5.4 (0.1)	5.6 (0.1)	0.96
determinantal vs uniform	92	RTX PRO 6000	6.2 (0.1)	4.8 (0.1)	1.30
determinantal vs uniform	93	RTX PRO 6000	5.5 (0.1)	8.2 (0.2)	0.67
memory on vs memory off	100	RTX PRO 6000	5.9 (0.1)	5.1 (0.1)	1.16
memory on vs memory off	101	RTX PRO 6000	7.9 (0.1)	4.9 (0.1)	1.61
memory on vs memory off	102	RTX PRO 6000	5.8 (0.1)	4.7 (0.0)	1.24
memory on vs memory off	103	RTX PRO 6000	4.5 (0.1)	7.6 (0.1)	0.59
memory on vs memory off	104	RTX PRO 6000	4.7 (0.1)	4.8 (0.1)	0.98

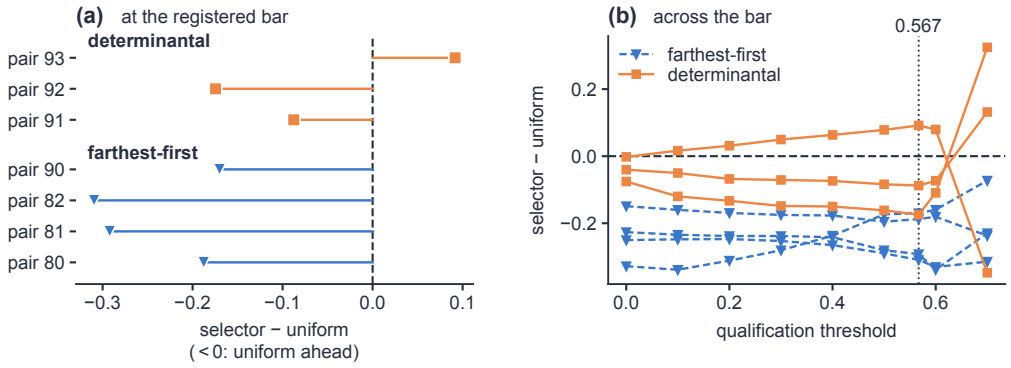


Figure B.1: **Farthest-first has lower yield than uniform selection in all four pairs, while the determinantal rule recommended in the cited study gives mixed results.** Each marker represents one training-seed pair. Within a turn, the selector and control each choose the same number of molecules from their own valid proposal lists at the same scoring cap. The trained arms follow different online trajectories, so these are not fixed-pool replays. Negative differences favour uniform selection. The published study chooses MaxMin-A for its balance across distance-based active-molecule and scaffold readings (Svensson et al., 2025). Its relative ranking of MaxMin and DPP varies across metrics. Here MaxMin has lower yield in all four pairs, at -0.188, -0.292, -0.309 and -0.170. The recommended determinantal rule has lower yield in two of three pairs, at -0.088 and -0.175, and higher yield in the third at $+0.092$. The reversal therefore varies across selectors. All seven predeclared pairs are complete.

are DPP minus random. Table B.9 gives each arm's endpoint and Table B.10 the coverage difference, random minus DPP. The common support ends at 239 calls, so the last inspected checkpoint is 224. Each row represents a pair, and checkpoints within it are correlated. Both uniform and DPP selection derive a draw from the ordered candidate tuple's process-dependent Python hash. Within one service process, the same ordered tuple produces the same draw rather than a fresh random sample at each request. The process hash seed was not retained, and an environment seed alone does not establish bitwise reproduction after a service restart. The full proposal pools and their recurrence were not recorded. These outcomes describe the implemented pool-conditioned procedures, not robustness to fresh per-request draws. The raw-rate differences are negative in all three. The qualified differences at 0.567 have mixed signs. At 0.7, DPP leads in two pairs. The selector ordering therefore depends on the qualification threshold.

Table B.9: Charged calls, calls per episode, raw rate and qualified discovery per charged call ($\tau = 0.567$) for each arm of the three determinantal pairs, in the final 128-episode pass.

seed	selector	calls	calls/ep.	raw rate	qualified rate
91	DPP	282	2.20	0.8050	0.7199
91	random	239	1.87	0.8452	0.8075
92	DPP	519	4.05	0.8324	0.6975
92	random	250	1.95	0.9080	0.8720
93	DPP	268	2.09	0.7948	0.7575
93	random	763	5.96	0.7969	0.6658

Table B.10: Coverage difference, random minus DPP, in distinct molecules discovered by the given number of charged calls, for the three determinantal pairs.

seed	32	64	96	128	160	192	224
91	-2	-2	$+1$	$+4$	$+10$	$+9$	$+13$
92	$+5$	$+2$	$+10$	$+5$	$+5$	$+14$	$+21$
93	-2	$+5$	$+4$	-4	-1	-4	-4

C Memory and multiplicity

C.1 Scaffold-level yield

The endpoint counts distinct molecules, raw in Figure A.1 and above the registered threshold everywhere a threshold is stated. Counting molecules is conservative in both cases. Screening campaigns operate on scaffolds, and the mechanism limits how many molecules may share a scaffold. Scaffold counts therefore evaluate the arms at the structural level targeted by memory. Scaffolds are Bemis–Murcko as computed by RDKit, the same definition the filter uses.

arm	n	molecules per call	scaffolds per call	distinct scaffolds
scaffold memory (2020)	4	0.7222	**0.5721**	146
IPS-GRPO (2026)	6	0.1695	0.0322	10
plain GRPO	9	0.0579	0.0146	5
top-k reward	7	0.0548	0.0035	3

The gap widens by a factor of four when the unit changes. Against the strongest other arm the filter is ahead by $4.3\times$ on molecules and by $17.8\times$ on scaffolds. The arms are disjoint on scaffolds per call as they are on molecules, at a filter minimum of 0.4827 against a best other of 0.0725 and $U = 88/88$. Variations around one scaffold still count as distinct molecules and account for most outputs of the corrections.

The result as a screening budget. The screening budget is the number of oracle calls needed to reach a given number of distinct scaffolds worth following up.

arm	≥ 10	≥ 25	≥ 50	≥ 100
scaffold memory (2020)	14	36	75	168
IPS-GRPO (2026)	never	never	never	never
plain GRPO	never	never	never	never
top-k reward	never	never	never	never

The table reports median oracle calls over each arm's runs. *Never* means the target was not reached within the window. The filter's median run reaches ten distinct qualifying scaffolds in fourteen calls, and no other arm's median run reaches ten within the window. The median is over each arm's runs, so an arm counts as reaching a target only when more than half of its runs do. The 128-episode window defines a fixed screening budget. These results concern yield within that window, not eventual yield.

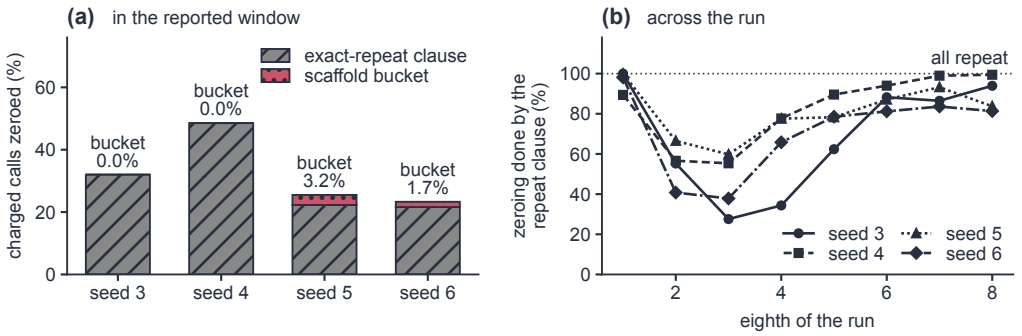


Figure C.1: **The scaffold bucket contributes little final-window suppression but is active during training.** (a) Charged oracle calls zeroed in the final validation pass, split by clause, with the scaffold bucket's own share of charged calls printed above each bar. Exact-repeat memory accounts for 87.4 to 100.0 per cent of final-window suppression. The scaffold bucket suppresses no molecules in two of four runs. (b) The same share over eighths of each run. This share follows a U-shaped trajectory. It begins near total suppression, when the untrained policy repeats molecules and all scaffold buckets are unfilled, falls to between 28 and 60 per cent during exploration, then rises as the policy collapses onto a few molecules. The bucket's suppression is concentrated in mid-training, when the policy trajectory develops. Final-window suppression alone therefore cannot establish the bucket's contribution. Training without the bucket tests that contribution. The completed ablation below uses a different commanded budget, so its interpretation remains specific to that operating point.

C.2 Suppression by the two memory clauses

Scaffold memory applies two rules sequentially. Exact-repeat memory first zeroes molecules already scored in the run, bypassing the scaffold check. The second rule applies the scaffold bucket, which gives the mechanism its name. The decomposition applies the deployed scoring routine twice to the same episode stream, once with both clauses and once with the repeat clause alone (Figure C.1). The zeroed molecules in the second condition are a subset of those in the first by construction. Their difference measures scaffold-bucket suppression. Containment is checked for every run using the deployed routine.

This result challenges the proposed structural mechanism. If exact-repeat removal alone explains the arm's effect on reward support, the structural mechanism inherited from the 2020 method is unnecessary. That would provide a simpler and more useful explanation. The training trajectory in panel (b) prevents that conclusion here.

C.3 Training without the scaffold bucket

The replay above quantifies suppression, not its effect on learning. A clause inactive in the reported window may still have shaped the preceding trajectory. Training without the bucket separates these effects and requires only a small change to the scoring service, enabling the existing exact-repeat clause alone.

The repeat-only control at commanded budget two. The budget-eight experiment below does not directly address the cohort at commanded budget two. Three further pairs were therefore run at a commanded budget of two on L20 against the same three memory seeds. No decision threshold was registered, and the budget-eight threshold was not reused. The results are therefore descriptive, without a threshold-based decision.

seed	repeat-only	memory	difference	repeat spend	memory spend
3	0.6736	0.7044	−0.0308	1.87	1.59
4	0.7460	0.7071	+0.0389	3.41	1.87
5	0.6585	0.7649	−0.1063	1.92	3.16
median difference			−0.0308		

Repeat-only is lower in two pairs of three and the median difference is -0.031, which is small relative to the primary effect. Both arms lie in 0.659–0.765, while every run without either mechanism lies in 0.003–0.234. Two limitations constrain this comparison. Realised spend is not matched inside a pair, most visibly at seed 4 where repeat-only spends 3.41 calls an episode against its partner's 1.87, so the differences describe endpoints at the arms' actual costs. Three pairs with mixed signs also cannot establish equivalence in either direction.

Registered design. Four repeat-only cells against matched scaffold-bucket cells, paired by seed, on one machine, at a commanded budget of eight. The evaluation metric is the paired difference $d = \text{murcko} - \text{repeat}$ read on the median of four, with a threshold of 0.15. Below this threshold the repeat clause carries the mechanism, whereas at or above it with positive differences across all four pairs the bucket does. Both interpretive branches and the threshold were specified before the paired results were observed. An earlier recorded prediction favoured the opposite outcome, leaving both predictions testable by this comparison.

An assumption check. The matched arm reuses five existing budget-8 scaffold cells, avoiding twenty-five GPU-hours of repeat training. Reuse assumes comparability between runs conducted a week apart, an assumption that failed in two earlier comparisons. Therefore, one fresh scaffold cell was evaluated at a seed unused by the previous five runs. It reaches 0.7221 against their band of 0.5528 to 0.7825, inside it and near the median. This measurement supports the reuse.

seed	scaffold bucket	repeat only	d	spend (bucket)	spend (repeat)
6	0.7577	0.6280	+0.1297	6.06	7.69
7	0.5528	0.6662	−0.1134	7.69	5.57
8	0.7053	0.7203	−0.0150	6.20	10.05
9	0.7825	0.6570	+0.1255	6.07	8.06
median d			+0.0553	positive in 2 of 4	

The median difference is $+0.0553$, below the registered 0.15 decision threshold. The four differences have mixed signs. Under the registered rule this did not support a bucket contribution of the specified size. Equivalence and the absence of a training effect remain unestablished. Every run executed the requested filter. Each repeat-only cell spent more than sixty per cent of its partner's calls, satisfying the registered minimum-spend condition. Three spent more than their partner.

Scope. All cells in this comparison retain exact-repeat suppression. There is no unfiltered or IPS-only arm at budget eight, so this ablation cannot establish that repeat-only memory beats either. The replay motivating it used the historical budget-two cohort. The training ablation uses budget eight. Their operating points differ and the result does not transport bucket redundancy to budget two. The scaffold arm's observed run-to-run range is 0.2297 over six cells. Four paired seeds and a decision threshold are not a powered non-inferiority or equivalence design. The original threshold is kept after observing the results, and both negative paired differences are reported.

Implications. Persistent exact-repeat suppression can retain a high endpoint in these budget-eight runs without the scaffold bucket. Which clause is necessary at the main budget-two operating point remains unresolved by this experiment.

C.4 Molecules emitted by each arm

The other displays in this paper report summary statistics. Figure C.2 shows individual molecules so that the endpoint's counted objects can be inspected directly. Figure C.3 does the same for the four farthest-first pairs.

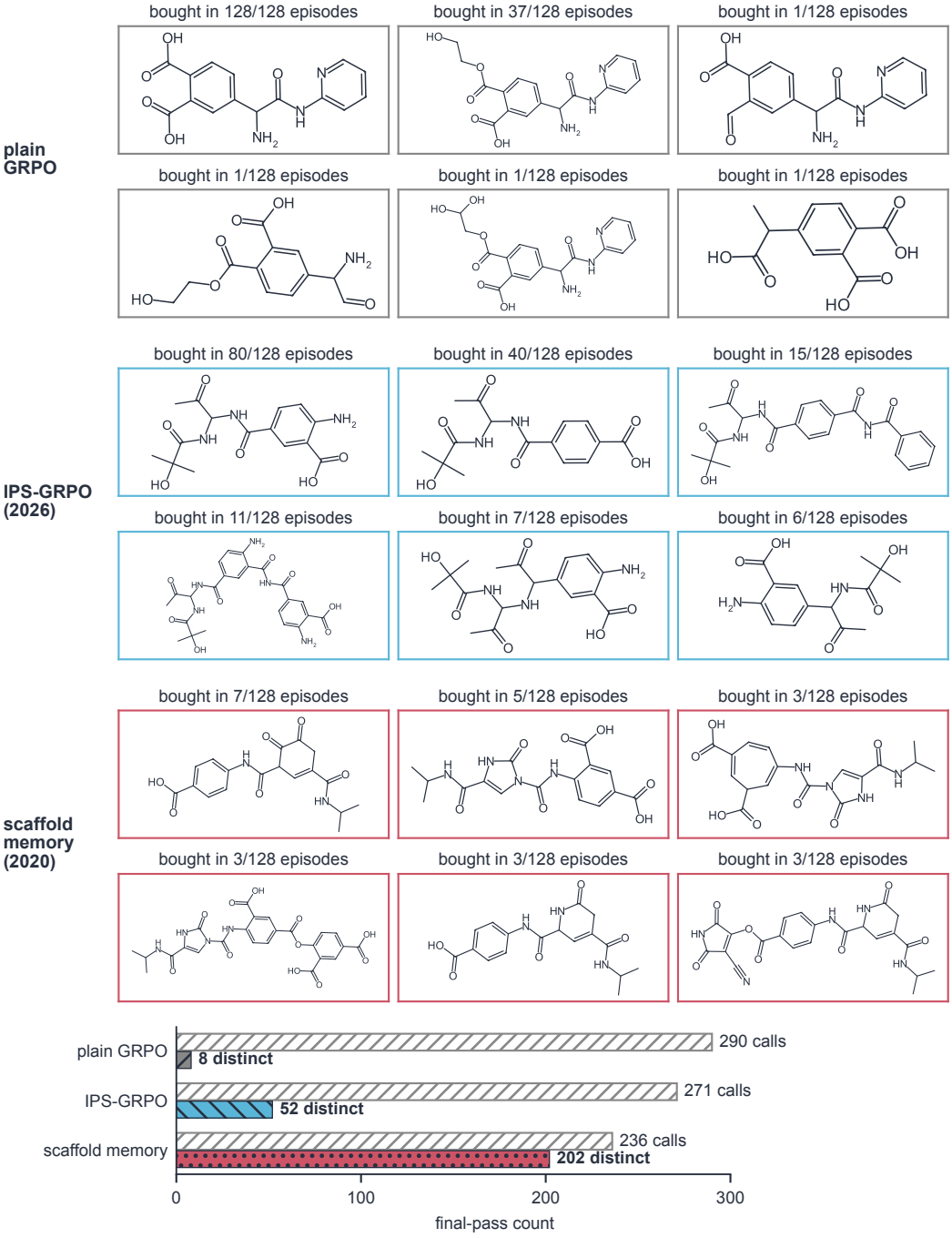


Figure C.2: **Individual molecular outputs show the collapse.** For one run per arm, the six molecules scored in the most episodes of the final validation pass are labelled by their frequency across 128 episodes. Below are the run's charged calls and distinct molecules in that pass. Plain GRPO uses 290 calls to find eight distinct molecules, scoring one of them in every episode. Inverse-probability scaling finds 52 distinct molecules for 271 calls, and its most-repeated molecule still appears in 80 episodes. Scaffold memory finds 202 for 236 calls, and its most-repeated molecule appears in seven. An episode that scores a molecule twice is charged twice but counted once here, so the labels count episodes, not calls. Structures are depicted from the recorded SMILES, each scaled to its own frame.

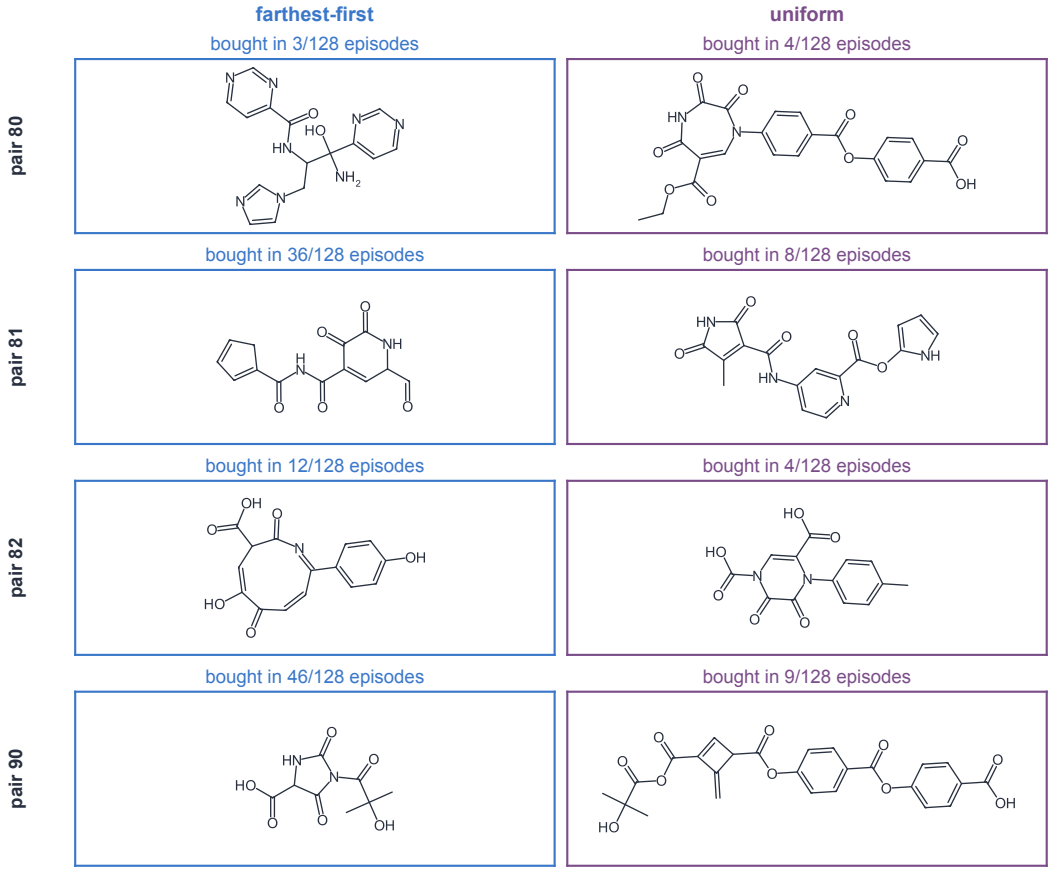


Figure C.3: **Farthest-first scores its most repeated qualifying compound in more episodes than uniform selection in three of four pairs.** For every cell of the four farthest-first pairs, the figure shows the qualifying compound purchased in the most episodes. Structures are drawn from recorded output at one common scale and labelled with the number of the 128 validation episodes that purchased each compound. Figure 3b uses the same counting rule.

C.5 Variety at a fixed subsample size

Deduplicating canonical SMILES rewards trivial variation, so an arm could in principle win the endpoint by emitting many near-identical molecules. Mean pairwise Tanimoto distance over fixed-size random subsamples removes the set-size confound. This measure differs from the filter's objective of limiting how many molecules share a scaffold, which does not maximise pairwise distance. On that quantity the filter's four runs give 0.735 to 0.792 against inverse-probability scaling's 0.473 to 0.668, again without overlap ($U = 24/24$, $p = 0.0048$).

C.6 Multiplicity against yield, by arm

D Robustness and scope

D.1 Threshold sensitivity and within-arm variation

We assess sensitivity to the qualification threshold and compare the between-arm gap with run-to-run variation.

The separation does not depend on the qualification threshold. Recomputing the endpoint across the original twenty-six completed runs and the expanded twenty-nine-run cohort gives the

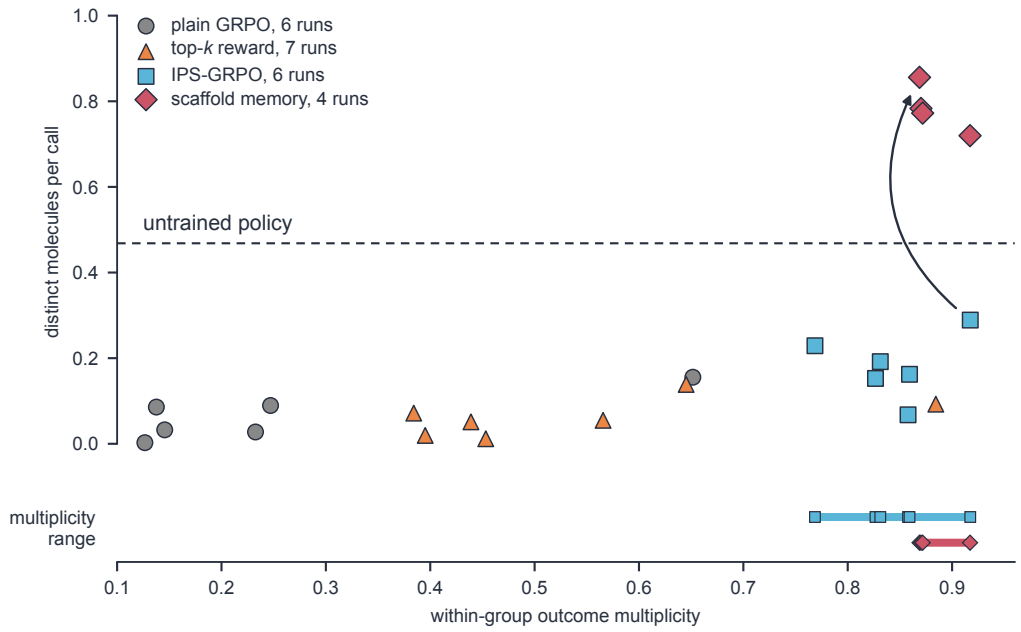


Figure C.4: **Within-group multiplicity does not order the endpoint across arms.** Each completed run is plotted by within-group outcome multiplicity, the quantity addressed by the 2026 corrections, and the common evaluation endpoint. The axes are computed separately from the trainer's per-update logs and the molecule logs, without reference to each other. The figure includes two populations differing by three runs. The rank correlation of $+0.800$ is computed across all twenty-six runs carrying both readings. The per-arm counts in the legend and the arm ranges below total twenty-three. The three cells whose filter never executed enter the correlation as control runs but are excluded from the arm ranges. Assigning them to the filter arm would misclassify untreated runs. Inverse-probability scaling preserves multiplicity most consistently, with all six runs at or above 0.769, compared with plain GRPO's median of 0.19. Its endpoint nevertheless remains below the untrained policy's. The lower panel displays each arm's multiplicity range, where the filter's range lies inside that of the correction. No correction run falls within the filter's interval and the filter's median is higher. The ranges and highlighted pair show that multiplicity does not monotonically order yield across these arms. They do not isolate a causal mechanism. The arrow compares individual runs without aggregation. The correction run has the highest multiplicity of either arm at 0.9174 and yields 0.289. The *lowest*-multiplicity filter run has multiplicity 0.8688 and yields 0.856. The three runs held out of the arm ranges are the three whose filter never ran (Appendix A.5).

same separation at all ten existing thresholds. The table below uses the expanded cohort consisting of seven memory runs and twenty-two others.

threshold	filter minimum	others maximum	separated	U	below own start
0.000	0.7197	0.2890	yes	154/154	22/22
0.100	0.7197	0.2727	yes	154/154	20/22
0.200	0.7197	0.2662	yes	154/154	17/22
0.300	0.7197	0.2630	yes	154/154	15/22
0.400	0.7155	0.2565	yes	154/154	15/22
0.500	0.7113	0.2435	yes	154/154	13/22
0.567	0.6840	0.2338	yes	154/154	8/22
0.600	0.6229	0.2273	yes	154/154	6/22
0.650	0.4703	0.2143	yes	154/154	3/22
0.700	0.2331	0.1818	yes	154/154	2/22

In the original twenty-six-run cohort, the two groups are disjoint at every threshold, and every one of the eighty-eight pairwise comparisons runs the same way at all ten. The calibrated threshold of

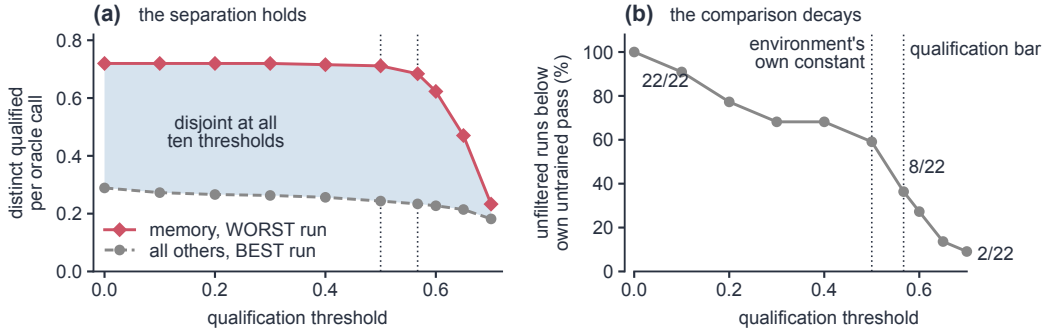


Figure D.1: **Arm separation persists across thresholds, while the comparison with untrained policies changes.** Both panels retain the original twenty-six-run cohort. (a) The filter arm's *worst* run against every other arm's *best* run, at ten qualification thresholds. The two groups never touch in either population, with all eighty-eight pairwise comparisons in the original twenty-six-run cohort, and all 154 in the expanded twenty-nine-run cohort, exhibiting consistent separation at every threshold. The shaded gap narrows from 0.431 to 0.051 as the threshold rises but remains positive. (b) The share of the twenty-two unfiltered runs ending below their own untrained validation pass. Counting distinct molecules at any score gives all twenty-two. At the calibrated threshold, the count is eight, falling to two at 0.700. The dotted rules mark the environment's own success constant of 0.500, which was fixed in the run configuration before any of this was measured, and the registered qualification threshold of 0.567. Both panels and the threshold table in Appendix D.1 use the same estimator.

0.567 is neither the most nor the least favourable. The gap narrows from 0.431 at threshold zero to 0.051 at 0.700, where few molecules qualify. Both values are reported in the table.

The last column reports a separate, threshold-dependent comparison. Counting every distinct molecule however poorly it scores, all twenty-two unfiltered runs end below their own untrained pass. That count falls monotonically as the threshold rises, to eight of twenty-two at the calibrated threshold and two at 0.700. The apparent decline from training depends on molecules the task does not count as successes. The between-arm separation persists across every threshold tested. The between-arm separation and the threshold-dependent comparison with the untrained policy are both reported.

The gap is nearly eight times the widest within-arm spread. The endpoint at the registered threshold is tabulated below for the expanded twenty-nine-run cohort. The following within-arm spread comparison retains the original twenty-six-run calculation.

arm	n	min	max	median	sd	sd / median
scaffold memory	7	0.6840	0.8647	0.7373	0.0677	0.09
IPS-GRPO	6	0.0639	0.2338	0.1695	0.0591	0.35
top-k reward	7	0.0117	0.1119	0.0548	0.0335	0.61
plain GRPO	9	0.0026	0.1555	0.0579	0.0491	0.85

In the original twenty-six-run cohort, the worst filtered run reaches 0.7044 and the best unfiltered run reaches 0.2338, a gap of 0.4707 against a widest within-arm standard deviation of 0.0591. A gap smaller than background variation is uninformative. Here, the gap is nearly eight times that variation. The filtered arm is also the steadiest of the four by a wide margin, at a relative spread of 0.04 against 0.35 to 0.85. These statistics describe the observed runs, not memory's failure probability. The cohort was not sampled from a defined population, and seven runs cannot estimate a failure rate.

D.2 VALIDATION TRAJECTORIES AND UNTRAINED REFERENCES

A complete run records 128 validation episodes before training, followed by five cycles of 640 training episodes and 128 validation episodes. The endpoint reported everywhere else is the sixth and final of those passes, not a tail slice of training. The other five passes provide the trajectory and a within-run untrained reference. Pass boundaries are identified from logged block sizes, with 64 episodes in validation blocks and 128 in training blocks. All twenty-six completed runs matched this layout. The expanded analysis also checks all twenty-nine complete runs. The trajectory table below includes the three confirmation re-executions.

arm	n	update 0	5	10	15	20	25
scaffold memory	7	0.0433	0.1393	0.2898	0.5915	0.7217	0.7373
IPS-GRPO	6	0.0617	0.2794	0.2500	0.2247	0.2116	0.1695
top-k reward	7	0.0617	0.2398	0.2050	0.0840	0.0821	0.0548
plain GRPO	9	0.0617	0.2440	0.1397	0.0383	0.0321	0.0579

The table gives expanded-cohort medians. The following trajectory description retains the original twenty-six-run medians. Only the filter arm shows a substantial rise, from 0.043 to 0.722. The corrections peak at update 5 and then decline, most slowly for inverse-probability scaling and fastest for plain GRPO. The three unfiltered arms start together at 0.0617 because at update 0 they represent the identical untrained policy. Fifteen of their cells, along with two of the three cells whose filter never executed, exhibit precisely that value. The filtered arm starts lower, at 0.0433, with three of its four cells at exactly that value, although its policy is the same untrained one. The filter is updated by every scored molecule, validation passes included, and marks a zeroed molecule in the agent's observation, so it acts in context before any update. The endpoint counts raw scores, so the difference lies in what the agent proposed after seeing those marks. The same in-context channel is open at the final pass, and we did not separate its share of memory's final-pass gain from the share the trained policy carries.

Each arm at its own best pass. The comparison above uses a fixed update, after the corrections have peaked and declined (table above). A best-of-six-passes analysis assesses whether the fixed endpoint inflates the gap. This analysis favours the corrections and has little effect on memory, as six of the seven memory cells' trajectories are monotonically increasing, so their best pass is update 25 and the reduction changes nothing for them. The seventh, one of the three confirmation cells re-executed on a verified service, peaks at update 20, 0.901 against 0.684 at update 25, and this increase leaves it well within the separated range. Every correction run peaks well before update 25 and receives its hindsight-optimal endpoint.

arm	n	update-25 median	best-pass median	best-pass ceiling
scaffold memory	7	0.7373	0.7649	0.9012
IPS-GRPO	6	0.1695	0.3273	0.3571
top-k reward	7	0.0548	0.2551	0.3309
plain GRPO	9	0.0579	0.2704	0.3363

Update-25 medians are repeated for comparison with best-pass medians and each arm's best-pass maximum. The best pass of any non-memory run, 0.357, still falls short of the lowest best pass among the memory runs, 0.704, and the rank separation stays complete at $U = 154/154$, unchanged from update 25. This hindsight-based checkpoint selection is unavailable in deployment because it uses the same six evaluation passes, not an independent criterion. It gives the corrections an upper bound, unnecessary for memory, that still leaves the gap open.

Change from each run's untrained pass. Every run records a validation pass before training, providing its own untrained reference. The raw and qualified readings differ. For the original twenty-six-run cohort, counting every distinct molecule however poorly it scores, all twenty-two unfiltered runs end below where they began and all four filtered runs end above. At the calibrated qualification threshold, all four filtered runs still end above their untrained references, and the unfiltered count falls to eight of twenty-two. Specifically, the median ratio of final pass to own pass 0 is 16.6 for the filter, 2.75 for inverse-probability scaling, 1.16 for the top-k reward and 0.94 for plain GRPO. Thus

both corrections improve on the untrained policy when qualification is required, but by far less than the filter. The between-arm ordering remains unchanged.

D.3 Descriptive stratification of the historical association

This post-hoc sensitivity analysis retains all 26 historical runs with both logged multiplicity and a complete final validation pass. Three additional memory re-executions enter the expanded 29-run endpoint cohort, but lack the corresponding training-log multiplicity summary and do not enter this correlation analysis. The pooled Spearman correlation remains 0.800. Stratification is structured by executed intervention, such that the three requested memory runs whose filter never executed join plain GRPO. Their original requested labels remain available in the sensitivity analysis below.

Executed arm	retained runs	within-arm Spearman ρ
Scaffold memory	4	-1.000
IPS-GRPO	6	$+0.143$
Top-k reward	7	$+0.500$
Plain GRPO, including three non-executed filters	9	$+0.750$

These small strata describe the observed runs and do not identify a within-arm mechanism. Omitting each executed arm in turn leaves pooled correlations between 0.729 and 0.863. Removing the three non-executed filter runs gives 0.786 over 23 runs. These exploratory analyses have no additional significance tests. No observed IPS run lies within memory's multiplicity interval of $[0.8688, 0.9172]$, yielding a count of zero among six runs. Range containment therefore provides neither matched multiplicity observations nor a causal effect at fixed multiplicity.

Population and checkpoint sensitivity. At $\tau = 0.567$, the original 26-run executed-arm comparison gives $U = 88/88$, and the expanded cohort gives $154/154$, both at update 25 and under the existing best-of-six-pass read. The final-pass minimum gaps are 0.4707 and 0.4502, respectively. The best-pass minimum gap is 0.3473 for both. Under the original *requested* labels, the 26-run comparison instead gives $96/133$ at the final pass and $110/133$ at the best pass, with overlapping ranges. All ten existing thresholds retain separation under the executed labels in each cohort. None does under the original requested labels. This sensitivity analysis shows how correcting the service assignments affects the main result and keeps the three untreated runs distinct from executed memory interventions.

Deleting all runs with each seed label in turn, for every label 0–6, preserves final-pass separation in both cohorts. This represents an influence analysis only, since identical seed labels across historical jobs do not establish matched machine–seed blocks. The minimum gap across these deletions ranges from 0.4707 to 0.5021 for the original cohort and from 0.4502 to 0.4817 for the expanded cohort. Runs supply the plotted observations. The six validation passes are repeated measurements within each run, and seed labels do not by themselves establish independent experimental units. Execution-history and service confounding prevent these analyses from replacing a balanced matched experiment or establishing an effective independent sample size. They also do not separate exact-repeat suppression from the scaffold-bucket mechanism.

D.4 Registered analysis and balanced extension on the second target

The original registered analysis measures raw novelty at threshold zero in two complete blocks, L20 seed 3 and RTX PRO 6000 seed 5. We keep this six-run population separate from the later extension. The extension adds L20 seed 6 and RTX PRO 6000 seed 7, and later the control run of L20 seed 4, giving five complete machine–seed blocks (three on L20, two on RTX PRO 6000) and five runs per arm. An automatic completeness check initially excluded L20 control seed 4 because it failed to find the objective name in the log. The retained log and configuration identify Perindopril MPO throughout the run, so it is included. All 15 completed Perindopril runs appear in Table D.2, each with its role.

The secondary qualification analysis uses the untrained policy's pooled-budget ninetieth percentile, computed from all available pre-existing baseline budget files. Its threshold is 0.142 for Perindopril and 0.567 for Osimertinib. The denominator for each target-relative ratio is the untrained *budget-*

Table D.1: Qualified discovery per charged call for each Osimertinib run at that target's pooled-baseline ninetieth percentile (0.567), in the final validation pass.

Osimertinib arm	seed	q90 endpoint	Osimertinib arm	seed	q90 endpoint
plain GRPO	0	0.0786	scaffold memory (2020)	5	0.7649
plain GRPO	1	0.0276	scaffold memory (2020)	6	0.7373
plain GRPO	2	0.0195	IPS-GRPO (2026)	0	0.1450
plain GRPO	4	0.0649	IPS-GRPO (2026)	1	0.1882
plain GRPO	5	0.1555	IPS-GRPO (2026)	2	0.1507
plain GRPO	6	0.0026	IPS-GRPO (2026)	3	0.0639
scaffold memory (2020)	3	0.7044	IPS-GRPO (2026)	5	0.2023
scaffold memory (2020)	4	0.7071	IPS-GRPO (2026)	6	0.2338

Table D.2: All 15 completed Perindopril cells with their role, arm and machine. Raw is discovery per charged call at threshold zero, and q90 is qualified discovery per charged call at the pooled-baseline ninetieth percentile (0.142).

seed	role	arm	machine	raw	q90
3	primary + ext.	GRPO	L20	0.1257	0.0963
3	primary + ext.	memory	L20	0.9027	0.7704
3	primary + ext.	IPS	L20	0.1115	0.1077
6	extension	GRPO	L20	0.0578	0.0325
6	extension	memory	L20	0.6944	0.6786
6	extension	IPS	L20	0.1993	0.1595
5	primary + ext.	GRPO	RTX PRO 6000	0.0109	0.0109
5	primary + ext.	memory	RTX PRO 6000	0.6816	0.5265
5	primary + ext.	IPS	RTX PRO 6000	0.1108	0.1082
7	extension	GRPO	RTX PRO 6000	0.0078	0.0078
7	extension	memory	RTX PRO 6000	0.9098	0.8798
7	extension	IPS	RTX PRO 6000	0.2212	0.2118
4	extension	GRPO	L20	0.0472	0.0472
4	extension	memory	L20	0.6914	0.5465
4	extension	IPS	L20	0.3387	0.3226

two endpoint at that threshold, not the pooled endpoint. This is a post-hoc target calibration, distinct from the registered threshold-zero analysis. The calibration remains unchanged after inclusion of the completed extension runs.

In the balanced extension, median qualified endpoints are 0.0325 for control, 0.6786 for memory and 0.1595 for IPS. Dividing by the untrained reference gives 0.47, 9.91 and 2.33. The memory/IPS median ratio is 4.26. Memory ranks first within each machine. Blocking limits the machine-mix explanation but does not equate the machines' training configurations or resolve uncertainty from only five pairs. Older runs without retained target configurations are labelled unverified.

Every endpoint uses the final validation pass, records 3840:3968, after checking the complete 3968-record layout. Earlier validation passes and training records do not enter its numerator or denominator. The same extraction procedure yields Tables D.1 and D.2.

D.5 Pre-registered replication at a second model size

Registration. The design, the criterion, a floor and both branches were recorded before any Qwen3-1.7B run existed. The first run was launched eleven hours later and produced its endpoint fourteen hours after the registration. The replication changes the model and nothing else. It uses three arms (Bemis–Murcko memory, plain GRPO and IPS), seeds 0, 1 and 2, the cohort's commanded budget of two calls with the same soft price, 25 updates and one machine for all nine runs. Each run reached one scoring service, which was checked before the first run against a reference score and a six-molecule test batch on which the filter must suppress scores. Each run's recorded configuration confirms that its model, filter, IPS switch and budget match its assigned arm before its endpoint is reported.

The floor. A smaller policy might produce too few qualifying molecules for an informative comparison. Such a result would be inconclusive, not negative. The registration therefore required the untrained policy, read from the first validation pass of the first run, to reach a qualified rate of 0.02. The observed rate is 0.0769.

The criterion. The statistic is rank separation, not an absolute level, because a smaller model is expected to produce lower rates everywhere. With three memory runs against six others, U ranges from 0 to 18. Complete separation, $U = 18$, replicates the cohort. $15 \leq U \leq 17$ was registered as ordering without disjointness and nothing more. $U < 15$ does not replicate.

Outcome. $U = 17/18$, with exact one-sided probability $P(U \geq 17) = 2/84 = 0.024$ over all labellings of the nine observed rates. This meets the registered middle branch. Memory endpoints span 0.461–0.951, and the others span 0.125–0.611. The single inversion places memory seed 0 below IPS seed 2. At this model size, the 4B cohort's ordering is reproduced, but complete separation is not.

Unregistered description. The commanded budget barely binds at this size. The untrained policy spends about half a call per episode, and eight of the nine trained runs spend at most one against the commanded two. The exception is memory seed 0 at 3.66. It found 387 distinct molecules in the final pass, more than the other two memory runs combined, and it is the run that falls below an unfiltered one on the per-call endpoint. This result does not change the registered branch. Plain GRPO's final pass contains 16 to 20 distinct molecules across 128 calls, one per episode, compared with 37 molecules in 64 calls in its untrained window. Five of the six runs without memory end below their own untrained raw distinct-per-call, and the exception is IPS seed 2. The untrained windows of the nine runs take only three distinct values, so they are not nine independent draws of the untrained policy.

arm	seed	qualified / calls	rate	calls / episode	distinct	untrained rate
memory (2020)	0	216 / 469	0.4606	3.66	387	0.0769
memory (2020)	1	116 / 122	0.9508	0.95	117	0.0615
memory (2020)	2	99 / 107	0.9252	0.84	101	0.0615
plain GRPO	0	16 / 128	0.1250	1.00	16	0.0625
plain GRPO	1	20 / 128	0.1562	1.00	20	0.0625
plain GRPO	2	17 / 128	0.1328	1.00	17	0.0625
IPS	0	53 / 128	0.4141	1.00	54	0.0625
IPS	1	45 / 127	0.3543	0.99	47	0.0625
IPS	2	77 / 126	0.6111	0.98	78	0.0625

D.6 Baseline parity and screening

The leading mechanism dates from 2020. We screened its setting to assess whether the inherited default gave it a favourable comparison.

The bucket size of the filter. Scaffold memory has a bucket size, the number of molecules allowed to share a scaffold before further instances earn zero. The incumbent uses twenty-five. Five, ten and fifty were screened at the operating point against the eighteen incumbent cells. The criterion was fixed before any treated number was read. An alternative was promoted only if it exceeded the control median by more than the control arm's own dispersion and also lay above the spend-adjusted line, a line of endpoint against spend fitted to the control runs before any treated run was read.

bucket	spend	endpoint	vs. control median 0.5742	vs. spend-adjusted line
5	6.35	0.5797	$+0.0055$	-0.0563
10	5.46	0.6852	$+0.1109$	$+0.0227$
25 (incumbent)	—	0.5742 median of 18	—	—

Neither alternative clears the control arm's own dispersion of 0.1344, and neither clears the spend-adjusted threshold. Bucket ten falls just short, and its improvement is reported without rounding to zero. The increase of $+0.1109$ represents measurable movement inside a floor of 0.1344, and one cell cannot separate it from the incumbent's spread. Bucket fifty failed twice on the same machine, at 128 and at 2432 episodes of 3968, and has no result. The screen finds no material headroom, so bucket twenty-five is reported as the best setting this screen observed, not as a tuned optimum. Two single-cell alternatives cannot establish that no better bucket exists. Both also ran at a commanded

budget of eight rather than the cohort's two, which is why their realised spend, 5.46 and 6.35 calls an episode, sits well above the cohort's median.

The constant of the top-k reward. The top-k reward is this paper's own construction, unlike IPS-GRPO and the selection rules. Its $k = 3$ setting therefore has no corresponding published constant. It was screened at the same operating point against the same criterion, under which an alternative is promoted only if it exceeds the incumbent's own median by more than the incumbent's own dispersion.

k	spend	endpoint	vs. control median 0.0548
2	5.09	0.0338	-0.0210
5	4.99	0.0188	-0.0361
8	15.49	0.0131	-0.0417
3 (incumbent)	7.61 median	0.0548 median of 7	—

All three alternatives, $k = 2$, 5 and 8, fall below the incumbent's seven-run median of 0.0548. Two of the three fall below it by more than the incumbent's own population standard deviation of 0.0310. The third, $k = 2$, falls short by 0.0210, which is less than that spread, so not all three clear it. A population standard deviation measures spread, not statistical significance. The screen therefore supports the common direction across all three settings, not an individual margin test. Realised spend varies widely across the three (4.99 to 15.49 calls an episode) with no accompanying gain, which is the same direction the historical cohort's own budget association points. The screen finds no material headroom here either, so $k = 3$ is reported as the best of the settings tried.

Unequal search across the compared arms. The arms did not all receive the same hyperparameter search. The filter's bucket size and the top-k reward's k are the two free constants this work constructs itself, and both were searched, over 5, 10 and 50 against a published 25 for the former and over 2, 5 and 8 against this paper's own 3 for the latter, with the existing default winning both times. IPS-GRPO's estimator and the selection rules' fingerprint choice remain at the values their own papers specify and were not searched here. The unequal search biases the comparison against IPS-GRPO and the selection rules. The screens rule out only a narrow set of alternatives. Specifically, memory's advantage does not come from a bucket chosen after screening, since no alternative was adopted. They do not exclude that the shared recipe suits some objectives better than others, or that IPS and the selectors would gain from a search they did not receive, so the comparisons in this paper are between default settings under one recipe. Within the four-arm comparison every cell shares one recipe exactly, differing only in the reward mode, the diversity filter and its two constants. The full key-by-key configuration comparison is in Appendix F.2.

Standing on the PMO benchmark. On PMO's (Gao et al., 2022) area under the top-ten curve at ten thousand oracle calls, the filtered arm reaches 0.750. The strongest published method reaches 0.837 and the weakest of the fourteen listed reaches 0.674. The filtered arm ranks below ten of them. The full range shows variation across published methods.

Performance on that measure and ours need not conflict. The budgets differ in two ways. The per-episode command, two calls, is three orders of magnitude smaller than PMO's ten-thousand-call total. The filtered arm's own total training spend, at a median 1.94 calls across 3200 episodes, is on the order of six thousand calls, comparable in scale to PMO's total rather than three orders of magnitude smaller. At that budget the statistic is maximised by finding ten good molecules early and exploiting them, which is the opposite of what a tight per-episode budget rewards. A method can therefore rank mid-table on PMO yet exceed its own untrained policy on raw novelty in every run here. Among the historical cohort's twenty-nine runs, only the filtered arm meets that condition.

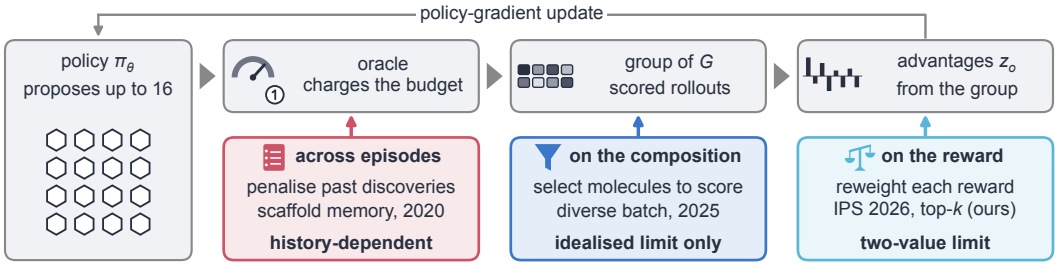


Figure E.1: **Each class of correction acts at a different stage of an update.** The policy proposes, the oracle charges the budget, the group is scored and its advantages drive the update. Scaffold memory (Blaschke et al., 2020) penalises past discoveries across episodes, before a rollout enters the group. Our adaptation of diverse mini-batch selection (Svensson et al., 2025) chooses which proposed molecules are scored within a turn, IPS (Sinha et al., 2026) reweights rollout rewards, and our top-k rule changes the episode objective. Section 5 gives conceptual limits, not a theorem for each deployed intervention. Proposition 1 concerns composition in an idealised bandit flow, and the two-value bound concerns rewards. Neither excludes memory, which uses history unavailable to a group calculation. Glyph counts are schematic.

E Theory

E.1 Proof of Proposition 1

With p a softmax over outcome logits z, fixed rewards r, group weights $\sum_o w_o = 1$, $A_o = r_o - \bar{r}_w$ and $\bar{r}_w = \sum_o w_o r_o$, the centred flow the main text refers to is

$$\dot{z}_i = \sum_o w_o A_o(\delta_{io} - p_i) = w_i(r_i - \bar{r}_w). \tag{E.1}$$

Proposition 1 (restatement). *The logit vector is stationary exactly when every outcome with positive weight has reward $\bar{r}_w$. If every weight is positive and rewards are not constant, the maximal-reward logit grows and the minimal-reward logit shrinks.*

The first display follows from $\nabla_{z_i} \log p_o = \delta_{io} - p_i$ and centring. Stationarity then reads $w_i(r_i - \bar{r}_w) = 0$ for every i. For the last part, $\bar{r}_w$ is a convex combination of the r_o with strictly positive weights, so $r_{\min} < \bar{r}_w < r_{\max}$ strictly, and the two signs follow. □

E.2 Scope of Proposition 1

Reweighting can change both the speed and direction of an outcome's logit drift, which defines the proposition's scope. Since $\bar{r}_w$ depends on w, the sign of $r_i - \bar{r}_w$ for an *intermediate* outcome can flip. Sampling two weightings at random reversed the middle outcome's direction in 1668 of 5000 trials, 33%. The main-text result concerns only the two extreme outcomes, whose signs are fixed by the convex-combination bound.

Divergence is not established. The proposition does not establish that the gap between the extreme outcomes grows without bound or that the lowest-reward outcome's probability tends to zero. The derivation establishes the two *signs* only. Positive derivatives at every finite time do not imply a divergent integral when w moves. On two outcomes with $r = (1, 0)$ the gap obeys $\dot{z}_{\max} - \dot{z}_{\min} = 2w_1(1 - w_1)$, so taking $w_1(t) = e^{-t}/2$ keeps every weight strictly positive at every finite time while the gap converges to $\int_0^\infty e^{-t}(1 - e^{-t}/2)\,\mathrm{d}t = 3/4$. Divergence needs w bounded *away* from zero uniformly in t, and in a run w is the group's empirical composition, which concentrates as the observed collapse proceeds. The stronger hypothesis needed for divergence therefore fails in the studied regime. Section 5 uses only the fixed-point characterisation, namely that a composition-side intervention cannot create an interior fixed point.

Four differences between the trained objective and the model. The proposition is a statement about a softmax over m outcomes under an unconstrained gradient flow. The trained objective differs from this model. We state each difference's consequence and report measurements where available.

Standardisation. The deployed advantage is $(s-\mu)/(\sigma+\epsilon)$ over a group, not the w-centred $r_i - \bar{r}_w$ above. The effect of standardisation can be determined exactly. An outcome appearing k_o times in a group of n contributes k_o/n to the mean, so μ *is* $\bar{r}_w$ with w the empirical composition, which motivates the analysis in terms of w. Dividing by $\sigma + \epsilon > 0$ rescales the whole drift by one positive scalar, preserving every sign and stationary point. The proposition therefore describes the standardised system as well as the centred one.

Clipping. The trained loss is dual-clipped, which the flow does not model and which can alter stationary points. We measure its prevalence but cannot bound its gradient contribution from a clipped-token fraction. Across 683 update steps in the 28 cells carrying per-update traces, that fraction has a median of 0.016 and never exceeds 0.029 at any observed step. No individual cell median exceeds 0.05. Even at the worst observed step, the remaining 97.1% of tokens do not establish equivalence to the outcome-logit flow.

A bandit, not a sequence model. The drift is the softmax policy-gradient flow specialised to a group-weighted bandit, and its global-convergence behaviour is standard (Mei et al., 2020). The trained policy is autoregressive over tokens and the map from logits to outcome probabilities is not one softmax. The result is limited to the bandit model. Its token-level extension remains open.

Entropy bonus but no KL term. No cell carries a KL term on the loss or in the reward. The trained loss does include an entropy bonus with coefficient 0.001 (Appendix F.2). Equation E.1 omits that bonus, so it does not characterise the stationary points of the deployed regularised objective.

The IPS floor and the determinantal kernel. The implementation floors the empirical frequency at $\epsilon = 0.001$ rather than adding it to the count, and the smallest frequency a group of eight can produce is $1/8$, so the floor never binds and the weight is exactly n/k_o at every count these runs contain. The mean IPS factor equals n times group outcome multiplicity to printed precision at every logged step of every IPS run. This is the implementation check required before endpoint inspection in the memory-plus-IPS confirmation (Appendix A.2). The constant therefore has no detectable effect here, and it was not tuned. The IPS weight itself is not clipped. The determinantal kernel's similarity, $L_{ij} = 1 - \text{dist}(i,j)$, is the average of a Tanimoto similarity and a Dice similarity, each positive semi-definite on binary fingerprints. An average of positive semi-definite kernels is positive semi-definite, so L is a valid DPP kernel and $L_{ii} = 1$ because a molecule is identical to itself.

E.3 Enumeration of the two-value trichotomy

Adjacent to the boundary at which inverse-probability scaling annihilates a two-value group, the standardised group has a spread at the floating-point floor, so the signs are decided by rounding. We verified the trichotomy, the boundary condition and the rounding neighbourhood by complete enumeration at group size $G = 8$. Of these cases, 168 preserve the ordering, 26 reverse it and 2 annihilate the group. The enumeration is exhaustive in the counts and sampled in the rewards, running over all seven count splits and a six-by-six grid of reward pairs, 196 cases after the ordering exclusion. This enumeration tests the result. The analytical derivation in the text covers the full parameter space. Specifically, the boundary reduces to the single comparison between a/b and n_a/n_b, so the computational grid checks, but does not replace, the algebra.

The exact magnitude invariance uses ideal standardisation without an additive stabiliser. With the deployed $\sigma+\epsilon$ denominator, changing the reward gap also changes its size relative to ϵ, so advantage magnitudes need not be exactly invariant. The preserve/reverse/tie ordering remains the relevant qualitative distinction away from numerical rounding. We do not claim an exact token-level gradient identity from this ideal calculation.

F IMPLEMENTATION AND REPRODUCIBILITY

F.1 IMPLEMENTATION SPECIFICATION

The released code is the authoritative record. This subsection consolidates the choices required for reimplementation, including details not specified elsewhere. All are taken from the code and the resolved configurations of the reported runs.

Training and generation. Qwen3-4B is trained under FSDP, with vLLM generating bfloat16 rollouts. Gradient checkpointing is enabled. Tensor parallelism uses degree one on the RTX PRO 6000 machine and two on the L20 machine. The advantage estimator is group-relative, standardised within the group, with $\gamma = 1$ and $\lambda = 1$. A group is eight rollouts of one prompt, a training batch is sixteen episodes, the PPO mini-batch is the same sixteen and one PPO epoch runs per batch, so each batch is one gradient step, at a micro-batch of two per device. The optimiser is AdamW at a constant learning rate $1\text{e}{-6}$ with weight decay 0.01, no warmup and gradient clipping at 1.0. The policy loss is dual-clipped, 0.2 on both sides with the dual-clip constant 3.0, aggregated token-mean. An entropy bonus with coefficient 0.001 is subtracted from the policy loss. There is no KL term on the loss or in the reward. Training rollouts are sampled at temperature 1.0 with nucleus sampling disabled (top-p 1.0, top-k off) and one sample per prompt. Validation passes sample at top-p 0.95 at the same temperature, a deliberate difference from training. Prompts are capped at 4096 tokens and responses at 512. An episode allows at most four turns and at most sixteen proposals a turn. The validation batch is sixty-four episodes, which is why a 128-episode pass is written as two blocks.

Selection rules. A turn's candidate list contains the valid parsed proposals in policy-generated order. All candidates are scored when a turn contains at most k valid proposals. Otherwise the rule picks k, and only those k are sent to the oracle and charged. Candidates are not deduplicated first, so a molecule proposed twice in one turn can be selected twice and charged twice, and the count of within-turn duplicates is recorded.

Farthest-first (MaxMin-A) starts from the first valid candidate, where the published rule draws it at random (Appendix B.1), then repeatedly adds the candidate whose minimum distance to the picked set is largest, breaking ties by the earlier position. The distance is the average of one minus the Tanimoto similarity on a Morgan fingerprint (radius 2, 2048 bits) and one minus the Dice similarity on the atom-pair fingerprint of the Bemis–Murcko scaffold core, falling back to the whole molecule when the core has no atoms.

Uniform draws k of the same candidates without replacement.

Determinantal enumerates every k-subset, scores it by the determinant of the similarity submatrix (L_{ij} one minus that distance, $L_{ii} = 1$), and draws one subset with probability proportional to that determinant. It falls back to the first k when every determinant is zero, which happens only when the pool is one molecule repeated, and rejects requests above $100{,}000$ subsets without approximate sampling. Both stochastic rules seed a local generator from the ordered candidate tuple's hash. Within one service process, the same tuple therefore returns the same subset, not an independent draw at each request. The process hash seed was not retained, so a restart need not reproduce the subset. Full candidate tuples and their recurrence were not logged (Appendix B.9).

Objectives and weighting by arm. Plain GRPO's episode reward is the best raw score. The top-k reward sums the three highest scores among distinct canonical molecules and divides by a fixed three. The memory arm's reward is the filtered score, the raw score times a multiplier that is zero for a suppressed molecule and one otherwise. Inverse-probability scaling is applied within the trainer, where each trajectory's token-level reward, after the zero budget penalty in every cell here, is multiplied by $1/\max(\hat{p}, \epsilon)$, where $\hat{p}$ is the frequency of that trajectory's episode reward within its group of eight and $\epsilon = 0.001$. Rewards are rounded to six decimals before frequencies are counted, and that rounding defines what counts as the same outcome. Group standardisation follows reward scaling. The deployed objective is $\mathrm{standardise}(w\, r)$ rather than $w\, \mathrm{standardise}(r)$. These are different algorithms, and the reported experiments use the former. The use of episode reward as outcome identity is supported by an empirical check. Across 768 recorded episodes, every distinct best score corresponded to exactly one distinct canonical molecule, a property that holds because these molecular property objectives are continuous.

Memory state and updates. The memory is scoped to a run identifier, so a service shared by concurrent runs cannot leak one run's state into another, and it is never reset mid-run. Exact-repeat suppression runs first and applies to any molecule the run has already been charged for, whatever it scored. The scaffold bucket operates subsequently and applies only to molecules meeting or exceeding the minimum score threshold of 0.4. The scaffold count is incremented and subsequently evaluated, so a bucket admits twenty-five molecules without penalty while the twenty-sixth proposal is the first to be zeroed. Every molecule the oracle scores updates this state, in validation passes as well as in training, and a zeroed molecule is marked in the agent's observation. Qualification is computed from raw scores, so the filter changes what the agent is shown and never what the endpoint counts.

Measured quantities. The endpoint follows Equation 1 on the final 128-episode validation pass, defined as the number of distinct canonical molecules whose best raw score reaches τ, divided by the charged oracle calls during that pass. The threshold τ is the 0.90 quantile, by linear interpolation between order statistics, of the per-molecule best scores pooled over the untrained policy's baseline runs at every measured budget, giving 0.567 on the primary target and 0.142 on the second. Within-episode spread is the mean pairwise selector distance among the distinct molecules scored in an episode, averaged over the episodes of the final pass that scored at least two, and the number of episodes used and skipped is printed for every cell, here 121 to 128 of 128. Within-group outcome multiplicity is the number of distinct episode rewards, rounded to six decimals, divided by the group size, averaged over an update's groups, and it is logged by the trainer for every arm whether or not the correction is on. It cannot be recovered from the molecule logs, which omit group membership. Fixed-size structural diversity is the mean pairwise Tanimoto distance on Morgan fingerprints over 200 random subsamples of twenty distinct molecules under one fixed seed. Common-call coverage is distinct molecules by the time a given number of calls has been charged, read at episode boundaries and bracketed in Appendix B.5.

F.2 Hyperparameters and cross-machine configuration differences

Qwen3-4B is trained with GRPO through verl-agent for 25 updates. Group size is 8, and training and mini-batch sizes are both 16, so each batch gives one gradient step. The learning rate is $1\mathrm{e}{-6}$, with $\gamma = 1$. There is no KL penalty on the loss or in the reward, as every one of the seventy-one cells has KL regularisation explicitly disabled in both locations. The configuration record contains a coefficient of 0.001 and a low-variance estimator, but the disabling switch leaves both inactive, so the trained objective excludes that term. Prompts are capped at 4096 tokens and responses at 512. Rollouts are generated with vLLM.

The budget-two recipe fixes the call-penalty coefficient at zero and disables enforcement during validation. The budget is stated in the prompt but enters the objective neither as a penalty nor as a constraint.

Commanded budget by experiment. Budgets are taken from each run's resolved configuration, not the launch scripts. The historical cohort, the three verified re-executions, the second-size replication, the prospective memory pairs, the memory-plus-IPS pairs and the budget-two repeat-only pairs are commanded two calls an episode. The seven selection pairs of Section 4.1, the two bucket-size alternatives of Appendix D.6 and the budget-eight memory-clause ablation are commanded eight, which is the value both arms of each pair carry, so every comparison this paper makes is within one commanded budget and none crosses the two.

The measured cross-machine boundary. Every run in Figure A.1 uses the same configuration, with no cross-machine comparison. The relevant cross-machine boundary concerns the selection cells of Section 4.1, which ran on the second machine after the pairing arm was set aside. We compare their resolved configurations key by key. Of 309 keys, eight identify replicates, comprising two seeds, the run identifier, the experiment name and four absolute paths. Setting those aside, 291 agree and ten differ. Five hardware differences comprise tensor-parallel degree, KV-cache fraction, eager execution, GPUs per node and the checkpoint-retention count. Five experimental differences comprise the dual step size η (0.001 against 0.02), whether a fixed multiplier is pinned on the token axis, and the three keys defining the token-budget grid (384 alone against 512 to 2048 in three steps).

The five experimental differences do not affect any reported contrast, because every selection comparison is paired within a single machine. Specifically, the two arms of each pair share identical configurations and differ only in the selection rule, ensuring that any machine-level constant cancels when computing differences.

Two scoring services within pair 90. The two arms of pair 90 ran concurrently on one machine and used separate instances of the scoring service. Both instances return the same reference score and identical scores for a second molecule, and select the same molecules for the same farthest-first request. The other three pairs share one service.