VERIFIED SUPERVISION AND COMPOSITIONAL GENERALIZATION IN A FINITE PROGRAM SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

This work studies whether low-rank fine-tuning on exhaustively enumerated, automatically verified composition trajectories can change a small language model's probability distribution so that correct programs rank higher on new tasks under unchanged evaluation. The experiments use Qwen3-0.6B for polynomial transformations over finite fields, where a solution is a sequence of three atomic operations chosen from five. All 125 distinct programs were enumerated and scored exactly, so Hit@32 uses no sampling or repeated candidates. With matched optimiser-step and loss-bearing target-token budgets, mean Hit@32 rose from 45.32% for atomic control to 72.72% for composition fine-tuning, a gain of 27.40 percentage points (95% t interval [25.40, 29.40], $p = 3.45 \cdot 10^{-7}$), positive in all six paired continuations from one atomic checkpoint. This contrast measures the full training recipe. Prespecified secondary comparisons showed a gain of 26.13 points on familiar compositions in new fields and losses of 16.02 and 11.80 points on the two withheld-pair families. Atomic APPLY accuracy decreased in all six composition branches relative to their paired controls. Selective pair exclusion also lowered withheld-stratum Hit@32 across 13 completed subsets. Experiments on Qwen3-0.6B-Base found no clear advantage of structured exploration over independent sampling at comparable candidate diversity, and 400 steps of training with a verifiable reward changed held-out coverage by +0.22 points (95% bootstrap interval $[-3.11, 3.67]$). A development-set follow-up with three continuation seeds from one checkpoint showed the same transfer pattern with higher mean program entropy. The positive effect supports consolidation of familiar composition families and transfer to new tasks.

1 INTRODUCTION

We study how verified composition training changes the ranking of correct programs on familiar compositions and on tasks with withheld operation pairs. In our controlled environment, gains concentrate in familiar composition families and vary with training sets selected by pair membership. Verified training combines candidate generation, automatic checking, and parameter updates. Related approaches include reasoning-trace self-training (Zelikman et al., 2022), reward-filtered self-training (Gulcehre et al., 2023), aggregation of sampled answers (Wang et al., 2023), and step-level reward modelling (Lightman et al., 2024). Here, the generation bottleneck denotes the scarcity of correct proposals and the concentration of sampled candidates at the tested budgets.

The main question is how the ranking changes across task families. A correct program can rise through a shared preference for its operation sequence or through a score that depends on the task. Distinguishing these effects matters for the theory of self-improvement and for practical systems of mathematical inference (Liu et al., 2025; Yue et al., 2025).

For each task, both branches rank the same complete set of programs. This holds evaluation fixed as the parameters change. The supervised trajectories come from exhaustive external enumeration. The reward-training experiments in Section 6.3 use candidates sampled by the current policy.

The six-run confirmatory comparison estimates the effect of the full composition training recipe. Controlled interventions vary the accepted training set and the generation procedure. Development diagnostics examine shared program preferences and atomic retention. The paper makes four contributions:

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.

1. Full-space evaluation scores all 125 programs, with a uniform-ranking reference and complete Hit@K curves. Six paired continuations establish a ranking gain from the combined composition-training recipe under this fixed evaluation.
2. A prespecified pair-exclusion study compares selective and matched random removal on withheld and control strata. Its 13 completed subsets show lower withheld-stratum Hit@32 and usually higher control-stratum scores after exclusion.
3. A development-set decomposition separates score changes into a shared program term and residual changes, clarifying what exact ranking measures alongside crossed START/TARGET diagnostics.
4. A search comparison calibrates IID temperature to approximate prefix-balanced diversity on development tasks, then fixes both sampling policies. Reward-group and entropy measurements describe the accompanying changes in candidate concentration.

Table 1 maps these claims to their scoring rules, replication units, and protocol status.

2 Related Work

Self-training with automatic checking has four steps: the model forms candidates, an external algorithm checks each of them, the correct ones are stored as positive trajectories, and the parameters are updated so that their probability grows. Zelikman et al. (2022) filter reasoning traces by final-answer correctness. Gulcehre et al. (2023) alternate generation and improvement using a learned reward model. ReST-EM (Singh et al., 2024) repeats policy sampling, binary correctness filtering, and supervised updates. Our main treatment obtains its trajectories by external enumeration. In rejection-sampling fine-tuning, Yuan et al. (2023) relate gains to distinct correct reasoning paths, motivating our separate controls for search diversity and training-set composition. At inference, Cobbe et al. (2021) use learned verifiers to rank sampled solutions, whereas our verifier labels the complete candidate space exactly. Wang et al. (2023) aggregate sampled answers, while Lightman et al. (2024) use human step-level labels to train a process reward model. V-STaR (Hosseini et al., 2024) trains a verifier on both correct and incorrect generated solutions. Our checking rule is fixed. Appendix E relates verifier error rates to the contamination of accepted training data.

The extent of the resulting generalisation remains contested. Yue et al. (2025) report that a verifiable reward raises single-sample accuracy while leaving coverage at large K at or below the base model, and ProRL (Liu et al., 2025) reports expanded coverage with a longer schedule, KL regularisation, and reference-policy resets. Both measure coverage by sampling. Wen et al. (2025) additionally evaluate the validity of sampled reasoning with CoT-Pass@K. Our executable programs permit exact checking of every candidate. Wu et al. (2026) examine budget and evaluation confounds in RLVR. We hold the candidate space fixed and compare structured search (Yao et al., 2023) with development-calibrated IID sampling. Zhou et al. (2026) analyse group-size effects through a U-statistic formulation, and Mishra et al. (2026) discuss shared sequence-level credit and zero-contrast groups. Our reward-group audit measures the latter in the saved training logs. Rad et al. (2026) study verifier noise in reward training. Our contamination control instead replaces supervised trajectories with verified incorrect examples.

Setlur et al. (2024) show that positive synthetic traces can reinforce spurious steps and that carefully constructed negatives support per-step credit assignment. Here even fully correct trajectories induce exposure-dependent ranking changes. Razeghi et al. (2022) relate numerical reasoning accuracy to pretraining term frequencies. Our exclusion study directly varies adjacent-pair exposure during fine-tuning.

Compositional generalisation means combining mastered transformations into a sequential program. Separate command of A and B does not guarantee correct execution of A after B: order changes the result and an early error propagates. Lake & Baroni (2018) show that sequence models generalise poorly to held-out combinations of familiar primitives, Zhao et al. (2024) study skill composition learned from examples, and Dziri et al. (2023) trace compositional failures to the structure of the task graph. ExeDec (Shi et al., 2024) tests compositional splits in RobustFill and DeepCoder and predicts execution subgoals before synthesising each program part. Our pair-exclusion split varies adjacent-operation exposure, and exact ranking evaluates the complete program space under a fixed scoring rule. Intermediate states enter our supervised targets, while inference scores operation se-

quences. Expert Iteration (Anthony et al., 2017) alternates tree-search targets and policy learning in games. Here the target-generating search is exhaustive and fixed: enumeration and an exact checking program determine the correct trajectories, and the model is fine-tuned on them with a low-rank adapter (Hu et al., 2022), receiving the full target answer including the operations and the verifiable intermediate states. Repeated training on recursively generated data can erode distributional coverage (Shumailov et al., 2024). Here the exact checking program controls trajectory correctness. The transfer and retention tests measure the consequences of training on the accepted data. GeoRA (Zhang et al., 2026) studies geometry-aware low-rank updates and retention on out-of-domain RLVR tasks. Our retention interventions keep the adapter parameterisation fixed and vary atomic replay and learning rate.

3 Environment and Data

The state is a polynomial of degree at most D over the finite field F_p, stored as the vector of its coefficients $[c_0, \ldots, c_D]$ taken modulo p, with the length preserved including leading zeros. Five exact atomic operations are defined over this state: SH1 shifts the argument by one, SC2 scales it by two, REV reverses the coefficient vector, and AC1 and AX1 add one to the free coefficient and to the coefficient at x. All are computed deterministically, do not increase the size of the state and admit exhaustive enumeration of programs of small depth (Table 5).

Composition and reward training use the fields 5, 7, 13, 19, 23 and 29. The fields 11 and 17 are held back for transfer, and three *motifs* are held back from supervision, a motif being an adjacent ordered pair of operations. The three are AX1 followed by SH1, AC1 followed by REV, and SC2 followed by AX1. At depth two the space contains $5^2 = 25$ programs, at depth three 125, and at depth four 625. Appendix A works through one program of each depth and gives a target that two different programs reach.

In PLAN mode the model receives the field, the start and target vectors, the list of admissible operations and the fixed program length, and must return a sequence of operations. In APPLY mode it is given an explicit sequence and must compute the final vector. APPLY was used to control the preservation of atomic skills, and the main compositional result was measured in PLAN mode.

A pool of 4000 composition tasks supplied 2250 selected training trajectories, fixed across the six continuations (Appendix A). The final sets were generated separately and frozen before evaluation. Four families separate different kinds of transfer: A tests new input states under familiar compositional signatures and familiar fields, B holds operation pairs under familiar fields, C uses familiar signatures in new fields, and D combines the two, with 1000 tasks each in A and B and 500 each in C and D. In B/D, every shortest correct program contains exactly one withheld-pair occurrence. In A/C, every shortest correct program avoids all three pairs. Appendix B gives their structure and the integrity checks.

4 Method

4.1 Compared branches

The six-run series uses Qwen3-0.6B (Yang et al., 2025). Each branch loads the same merged atomic checkpoint and initialises a fresh rank-32 LoRA adapter and optimiser. The paired run seed controls adapter initialisation, atomic-example selection, shuffling, and dropout. The composition branch receives full correct trajectories and 20 percent atomic replay by examples. The atomic-only control matches optimiser steps and loss-bearing target tokens. Prompt-token counts and total computation can differ. The contrast measures the combined recipe under a matched supervised-target budget. Figure 3 and Table 11 in Appendix B state the design and the settings. The pair-exclusion study holds prompt format, final training quotas, and atomic replay fixed while changing the removal rule (Section 6.1).

4.2 EXACT RANKING AND METRICS

Evaluation enumerated all distinct programs of the fixed depth for each task. The model score sums the conditional log-probabilities of its actions,

$$L(o_1, \ldots, o_\ell \mid x) = \sum_{i=1}^{\ell} \log p_\theta(o_i \mid x, o_1, \ldots, o_{i-1}) \,. \tag{1}$$

Here ℓ is program length. In the six-run series, each operation is one added token, `<OP0>`–`<OP4>`. Tokenisation checks enforce this even after the separating space, which the added token consumes. Each factor uses the full vocabulary. The prompt ends at `PROGRAM:`. Prompt tokens, the subsequent `TRACE:` text, and EOS contribute no score. The metric therefore measures the operation-sequence prefix of a completion at a fixed program length. Training supervises both this sequence and its intermediate-state continuation (Appendix A).

Programs were sorted by decreasing score, ties were broken by the lexicographic order of the operation signatures, and the checking program marked all correct variants. The main measure Hit@32 is the share of tasks for which at least one correct program enters the first 32 positions. Hit@K indexes the retained candidates after scoring the full space. Sampled pass@K (Chen et al., 2021) instead evaluates K generated candidates. Depth two has only 25 candidate programs, so Hit@32 is structurally saturated there and a zero difference on it carries no information about the absence of an effect.

The search and reward experiments spell operations as text, such as `SH1`. For each prefix, the scorer sums full-vocabulary token log-probabilities for the continuation `␣SH1`, including its leading space, and applies a softmax to the five operation scores. Equation 1 then sums these local action log-probabilities. Trace and EOS scores are again excluded. This rule and the added-token rule evaluate distinct distributions, so numerical contrasts are compared within each series. Prefix-dependent normalisers can change ranks. Across 96 development continuations, the rules differ by at most 0.75 points in mean B/D Hit@32 (Appendix I). This sensitivity check concerns those checkpoints, as the six-run per-prefix logits were not retained in the analysis release. Exact enumeration fixes sampling variability for a specified score. A program-only score term is constant across tasks for that program. The residual and crossed START/TARGET interaction describe task dependence.

A uniform random ordering of the 125 programs provides a reference based on the number of correct programs per task (Appendix F). The confirmatory contrast compares composition training with atomic control. The pair-exclusion study uses matched random removal.

4.3 STATISTICAL PLAN

The replication unit is a paired continuation run, conditional on the shared atomic checkpoint, base model, and fixed datasets. For each of the six runs the difference in Hit@32 between the branches was computed on the same 1000 tasks of family A, and over the six differences the mean, the 95 percent t interval and the two-sided one-sample t test were computed. An exact sign-permutation test over all $2^6 = 64$ sign combinations and a two-level bootstrap over both runs and tasks served as robustness checks, and the Holm correction was applied to five secondary measures (Holm, 1979). The conditions for a positive result were fixed before the final set was opened. Appendix B gives them with the resampling detail. Intervention comparisons are reported separately, with unadjusted descriptive intervals conditional on each study's shared starting checkpoint. The Holm correction applies only to the five confirmatory-series secondary measures.

5 CONFIRMATORY COMPARISON

This section reports the prespecified six-run comparison and its secondary transfer and retention outcomes. Later sections separate controlled interventions from exploratory score diagnostics (Table 1).

Comparison	Score and paired unit	Status and interpretation
Composition vs atomic (§5.1)	Full vocabulary, 6 paired continuations	Prespecified recipe contrast, shared atomic checkpoint.
Pair exclusion (§6.1)	Full vocabulary, 3 seeds/subset	Fixed design, 13/20 subsets. Descriptive training-set contrast.
Search / reward (E1/E2)	Sampled pass@32, 3 seeds	Prespecified fixed-budget contrasts.
Supervision / motifs (E3/E4)	Five actions, 3 seeds	Exploratory exact-ranking and motif analyses.
Selection / entropy	Full vocabulary / five actions, 6 / 3 seed pairs	Fixed endpoints, checkpoint-conditional intervals.
0.6B / 8B	Full vocabulary, 3 pairs/size	Development, one checkpoint/size.
CF/H/replay	Five actions, 3 seeds/mask/recipe	96 development continuations, post-hoc score diagnostics.

Table 1: Evidence map. Both exact scorers rank all 125 programs. E1/E2 intervals resample paired tasks within each fixed seed. E3/E4 use crossed seed–task resampling. Pair subsets share a pool and reused seeds. Development continuations share their initial checkpoint. Appendix D details the scope of each comparison.

5.1 Ranking gain

The starting atomic adapter reached 99.8 percent in PLAN mode and 91.5 percent in APPLY mode, its weakest atomic operation being the shift SH1 at 57.5 percent.

On the main set A of depth three, mean Hit@32 rises from 45.32 percent under atomic control to 72.72 after composition training. The paired gain is 27.40 percentage points, with a 95 percent t interval of $[25.40, 29.40]$. All six differences are positive (Figure 10). The two-sided t test gives $p = 3.445 \times 10^{-7}$, and the exact sign-permutation test gives $p = 0.03125$. Table 12 reports the bootstrap robustness check. The treatment adds multi-operation prompts, sequences, and intermediate states to a depth-one atomic control. The 27.40-point estimate measures this combined training change. Shared program preferences can contribute to exact-rank gains. The task-level diagnostics in Appendix F describe the accompanying mass and entropy changes. Appendix I decomposes program scores on a separate development series.

The effect appeared in a continuous measure as well. In all six runs the normalised probability mass of the correct set grew, by 4.41 percentage points on average with a 95 percent t interval of $[4.25, 4.57]$ (Figure 5). This mass is conditional on the enumerated 125-program space.

Figure 11 in Appendix F extends the same comparison to every $K = 1, \ldots, 125$. Uniform ranking gives 39.03 percent at $K = 32$, compared with 45.32 for atomic control and 72.72 for composition training. At $K = 1$, these values are 1.43, 2.30, and 10.00 percent. The endpoint distributions also differ in entropy: after normalising the scores over the 125 programs, mean entropy is 0.765 nats in the atomic control and 3.892 after composition training. This is entropy conditional on the finite candidate set. On unique-solution tasks, mean correct mass falls while Hit@8 rises. On multiple-solution tasks, both improve. Table 25 shows this distinction: higher entropy can accompany better ranks even when the mean correct probability decreases. Table 2 gives selected ranking budgets, and Tables 23–24 give paired task transitions and a stratification by the number of correct programs. The mean Hit@32 gain is 16.17 points on the 566 tasks with one correct program and 42.05 points on the 434 tasks with multiple correct programs. Composition training exceeds the corresponding uniform references by 39.36 and 26.29 points.

5.2 Boundaries of transfer

The secondary sets determine the boundaries of the effect, and Figure 6 in Appendix C collects them. On family C, where the composition families stayed familiar but new finite fields were used, the gain was 26.13 percentage points across tasks over new fields and coefficients. On family B with operation pairs held back in advance the result fell by 16.02 points, and on family D, which combined withheld pairs and new fields, by 11.80. Thus correct-program ranks improve on new fields within familiar composition families and fall on families containing withheld pairs. On the

K	Uniform (%)	Atomic (%)	Comp. (%)	Δ (pp)	95% t interval
1	1.43	2.30	10.00	+7.70	[7.05, 8.35]
8	11.04	12.02	36.73	+24.72	[22.93, 26.51]
16	21.19	22.60	52.93	+30.33	[29.56, 31.11]
32	39.03	45.32	72.72	+27.40	[25.40, 29.40]
64	66.49	75.80	94.52	+18.72	[18.06, 19.37]
125	100.00	100.00	100.00	+0.00	[0.00, 0.00]

Table 2: Exact Hit@K on the six-run family-A test. Uniform is the expected value for a random permutation of all 125 programs. Differences and pointwise intervals compare composition with atomic control across paired training runs. Hit@125 is fixed by exhaustive enumeration.

combined set of depth four, the gain is 5.20 points. All four secondary contrasts remain significant under the prespecified Holm correction (Appendix B).

5.3 Preservation of atomic skills

The positive compositional result was accompanied by partial forgetting. After the control atomic training the mean PLAN was 99.72 percent and the mean APPLY 90.28. After composition fine-tuning PLAN was 97.78 and APPLY 85.70 (Figure 4). The main contribution came from SH1, which the control branch left at 51.5 percent and the composition branch at 29.08, while the remaining four operations stayed at 99 to 100 percent. The descriptive retention rule limits the largest loss from the shared starting checkpoint to two points across overall PLAN and each operation's PLAN/APPLY accuracy. Two atomic-control branches among the twelve trained branches meet this rule. A post-hoc joint check requires a family-A Hit@32 gain of at least five points and an overall APPLY loss of at most two points against the paired atomic control. All six ranking gains pass, while APPLY losses of 2.8–5.8 points leave zero of six runs meeting both requirements (Table 23).

6 Prespecified interventions

These studies fixed their designs and endpoints before the corresponding outcome evaluation. Their seed pairs share starting checkpoints, so intervals are conditional on each study's initialisation. The pair-exclusion matrix is partial.

6.1 Pair exclusion and matched random removal

The pair-exclusion study compares selective removal of operation pairs with random removal of the same number of trajectories, matched by depth and field. Both arms use multi-operation prompts, match final target-token quotas, and share atomic replay. It contains 13 completed subsets: five with one excluded pair, five with two, and three with three. Each subset has three paired training seeds and 1000 evaluation tasks with at least one shortest correct program containing an excluded pair. All 39 paired differences favour random removal. The 13 subset means range from -43.43 to -5.37 points. Results are reported within subset and GPU stratum (Table 27). With three seeds, the smallest attainable two-sided sign-flip p value is 0.25. The pattern is descriptive across these completed subsets. The intervention changes operation marginals, signatures, and intermediate states along with pair membership. Its estimand is the resulting training-set contrast. In subsets 1/4 and 2/4, every task has a correct alternative avoiding the excluded pairs, so the strict all-solutions-require-a-pair stratum is empty.

On the prespecified control stratum, whose correct programs all avoid the excluded pairs, selective exclusion improves mean Hit@32 in 11 of 13 subsets and 35 of 39 seed pairs. Subset means range from -0.80 to $+9.50$ points (Table 28). Figure 1 shows both contrasts. Saved secondary endpoints also give lower withheld-stratum Hit@8 and correct-set mass in all 13 subsets. Hit@1 falls in ten (Table 30). Appendix G.2 gives the design and completion accounting: 78 of 120 planned trainings are represented. For subsets 1/1, 1/2, 1/5, and 2/3, a count-based sensitivity analysis also bounds the effect below zero on tasks whose *every* correct program requires an excluded pair, in each of their twelve seed comparisons (Table 29). These are post-hoc bounds on observed outcomes.

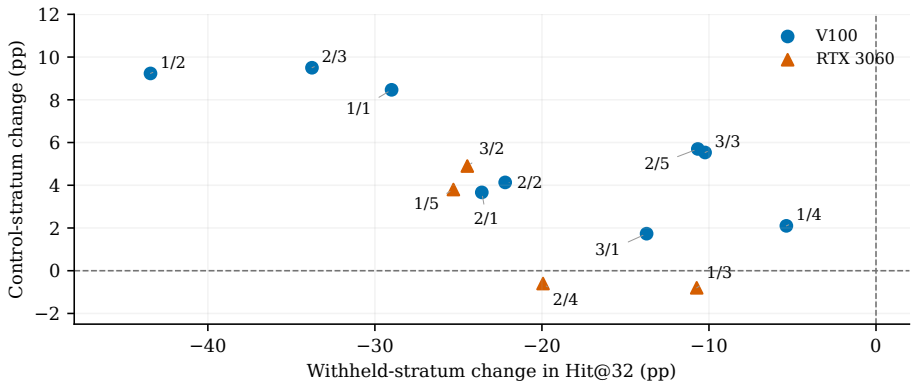

Figure 1: Selective pair exclusion minus matched random removal on the withheld and control strata. Each point averages three paired seeds. Labels give the number of excluded pairs and subset index. Colour and shape identify the GPU stratum. The comparison is descriptive. Per-seed results and within-subset intervals are in Tables 27 and 28.

Comparison	Pairs	Control (%)	Test (%)	Δ (pp)	95% t interval
Trajectory selection	6	61.55	61.30	-0.25	[-2.3, 1.8]
Entropy bonus	3	49.43	48.87	-0.57	[-4.1, 2.9]

Table 3: Paired trajectory-selection and entropy-bonus contrasts on separate 1000-task holdouts. Selection uses full-vocabulary scores. The entropy comparison uses the five-action policy at temperature 0.7. Intervals are descriptive, conditional on each study's shared initialisation. Exact sign-flip p values are 0.8125 and 0.75, respectively.

6.2 Trajectory selection

Six paired runs compare random selection with a rule that increases adjacent-pair balance and intermediate-state coverage. Both arms select 750 verified trajectories from one pool of 2250, with matched depth, field, and target-token quotas, and use 20 percent atomic replay. The structured rule exceeds the frozen diversity threshold in every pair. Its Hit@32 contrast on a new holdout is close to zero (Table 3). The intervention changes the selected tasks together with their trajectories (Appendix G.1).

6.3 Search and reward training

Search and reward experiments use Qwen3-0.6B-Base with its own atomic checkpoint and three training seeds. Its splits and five-action scoring rule are separate from the confirmatory comparison. Sampled-coverage intervals resample paired task outcomes within each seed, conditional on the three observed runs (Appendix E).

The IID temperature approximately matches prefix-balanced diversity on development tasks and is frozen for holdout evaluation. Under these policies, the coverage contrast has an interval spanning zero (Table 4, Figure 8). The 400-step GRPO comparison (Shao et al., 2024) uses IID training and evaluation at $T = 8$ on 300 held-out tasks, averaged over three seeds. Its interval permits changes of several points. Evaluating the same checkpoints at $T = 0.7$, per-candidate accuracy rises from 0.0151 to 0.0358, while distinct programs per task fall from 3.80 to 1.81 (Table 18).

6.4 Entropy intervention and reward-group diagnostics

A post-hoc audit of the six E1 reward-training runs at $T_{\text{train}} = 0.7$ finds mixed rewards in 3.17 percent of IID groups and 25.33 percent of prefix-balanced groups, each containing eight answers (Table 31). Group-entropy means are similar in the first and last training blocks (Figure 12). These diagnostics characterise the E1 runs. The Q5 coverage contrast uses separate training runs at $T = 8$.

Comparison	Endpoint	Change (pp)
Search, dev-calibrated (E1)	Sampled pass@32	-1.11 $[-4.22, 1.89]$
Reward, 400 steps (E2)	Sampled pass@32	$+0.22$ $[-3.11, 3.67]$

Table 4: Prespecified search and reward contrasts. E1 calibrates mean candidate diversity on development tasks. E2 compares reward training with the shared starting checkpoint. Brackets give 95 percent within-seed paired-task bootstrap intervals. Appendix E reports all settings and exploratory extensions.

A separate prespecified experiment adds an operation-choice entropy bonus for 150 steps, using three paired seeds and a control retrained with zero bonus. Candidate-conditional entropy rises from 1.435 to 1.600 nats, and the Hit@32 interval spans zero (Table 3). Mixed rewards occur in 55 of 450 control groups and 62 of 450 bonus groups. The comparison estimates the total effect of the bonus, including exploration and policy smoothing. Appendix G.3 gives the objective, group counts, and sensitivity calibration.

7 Exploratory and development analyses

These analyses describe additional endpoints, continuations, and post-hoc score transformations. Their estimates remain conditional on the tested checkpoints.

7.1 Supervision and motif transfer

Experiments E3 and E4 evaluate externally enumerated supervision and motif transfer on Qwen3-0.6B-Base. The exact-ranking endpoint and motif hypothesis have uncertain prespecification status and are treated as exploratory. Supervision raises Hit@32 on familiar compositions across known and new fields. It lowers B/D scores and fails that study's atomic-retention rule. An atomic-only control and adapters trained with the excluded pairs provide boundary comparisons (Appendix E, Figure 7).

7.2 Development continuations at two model sizes

Three paired continuations per size compare full-trajectory training with matched atomic control, conditional on one atomic checkpoint at each size. Both sizes improve familiar-family ranking and place B/D below their uniform references (Table 32). Family-A composition Hit@1 is 15.50 percent at 0.6B and 25.83 at 8B. Appendix H gives initial scores, entropy, and retention for these development tasks.

7.3 Program preferences and atomic replay

A separate matrix of 96 continuations shares one starting checkpoint and crosses four objectives, two learning rates, replay, and two pair masks. A shared program term accounts for a mean 52.1 percent of in-sample score-change variance on reference A. Strict joint accuracy is a two-program diagnostic: the designated program must outrank its alternative in all four START/TARGET combinations. Subtracting the mean training-induced change per program raises B/D Hit@32 from 22.43 to 57.56 percent and changes strict joint accuracy from 6.38 to 3.97 percent (Table 36 and Appendix I.2). This reference-dependent decomposition retains the initial score structure. The correction leaves the signed crossed interaction unchanged and can alter individual cell comparisons and strict joint accuracy.

Explicit replay raises atomic APPLY in all 48 matched comparisons. At learning rate 3×10^{-5} and replay loss weight 0.25, mean APPLY is 61.80–62.00 percent across the four objectives, compared with 62.55 initially. Family-A Hit@32 improves in those conditions, while B/D remains below its initial level. Replay here adds atomic data and computation at a fixed composition budget. The full matrix, paired effects, and retention trade-offs are reported in Figure 13 and Table 39.

8 DISCUSSION

The measurements support an exposure-dependent reading of verified supervision in this finite program space. The six-run comparison establishes a gain for the full composition-training recipe. Under matched multi-operation prompts, selective pair removal shifts performance away from the withheld stratum and towards the control stratum. Our interpretation is that the accepted training set shapes shared program preferences alongside task-dependent scores. Pair removal changes operation frequencies, signatures, and intermediate states together. Adaptation to multi-operation prompts is an additional explanation for the six-run A/C gain. A format- and budget-matched control reproducing that gain without exposure to the relevant pairs would weaken its pair-exposure interpretation.

The generation bottleneck concerns the policy-generated part of the pipeline: correct candidates are scarce, and sampled programs are concentrated at the tested budgets. Search, reward, selection, and entropy contrasts have intervals spanning zero. Their widths permit effects of several points, as quantified in Appendix G.3. The supervised treatment receives its correct trajectories through external enumeration. With exact verification, the accepted set's composition and coverage remain distinct constraints on training. Appendix E relates proposer accuracy to accepted-set contamination and reports the separate single-seed replacement control.

The practical implications concern measurement and retention. Report full ranking curves against a task-specific uniform reference, audit exposure in accepted trajectories, and evaluate atomic execution alongside composition scores. The development decomposition shows how shared program preferences can enter exact ranking. Atomic replay at the lower learning rate retains more APPLY accuracy in that development setting. The six-run SH1 losses are consistent with forgetting under sequential fine-tuning (Kirkpatrick et al., 2017). Verified incorrect programs also provide possible training negatives, a direction motivated by the per-step credit-assignment results of Setlur et al. (2024).

8.1 LIMITATIONS

The findings concern five deterministic operations, principally depth-three programs, and the tested Qwen checkpoints. The six-run comparison estimates a combined training recipe, and the exclusion contrast estimates a selection rule. The completed subsets share a pool and reused seeds, with one GPU per subset. Development continuations are conditional on their starting checkpoints. The score decomposition requires per-program scores, while changing the six-run normaliser requires per-prefix logits. The six-run task summaries support the ranking and distributional analyses reported here. Crossed diagnostics and normaliser comparisons use the development checkpoints. Appendix D details these identification limits and protocol scopes.

9 CONCLUSION

Verified composition training raises correct-program ranks under fixed exhaustive evaluation in six paired continuations. The gain measures the full continuation recipe. Prespecified pair-exclusion experiments show a consistent loss on the withheld stratum across the completed subsets, with gains on the control stratum in most subsets. At the tested budgets, search, reward training, trajectory selection, and a reward-entropy bonus have coverage intervals spanning zero. The score decomposition quantifies shared program preferences in development continuations. The transfer and atomic APPLY results delimit the ranking gain for these training recipes in the tested DSL.

AI USE STATEMENT

Generative AI tools assisted with experiment design and critique, implementation of experiment and analysis code, literature search, manuscript drafting and editing, and LaTeX preparation. Tasks and reference answers were generated by the DSL code. The authors checked data provenance, file hashes, derivations, paired statistics, ranking diagnostics, and analysis code. Experimental outcomes are derived from recorded runs. The authors are responsible for the content of the submission.

Reproducibility statement

Appendices A, B, and C give the environment, prompts, training settings, statistical procedures, and integrity checks. Appendices D and E give limitations and search, reward, and supervision details. Appendix F gives full ranking curves and paired outcomes. Appendices G, H, and I document the training-data experiments, model-size comparisons, and score diagnostics. The supplementary material contains numerical inputs and CPU analysis scripts.

References

Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1705.08439.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, et al. Evaluating large language models trained on code. arXiv:2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv:2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, et al. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18654.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, et al. Reinforced self-training (ReST) for language modeling. arXiv:2308.08998, 2023. URL https://arxiv.org/abs/2308.08998.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7038–7051, 2021. doi: 10.18653/v1/2021.emnlp-main.564. URL https://arxiv.org/abs/2104.08315.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-STaR: Training verifiers for self-taught reasoners. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2402.06457.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://arxiv.org/abs/1612.00796.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pp. 2873–2882, 2018. URL https://arxiv.org/abs/1711.00350.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.20050.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-0608. URL https://arxiv.org/abs/2505.24864.

Amritansh Mishra, Supriyo Chakraborty, and Berkcan Kapusuzoglu. On the policy gradient foundations of Group Relative Policy Optimization: Credit assignment, gradient sparsity, and rank collapse. *arXiv:2606.29238*, 2026. URL https://arxiv.org/abs/2606.29238.

Ali Rad, Khashayar Filom, Darioush Keivan, Peyman Mohajerin Esfahani, and Ehsan Kamalinejad. Rate or fate? rlv$^{\varepsilon}$r: Reinforcement learning with verifiable noisy rewards. *arXiv:2601.04411*, 2026. URL https://arxiv.org/abs/2601.04411.

Yasaman Razeghi, Robert L. Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 840–854, 2022. doi: 10.18653/v1/2022.findings-emnlp.59. URL https://aclanthology.org/2022.findings-emnlp.59/.

Amrith Setlur, Saurabh Garg, Xinyang Geng, Naman Garg, Virginia Smith, and Aviral Kumar. RL on incorrect synthetic data scales the efficiency of LLM math reasoning by eight-fold. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-1361. URL https://arxiv.org/abs/2406.14532.

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:2402.03300, 2024. URL https://arxiv.org/abs/2402.03300.

Kensen Shi, Joey Hong, Yinlin Deng, Pengcheng Yin, Manzil Zaheer, and Charles Sutton. ExeDec: Execution decomposition for compositional generalization in neural program synthesis. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.13883v2.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024. doi: 10.1038/s41586-024-07566-y. URL https://www.nature.com/articles/s41586-024-07566-y.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, et al. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2312.06585.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. *arXiv:2506.14245*, 2025. URL https://arxiv.org/abs/2506.14245.

Fang Wu, Aaron Tu, Weihao Xuan, Heli Qi, Xu Huang, et al. Position: The hidden costs and measurement gaps of reinforcement learning with verifiable rewards. *arXiv:2509.21882*, 2026. URL https://arxiv.org/abs/2509.21882v3.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.10601.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. arXiv:2308.01825, 2023. URL https://arxiv.org/abs/2308.01825.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2504.13837.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2203.14465.

Jiaying Zhang, Lei Shi, Jiguo Li, Jun Xu, Jiuchong Gao, et al. GeoRA: Geometry-aware low-rank adaptation for RLVR. *arXiv:2601.09361*, 2026. URL https://arxiv.org/abs/2601.09361v4.

Haoyu Zhao, Simran Kaur, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Can models learn skill composition from examples? In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2409.19808.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Hongyi Zhou, Kai Ye, Erhan Xu, Jin Zhu, Ying Yang, et al. Demystifying Group Relative Policy Optimization: Its policy gradient is a U-statistic. *arXiv:2603.01162*, 2026. URL https://arxiv.org/abs/2603.01162.

A Worked examples of the environment

Table 6 shows all five transformations on one starting polynomial over F_{11}, which makes it visible that they change different parts of the state. Figure 2 traces one shortest program of depth three and gives one target reached by two different programs of depth four. Table 7 collects further correct compositions across fields, degrees and depths. Every row was re-executed with the deterministic interpreter of the experiments, and every uniqueness claim was re-checked by exhaustive search over all programs of that depth and shorter.

Code	Transformation	Effect
SH1	$P(x) \mapsto P(x+1)$	shift of the argument by one
SC2	$P(x) \mapsto P(2x)$	scaling of the argument
REV	$[c_0, \ldots, c_D] \mapsto [c_D, \ldots, c_0]$	reversal of the vector
AC1	$c_0 \mapsto c_0 + 1$	change of the free coefficient
AX1	$c_1 \mapsto c_1 + 1$	change of the coefficient at x

Table 5: The five atomic transformations, all taken modulo p. Each one is computed deterministically and none of them changes the length of the state vector, so programs of small depth can be enumerated in full. AC1 and AX1 look similar and touch different coordinates, so after REV or SH1 they lead to different final states.

A.1 Exact prompts and supervision

The five added operation tokens map as follows: <OP0> = SH1, <OP1> = SC2, <OP2> = REV, <OP3> = AC1, and <OP4> = AX1. The PLAN prompt for Figure 2(a) is:

```
FIELD: 11
DEGREE_CAP: 3
START: [0, 2, 2, 5]
```

Code	Exact action on $P(x) = 2 + 5x + 3x^2 + 4x^3$	Result
SH1	shift of the argument, $P(x) \to P(x+1)$: $2 + 5(x+1) + 3(x+1)^2 + 4(x+1)^3 = 14 + 23x + 15x^2 + 4x^3$	$[3, 1, 4, 4]$
SC2	scale of the argument, $P(x) \to P(2x)$: the coefficient at x^i is multiplied by 2^i, giving $[2, 10, 12, 32]$	$[2, 10, 1, 10]$
REV	reversal of the coefficient vector. This is neither a reversal of digits nor a substitution of $-x$ for x: $[c_0, c_1, c_2, c_3] \to [c_3, c_2, c_1, c_0]$	$[4, 3, 5, 2]$
AC1	one is added to the free coefficient only: $[2 + 1, 5, 3, 4]$	$[3, 5, 3, 4]$
AX1	one is added to the coefficient at x only: $[2, 5 + 1, 3, 4]$	$[2, 6, 3, 4]$

Table 6: All five transformations on one starting polynomial over F_{11}, with the state $[2, 5, 3, 4]$. Coefficients are reduced modulo 11 after every operation. AC1 and AX1 look similar and change different coordinates, so after REV or SH1 they lead to different final states. Every row was re-executed with the same deterministic interpreter that the experiments use.

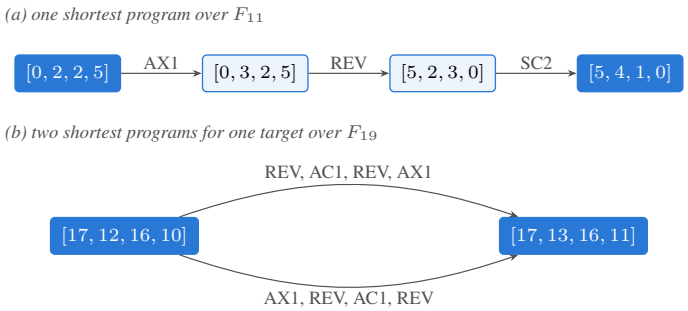

Figure 2: Panel (a) gives every intermediate state of one shortest program of depth three over F_{11}. Order matters: exchanging the first two operations sends the start to $[5, 6, 8, 0]$ instead of the target. Panel (b) gives one target over F_{19} that two different programs of depth four reach. Because several programs may reach one target, Hit@32 uses the best rank among all correct programs.

Depth	p	Start	Target	A correct program	Type
2	7	$[1, 2, 3]$	$[3, 2, 2]$	AC1, REV	unique shortest
3	11	$[0, 2, 2, 5]$	$[5, 4, 1, 0]$	AX1, REV, SC2	unique shortest
3	7	$[4, 4, 5, 6]$	$[0, 2, 4, 5]$	SH1, REV, AC1	unique shortest
4	11	$[10, 10, 2]$	$[0, 10, 7]$	SC2, REV, SC2, SH1	unique shortest
4	23	$[7, 8, 18, 5]$	$[16, 20, 1, 10]$	REV, AX1, SH1, SC2	one solution
4	19	$[17, 12, 16, 10]$	$[17, 13, 16, 11]$	the two of Fig. 2	two solutions

Table 7: Correct compositions across different fields, degrees and depths. Every row was re-executed with the deterministic interpreter of the experiments, and every uniqueness claim was re-checked by exhaustive search over all programs of that depth and shorter. Order matters throughout: on the second row the sequence REV, AX1, SC2 gives $[5, 6, 8, 0]$ instead of the target, and on the fourth row exchanging the first two operations gives $[4, 8, 6]$.

```
TARGET: [5, 4, 1, 0]
ALLOWED: <OP0> <OP1> <OP2> <OP3> <OP4>
MAX_STEPS: 3
Return exactly one line:
PROGRAM:
```

The full-trajectory training completion is:

```
 <OP4> <OP2> <OP1>
TRACE: [0, 3, 2, 5] -> [5, 2, 3, 0] -> [5, 4, 1, 0]
```

The executed PLAN prompt requests one line, while full-trajectory targets append a TRACE line. Exact ranking scores the operation-token prefix, independently of compliance with that instruction. The atomic APPLY prompt for SH1 on the same start is:

```
FIELD: 11
DEGREE_CAP: 3
```

```
START: [0, 2, 2, 5]
PROGRAM: <OP0>
Return exactly one line:
RESULT:
```

Its completion is `[9, 10, 6, 5]`. Training adds a leading space and EOS to each completion. Prompt tokens are masked out of the loss.

LoRA targets `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. The five added operation-token embedding rows are trainable, with tied input/output updates. Other base-model parameters are frozen.

The generator stores each task's witness and up to two shortest correct programs, taken first in enumeration order and deduplicated before selection. A deterministic selector draws 2250 trajectories from the 4000-task pool, using each selected task once. It cycles through depth/first-operation strata with 450 trajectories per first operation. Within a stratum it minimises, in order, mean position/operation counts, mean adjacent-pair counts, and full-program counts in the selected corpus. SHA-256 trajectory fingerprints break ties. This selection is frozen across all six continuations. Evaluation credits every correct program.

Each atomic state/operation pair supplies a one-operation PLAN and an APPLY example. The composition branch adds 562 atomic examples to the 2250 trajectories, approximately 20 percent replay by count. Replay cycles over ten mode/operation strata, sampling with replacement within each stratum. For the matched-target-budget control, atomic examples are concatenated into records that match each composition record's number of loss-bearing target tokens. Position indices restart at each segment, activating Transformers' block-diagonal causal mask. A runtime guard checks segment detection and attention compatibility. Excess final target labels are masked. The control matches target-token and optimiser-step budgets. Prompt tokens and total computation can differ. This procedure is the atomic-data control used in the main comparison.

B Design, frozen protocol and statistical detail

Tables 8, 9, and 10 specify the evaluation families, decision rules, and training branches. Tables 12 and 13 report the run-level statistics and paired task counts.

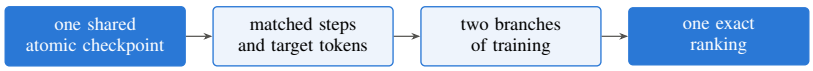

Figure 3: Six paired continuations from one merged atomic checkpoint. The branches match optimiser steps, loss-bearing target tokens, epochs, and effective batch size. Prompt-token counts and total computation can differ. Both branches use the same exhaustive evaluation.

Set	Signatures	Fields	Tasks	Role
A	familiar	familiar	1000	main test
B	held motifs	familiar	1000	transfer to new motifs
C	familiar	new	500	transfer between fields
D	held motifs	new	500	double transfer

Table 8: Structure of the final sets of the confirmatory series. All specific input states and tasks were new, the sets were frozen before the confirmatory series and were not used to select any hyperparameter, and the leak audit found no intersection with the atomic, training and development sets. Check sets of 125 tasks at depths 2 and 4 were built for each family in addition.

For a run s the difference of the main measure was defined as $\Delta_s = \text{Hit@32}_s(\text{composition}) - \text{Hit@32}_s(\text{control})$, and over the six differences

$$\bar{\Delta} = \frac{1}{6}\sum_s \Delta_s = 0.274, \qquad \bar{\Delta} \pm t_{0.975,5}\frac{s_\Delta}{\sqrt{6}} = [0.25403, 0.29397]. \tag{2}$$

Condition	Threshold
mean gain in Hit@32	at least 5 percentage points
lower bound of the 95 percent t interval	strictly above zero
two-sided p value	below 0.05
number of positive runs	at least 5 of 6
data integrity	full check, matched targets/steps, no leaks

Table 9: Conditions for a positive result, fixed before the final set was opened, together with the main endpoint, set A, depth, and number of paired continuation runs. All five were met.

Branch	Training data	Purpose
atomic control	fresh atomic tasks	control for further optimiser steps and supervised target tokens
composition branch	full correct composition trajectories and 20 percent atomic replay	consolidation of verified operation sequences

Table 10: The two branches of each paired continuation from the shared atomic checkpoint. Exhaustive search and exact verification supply the composition trajectories. The treatment supervises program and intermediate-state tokens. The atomic control matches loss-bearing target tokens and optimiser steps. Prompt-token counts and total computation can differ.

Parameter	Value
LoRA rank	32
scaling factor	64
dropout probability	0.05
learning rate	1×10^{-4}
epochs	2
effective batch size	64
warm-up fraction	3 percent
gradient-norm clip	1.0
arithmetic format	bfloat16
atomic replay in the main branch	20 percent by examples

Table 11: Fine-tuning parameters of the confirmatory series, from its frozen protocol. Each branch starts with the same merged atomic weights, a freshly initialised LoRA adapter, and a new optimiser. The two branches share the run seed.

Quantity	Value
mean gain	$+27.40$ pp
95 percent t interval	$[25.403, 29.397]$ pp
t, degrees of freedom	35.275, $\mathrm{df} = 5$
two-sided p from the t test	3.445×10^{-7}
positive runs	6 of 6
exact sign-permutation test	$p = 0.03125$
two-level bootstrap 95 percent interval	$[23.117, 31.667]$ pp

Table 12: Statistical confirmation of the main effect on set A at depth 3. The unit of replication is the training run, so the interval and the p value are computed over the six differences between branches. The mean gain, the t interval and the t statistic in this table were recomputed here from the per-run values of Fig. 10 and agree with the values reported by that series to three decimals. With six runs the smallest attainable two-sided p of the exact sign-permutation test is 0.03125, which is the value obtained.

The two-level resampling redrew both runs and tasks, used the same resampled sequence of task indices in all selected runs so that the paired structure was preserved, and ran 20000 repetitions, giving the percentile interval $[0.23117, 0.31667]$.

Run	Both	Control only	Composition only	Neither	Δ, pp
1	305	130	432	133	+30.2
2	316	141	408	135	+26.7
3	305	149	409	137	+26.0
4	343	117	409	131	+29.2
5	321	138	408	133	+27.0
6	325	129	382	164	+25.3

Table 13: Number of the 1000 tasks of set A solved by both branches, by the control branch only, by the composition branch only, and by neither. The difference in Hit@32 equals the count solved only by the main branch minus the count solved only by the control, divided by 1000. Each row sums to 1000 and reproduces the per-run values of Fig. 10, which was checked here and not assumed.

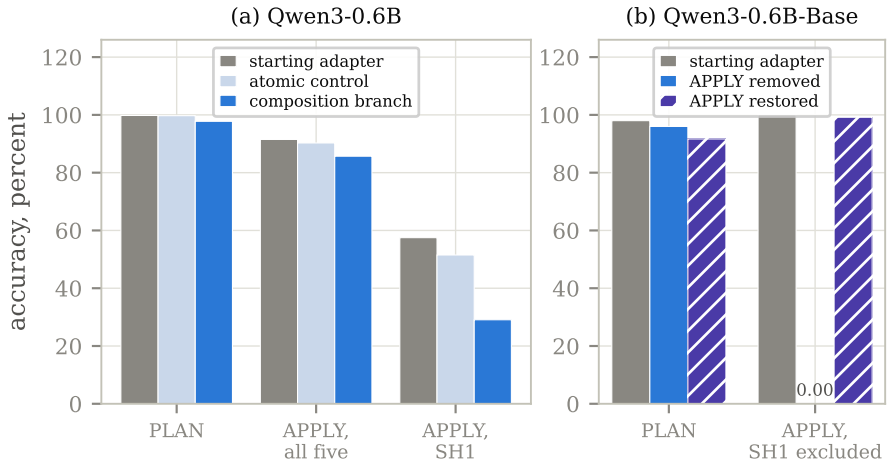

Figure 4: Atomic retention. Panel (a) averages APPLY over all five operations. Panel (b) uses a Qwen3-0.6B-Base split excluding SH1 and a different training mixture. Each panel supports comparisons within its own setting.

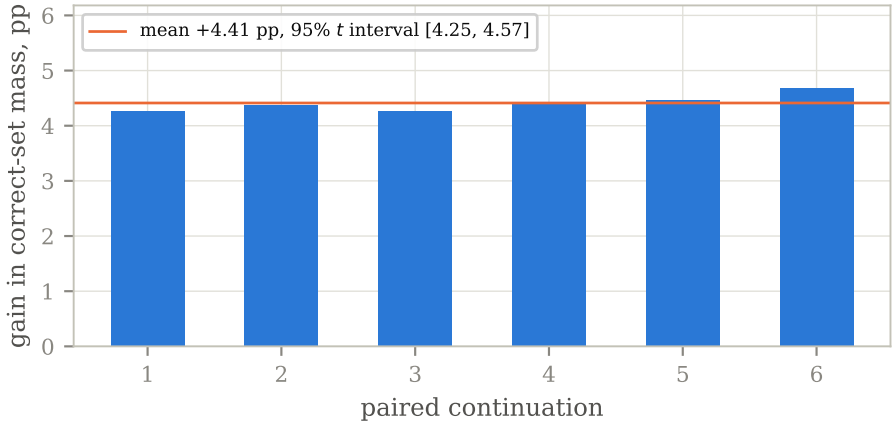

Figure 5: Correct-set mass gain within the 125-program space on 1000 family-A tasks. The mean and interval use six paired continuations from one atomic checkpoint.

C ATOMIC SKILLS AND DATA INTEGRITY

The full re-check covered 2080 shard files of rankings, verifying for every task the completeness of the enumeration of 25, 125 or 625 programs, the correctness of the labels and the sort order. The leak audit gave zero intersections, and the training budgets inside each pair of branches were equal in

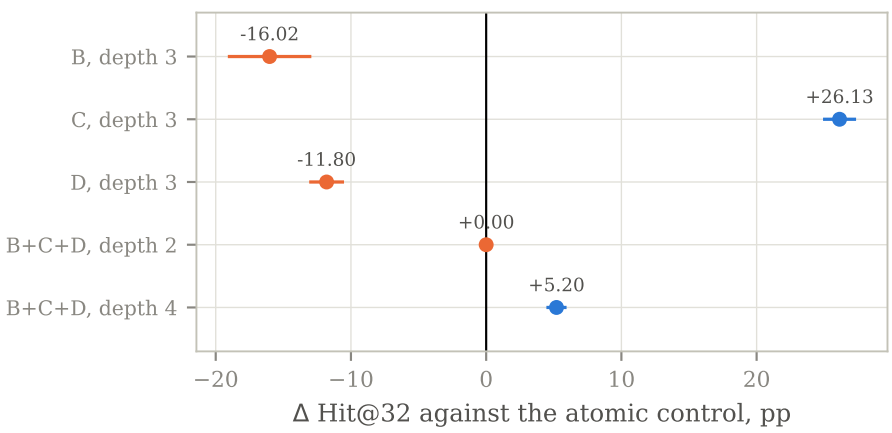

Figure 6: Secondary measures of transfer with their 95 percent intervals, as differences against the atomic control. Family B holds local motifs under familiar fields, C uses familiar signatures in new fields, and D combines the two. The depth-2 row is structurally saturated, so its zero carries no information. After the Holm correction the three depth-3 rows and the depth-4 row have p values of 8.487×10^{-5}, 1.879×10^{-7}, 1.024×10^{-5} and 2.976×10^{-5}.

Check	**Result**
six paired continuations completed	yes
matched optimiser steps and target-token counts	confirmed
identical tasks for branches and runs	confirmed
number of leaks found	0
re-checked ranking shards	2080
completeness of the depth 2, 3 and 4 spaces	25, 125 and 625 programs
re-check of labels and of the sort order	passed
analysis plan fixed before final evaluation	yes
pilot adapters used in the confirmation	no
random generation at the main evaluation	not used
correction for multiple secondary tests	Holm
negative secondary results published	yes

Table 14: Technical audit and reproducibility control of the confirmatory series. The full re-check covered 2080 shard files of raw rankings, and for every task it re-verified the completeness of the enumeration, the labels of the verifier and the sort order.

optimiser steps, epochs, effective batch size and loss-bearing target tokens (Table 14). Acceleration on three RTX 5080 cards was applied only to a part of the secondary rankings, and those data were reprocessed by the frozen evaluation code. Task-level metrics from all twelve family-A checkpoints reproduce the six paired Hit@32 results and provide the full curves and paired task summaries in Appendix F.

D Full list of limitations

1. The main six-run result concerns Qwen3-0.6B and full trajectory supervision, conditional on one shared atomic checkpoint. The 8B comparison uses a shared development set and three continuation seeds from one checkpoint per model size.
2. Transfer is defined by the tested shifts: new states, arithmetic fields, local motifs, and program depth. The synthetic environment has a fixed small operation vocabulary and an exact interpreter.
3. The full-trajectory treatment jointly changes prompt depth, operation sequences, and intermediate-state supervision. Its measured effect applies to that combined intervention. The uniform-ranking reference describes task difficulty. Isolating format exposure requires a control matched on multi-operation prompts.

4. Atomic retention depends on operation and recipe. SH1 is the weakest APPLY operation in the six-run comparison. The development checkpoints also vary in their starting atomic accuracy. Intermediate-state supervision, distribution shift, and replay frequency remain candidate explanations for the retention losses.
5. Six main runs and three seeds per intervention give coarse exact sign-flip tests. Seed-level intervals condition on the shared initialisation and data.
6. The reward-training evidence concerns the reported budgets and algorithms. Empirical entropy of eight sampled answers and entropy of the distribution over 125 programs are separate measurements.
7. The withheld-pair study reports the 13 completed subsets, with three seed pairs per subset. GPU and subset composition vary across blocks. The task stratum includes tasks with an alternative correct program that avoids the withheld pairs. Appendix G.2 gives strict-stratum bounds. The removal rule also changes operation marginals, program signatures, and intermediate states.
8. The trajectory-contamination levels use one training seed each. The boundary controls of Appendix E reuse already inspected sets and are post-hoc.
9. The score-prior correction uses development tasks and a post-hoc choice of ranking rule. The paired START/TARGET contrasts provide a separate check of input dependence. The normalisation check uses saved scores from the development matrix. Re-scoring the six-run checkpoints with a five-action normaliser requires per-prefix logits beyond their stored program scores.

E SEARCH, REWARD, AND SUPERVISION EXPERIMENTS

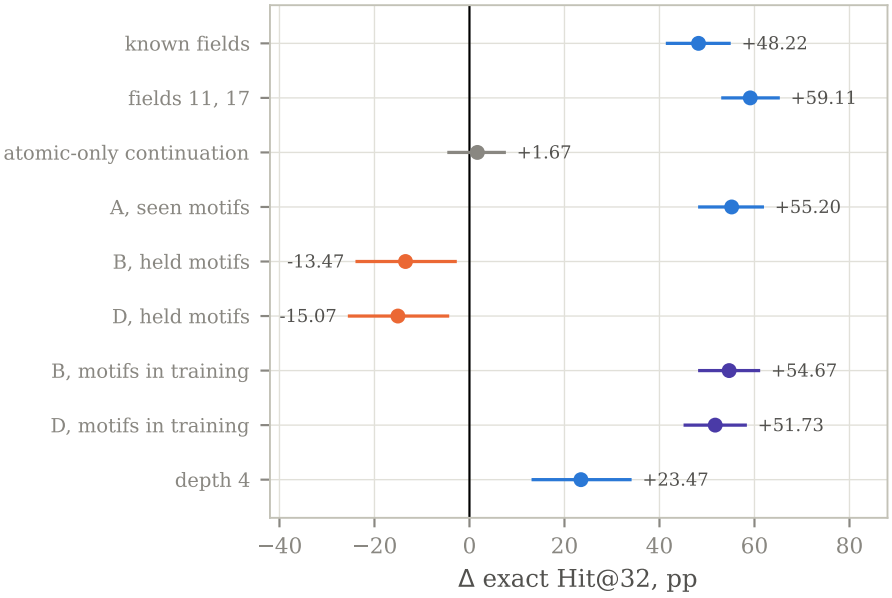

Figure 7: Qwen3-0.6B-Base supervision contrasts in exact Hit@32 with their 95 percent crossed seed-by-task intervals. Rows in the darker colour are measured against the shared starting checkpoint. The grey row compares atomic-only continuation with that start. Rows labelled "B, motifs in training" and "D, motifs in training" evaluate adapters whose supervision contained the withheld pairs.

These experiments use Qwen3-0.6B-Base, three training seeds, 300-task depth/field splits, and 250-task motif splits. E1/E2 intervals use 10000 bootstrap draws: paired task differences are resampled independently within each seed and then averaged over the three fixed seeds. They quantify within-seed task variation. The exact-ranking analyses in E3–E5 use crossed resampling of seeds and task

indices, with one common task draw across selected seeds. Supervised runs take 1875 optimiser steps at batch size 8 and learning rate 2×10^{-4}. Policy-gradient runs take 400 steps with group size 8, learning rate 10^{-5}, and Kullback–Leibler coefficient 0.01. Atomic SFT uses fields 5 and 7, while reward training uses all six training fields listed in Section 3.

The scoring rules differ: the six-run comparison reads the log-probability of an operation off a softmax over the full vocabulary with five added operation tokens, while these experiments take a softmax over the five admissible continuations alone. Both sum log-probabilities along a program. Restricting the normaliser changes prefix-dependent terms and can reorder programs, so the two scores define distinct evaluation rules. Temperature divides each next-operation token-score sum before the five-action softmax at each prefix. IID samples 32 programs with replacement. Prefix-balanced evaluation forces seven starts with SH1 and six with each other operation, leaving one candidate unconstrained. Suffixes use the local policy at $T = 0.7$. Prefix-balanced GRPO groups force one start per operation and leave three of eight candidates unconstrained.

Q1 compares prefix-balanced search with IID at $T^* = 2$, selected to approximate development diversity (9.63 versus 9.88 distinct programs). These policies are frozen for holdout evaluation, where the respective means are 10.69 and 12.03. Q3 changes training and search together: prefix-trained with prefix-balanced search versus IID-trained with IID search, all at $T = 0.7$. Q5 uses IID at $T = 8$ during training and evaluation, averaging coverage over three seeds on a separate 300-task holdout. Q6 compares training at 8 and 0.7 under evaluation at 8. Q4 and Q8 compare initial-checkpoint IID at $T = 3$ and $T = 8$, respectively, with prefix-balanced search at $T = 0.7$. Both IID temperatures were selected on development tasks before holdout evaluation.

Exact-key training overlaps affect 14/100 atomic APPLY validation items (eight of 80 non-SH1) and 11/100 PLAN items. Filtering them changes starting-checkpoint full APPLY from 83/100 to 73/86 (84.88 percent), non-SH1 APPLY from 80/80 to 72/72, and PLAN from 98/100 to 87/89 (97.75 percent). Retention tables use the original sets. One of 100 search-development tasks overlaps reward training. Removing it changes mean distinct candidates from 9.630 to 9.646 and preserves $T^* = 2$. Both 300-task E1/E2 holdouts have zero exact-key overlap with atomic and reward training and are used in full.

Studies E1 and E2 used protocols fixed before their holdout evaluation. E3 prespecified sampled pass@32. Its exact-ranking endpoint has uncertain prespecification status in the protocol records. The sampled contrasts are $+63.33$ and $+62.56$ percentage points, and the exact-ranking contrasts are $+48.22$ and $+59.11$. E4 evaluates three withheld pairs chosen using observed results, on independently generated tasks. Both exact-ranking analyses are exploratory. The boundary controls reuse their evaluation sets and are post-hoc.

Table 15 gives protocol status, Tables 16–18 search and reward results, Tables 19–20 supervision and pair-exposure results, and Tables 21–22 retention and depth-four transfer.

	Intervention	**Evaluation**	**Status**
E1	search with development-calibrated diversity	300 tasks, 3 seeds	confirmatory
E2	400 steps of GRPO with an exploring sampler	300 tasks, 3 seeds	confirmatory
E3	supervision on correct depth-2 and depth-3 programs	4×300, 3 seeds	exploratory
E4	removal of three operation pairs from supervision	4×250, 3 seeds	exploratory
E5	step- and example-matched control, motif-present control, APPLY mixture, depth 4, trajectory contamination	reused sets, 125 new tasks	post-hoc

Table 15: The five studies and their protocol status. E1 and E2 used prespecified endpoints and decision rules. An amendment changed E3's primary endpoint from sampled pass@32 to exact Hit@32 and added E4. Its position in time relative to the opening of E3's data cannot be established from the artefacts. E4's pair selection used observed outcomes. E5 collects boundary conditions on sets that E3 and E4 had already opened, together with one newly generated depth-4 split, and is therefore reported as post-hoc. The contamination levels of E5 use one training seed each and the rest of E5 uses three.

For correctness prevalence c, true-accept rate α, and false-accept rate β, acceptance and correctness have covariance $c(1-c)(\alpha-\beta)$. A mixture calculation relates accepted-set contamination to these rates. For proposer accuracy c, true-accept rate $\alpha = 1$, and false-accept rate β, the incorrect fraction

Search arm	pass@32	Distinct
iid, $T = 0.7$, reference control	0.060	3.11
iid, $T = 1.6$	0.190	8.04
iid, $T = 2.0$	0.250	9.88
iid, $T = 3.0$	0.390	14.57
iid, $T = 5.0$	0.460	19.17
iid, $T = 8.0$	0.500	21.92
iid, $T = 100.0$, near uniform	0.420	23.74
temperature mixture	0.090	4.25
prefix-balanced, $T = 0.7$	0.290	9.63

Table 16: Development search sweep, one seed: 100 PLAN tasks over fields 5 and 7, 46 at depth two and 54 at depth three. Distinct averages the number of different programs among 32 draws. Temperature scales each operation's token-score sum before the local softmax. The mixture allocates eight draws each to $T = 0.2, 0.6, 1.0, 1.4$. Temperature 2 approximates prefix-balanced diversity and fixes T^*, while 8 maximises development coverage and fixes the E2 training sampler. Both choices precede holdout evaluation.

Contrast	Δ, pp	95% CI	Positive seeds
E1, structure against diversity			
Q1, prefix-balanced − iid at T^*	-1.11	$[-4.22, +1.89]$	0/3
Q2, GRPO effect, iid branch	-0.11	$[-0.89, +0.67]$	1/3
Q2, GRPO effect, prefix branch	-1.89	$[-3.78, +0.00]$	0/3
Q3 joint training/search contrast, 300 tasks	$+9.00$	$[+6.89, +11.11]$	3/3
Q3 joint training/search contrast, 200 tasks	$+16.33$	$[+13.17, +19.50]$	3/3
Q4, iid at $T = 3.0$ − prefix-balanced	$+9.78$	$[+6.44, +13.00]$	3/3
E2, self-training			
Q5, after GRPO − its starting point	$+0.22$	$[-3.11, +3.67]$	2/3
Q6, exploring − cold training sampler	$+0.89$	$[-2.45, +4.22]$	2/3
Q7, iid at $T = 100$ − prefix-balanced	$+11.22$	$[+7.33, +15.11]$	3/3
Q8, iid at $T = 8.0$ − prefix-balanced	$+12.67$	$[+8.89, +16.56]$	3/3

Table 17: Prespecified contrasts in sampled pass@32, three training seeds. The two frozen exploration sets carry 300 depth-3 PLAN tasks each over fields 11 and 17. The row marked as the 200-task set uses a separate frozen set, re-analysed and not re-collected. Intervals use 10000 bootstrap draws of paired task differences, independently within each seed, averaged over the three fixed seeds. The final column counts positive treatment-minus-control seed differences, irrespective of the pooled estimate's sign. Q1 and Q5 are the primary endpoints. Both intervals contain zero.

Adapter	Accuracy	Distinct	pass@32
before GRPO	0.0151	3.80	0.044
after GRPO, $T_{\text{train}} = 0.7$	0.0139	2.49	0.038
after GRPO, $T_{\text{train}} = 8.0$	0.0358	1.81	0.061

Table 18: Candidate accuracy, diversity, and coverage under IID evaluation at $T = 0.7$, averaged over three seeds on the 300-task E2 holdout. Accuracy is the fraction of verified correct candidates, and Distinct counts different programs among 32 draws. Q5 compares the initial and $T_{\text{train}} = 8$ checkpoints at $T_{\text{eval}} = 8$, with coverage 32.22 versus 32.44 percent. Their change under cold evaluation here is 1.67 percentage points before rounding.

among accepted candidates is

$$\rho = \frac{\beta(1-c)}{c + \beta(1-c)}, \qquad \beta = \frac{\rho c}{(1-c)(1-\rho)}. \tag{3}$$

At the pre-reward IID sampler's measured $c = 0.0151$ at temperature 0.7 (Table 18), the three observed replacement fractions correspond to $\beta \approx 0.0017, 0.0052, 0.0152$ in this calculation. Uniform proposals on single-solution tasks have $c = 1/125$, giving $0.0009, 0.0027, 0.0080$. Thus a weaker proposer requires a smaller false-accept rate to keep the accepted set equally clean. The mapping matches aggregate contamination. The distribution of incorrectly accepted trajectories would determine its training effect. Our exact verifier has $\beta = 0$, including for policy-generated candidates. For that setting, the reward-group measurements describe the separate constraint on obtaining a useful training signal from accepted candidates.

Contrast	Endpoint	Δ, pp	95% CI
E3, supervision minus shared starting checkpoint			
known fields	Hit@32	+48.22	[+41.33, +55.00]
fields 11 and 17	Hit@32	+59.11	[+53.00, +65.33]
E3, evaluation-condition contrasts after supervision			
known − held-out fields, depth 3	Hit@32	−2.22	[−6.00, +1.56]
depth 2 − depth 3, known fields	Hit@8	+42.89	[+38.00, +47.67]
E5, atomic continuation minus shared starting checkpoint			
known fields	Hit@32	+0.11	[−6.44, +6.44]
fields 11 and 17	Hit@32	+1.67	[−4.67, +7.67]
E5, composition supervision minus atomic continuation			
known fields	Hit@32	+48.11	[+38.67, +57.11]
fields 11 and 17	Hit@32	+57.44	[+49.67, +65.33]

Table 19: Exact-ranking contrasts for E3/E5, with 300 tasks per split and three training seeds. Depth three ranks all 125 programs, depth two all 25. The depth contrast uses Hit@8, all other rows Hit@32. E5 matches 1875 optimiser steps, learning rate, and example count. Target-token counts can differ. Intervals use crossed seed–task resampling.

Family	Fields, motifs	Hit@32	Δ, pp	95% CI
supervision with the three operation pairs withheld				
A	known, seen	0.940	+55.20	[+48.13, +62.00]
B	known, held	0.253	−13.47	[−24.00, −2.67]
C	11 and 17, seen	0.936	+58.80	[+51.47, +65.73]
D	11 and 17, held	0.265	−15.07	[−25.60, −4.27]
supervision that contained those pairs				
B	known, held	0.935	+54.67	[+48.13, +61.20]
D	11 and 17, held	0.933	+51.73	[+45.07, +58.40]

Table 20: The motif experiment, 250 depth-3 tasks per family and three training seeds. A task enters a held family only when every minimal-depth correct program for it contains one of the three withheld pairs, so an unheld route cannot solve it. Differences are taken against the shared starting checkpoint, which reaches Hit@32 of 0.388, 0.388, 0.348 and 0.416 on families A to D. The upper block is the E4 extension and the lower block is the control of E5 on the same two evaluation families. The upper-block changes give the descriptive contrasts $(\Delta_A + \Delta_C - \Delta_B - \Delta_D)/2 = +71.27$ pp for seen minus held motifs (95% CI $[+58.53, +83.60]$) and $(\Delta_A + \Delta_B - \Delta_C - \Delta_D)/2 = -1.00$ pp for known minus new fields (95% CI $[-8.27, +6.33]$).

Training mixture	APPLY	Parse	PLAN	Vectors
no composition supervision	1.000	1.000	0.980	80/80
APPLY absent from the stream	0.000	0.000	0.910–0.980	0/480
APPLY at 17 percent of the stream	0.975–1.000	1.000	0.850–0.970	240/240

Table 21: Atomic retention on the original validation panels: 80 non-SH1 APPLY items and 100 PLAN items, with eight and eleven exact-key training overlaps, respectively (Appendix E). APPLY and PLAN give greedy pass@1, and Parse gives the APPLY parse rate. Vectors counts APPLY answers containing a coefficient vector, pooled over six adapters in the middle row and three in the last row. These fixed-panel scores measure retention.

Adapter	Hit@1	Hit@8	Hit@32	Best rank
no supervision	0.000	0.016	0.056	237.6
supervision, seed 0	0.040	0.128	0.336	92.8
supervision, seed 1	0.024	0.104	0.336	88.9
supervision, seed 2	0.016	0.072	0.200	143.2

Table 22: Exhaustive ranking of all 625 programs of depth 4 on 125 tasks over known fields, where supervision had covered depths 2 and 3 only. The last column is the mean rank of the best correct program. The pooled difference in Hit@32 is +23.47 pp, 95% CI $[+13.07, +34.13]$, and the three seeds are listed separately because they do not agree closely.

F Full ranking curves and paired task outcomes

The task-level reanalysis uses all six atomic-control checkpoints and all six composition checkpoints. The source audit checked 480 ranking shards against their hashes and reproduced the reported Hit@32 for each checkpoint. The released task metrics contain the best correct rank, the

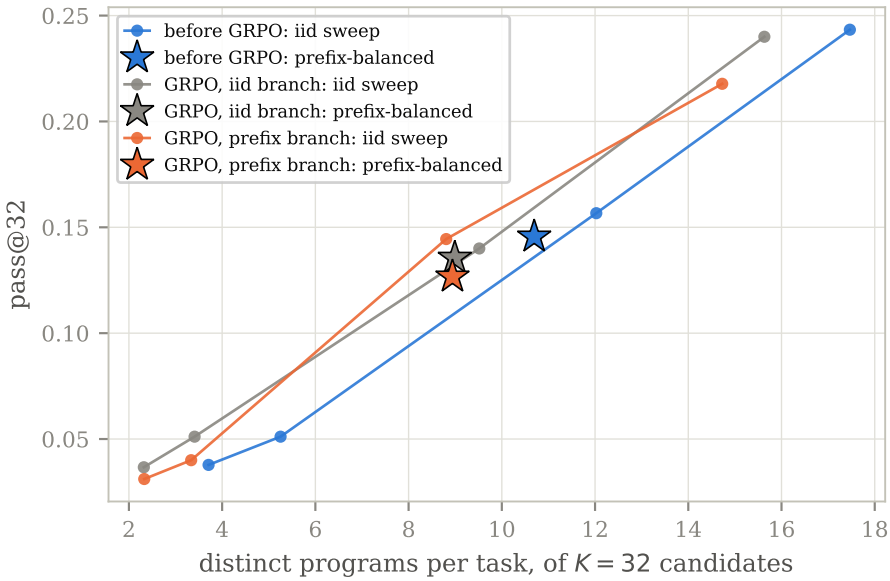

Figure 8: Coverage of the prespecified search holdout against realised candidate diversity, 300 depth-3 PLAN tasks over fields 11 and 17 at $K = 32$, averaged over three seeds. The prefix-balanced points (stars) occupy a similar coverage–diversity region to the IID temperature sweeps.

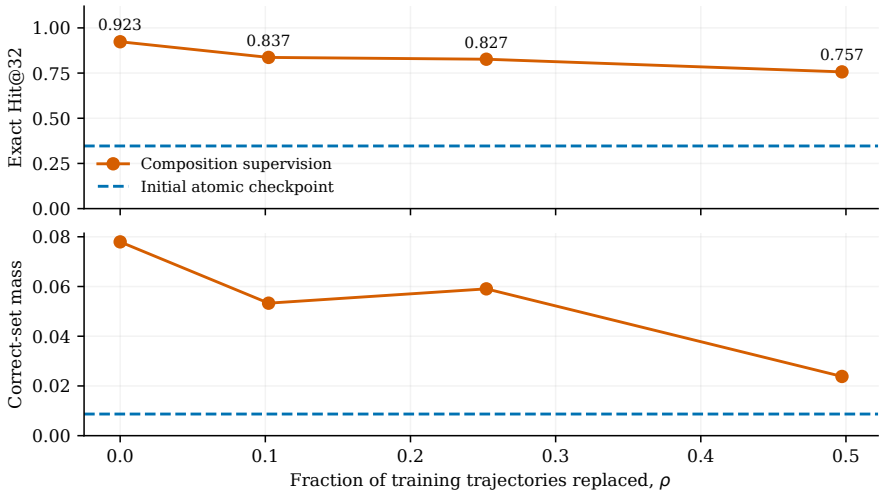

Figure 9: Composition supervision with a fraction ρ of correct training trajectories replaced by verified incorrect programs of the same length. Realised replacement fractions are 0, 10.225, 25.225, and 49.725 percent of 4000 examples. The task and example counts stay fixed. Evaluation uses 300 depth-3 tasks over fields 11 and 17. At the highest replacement level, Hit@32 is 75.67 percent, versus 92.33 with correct labels and 34.67 at initialisation. The intervention controls training-set contamination, not the conditional false-accept probability β of a verifier. One training seed per level. This figure is exploratory.

number of correct programs, and distributional summaries for each of 12000 task–checkpoint pairs. The tables below are recomputed from those task metrics. The complete 125-point curves are included in the source package.

Run	Atomic (%)	Comp. (%)	Gained	Lost	ΔAPPLY (pp)
1	43.5	73.7	432	130	-4.9
2	45.7	72.4	408	141	-5.8
3	45.4	71.4	409	149	-5.1
4	46.0	75.2	409	117	-2.8
5	45.9	72.9	408	138	-4.7
6	45.4	70.7	382	129	-4.2

Table 23: Task transitions at Hit@32 on the same 1000 family-A tasks in each paired continuation. The final column is composition minus atomic-control APPLY accuracy. The post-hoc joint check requires at least a five-point Hit@32 gain and an APPLY loss of at most two points relative to that control. All six gains exceed five points, and all six APPLY losses exceed two.

Correct programs	Tasks	Uniform (%)	Atomic (%)	Comp. (%)	Δ (pp)
One	566	25.60	48.79	64.96	+16.17
Multiple	434	56.54	40.78	82.83	+42.05

Table 24: Exploratory stratification of the six-run family-A result by the number of exact correct programs. Values average within a run and then across six runs. The difference is composition minus atomic control. Subtracting the same uniform reference from both arms leaves it unchanged.

With m_x correct programs for task x, a uniform random permutation of all 125 programs has expected Hit@K, averaged over the same task set T,

$$H_{\text{unif}}(K) = \frac{1}{|T|} \sum_{x \in T} \left[1 - \frac{\binom{125-m_x}{K}}{\binom{125}{K}} \right], \tag{4}$$

with $\binom{a}{K} = 0$ for $K > a$. The expected best-correct rank is the task mean of $126/(m_x + 1)$: 50.16 for this test, compared with observed means of 44.13 for atomic control and 22.25 after composition training.

For tasks with one or multiple correct programs, the uniform Hit@32 references are 25.60 and 56.54 percent. Composition scores exceed them by 39.36 and 26.29 points. These comparisons describe absolute ranking performance. The paired composition-minus-atomic differences remain 16.17 and 42.05 points.

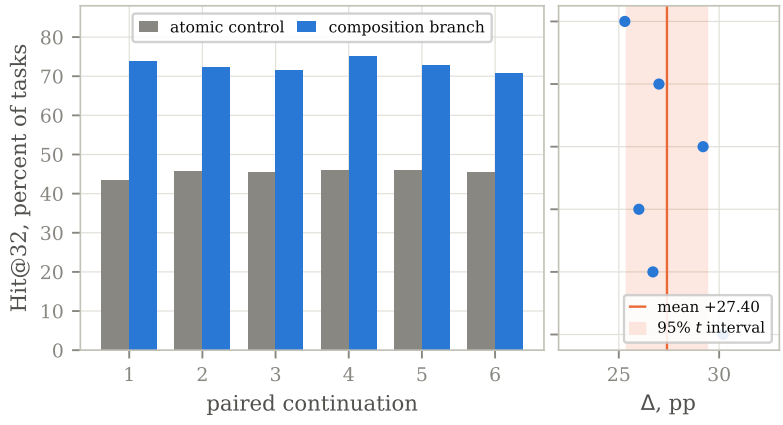

Figure 10: Left, Hit@32 of the two branches in each of the six paired continuations, on the same 1000 tasks of family A at depth three. Right, the six per-run differences, the mean of 27.40 percentage points and the 95 percent t interval computed over them. The spread runs from 25.3 to 30.2 percentage points.

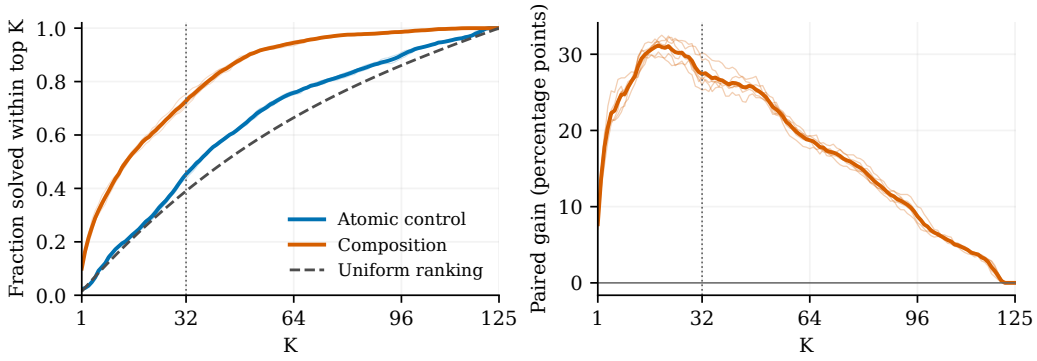

Figure 11: Complete Hit@K curves for the six-run family-A test. Left: fraction of tasks with a correct program in the first K positions. Right: composition minus atomic control, in percentage points. Thick lines are means. Thin lines are individual paired runs. The dashed curve is the exact uniform-ranking reference. The dotted line marks the prespecified $K = 32$. All tasks have an exact solution, so all curves reach one at $K = 125$.

Solutions	Arm	Hit@1	Hit@8	Mass	Median mass	exp H	Rank
One	Atomic	4.06	15.93	3.960	2.44×10^{-6}	2.46	42.93
One	Comp.	6.89	25.53	3.276	1.48	56.84	27.68
Multiple	Atomic	0.00	6.91	0.009	5.51×10^{-7}	2.27	45.70
Multiple	Comp.	14.06	51.34	11.068	5.08	52.53	15.17

Table 25: Six-run family-A distributional diagnostics by solution multiplicity. Hit rates and mass are percentages. Means weight the six runs equally. Medians pool task–run pairs. Rank is best correct rank, and exp H is the effective number of programs. Each run has 566 unique-solution and 434 multiple-solution tasks. All summaries are post-hoc.

For scores $L_\theta(p \mid x)$, the candidate-conditional distribution is

$$q_\theta(p \mid x) = \frac{\exp L_\theta(p \mid x)}{\sum_{p' \in \mathcal{P}_3} \exp L_\theta(p' \mid x)}, \qquad H_\theta(x) = -\sum_{p \in \mathcal{P}_3} q_\theta(p \mid x) \log q_\theta(p \mid x), \tag{5}$$

where $|\mathcal{P}_3| = 125$. Entropy and total correct mass refer to this normalised finite set. Mean exp H is 2.38 for atomic control and 54.97 for composition training. Averaging exp H over tasks differs from exponentiating the average entropy. These are endpoint measurements.

Table 25 separates unique- and multiple-solution tasks. For unique solutions, mean correct mass decreases from 3.96 to 3.28 percent, while Hit@8 rises from 15.93 to 25.53 percent. Mean mass is sensitive to a small number of concentrated correct predictions: its median is 2.44×10^{-6} percent under atomic control and 1.48 percent after composition training. For multiple solutions, mean correct mass rises from 0.009 to 11.07 percent. Atomic-control Hit@1 is zero in the multiple-solution stratum, while composition Hit@1 is 14.06 percent.

Thus increased entropy and improved ranking coexist in both strata, with different changes in mean correct mass. A more diffuse distribution can rank a correct program higher while assigning it less probability. These task-level summaries measure concentration within a task. Identifying a shared top program across tasks or estimating a program-specific score shift requires the saved program identities and scores. Those arrays are separate from the task-summary release used for this analysis.

G Training-data and reward experiments

The training-data and reward studies comprise 96 completed trainings: twelve for trajectory selection, 78 for withheld pairs, and six for the entropy bonus. Each study froze its data selection, endpoint, and analysis procedure before the corresponding outcome evaluation.

Study	Seed	Control (%)	Test (%)	Δ (pp)
Selection	83000	61.9	63.6	+1.7
Selection	83001	61.9	59.9	-2.0
Selection	83002	61.4	60.9	-0.5
Selection	83003	61.4	59.4	-2.0
Selection	83004	62.1	60.9	-1.2
Selection	83005	60.6	63.1	+2.5
Entropy	0	49.6	47.7	-1.9
Entropy	1	50.3	49.6	-0.7
Entropy	2	48.4	49.3	+0.9

Table 26: All paired Hit@32 outcomes for the trajectory-selection and entropy-bonus experiments.

G.1 Trajectory selection

Seeds 83000–83005 each select two 750-trajectory sets from a frozen pool of 2250. Each selected set contains 250 examples at depths two, three, and four. Quotas also match field and target-answer length under the frozen tokenizer. The structured rule greedily maximises an equally weighted pair-balance and state-coverage objective. Pair balance is

$$J_{\text{pair}} = \frac{\sum_{a,b} \log(1 + n_{ab})}{P \log(1 + 1500/P)}, \tag{6}$$

where P is the number of pair types in the pool and 1500 is the number of adjacent-pair occurrences in a balanced selection. State coverage counts classes $(p, \lfloor 4c_0/p \rfloor, \ldots, \lfloor 4c_d/p \rfloor)$ among intermediate states, normalised by the number of classes in the pool. The structured-minus-random objective differences are 0.0630–0.0684, above the frozen 0.03 threshold.

Both arms use the shared Qwen3-0.6B atomic checkpoint, two epochs, learning rate 10^{-4}, LoRA rank 32, scale 64, dropout 0.05, and effective batch 64. With replay, each training set contains 938 records and yields 30 optimiser steps. Actual loss-bearing target tokens match within a pair (87302–87480 across seeds). A new family-A holdout has 1000 depth-three tasks. The first four pairs were evaluated on RTX 5060 Ti and the last two on V100, with both arms of each pair evaluated in FP32 on the same platform. A repeated evaluation of one random-selection checkpoint agreed in Hit@32 across platforms.

Mean candidate-conditional entropy is 3.946 nats for random and 3.970 for structured selection. Total correct mass is 0.0392 and 0.0420. These are exploratory endpoints. The mean atomic PLAN drop from initialisation is largest for REV (51.0 and 51.4 points), and SH1 APPLY drops by 26.2 and 27.0 points. Table 26 reports all paired primary outcomes.

G.2 Withheld-pair study

The frozen design selected five subsets for each $k \in \{1, 2, 3, 5\}$, using draw seed 84000. For every subset, the selective arm removes trajectories containing an excluded pair. The control removes the same number at random within each depth/field cell. Both arms then select 250 trajectories at each depth, matching field and target-token quotas, and use identical seed-specific atomic replay. The training seeds are 85000–85002. Training uses FP16 autocast, FP32 master weights, and dynamic loss scaling, with the same learning rate, LoRA settings, epoch count, and effective batch as the trajectory-selection study.

The report includes the 13 completed subsets, totaling 78 trainings out of the 120-row planned matrix. A further six rows correspond to the subset $k = 5$, set 5, which was infeasible under the fixed matching rule. The remaining 36 rows were unstarted at the cutoff. This accounts for the denominator of the partial series. Subset completion followed the frozen queue and GPU availability. The prospective hardware amendment bound both arms and all three seeds of a subset to one physical GPU. Table 27 identifies the V100 and RTX 3060 blocks.

Each subset has two new 1000-task strata. The withheld stratum requires at least one shortest correct depth-three program containing an excluded pair. The control stratum requires every correct program to avoid those pairs. Tasks may therefore have alternative correct programs in the withheld

k/set	GPU	Seed 0	Seed 1	Seed 2	Mean	95% t interval
1/1	V100	-36.9	-25.6	-24.5	-29.00	[-46.1, -11.9]
1/2	V100	-32.6	-45.8	-51.9	-43.43	[-67.9, -18.9]
1/3	RTX 3060	-16.3	-11.6	-4.3	-10.73	[-25.8, 4.3]
1/4	V100	-7.1	-3.1	-5.9	-5.37	[-10.5, -0.3]
1/5	RTX 3060	-24.3	-26.5	-25.1	-25.30	[-28.1, -22.5]
2/1	V100	-23.1	-24.9	-22.8	-23.60	[-26.4, -20.8]
2/2	V100	-19.6	-22.4	-24.6	-22.20	[-28.4, -16.0]
2/3	V100	-32.8	-38.0	-30.5	-33.77	[-43.3, -24.2]
2/4	RTX 3060	-21.3	-23.4	-15.1	-19.93	[-30.7, -9.2]
2/5	V100	-7.0	-10.9	-14.1	-10.67	[-19.5, -1.8]
3/1	V100	-9.6	-15.4	-16.2	-13.73	[-22.7, -4.8]
3/2	RTX 3060	-32.8	-32.3	-8.3	-24.47	[-59.3, 10.3]
3/3	V100	-6.8	-7.7	-16.2	-10.23	[-23.1, 2.7]

Table 27: Withheld minus matched-random-removal Hit@32, in percentage points, using full-vocabulary scores. Seeds are 85000–85002. All 39 paired differences are negative. Each subset has exact two-sided sign-flip $p = 0.25$. Intervals are descriptive, conditional on the shared initialisation. Set 1/1 precedes the hardware amendment. All arms within a subset use the same physical GPU.

k/set	GPU	Seed 0	Seed 1	Seed 2	Mean	95% t interval
1/1	V100	+10.6	+7.4	+7.4	+8.47	[3.9, 13.1]
1/2	V100	+8.7	+10.2	+8.8	+9.23	[7.2, 11.3]
1/3	RTX 3060	-2.0	-1.0	+0.6	-0.80	[-4.1, 2.5]
1/4	V100	+1.0	+1.1	+4.2	+2.10	[-2.4, 6.6]
1/5	RTX 3060	+3.8	+2.9	+4.7	+3.80	[1.6, 6.0]
2/1	V100	+5.5	+2.5	+3.0	+3.67	[-0.3, 7.7]
2/2	V100	+6.9	+2.0	+3.5	+4.13	[-2.1, 10.4]
2/3	V100	+10.1	+9.1	+9.3	+9.50	[8.2, 10.8]
2/4	RTX 3060	-1.4	-1.2	+0.8	-0.60	[-3.6, 2.4]
2/5	V100	+4.3	+6.6	+6.2	+5.70	[2.6, 8.8]
3/1	V100	+2.7	+1.3	+1.2	+1.73	[-0.3, 3.8]
3/2	RTX 3060	+7.2	+4.9	+2.6	+4.90	[-0.8, 10.6]
3/3	V100	+7.4	+3.3	+5.9	+5.53	[0.4, 10.7]

Table 28: Prespecified secondary endpoint on the 1000 control-stratum tasks per subset, whose correct programs all avoid the excluded pairs. Values are selective exclusion minus matched random removal, in Hit@32 percentage points. Full-vocabulary scores, seeds and GPUs match Table 27. Intervals and signs are descriptive, conditional on the shared initialisation.

stratum. Table 29 reports their frequency. A stricter all-solutions rule was relaxed before training after generation proved infeasible for a preselected pair. Pair sets and outcomes stayed fixed. Task and raw-state fingerprints are disjoint from training and previous evaluation files under the source audit.

The strict stratum contains only tasks whose every correct program requires an excluded pair. Its observed effect can be bounded using the total successes, even when successes are retained only in aggregate. Let $N = 1000$, let A be the number of tasks with an alternative correct program, and let H_b be the number of successes in arm b. For $A < N$, the strict success rate satisfies

$$\ell_b = \frac{\max(0, H_b - A)}{N - A} \le h_b^{\text{strict}} \le \frac{\min(H_b, N - A)}{N - A} = u_b, \quad \Delta^{\text{strict}} \in [\ell_{\text{sel}} - u_{\text{rnd}}, u_{\text{sel}} - \ell_{\text{rnd}}]. \tag{7}$$

These bounds allocate successes to the alternative-solution tasks in the most favourable and least favourable ways for each arm. They are sharp given these counts. Table 29 averages the bounds over seeds. In subsets 1/1, 1/2, 1/5, and 2/3 the upper bound is negative in every seed, so alternative solutions alone cannot account for the observed losses there. Across the 13 subsets, the descriptive Spearman correlation between alternative-solution share and the overall Hit@32 difference is 0.713. This association mixes pair sets and hardware strata.

k/set	Excluded adjacent pairs	Strict n	Strict Δ bounds (pp)
1/1	$AX1 \to REV$	965	[-33.68, -26.42]
1/2	$REV \to SH1$	884	[-62.25, -36.01]
1/3	$SH1 \to SC2$	722	[-53.37, 23.64]
1/4	$SH1 \to AC1$	0	Empty
1/5	$REV \to SC2$	862	[-45.36, -13.34]
2/1	$AC1 \to AX1$, $SH1 \to AX1$	541	[-99.94, 41.22]
2/2	$AX1 \to AC1$, $REV \to AC1$	489	[-100.00, 57.81]
2/3	$SC2 \to SC2$, $SH1 \to REV$	881	[-51.84, -24.82]
2/4	$AC1 \to AX1$, $SC2 \to AC1$	0	Empty
2/5	$AX1 \to AC1$, $REV \to AX1$	510	[-80.00, 59.08]
3/1	$REV \to AC1$, $SC2 \to AC1$, $SH1 \to AX1$	621	[-69.46, 38.92]
3/2	$SC2 \to AC1$, $SC2 \to SH1$, $SH1 \to SC2$	642	[-74.56, 16.67]
3/3	$AX1 \to AC1$, $SC2 \to REV$, $SH1 \to SC2$	633	[-73.09, 41.81]

Table 29: Frozen pair sets and post-hoc bounds on selective-minus-random Hit@32 when every correct program contains an excluded pair. Strict n is out of 1000 withheld tasks. The other $1000-n$ tasks have a correct program avoiding every excluded pair. Bounds average the sharp count-based bounds across three seeds (Equation 7). They describe these observed tasks, with no sampling or independence assumption. They are identification bounds, separate from the seed intervals in Table 27.

k/set	W: Δ@1	W: Δ@8	W: ΔM	C: Δ@1	C: Δ@8	C: ΔM
1/1	-3.30	-11.93	-1.13	+2.13	+5.07	+1.85
1/2	-4.70	-27.90	-1.32	+2.43	+6.77	+1.31
1/3	-2.73	-8.20	-0.50	+0.60	+0.33	+0.17
1/4	+0.53	-2.73	-0.51	-0.60	+0.03	-0.03
1/5	-0.03	-11.50	-0.62	+1.97	+3.70	+0.76
2/1	-11.13	-16.30	-3.83	-2.17	-3.87	-0.35
2/2	+0.77	-9.17	-1.00	-0.80	-0.13	-0.06
2/3	-0.60	-8.93	-0.91	+1.67	+4.57	+1.05
2/4	-7.27	-24.50	-3.76	-2.33	-1.70	-0.51
2/5	-4.20	-2.93	-1.28	+1.17	+1.50	+0.43
3/1	-0.80	-5.03	-0.61	+0.50	-0.27	+0.44
3/2	-1.53	-6.70	-0.85	-1.47	-0.10	+0.26
3/3	+0.13	-4.80	-0.24	+1.20	+4.07	+0.41

Table 30: Secondary endpoint differences: selective exclusion minus matched random removal, in percentage points, averaged over three paired seeds within each subset. W and C denote the withheld and control strata. M is total correct probability conditional on the 125-program space, from full-vocabulary scores. These endpoints are descriptive. Per-seed values accompany the source package.

The secondary endpoints in Table 30 use the same tasks and checkpoints. Withheld-stratum Hit@8 and correct-set mass fall in every completed subset. Hit@1 falls in ten. The control-stratum signs vary with the endpoint. Subsets 1/4 and 2/4 have an alternative correct program on every withheld task, so their strict all-solutions-require-a-pair strata are empty.

The sign pattern across 39 paired differences is descriptive. The paired t intervals assume the three seed differences are a useful sample within each subset. The corresponding exact sign-flip test has only eight sign assignments, and its minimum two-sided p value is 0.25. GPU and subset differences are retained in the reporting unit. Grouping by the number of excluded pairs and the GPU gives six observed blocks, with mean effects from -25.93 to -11.98 points, using the subset means in Table 27. Each pair set occurs on one GPU, confounding pair identity and hardware. The shared pool and reused seeds further couple the subsets. These are descriptive block means for the completed matrix.

Removing a pair also changes operation marginals, program signatures, and intermediate states. The frozen selection records identify the chosen trajectories by ID and hash. The released records omit the corresponding operation sequences, so they support selection-integrity checks but do not

Sampler	Groups	All wrong	Mixed	All correct	Entropy
IID	1200	1121	38	41	0.298
Prefix-balanced	1200	896	304	0	1.493

Table 31: E1 reward-training logs: IID and prefix-balanced arms at $T_{\text{train}} = 0.7$, each with seeds 0–2 and 400 steps. Each group contains eight answers. Entropy is empirical entropy within a sampled group, in nats.

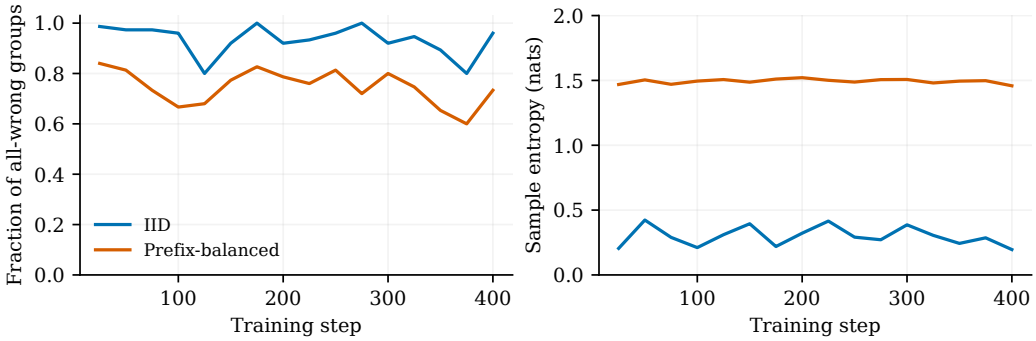

Figure 12: E1 reward-training runs from Table 31, both samplers at $T_{\text{train}} = 0.7$. Consecutive 25-step blocks are averaged over three seeds per sampler. Left: fraction of groups with eight incorrect answers. Right: empirical entropy of the eight sampled answers, grouping identical parsed programs. This entropy describes sampled groups. The candidate-conditional entropy is calculated separately from exact scores.

determine operation or signature frequencies. The experiment estimates the effect of the selection rule on the resulting training set.

G.3 Reward-group audit and entropy bonus

At the first and last 25-step blocks, IID all-wrong fractions are 0.987 and 0.960. Prefix-balanced fractions are 0.840 and 0.733. Mean group reward rises from 0.0017 to 0.0250 for IID and from 0.0200 to 0.0450 for prefix-balanced sampling. Empirical entropy at those endpoints is 0.205 and 0.196 nats for IID, and 1.469 and 1.458 for prefix-balanced sampling. These logs support a sparse within-group reward signal and a sampler-dependent diversity profile.

The entropy experiment uses Qwen3-0.6B-Base with its own atomic adapter, a 500-task reward-training set, and a new 1000-task depth-three holdout over fields 11 and 17. Seeds 0–2 each form a pair. The objective uses standardised within-group binary rewards, a KL coefficient of 0.01, and an operation-choice entropy bonus of zero or $0.01H/\log 5$. Learning rate is 10^{-5}, group size eight, and temperature 0.7. The implementation has no clipping. All six runs use the same physical RTX 2080 Ti and FP32 computation with activation checkpointing. A technical benchmark fixed the common 150-step duration before full training and holdout evaluation. Exact ranking uses the categorical action policy at temperature 0.7.

Across 450 groups per arm, the control has 388 all-wrong, 55 mixed, and seven all-correct groups. The bonus arm has 385, 62, and three. Mean empirical group entropy is 0.7857 and 0.8572 nats. Conditional entropy over all 125 programs is 1.4353 and 1.5999 nats. The paired Hit@32 differences are -1.9, -0.7, and $+0.9$ points (Table 26).

For a sensitivity calibration, hold the observed standard errors fixed and solve for a two-sided test with size 0.05 and power 0.8. The noncentral t calculation gives 2.76 points for trajectory selection ($n = 6$) and 4.59 for the entropy bonus ($n = 3$). Approximating the search and 400-step reward bootstrap intervals by normal intervals gives 4.37 and 4.85 points, conditional on the three observed runs and the within-seed resampling scheme. These are planning values based on the observed dis-

Model	Family	Uniform	Initial	Atomic	Comp.	Comp.@1	Δ (pp)	95% t interval
0.6B	A	40.08	61.50	58.50	86.83	15.50	+28.33	[23.3, 33.4]
0.6B	B	25.60	31.00	29.83	7.83	0.00	-22.00	[-27.0, -17.0]
0.6B	C	39.44	52.00	51.83	84.83	15.83	+33.00	[28.5, 37.5]
0.6B	D	25.70	31.00	26.17	4.50	0.00	-21.67	[-31.2, -12.2]
8B	A	40.08	48.00	50.67	91.17	25.83	+40.50	[37.2, 43.8]
8B	B	25.60	14.50	19.17	8.33	0.00	-10.83	[-17.7, -4.0]
8B	C	39.44	40.00	40.50	87.00	22.33	+46.50	[45.3, 47.7]
8B	D	25.70	20.00	21.33	7.50	0.00	-13.83	[-17.6, -10.0]

Table 32: Development continuations: three paired seeds per size, 200 tasks per family, full-vocabulary scores. Values are percentages. Comp.@1 is composition Hit@1, and Initial is the shared atomic checkpoint. Intervals condition on each initialisation and development set.

persion. Confidence intervals remain the basis for interpreting the measured effects. The calibration and the interval half-widths describe different quantities.

H Development continuations and model size

The continuation series uses Qwen3-0.6B and Qwen3-8B, with one frozen atomic checkpoint per size and continuation seeds 0, 1, and 2. Training uses BF16, two epochs, LoRA rank 32, scale 64, dropout 0.05, learning rate 10^{-4}, effective batch 64, and 158 optimiser steps. Full-trace composition training uses 20 percent atomic replay by examples. Its paired atomic-only control matches the actual loss-bearing target-token count and optimiser steps. The common development set has 200 tasks in each family A–D. Evaluation uses full-vocabulary operation-token scores and lexicographic tie-breaking.

In the 8B runs, the packed atomic control processes 2.58 times as many forward tokens, including prompts and padding, at the matched target-token budget. Mean initial atomic PLAN/APPLY is 99.20/63.00 percent for 0.6B and 99.10/69.80 for 8B. Final atomic-control versus composition APPLY is 72.33 versus 62.60 percent for 0.6B and 76.20 versus 71.87 for 8B. The source package gives the complete breakdown by operation and checkpoint. Applying the joint criterion of a five-point family-A Hit@32 gain and at most a two-point APPLY loss against the paired atomic control gives zero of three continuations at each size. The paired APPLY losses are 9.4–10.2 points at 0.6B and 3.8–4.9 at 8B. This joint criterion is a post-hoc diagnostic.

At the final checkpoints, candidate-conditional entropy over all 125 programs averages 0.47 nats for atomic control and 3.78 for composition at 0.6B, and 1.43 and 3.57 nats at 8B, respectively, with equal weight for the three continuation seeds and four families A–D.

The audit re-executed correctness labels and checked complete rankings, task identities, file receipts, and budgets for thirteen checkpoints per size: one initial, six midpoint, and six final models, totalling 1.3 million candidate records per size. Table 32 uses three final seed differences. Ranking changes have the same direction at both sizes. The two continuation settings differ in initialisation and starting atomic performance. The initial 8B checkpoint already scores below the uniform reference on B and D (14.50 and 20.00 percent Hit@32). Atomic continuation changes Hit@32 from initialisation by -4.83 to $+4.67$ points across the eight size–family cells. All three arms have zero Hit@1 on B and D at both sizes.

I Program-score priors and conditional diagnostics

I.1 Design and analysis scope

This development series crosses four objectives with four training recipes, two fixed pair masks, and three seeds, giving 96 continuations of one checkpoint. The objectives are operation-token cross-entropy (CE), CE plus a counterfactual START/TARGET loss (CF), CE plus an entropy bonus (H), and their combination. CE uses the full-vocabulary token normaliser during training. CF, H, and the primary evaluation use the local five-action policy. For H, the loss subtracts $0.01H(q)$, where

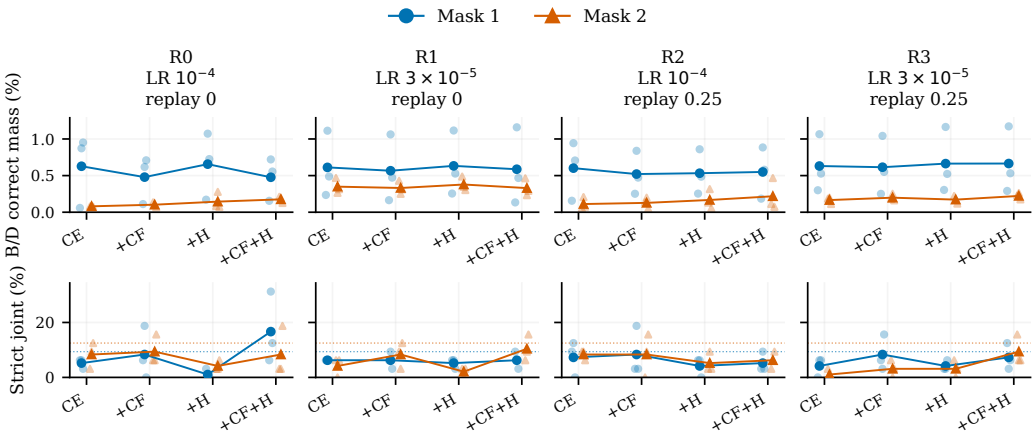

Figure 13: Development continuations at step 128. Solid points average three seeds per mask, with faint points for individual seeds. CE denotes cross-entropy, CF the START/TARGET term, and H the entropy bonus. R0–R3 vary learning rate and replay weight. Dotted lines show initial joint accuracy. Initial mean B/D correct mass is 2.97 percent, above the mass-axis range. Tables 34 and 35 give all paired contrasts and seed intervals. The 96 continuations share one starting checkpoint.

q is the normalised distribution over all 125 programs. CF adds $0.1\,\mathrm{softplus}(-I)$ for the crossed interaction defined below.

R0 and R1 use learning rates 10^{-4} and 3×10^{-5}, respectively. R2 and R3 repeat those rates with an atomic-replay loss weighted by 0.25. All conditions take 128 optimiser steps with the same 512 composition exposures and 1536 supervised operation tokens within each mask–seed pair. Replay adds one PLAN and one APPLY example per step, or 256 atomic exposures. The weight 0.25 is a loss coefficient. Replay therefore adds data and computation at a fixed composition budget. The two masks exclude SC2–REV, SC2–SC2, SH1–SC2 and AX1–REV, SH1–REV, SH1–SC2, respectively. Exclusion is verified in the continuation data. The initial checkpoint's pre-continuation pair exposure is documented by its provenance declaration. The supplement records the frozen streams and per-run budgets.

The analysis plan was fixed and hashed with 36 of the 96 continuation outcomes already available. The designated development contrasts are CF at each H setting, H at each CF setting, their interaction, and replay with the other settings fixed. Designated endpoints are equally weighted B/D correct mass, B/D Hit@32, and strict crossed joint accuracy, using local scores at temperature 1 and step 128. Hit@1/8, A/C, full-vocabulary scores, score corrections, temperatures, and intermediate checkpoints are sensitivity analyses and diagnostics. The nominal 95 percent paired t intervals use three continuation seeds within each mask, with two degrees of freedom (Tables 34 and 35). The supplement gives all seed differences, signs over six mask–seed pairs per recipe, and 5000-draw paired bootstrap intervals conditional on the checkpoints. Resampling uses START clusters within field/family for ordinary tasks and whole crossed panels. The checkpoint, masks, and seeds are shared across conditions.

Each mask has 200 ordinary tasks in each family A–D and 32 B/D diagnostic panels that cross two START states with two TARGET states. In a panel, the diagonal cells have one designated correct program p_0, and the off-diagonal cells have p_1. Strict joint accuracy requires the correct choice between these programs in all four cells. The 128 cells share inputs within panels. Joint accuracy changes in increments of 3.125 percentage points per checkpoint. Cell accuracy and signed interaction accompany this coarse measure (Table 38). All designated joint-contrast seed intervals span zero. Ordinary B/D Hit@1 is zero in all 96 continuations (Table 33).

Switching to full-vocabulary scores across the 96 trained checkpoints shifts mean B/D Hit@32 by at most 0.75 points and strict joint accuracy by at most one panel (3.125 points). For the initial checkpoint, the largest mask-specific changes are 6.50 and 12.50 points, respectively.

Recipe	Loss	Mass	Hit@32	Joint	PLAN	APPLY	R^2_μ
Initial	—	2.97	50.63	10.94	99.45	62.55	—
R0	CE	0.35	18.67	6.77	66.47	18.95	0.539
R0	CE+CF	0.29	18.25	8.85	70.43	18.92	0.555
R0	CE+H	0.40	18.58	2.60	65.35	18.23	0.547
R0	CE+CF+H	0.33	22.54	12.50	66.00	19.35	0.546
R1	CE	0.48	25.17	5.21	82.12	55.52	0.498
R1	CE+CF	0.45	25.04	7.29	81.92	55.92	0.505
R1	CE+H	0.51	25.25	3.65	80.48	57.22	0.501
R1	CE+CF+H	0.46	24.46	8.33	81.95	56.35	0.507
R2	CE	0.36	20.71	7.81	92.60	61.15	0.536
R2	CE+CF	0.32	21.79	8.33	94.10	60.38	0.526
R2	CE+H	0.35	23.04	4.69	95.80	61.10	0.515
R2	CE+CF+H	0.38	24.21	5.73	91.33	60.42	0.508
R3	CE	0.40	21.21	2.60	96.22	61.82	0.516
R3	CE+CF	0.41	23.63	5.73	95.53	61.80	0.517
R3	CE+H	0.42	21.96	3.65	96.27	61.92	0.513
R3	CE+CF+H	0.44	24.33	8.33	95.97	62.00	0.516

Table 33: Development diagnostics for all 96 continuations, averaged over two masks and three seeds per arm. Mass and accuracies are percentages. Ordinary B/D Hit@1 is zero in every trained continuation. Joint is strict success over all four cells of a crossed START/TARGET panel. PLAN and APPLY are atomic accuracies. The post-hoc R^2_μ is the fraction of reference-set score-change variance accounted for by the program-only term (Equation 8). Recipes are defined in Appendix I.1.

I.2 Shared program-score shifts

Let $s_0(x,p)$ and $s_1(x,p)$ be the saved initial and trained local-policy scores. Over 200 ordinary family-A reference tasks, estimate

$$\mu_A(p) = \frac{1}{|A|}\sum_{x\in A}\big[s_1(x,p) - s_0(x,p)\big], \qquad s_{\mathrm{corr}}(x,p) = s_1(x,p) - \mu_A(p). \tag{8}$$

The correction removes each program's mean training-induced change on reference A and retains the initial score structure: $s_{\mathrm{corr}} = s_0 + [s_1 - s_0 - \mu_A]$. Correctness labels enter only the evaluation of the resulting ranks. A second rule centres both checkpoints separately, $s^{\mathrm{ctr}}_\theta(x,p) = s_\theta(x,p) - |A|^{-1}\sum_{x'\in A} s_\theta(x',p)$. The use of a program-specific reference term is related to contextual calibration and domain-conditional PMI scoring (Zhao et al., 2021; Holtzman et al., 2021). Here the reference is an average log score over actual development tasks. For $D(x,p) = s_1(x,p) - s_0(x,p)$, the reference-set variance share is $R^2_\mu = 1 - \mathrm{Var}_{x\in A,p}[D(x,p) - \mu_A(p)]/\mathrm{Var}_{x\in A,p}[D(x,p)]$. It averages 0.521 and ranges from 0.419 to 0.615 across the 96 continuations. Table 33 reports recipe means. This is an in-sample score decomposition.

The initial checkpoint ranks SH1–SC2–REV first on 582 of the 800 ordinary tasks of each mask. Centring removes a program-specific reference term from both checkpoints.

Table 36 compares raw, corrected, and centred B/D scores using references A, its 100-task subsamples, and C, separately from the target families. Means pool masks and recipes. Corrected Hit@32 rises while Hit@1 remains below its initial value.

For further decomposition, regress the mean score change on position-specific operation indicators, then add indicators for the two adjacent operation pairs. Both fits use the same family-A scores and give different B/D rankings (Table 36). The reference variance share R^2_μ measures score changes on A and has no fixed mapping to Hit@32.

Post-hoc centring with C as reference for A/B and A for C/D improves trained B/D Hit@32 and lowers A/C (Table 37). The four-family mean rises for trained checkpoints and falls for the initial checkpoint. Effects depend on checkpoint, target family, and ranking budget.

For a crossed panel, let $d_{ij} = s(x_{ij},p_0) - s(x_{ij},p_1)$. The signed interaction $I = d_{00} - d_{01} - d_{10} + d_{11}$ is invariant to any added program-only score term: its coefficients sum to zero. Positive values favour the designated diagonal assignment. Its mean falls from 2.25 initially to 0.74–1.65 across trained

Recipe	Contrast	Δ mass	Δ Hit@32	Δ joint
R0	CF, H=0	-0.148 [-0.581, +0.284]	-1.08 [-16.55, +14.38]	+3.13 [-17.41, +23.66]
R0	CF, H=0.01	-0.179 [-0.597, +0.239]	+5.17 [-17.26, +27.60]	+15.63 [-19.95, +51.20]
R0	H, CF=0	+0.029 [-0.530, +0.588]	-2.92 [-14.10, +8.27]	-4.17 [-8.65, +0.32]
R0	H, CF=0.1	-0.002 [-0.335, +0.331]	+3.33 [-3.51, +10.17]	+8.33 [-9.59, +26.26]
R0	CF × H	-0.031 [-0.257, +0.195]	+6.25 [-7.20, +19.70]	+12.50 [-8.04, +33.04]
R1	CF, H=0	-0.046 [-0.116, +0.024]	+0.17 [-9.50, +9.83]	+0.00 [-7.76, +7.76]
R1	CF, H=0.01	-0.046 [-0.253, +0.161]	-1.67 [-11.29, +7.96]	+1.04 [-3.44, +5.52]
R1	H, CF=0	+0.021 [-0.025, +0.067]	+0.50 [-1.14, +2.14]	-1.04 [-5.52, +3.44]
R1	H, CF=0.1	+0.021 [-0.147, +0.188]	-1.33 [-2.28, -0.38]	+0.00 [-7.76, +7.76]
R1	CF × H	+0.000 [-0.203, +0.203]	-1.83 [-2.55, -1.12]	+1.04 [-7.92, +10.01]
R2	CF, H=0	-0.081 [-0.493, +0.331]	+1.67 [-8.90, +12.23]	+1.04 [-15.12, +17.20]
R2	CF, H=0.01	+0.017 [-0.187, +0.221]	-0.42 [-28.47, +27.64]	+1.04 [-7.92, +10.01]
R2	H, CF=0	-0.069 [-0.466, +0.328]	+1.08 [-36.47, +38.64]	-3.13 [-10.89, +4.64]
R2	H, CF=0.1	+0.029 [-0.191, +0.248]	-1.00 [-2.64, +0.64]	-3.13 [-16.57, +10.32]
R2	CF × H	+0.098 [-0.518, +0.714]	-2.08 [-40.33, +36.16]	+0.00 [-7.76, +7.76]
R3	CF, H=0	-0.016 [-0.108, +0.076]	+2.92 [-6.41, +12.24]	+4.17 [-11.99, +20.33]
R3	CF, H=0.01	+0.001 [-0.030, +0.033]	+3.00 [-0.73, +6.73]	+3.13 [-10.32, +16.57]
R3	H, CF=0	+0.033 [-0.109, +0.175]	+0.58 [-5.19, +6.35]	+0.00 [-7.76, +7.76]
R3	H, CF=0.1	+0.050 [-0.136, +0.236]	+0.67 [-11.09, +12.42]	-1.04 [-20.58, +18.49]
R3	CF × H	+0.017 [-0.052, +0.087]	+0.08 [-9.96, +10.12]	-1.04 [-18.97, +16.89]
R2−R0	Replay CE	-0.026 [-0.499, +0.448]	+1.92 [-12.39, +16.22]	+2.08 [-9.77, +13.94]
R2−R0	Replay CE+CF	+0.042 [-0.564, +0.648]	+4.67 [-0.31, +9.65]	+0.00 [-7.76, +7.76]
R2−R0	Replay CE+H	-0.124 [-0.569, +0.322]	+5.92 [-28.25, +40.08]	+3.13 [-4.64, +10.89]
R2−R0	Replay CE+CF+H	+0.072 [-0.124, +0.269]	+0.33 [-1.56, +2.23]	-11.46 [-35.17, +12.26]
R3−R1	Replay CE	+0.019 [-0.127, +0.165]	-1.17 [-6.73, +4.40]	-2.08 [-11.05, +6.88]
R3−R1	Replay CE+CF	+0.048 [-0.100, +0.196]	+1.58 [-17.78, +20.95]	+2.08 [-21.63, +25.80]
R3−R1	Replay CE+H	+0.031 [-0.046, +0.108]	-1.08 [-2.88, +0.71]	-1.04 [-10.01, +7.92]
R3−R1	Replay CE+CF+H	+0.078 [-0.103, +0.259]	+3.58 [-9.91, +17.08]	+1.04 [-7.92, +10.01]

Table 34: Designated development contrasts for mask 1. Entries use local scores at step 128 and give percentage points with nominal 95 percent paired t intervals over three continuation seeds. B/D have equal weight. Replay adds atomic data at fixed composition budget. Appendix I.1 gives the analysis-plan timing.

recipe–objective means. Mean strict joint accuracy is 10.94 percent initially, 6.38 after training, and 3.97 after the correction. The corrected scores give higher B/D Hit@32 and lower strict joint accuracy.

I.3 Objective contrasts and retention

CF raises mean joint accuracy in all eight recipe–H comparisons by 0.52–9.90 points, averaged over the masks and seeds. The mask-specific seed intervals span zero, and correct-mass changes vary by condition (Tables 34 and 35).

H raises mean B/D entropy in all eight recipe–CF comparisons. Mean correct mass across the 16 conditions is 0.29–0.51 percent, against 2.97 initially. Each B/D task has one correct program, so its uniform mass is 0.80 percent. For correct mass m, the decomposition $H(q) = h_2(m) + mH(q \mid \text{correct}) + (1-m)H(q \mid \text{incorrect})$ has $H(q \mid \text{correct}) = 0$ on these tasks. Most of the remaining entropy comes from incorrect programs. Entropy and rank concentration can move differently: in R0, adding H to CE changes the most frequent top-1 share from 39.13 to 59.19 percent and the top-8 union from 41.3 to 37.3 programs across 800 tasks. Table 38 gives all conditions.

Replay improves atomic APPLY in all 48 matched mask–seed–objective comparisons. It reduces the unigram share of score-shift variance and raises the added pair share (Table 39). At the lower learning rate, replay gives mean APPLY accuracies of 61.80–62.00 percent across the four objectives, compared with 62.55 at initialisation. Family-A Hit@32 is 72.50–75.67 percent, against 49.25 initially. B/D Hit@32 is 21.21–24.33 percent, against 50.63 initially. These ranges describe condition means at a fixed composition budget with added atomic data.

Recipe	Contrast	Δ mass	Δ Hit@32	Δ joint
R0	CF, H=0	+0.022 [-0.059, +0.104]	+0.25 [-3.04, +3.54]	+1.04 [-7.92, +10.01]
R0	CF, H=0.01	+0.031 [-0.238, +0.301]	+2.75 [-13.40, +18.90]	+4.17 [-13.76, +22.09]
R0	H, CF=0	+0.064 [-0.207, +0.335]	+2.75 [-15.13, +20.63]	-4.17 [-13.13, +4.80]
R0	H, CF=0.1	+0.073 [+0.044, +0.102]	+5.25 [+3.39, +7.11]	-1.04 [-10.01, +7.92]
R0	CF × H	+0.009 [-0.240, +0.258]	+2.50 [-16.90, +21.90]	+3.13 [-12.40, +18.65]
R1	CF, H=0	-0.018 [-0.063, +0.027]	-0.42 [-6.94, +6.11]	+4.17 [-0.32, +8.65]
R1	CF, H=0.01	-0.049 [-0.186, +0.088]	+0.08 [-8.62, +8.79]	+8.33 [-3.52, +20.19]
R1	H, CF=0	+0.029 [-0.088, +0.147]	-0.33 [-6.10, +5.44]	-2.08 [-6.57, +2.40]
R1	H, CF=0.1	-0.002 [-0.078, +0.075]	+0.17 [-4.33, +4.66]	+2.08 [-9.77, +13.94]
R1	CF × H	-0.031 [-0.180, +0.118]	+0.50 [-7.43, +8.43]	+4.17 [-4.80, +13.13]
R2	CF, H=0	+0.016 [-0.133, +0.166]	+0.50 [-10.68, +11.68]	+0.00 [-23.29, +23.29]
R2	CF, H=0.01	+0.050 [-0.171, +0.272]	+2.75 [+0.60, +4.90]	+1.04 [-10.82, +12.90]
R2	H, CF=0	+0.056 [-0.113, +0.225]	+3.58 [-12.57, +19.73]	-3.13 [-10.89, +4.64]
R2	H, CF=0.1	+0.090 [-0.297, +0.477]	+5.83 [-18.75, +30.41]	-2.08 [-13.94, +9.77]
R2	CF × H	+0.034 [-0.294, +0.362]	+2.25 [-10.84, +15.34]	+1.04 [-16.89, +18.97]
R3	CF, H=0	+0.033 [-0.047, +0.112]	+1.92 [-3.68, +7.52]	+2.08 [-2.40, +6.57]
R3	CF, H=0.01	+0.050 [-0.050, +0.150]	+1.75 [-0.96, +4.46]	+6.25 [-1.51, +14.01]
R3	H, CF=0	+0.006 [-0.069, +0.082]	+0.92 [-4.68, +6.52]	+2.08 [-2.40, +6.57]
R3	H, CF=0.1	+0.024 [-0.029, +0.077]	+0.75 [-3.13, +4.63]	+6.25 [-1.51, +14.01]
R3	CF × H	+0.018 [-0.018, +0.053]	-0.17 [-3.81, +3.47]	+4.17 [-4.80, +13.13]
R2−R0	Replay CE	+0.031 [-0.163, +0.225]	+2.17 [-11.68, +16.02]	+0.00 [-15.53, +15.53]
R2−R0	Replay CE+CF	+0.025 [-0.120, +0.170]	+2.42 [-1.38, +6.21]	-1.04 [-12.90, +10.82]
R2−R0	Replay CE+H	+0.023 [-0.062, +0.108]	+3.00 [-9.61, +15.61]	+1.04 [-10.82, +12.90]
R2−R0	Replay CE+CF+H	+0.042 [-0.465, +0.549]	+3.00 [-24.47, +30.47]	-2.08 [-18.24, +14.08]
R3−R1	Replay CE	-0.183 [-0.457, +0.091]	-6.75 [-26.77, +13.27]	-3.13 [-10.89, +4.64]
R3−R1	Replay CE+CF	-0.132 [-0.446, +0.181]	-4.42 [-19.88, +11.05]	-5.21 [-14.17, +3.76]
R3−R1	Replay CE+H	-0.206 [-0.334, -0.078]	-5.50 [-15.20, +4.20]	+1.04 [-3.44, +5.52]
R3−R1	Replay CE+CF+H	-0.107 [-0.436, +0.222]	-3.83 [-20.17, +12.51]	-1.04 [-24.76, +22.67]

Table 35: Designated development contrasts for mask 2. Entries use local scores at step 128 and give percentage points with nominal 95 percent paired t intervals over three continuation seeds. B/D have equal weight. Replay adds atomic data at fixed composition budget. Appendix I.1 gives the analysis-plan timing.

Ranking rule	Hit@1	Hit@8	Hit@32	Mean rank
Initial	3.13	15.00	50.63	39.32
Trained	0.00	1.10	22.43	59.19
Trained, shift from A	2.50	12.80	57.56	35.19
Trained, shift from half of A	2.50	12.89	57.42	35.18
Trained, shift from C	2.50	12.18	59.36	35.18
Trained, bigram shift from A	0.23	18.17	62.69	31.81
Trained, unigram shift from A	0.79	12.26	51.45	39.85
Initial, centred scores	1.13	10.63	33.88	49.84
Trained, centred scores	1.08	12.24	44.89	43.74

Table 36: Post-hoc B/D diagnosis in the 96 development continuations. Hit values are percentages, averaged over the two masks and continuation recipes. A and C denote unlabelled reference tasks in the corresponding families. Initial checkpoints are shared within mask.

I.4 Pair preferences and saved training trajectories

For an ordered pair a, b, define its reference preference as the mean score of programs containing that pair minus the mean score of programs that avoid it, averaged over A. The standard deviation across the 25 preferences is 3.075 nats initially and 1.784–2.091 across trained condition means. To describe residuals, fit each pair's shift to its initial preference and $\log(1 + n_{ab})$ among unmasked pairs, with an intercept. Mean residuals for masked pairs range from -1.790 to -0.914 nats across conditions (Table 38). Pair indicators share programs, so this is a descriptive regression. Supplementary tables and a heatmap give all 25 pairs per continuation under references A and C.

Checkpoint	Family	Raw@1	Ctr.@1	Raw@8	Ctr.@8	Raw@32	Ctr.@32
Initial	A	1.75	5.50	13.75	25.00	49.25	51.50
Initial	B	2.75	2.50	15.00	16.25	52.00	40.50
Initial	C	2.00	5.25	12.25	17.50	53.75	51.00
Initial	D	3.50	0.75	15.00	8.25	49.25	28.50
Trained	A	9.40	8.39	32.97	30.82	71.39	62.89
Trained	B	0.00	1.20	1.56	14.56	25.44	49.68
Trained	C	8.14	7.27	32.95	26.45	75.45	60.74
Trained	D	0.00	1.07	0.63	10.51	19.41	40.90

Table 37: Disjoint-reference centring in percent. Family C supplies the unlabelled reference for A/B, and A supplies it for C/D. The rule is applied separately to each checkpoint. Initial rows average two distinct mask evaluations. Trained rows average 96 continuations sharing those initial evaluations. Each target and reference family has 200 tasks per mask. All comparisons are descriptive and post-hoc.

Recipe	Loss	H	Mode (%)	Top-8 union	Cell (%)	I	Pair SD	Residual
R0	CE	3.205	39.13	41.3	55.60	0.888	2.091	-1.398
R0	CE+CF	3.238	44.21	42.3	55.86	1.373	2.083	-1.370
R0	CE+H	3.369	59.19	37.3	52.34	0.739	1.949	-1.537
R0	CE+CF+H	3.251	45.31	45.0	59.64	1.645	1.784	-0.971
R1	CE	3.373	42.40	56.3	54.17	1.062	2.009	-1.475
R1	CE+CF	3.343	43.58	56.3	54.82	1.173	2.052	-1.475
R1	CE+H	3.423	43.06	56.5	52.73	0.963	2.028	-1.440
R1	CE+CF+H	3.356	38.71	56.5	55.73	1.150	1.995	-1.446
R2	CE	3.248	51.92	48.8	56.51	0.777	2.027	-1.066
R2	CE+CF	3.223	50.63	48.2	55.86	0.904	1.945	-0.914
R2	CE+H	3.264	46.69	49.2	54.43	1.157	1.870	-0.962
R2	CE+CF+H	3.281	52.83	50.0	54.43	0.933	1.925	-1.027
R3	CE	3.287	38.98	55.8	52.21	0.865	2.087	-1.723
R3	CE+CF	3.316	38.38	56.0	54.69	1.499	2.034	-1.790
R3	CE+H	3.330	37.27	57.5	53.65	0.990	1.984	-1.588
R3	CE+CF+H	3.351	33.13	57.8	55.60	1.532	1.951	-1.682

Table 38: Secondary diagnostics, averaged over masks and seeds. H is B/D entropy in nats. Mode is the fraction of 800 ordinary A–D tasks sharing the most frequent top-1 program. Top-8 union counts distinct programs across those tasks. Cell and signed interaction I use B/D crossed panels. Pair SD is the dispersion of 25 pair preferences on reference A. Residual is the mean masked-pair residual after fitting score shift on initial preference and log frequency among unmasked pairs. These score descriptors have overlapping program support.

Recipe	Loss	Δ PLAN	Δ APPLY	ΔR^2_U	$\Delta(R^2_{UB} - R^2_U)$	APPLY +
R2−R0	CE	+26.13	+42.20	-0.31	+0.10	6/6
R2−R0	CE+CF	+23.67	+41.47	-3.85	+0.54	6/6
R2−R0	CE+H	+30.45	+42.87	-3.63	+0.02	6/6
R2−R0	CE+CF+H	+25.33	+41.07	-4.52	+0.64	6/6
R3−R1	CE	+14.10	+6.30	-0.73	+1.98	6/6
R3−R1	CE+CF	+13.62	+5.88	-0.86	+1.48	6/6
R3−R1	CE+H	+15.78	+4.70	-1.08	+1.79	6/6
R3−R1	CE+CF+H	+14.02	+5.65	-1.04	+1.38	6/6

Table 39: Replay contrasts in percentage points for atomic accuracy and reference-A score-shift variance. R^2_U uses position-specific operation indicators, and $R^2_{UB} - R^2_U$ adds adjacent pairs. APPLY + counts positive changes in six mask–seed pairs. Each arm uses 128 composition updates.

The saved monitors cover steps 0, 8, 16, 32, 64, and 128 on fixed tasks: 64 ordinary A–D tasks, 32 B/D crossed panels, four TRAIN panels, and 100 atomic tasks per mode and mask. Figure 14 uses the B/D development tasks. At step 128, strict joint accuracy on the four TRAIN panels averages

58.33–79.17 percent across recipe–objective conditions. These probes measure training-panel fit. Figure 15 gives full-set B/D ranking curves. The supplement includes per-family curves, temperature and scoring checks, atomic results, and all monitor steps.

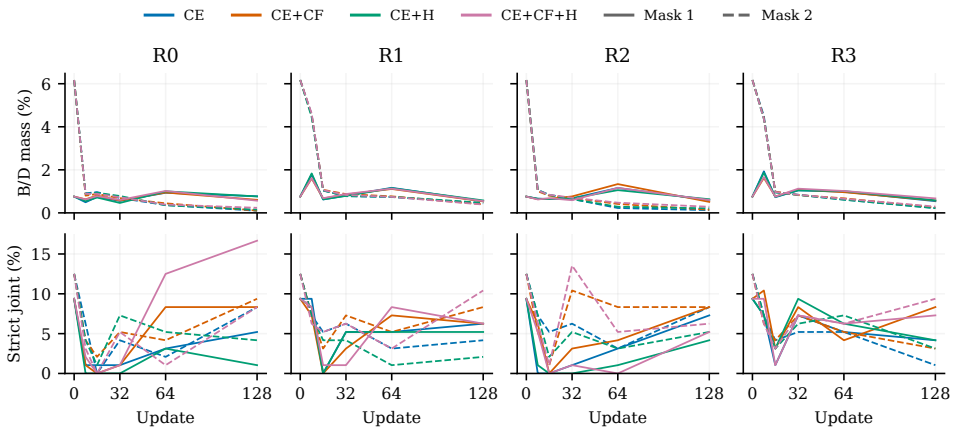

Figure 14: Development monitors averaged over three seeds per condition, with lines distinguishing masks. Mass uses B/D tasks within a 64-task A–D set, and joint accuracy uses 32 B/D panels per mask. The final endpoint is step 128.

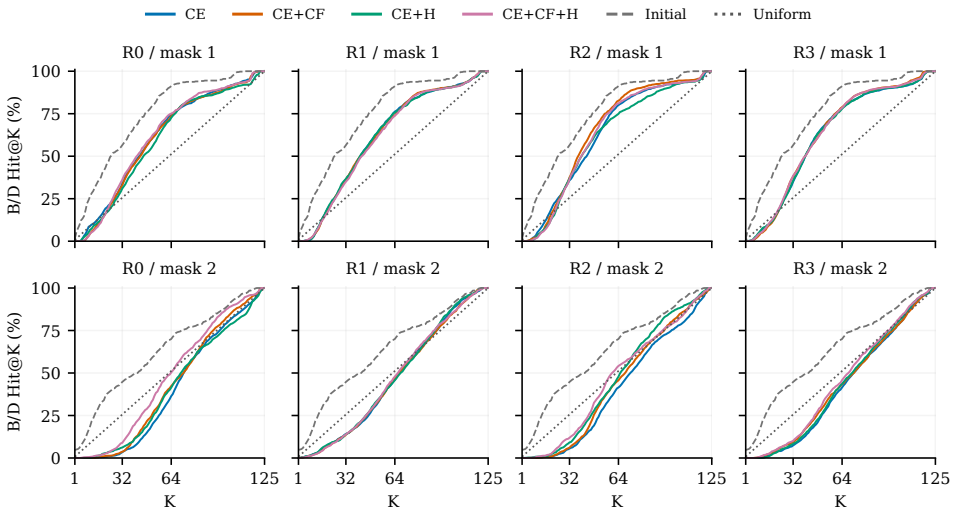

Figure 15: Local-score B/D Hit@K at step 128, averaged over three seeds per recipe–mask panel. Grey dashed: shared initial checkpoint. Dotted: uniform ranking with one correct program per task.

I.5 Additional checkpoint diagnostics

Three composition and three atomic-control continuations under the three-pair mask score 9.58 versus 25.58 percent on ordinary B/D Hit@32 and 42.71 versus 9.38 on strict joint accuracy, using local scores. Their recipe and mask differ from the factorial matrix, and pre-continuation pair exposure is uncertain.

A one-seed, one-mask study compared CE, CE+H, CE+CF, and CE+CF+H. All had zero B/D Hit@1/8. Strict joint gains over CE were 3.13 points, with panel-bootstrap intervals spanning zero. CE+CF+H raised the secondary four-TARGET pair score by 10.00 points [5.83, 13.92]. Its atomic PLAN/APPLY was 32.8/29.1 percent, versus 99.2/62.7 initially.