G^{2}ARL: GROUP GFLOWNETS FOR AGENTIC REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many agentic tasks admit multiple successful trajectories, yet reward maximization does not explicitly preserve their diversity. GFlowNets match a reward-proportional trajectory distribution, but a sparse reward over long interactions makes such a distribution difficult to learn. We introduce Group GFlowNet for Agentic RL (G^{2}ARL), which combines group level flow reallocation with coverage-guided dynamic rebalancing for multi-turn language model agents. The flow reallocation loss learns relative probability allocation from pairwise trajectory comparisons, while dynamic gating and submodular replay retain verified successes selected for coverage and return them to groups that have none. Across ALFWorld, WebShop, and ScienceWorld, G^{2}ARL achieves the highest success among all retrained baselines in every main benchmark setting. Across benchmark settings, G^{2}ARL achieves 4.2 to 20.7 percentage points higher success than the strongest retrained comparator, while also producing shorter solutions with fewer repeated actions and loops.

1 INTRODUCTION

Reinforcement learning for language models has moved to agentic settings. A policy completes a task through a sequence of actions in an environment and receives a terminal reward (Yao et al., 2022b). ALFWorld, WebShop, and ScienceWorld are representative benchmarks (Shridhar et al., 2020; Yao et al., 2022a; Wang et al., 2022). Agents are commonly post-trained with reward-maximizing policy-gradient methods such as PPO, GRPO, and GiGPO (Schulman et al., 2017; Shao et al., 2024; Feng et al., 2026). Prior work reports that such training reduces entropy and concentrates outputs on a limited set (Cui et al., 2025; Kirk et al., 2024; Chen et al., 2026), and RAGEN identifies a related Echo Trap in multi-turn agents (Wang et al., 2025b).

Generative Flow Networks (GFlowNets) offer a different objective. They learn a policy that samples complete trajectories in proportion to a reward (Bengio et al., 2021; 2023), so successful trajectories share probability instead of concentrating on one of them. On the prefix tree of a language model this is the optimal policy of entropy-regularized reinforcement learning (Tiapkin et al., 2024; Mohammadpour et al., 2024; Deleu et al., 2024). Trajectory balance (TB) and sub-trajectory balance (SubTB) train such policies (Malkin et al., 2022a; Madan et al., 2023), and GFlowVLM applies Var-TB and SubTB to a vision-language agent on ALFWorld against supervised and PPO baselines (Kang et al., 2025). However, the sparse reward and long horizon make such a distribution hard to learn. Further, TB constrains each trajectory separately through a shared learned partition function, SubTB each sub-trajectory through learned state flows, and group-centered variants such as Var-TB weight every trajectory of a group alike (Kang et al., 2025; Richter et al., 2020). Neither singles out the informative comparisons within a task. A success and a failure of the same task show where the probability should move. With binary scores, however, trajectories that all succeed or all fail have equal scores, so their comparisons contain no score difference. As the policy improves, more and more groups consist only of successes (Figure 4).

We introduce Group GFlowNet for Agentic RL (G^{2}ARL, Figure 1), which couples two mechanisms. Group flow reallocation learns from all trajectories of a task through pairwise comparisons and gives the largest weight to comparisons between a success and a failure. Coverage-guided dynamic rebalancing keeps verified successes in a replay buffer and returns them to groups that have none, so that the comparisons keep carrying information about the score. The two mechanisms are co-

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

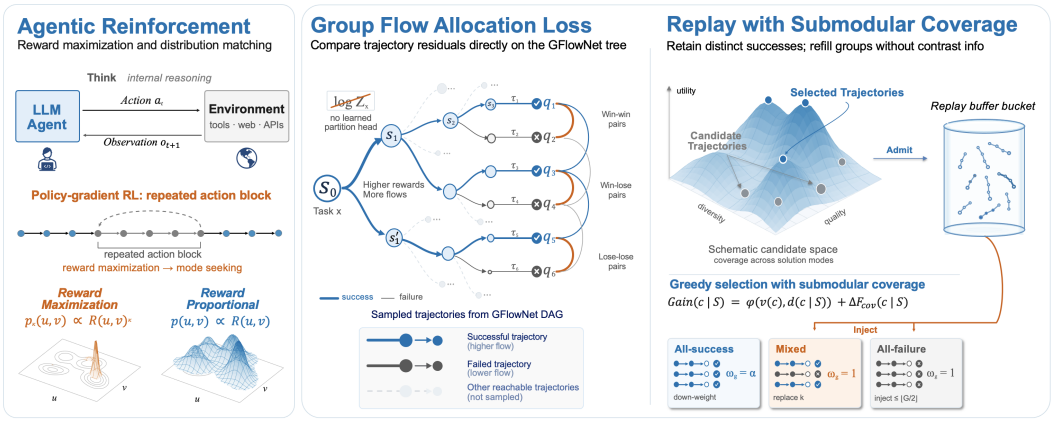

Figure 1: **Overview of G^{2}ARL: group flow reallocation and coverage-guided dynamic rebalancing for agentic reinforcement learning.** Within groups that contain both successful and failed trajectories, pairwise residuals weighted by success and failure allocate flow between them without learning a partition function. The replay buffer keeps, by greedy selection, stored successes scored on quality (v), dissimilarity (d) and marginal submodular coverage, and injects them into all-failure groups when a same-task success is stored and into mixed groups with a decreasing probability.

designed: the loss sets how flow is allocated within a group, and replay sets which trajectories the group contains. Our contributions are as follows.

(i) **Distribution matching for multi-turn language-model agents.** We integrate GFlowNet distribution matching with the reasoning and instruction-following capabilities of language models for multi-turn decision making, and compare it with policy-gradient and GFlowNet baselines on ALFWorld, WebShop and ScienceWorld at different scales.

(ii) **Group-level flow reallocation.** We introduce a pairwise residual loss that matches the relative probabilities of trajectories to a reward-weighted target. By comparing every pair of trajectories of a task, the loss learns from the whole group at once without a learned partition function.

(iii) **Coverage-guided dynamic rebalancing.** We combine dynamic group gating with submodular replay to retain and reuse verified successes selected for coverage (Wang et al., 2026). The mechanism reweights training groups and restores missing success-failure contrast. Across the benchmark and model-scale settings, G^{2}ARL achieves higher success than baselines. It also produces shorter successful trajectories on average, with lower repeat-action rates and loop shares.

2 Related Work

GFlowNet objectives and language-model post-training. Trajectory balance and sub-trajectory balance train GFlowNets to sample from reward-proportional distributions (Malkin et al., 2022a; Madan et al., 2023). Theoretical work connects these objectives to entropy-regularized reinforcement learning and reference-policy adaptation (Tiapkin et al., 2024; Deleu et al., 2024; Venkatraman et al., 2024). Language-model applications span amortized inference (Hu et al., 2024), diverse reasoning (Yu et al., 2024; Takase et al., 2024), and red-teaming (Lee et al., 2025). FlowRL studies reward-distribution matching, while TBA scales post-training through asynchronous sampling and learning (Zhu et al., 2026; Bartoldson et al., 2026). Several objectives replace the learned partition function with a statistic of the sampled group: Var-TB in GFlowVLM (Kang et al., 2025), an in-group VarGrad estimate in TBA (Bartoldson et al., 2026), and group-centered log-ratio errors in GVPO and GFlowRL (Zhang et al., 2026a; Liu et al., 2026a). EP-GFlowNets introduce a contrastive balance condition and build their aggregating balance condition on it (Da Silva et al., 2024). Stable-GFlowNet masks low-reward-contrast pairs and uses high-reward replay as comparison anchors for single-turn red-teaming (Kwon et al., 2026). With equal weights the G^{2}ARL loss degenerates to the group-centered Var-TB/VarGrad form (Lemma 1). G^{2}ARL learns from every pair in a same-task group, weights pairs by success and failure, keeps within gated groups the same-score constraints that a reward-contrast mask removes, and couples the loss with group gating and coverage-selected replay in multi-turn environments.

Reinforcement learning for interactive agents. WebRL and Search-R1 train language models through web navigation and search interactions (Qi et al., 2025; Jin et al., 2025). GiGPO, RLVMR, and iStar address multi-turn credit assignment through step-level learning signals (Feng et al., 2026; Zhang et al., 2026c; Liu et al., 2026b). AgentFlow optimizes planning within a multi-module workflow (Li et al., 2026). AgentGym-RL and Agent Lightning provide reusable frameworks for training agents across environments and implementations (Xi et al., 2025b; Luo et al., 2025). GFlowVLM applies GFlowNet training to vision-language agents (Kang et al., 2025), while SkillFlow trains an orchestrator that invokes a frozen executor and maintains an evolving skill library (Zhang et al., 2026b).

Experience replay and training diversity. Experience reuse underlies expert iteration, self-imitation, and reinforced self-training (Anthony et al., 2017; Oh et al., 2018; Gulcehre et al., 2023). In group-based policy optimization, DAPO resamples groups whose trajectories all succeed or all fail, whereas RePO, ExGRPO, RLEP, and LUFFY incorporate past trajectories or expert demonstrations (Yu et al., 2026; Li et al., 2025; Zhan et al., 2026; Zhang et al., 2025; Yan et al., 2026). EFRame combines exploration, filtering, and replay, while Buffer Matters adapts training batches using historical experience (Wang et al., 2025a; Wan et al., 2026). Headroom-Drift Replay selects stored groups by learning value and compatibility with the current policy (Park & Chang, 2026). Diversity-aware objectives and skill reuse provide complementary approaches (Yao et al., 2026; Xu et al., 2026). GFlowNet studies examine reward-prioritized replay for sample efficiency and mode discovery (Shen et al., 2023; Vemgal et al., 2023), while RapTB combines trajectory quality, dissimilarity, and submodular coverage for replay selection (Wang et al., 2026). G^{2}ARL combines coverage-selected replay with group reweighting and same-task trajectory injection for reward-aware flow reallocation in multi-turn agents.

Benchmarks and behavior evaluation. ALFWorld, WebShop, and ScienceWorld provide task-specific environments for household manipulation, online shopping, and scientific experimentation (Shridhar et al., 2020; Yao et al., 2022a; Wang et al., 2022). AgentBench, AgentBoard, and AgentGym broaden coverage through multi-environment evaluation and training (Liu et al., 2024; Ma et al., 2024; Xi et al., 2025a). WebArena and the τ-bench family extend evaluation to realistic web tasks and interactions between agents, users, and tools (Zhou et al., 2024; Yao et al., 2024; Barres et al., 2025). AgentQuest measures progress and repetition beyond task success (Gioacchini et al., 2024), while RAGEN examines training instability through the Echo Trap (Wang et al., 2025b).

3 G^{2}ARL

3.1 Setting and target distribution

A language-model policy π_θ completes task x by generating a trajectory $\tau = (a_0, \ldots, a_{T-1})$. At each turn, s_t contains the task and interaction history, and a_t contains the generated reasoning and command. The environment executes the command and returns a terminal score $y(\tau)$ when the episode ends. Let $n(\tau)$ count invalid actions and π_{ref} denote the frozen starting policy. We target

$$R_x(\tau) = \pi_{\text{ref}}(\tau \mid x)^\rho \exp\big[\beta y(\tau) - \lambda n(\tau)\big], \quad \pi^*(\tau \mid x) = \frac{R_x(\tau)}{Z_x}, \tag{1}$$

where $Z_x = \sum_\tau R_x(\tau)$ normalizes the target distribution. The score weight β favors higher-scoring trajectories, while λ penalizes invalid actions. We use $\rho = 1$. On ALFWorld and WebShop, y is the binary success bit. On ScienceWorld, $y \in [-1, 1]$ is the terminal score divided by 100, with a separate binary success flag.

A *group* contains G trajectories of the same task. For trajectory i, define

$$D_i = \log \frac{\pi_\theta(\tau_i \mid x)}{\pi_{\text{ref}}(\tau_i \mid x)^\rho}, \qquad q_i = D_i - \beta y_i + \lambda n_i = \log \frac{\pi_\theta(\tau_i \mid x)}{R_x(\tau_i)}. \tag{2}$$

Trajectory log-probabilities sum the generated-token log-probabilities over turns. Thus, q_i measures the policy's log-probability relative to the unnormalized target. TB learns $\log Z_x$ by minimizing $G^{-1}\sum_i(\log Z_x + q_i)^2$ (Malkin et al., 2022a). We instead compare residuals within each task to learn relative probability allocation.

3.2 GROUP FLOW REALLOCATION

On the trajectory-prefix tree, $\pi_\theta(\tau \mid x)$ is the fraction of total flow assigned to trajectory τ (Bengio et al., 2023). We learn this allocation by matching policy probability ratios to target ratios. Let W and L contain the successful and failed trajectories of task x, with sizes n_W and n_L. The group flow reallocation loss is

$$\begin{aligned} \mathcal{L}_{\text{pairs}} &= \frac{1}{\mathcal{N}_g} \sum_{i<j} w_{ij} e_{ij}^2, \\ e_{ij} &:= q_i - q_j = \log \frac{\pi_\theta(\tau_i \mid x)}{\pi_\theta(\tau_j \mid x)} - \log \frac{R_x(\tau_i)}{R_x(\tau_j)}. \end{aligned} \tag{3}$$

Here w_{ij} takes the value w_{WW}, w_{LL}, or w_{WL} for two successes, two failures, or a success and a failure, and $\mathcal{N}_g = \frac{2G}{G-1} \sum_{i<j} w_{ij} > 0$. Weights and the normalizer are held fixed during differentiation. Each comparison is a contrastive balance residual, in which the shared partition function cancels (Da Silva et al., 2024).

Learning relative flow. The error e_{ij} measures whether trajectory i receives too much or too little probability relative to j, compared with the target. Its contribution to the policy gradient is

$$\nabla_\theta \left(\frac{w_{ij}}{\mathcal{N}_g} e_{ij}^2 \right) = \frac{2 w_{ij}}{\mathcal{N}_g} e_{ij} \left[\nabla_\theta \log \pi_\theta(\tau_i \mid x) - \nabla_\theta \log \pi_\theta(\tau_j \mid x) \right]. \tag{4}$$

When $e_{ij} > 0$, the pair term favors decreasing the log-probability ratio. When $e_{ij} < 0$, it favors increasing it. The comparison therefore trains a probability ratio, rather than constructing a group-normalized advantage. For binary scores, a successful trajectory i and a failed trajectory j have zero pair error when

$$\frac{\pi_\theta(\tau_i \mid x)}{\pi_\theta(\tau_j \mid x)} = \left[\frac{\pi_{\text{ref}}(\tau_i \mid x)}{\pi_{\text{ref}}(\tau_j \mid x)} \right]^\rho \exp\!\big[\beta - \lambda(n_i - n_j)\big]. \tag{5}$$

Thus, success multiplies the cost-adjusted reference ratio by e^β. Between two successes or two failures with equal scores, the score term cancels, leaving the reference ratio and invalid-action costs to determine allocation. For binary scores, successful-failed pairs carry the score term (Lemma 1(ii)). We weight these pairs highest and failure-failure pairs lowest, making success-failure contrast the stiffest direction in residual coordinates for mixed groups (Appendix A.5). With equal weights the loss degenerates to group-centered TB, the Var-TB/VarGrad form, which weights every comparison alike (Kang et al., 2025; Richter et al., 2020). The derivation, class weights, and efficient evaluation are given in Appendix A.

Proposition 1 (Flow allocation and target preservation)**.** *Assume the policy and target are positive on the same finite trajectory support. For a mixed group with non-negative weights and $w_{WL} > 0$, $\mathcal{L}_{\text{pairs}} = 0$ if and only if $\pi_\theta(\tau_i \mid x)/\pi_\theta(\tau_j \mid x) = R_x(\tau_i)/R_x(\tau_j)$ for every pair of the group. If the positive-weight comparisons connect the full support, the unique zero-loss distribution is $\pi^* = R_x/Z_x$, independently of the weight values. This distribution is a stationary point of every such pairwise objective.*

Proof. Zero loss forces $q_i = q_j$ on every positive-weight pair. In a mixed group the successful-failed pairs connect all members, so all residuals agree, which is equivalent to matching every target ratio. If the comparison graph connects the full support, $q(\tau) = c$ everywhere, and normalization gives $e^c = 1/Z_x$. At π^* every e_{ij} vanishes. Appendix A.4 gives the full statement. □

Optimality of the recovered policy. For $\rho = 1$, the recovered distribution also uniquely maximizes the expected cost-adjusted score $\beta y - \lambda n$ with a KL penalty to the reference (Deleu et al., 2025), because

$$\begin{aligned} \mathcal{J}(\pi) &:= \mathbb{E}_{\tau \sim \pi}[\beta y(\tau) - \lambda n(\tau)] - \mathrm{KL}(\pi \| \pi_{\text{ref}}) \\ &= \log Z_x - \mathrm{KL}(\pi \| \pi^*). \end{aligned} \tag{6}$$

For equal weights, uniqueness and this identification are known for GVPO's group-centered objective (Zhang et al., 2026a). Appendix A.4 strengthens this result to stationary distributions in unrestricted trajectory-logit coordinates. These statements concern the pairwise objective, not convergence of the parameterized policy or the additional replay penalty (Appendix A.8).

3.3 COVERAGE-GUIDED DYNAMIC REBALANCING

We rebalance training groups using a task-specific replay buffer.

For each task x, the buffer $\mathcal{B}_x$ stores verified successful trajectories. When the buffer exceeds its capacity, we retain a subset greedily according to

$$c^* = \arg\max_{c \notin S} \left[\phi\big(v(c), d(c \mid S)\big) + \Delta F_{\mathrm{cov}}(c \mid S) \right], \qquad d(c \mid S) = 1 - \max_{s \in S} \mathrm{sim}(c, s), \tag{7}$$

where $v(c)$ measures trajectory quality, $d(c \mid S)$ measures dissimilarity, and ΔF_{cov} is the marginal gain of a monotone submodular coverage function. The coverage term rewards new solution-length and tool-use patterns with diminishing returns, so that repeatedly observed behaviors gain less.

Given a sampled group $\mathcal{G}_x$, we define a dynamic rebalancing operator

$$\widetilde{\mathcal{G}}_x = \mathcal{R}_u(\mathcal{G}_x, \mathcal{B}_x), \tag{8}$$

where u denotes training progress. For an all-failure group, $\mathcal{R}_u$ replaces part of the group with stored successes when available, creating success-failure comparisons. For a mixed group, stored successes are injected with a probability that decreases during training while retaining at least one failure. All-success groups supply the buffer, remain unchanged, and enter the loss with weight α. Injected trajectories are replayed in the environment and re-scored under the current policy before entering the loss.

The final objective is

$$\mathcal{L}_{\mathrm{G^2ARL}}(x) = \omega_x \left[\mathcal{L}_{\mathrm{pairs}}\left(\widetilde{\mathcal{G}}_x\right) + \eta\, \widehat{\mathrm{KL}}_x \right], \tag{9}$$

where $\omega_x = \alpha$ for an all-success group, $\omega_x = 1$ for a mixed group or an all-failure group successfully rebalanced with stored successes and $\omega_x = 0$ otherwise. $\widehat{\mathrm{KL}}_x$ is a per-turn log-probability displacement penalty on replayed trajectories, with coefficient η (Appendix A.7). ScienceWorld additionally uses a partial-score fallback, under which an all-failure group refilled only with stored partial-score examples enters the loss with weight 0.25 (Appendix A.7).

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Tasks and models. We evaluate household manipulation on ALFWorld, online shopping on WebShop, and experimental planning on ScienceWorld (Shridhar et al., 2020; Yao et al., 2022a; Wang et al., 2022). Evaluation covers 134 ALFWorld tasks in unseen rooms, 500 held-out shopping instructions, and $1{,}684$ held-out ScienceWorld tasks. We use Qwen2.5-Instruct at 1.5B and 7B on ALFWorld and WebShop, and at 1.5B on ScienceWorld.

Baseline training. We retrain seven baselines: PPO, GRPO, GiGPO, Flow-GRPO, FlowRL, TB, and the modified SubTB for LLM (Schulman et al., 2017; Shao et al., 2024; Feng et al., 2026; Li et al., 2026; Zhu et al., 2026; Malkin et al., 2022a; Hu et al., 2024). Within each benchmark and model size, all methods share the starting checkpoint, prompt, policy LoRA configuration, and evaluation procedure. ALFWorld and ScienceWorld use shared supervised starts. WebShop starts directly from the instruction-tuned model. Algorithm-specific training settings are documented in Appendix D.4. In the main ALFWorld comparison, G^{2}ARL, TB, and SubTB train on 64 tasks. The other methods use all $3{,}553$ training tasks. At 1.5B, training G^{2}ARL on the full set gives 87.34% success against 87.55% with the subset (Table 17).

Comparison with published ALFWorld results. Our evaluation tests generalization to rooms absent from training. GiGPO's original released code instead evaluates held-out configurations in familiar rooms (Feng et al., 2026). Our supervised initialization and LoRA training also differ from its original instruction-tuned training setup. Published and retrained scores are therefore not directly comparable. Appendix D.4 documents these differences and the familiar-room comparison.

Table 1: **Task success and behavior on ALFWorld and WebShop.** Evaluation uses our retrained models at decoding temperature 0.4. Rep is the repeat-action rate over successful rollouts. Appendix B.3 gives both bases. Bold marks the best value in each column and model-size setting. Init. is each block's starting policy (Appendix D.2). Metric definitions are in Section 4.2.

	ALFWorld				WebShop				
Method	Succ ↑(%)	Rep (%) ↓	Len ↓	UV ↑	Score ↑	Succ ↑(%)	Rep (%) ↓	Len ↓	UV ↑
Qwen2.5-1.5B-Instruct									
Init.	$48.89_{\pm 0.52}$	$23.52_{\pm 0.55}$	$19.91_{\pm 0.27}$	$\mathbf{7.321_{\pm 0.129}}$	$22.27_{\pm 0.07}$	$5.26_{\pm 0.15}$	$34.08_{\pm 2.51}$	$8.89_{\pm 0.35}$	$2.524_{\pm 0.078}$
PPO	$67.34_{\pm 0.87}$	$22.44_{\pm 0.40}$	$16.59_{\pm 0.19}$	$5.749_{\pm 0.295}$	$75.88_{\pm 0.26}$	$50.66_{\pm 0.23}$	$12.16_{\pm 0.22}$	$6.25_{\pm 0.04}$	$9.323_{\pm 0.009}$
GRPO	$65.71_{\pm 1.00}$	$21.57_{\pm 0.55}$	$18.16_{\pm 0.35}$	$6.289_{\pm 0.164}$	$71.49_{\pm 0.09}$	$55.99_{\pm 0.10}$	$16.05_{\pm 0.19}$	$6.51_{\pm 0.03}$	$9.503_{\pm 0.147}$
GiGPO	$69.62_{\pm 0.66}$	$20.64_{\pm 0.26}$	$15.62_{\pm 0.17}$	$5.316_{\pm 0.082}$	$77.39_{\pm 0.30}$	$64.19_{\pm 0.34}$	$13.61_{\pm 0.12}$	$6.37_{\pm 0.01}$	$7.332_{\pm 0.057}$
Flow-GRPO	$62.69_{\pm 0.68}$	$20.38_{\pm 0.16}$	$18.18_{\pm 0.16}$	$6.682_{\pm 0.136}$	$73.22_{\pm 0.21}$	$60.98_{\pm 0.19}$	$13.21_{\pm 0.04}$	$6.33_{\pm 0.03}$	$\mathbf{11.310_{\pm 0.132}}$
FlowRL	$56.11_{\pm 0.26}$	$22.09_{\pm 0.57}$	$19.04_{\pm 0.29}$	$6.936_{\pm 0.032}$	$14.61_{\pm 0.29}$	$3.45_{\pm 0.14}$	$33.59_{\pm 0.48}$	$6.35_{\pm 0.11}$	$3.965_{\pm 0.284}$
TB	$54.41_{\pm 0.21}$	$22.96_{\pm 0.48}$	$19.28_{\pm 0.27}$	$7.296_{\pm 0.143}$	$49.04_{\pm 0.35}$	$13.34_{\pm 0.37}$	$29.07_{\pm 0.35}$	$7.99_{\pm 0.02}$	$3.647_{\pm 0.180}$
SubTB (Hu et al.)	$51.11_{\pm 0.79}$	$23.07_{\pm 0.08}$	$19.53_{\pm 0.05}$	$7.253_{\pm 0.113}$	$40.81_{\pm 0.45}$	$8.94_{\pm 0.16}$	$28.98_{\pm 0.57}$	$8.02_{\pm 0.13}$	$3.159_{\pm 0.101}$
G^2ARL (ours)	$\mathbf{87.55_{\pm 0.17}}$	$\mathbf{15.71_{\pm 0.13}}$	$\mathbf{13.41_{\pm 0.09}}$	$7.037_{\pm 0.094}$	$\mathbf{91.79_{\pm 0.07}}$	$\mathbf{84.92_{\pm 0.03}}$	$\mathbf{0.00_{\pm 0.00}}$	$\mathbf{5.06_{\pm 0.00}}$	$6.300_{\pm 0.084}$
Qwen2.5-7B-Instruct									
Init.	$64.47_{\pm 0.69}$	$12.90_{\pm 0.31}$	$19.61_{\pm 0.16}$	$4.426_{\pm 0.168}$	–	–	–	–	–
PPO	$80.88_{\pm 0.31}$	$12.27_{\pm 0.04}$	$16.91_{\pm 0.04}$	$3.614_{\pm 0.078}$	$78.29_{\pm 0.07}$	$63.51_{\pm 0.16}$	$4.95_{\pm 0.06}$	$5.94_{\pm 0.02}$	$7.967_{\pm 0.104}$
GRPO	$81.44_{\pm 0.13}$	$13.48_{\pm 0.22}$	$16.91_{\pm 0.08}$	$3.930_{\pm 0.100}$	$71.23_{\pm 0.29}$	$59.08_{\pm 0.20}$	$9.65_{\pm 0.05}$	$6.31_{\pm 0.01}$	$8.168_{\pm 0.067}$
GiGPO	$88.61_{\pm 0.21}$	$13.17_{\pm 0.11}$	$14.50_{\pm 0.04}$	$3.448_{\pm 0.081}$	$81.32_{\pm 0.14}$	$70.50_{\pm 0.26}$	$3.91_{\pm 0.02}$	$5.58_{\pm 0.01}$	$6.409_{\pm 0.017}$
Flow-GRPO	$81.72_{\pm 0.22}$	$13.13_{\pm 0.13}$	$16.63_{\pm 0.06}$	$3.982_{\pm 0.065}$	$81.14_{\pm 0.15}$	$71.48_{\pm 0.16}$	$9.98_{\pm 0.25}$	$6.17_{\pm 0.03}$	$\mathbf{10.104_{\pm 0.052}}$
FlowRL	$65.32_{\pm 0.50}$	$12.85_{\pm 0.34}$	$19.47_{\pm 0.15}$	$4.367_{\pm 0.067}$	–	–	–	–	–
TB	$73.24_{\pm 0.50}$	$15.28_{\pm 0.17}$	$19.58_{\pm 0.13}$	$\mathbf{5.294_{\pm 0.070}}$	$4.48_{\pm 0.23}$	$1.04_{\pm 0.04}$	$21.57_{\pm 1.13}$	$7.55_{\pm 0.39}$	$2.917_{\pm 0.402}$
SubTB (Hu et al.)	$68.85_{\pm 0.58}$	$13.46_{\pm 0.25}$	$19.49_{\pm 0.15}$	$4.472_{\pm 0.072}$	$11.71_{\pm 0.09}$	$2.35_{\pm 0.08}$	$15.54_{\pm 0.39}$	$7.87_{\pm 0.30}$	$2.357_{\pm 0.184}$
G^2ARL (ours)	$\mathbf{92.79_{\pm 0.22}}$	$\mathbf{11.57_{\pm 0.18}}$	$\mathbf{13.86_{\pm 0.07}}$	$5.217_{\pm 0.071}$	$\mathbf{89.97_{\pm 0.07}}$	$\mathbf{82.15_{\pm 0.05}}$	$\mathbf{0.16_{\pm 0.02}}$	$\mathbf{5.05_{\pm 0.00}}$	$9.199_{\pm 0.076}$

Subscripts are standard deviations over decoding seeds. –: not evaluated or no successful rollouts.

Evaluation. Main results use decoding temperature 0.4, with 16 rollouts per task on ALFWorld and WebShop and one on ScienceWorld. We additionally evaluate the 1.5B models, and 7B models on ALFWorld and WebShop, at temperatures 0.7, 1.0, and 1.3 (Appendix B.1). On WebShop the 1.5B Init. row and the 7B TB and SubTB (Hu et al.) configurations are read at t=0.4 only (Table 1). Each method is evaluated on the same tasks and with the same rollout budget within a setting.

4.2 Evaluation metrics

We evaluate task performance, efficiency, repetition, and solution diversity. **Success** is the percentage of individual rollouts accepted by the environment verifier, rather than the fraction of tasks solved at least once. **Score** is the mean environment score on the native WebShop and ScienceWorld scales. **Len** is the mean number of turns among successful rollouts.

For a trajectory τ, let ℓ_τ be its number of commands and U_τ the number of verbatim-distinct commands. The repeat-action rate over a set of rollouts A is

$$\mathrm{Rep}(A) = \frac{100}{|A|} \sum_{\tau \in A} \left(1 - \frac{U_\tau}{\ell_\tau}\right). \tag{10}$$

We report Rep over successful rollouts, and over all rollouts in Appendix B.3. **Loop share** is the percentage of rollouts containing an immediately repeated block of 2 to 12 commands whose repeated span reaches at least six actions. To measure variation across solutions, let $\mathcal{X}^+$ contain tasks solved by at least one rollout, $\mathcal{E}^+$ the set of successful rollouts and $\mathbf{c}(\tau)$ the command sequence of τ. **UV** is the mean number of verbatim-distinct successful command sequences per solved task:

$$\mathrm{UV} = \frac{1}{|\mathcal{X}^+|} \sum_{x \in \mathcal{X}^+} \left|\left\{\mathbf{c}(\tau) : \tau \in \mathcal{E}^+,\ x_\tau = x\right\}\right|. \tag{11}$$

Metrics are first computed within each evaluation run and then averaged across runs. We report standard deviations across decoding seeds unless otherwise noted. The ALFWorld-1.5B G^2ARL and GiGPO rows average different training runs (Table 18), and the standard deviation over runs is 2.48 points of success for G^2ARL. Paired 95% confidence intervals use a task-level cluster bootstrap with 10,000 resamples, keeping all rollouts of a task together (Appendix D.1).

Table 2: **Task success and behavior on ScienceWorld at 1.5B.** Evaluation uses temperature 0.4 and one rollout per task. Notation follows Table 1. SFT init. is the shared starting policy.

Method	Score ↑	Succ (%) ↑	Rep (%) ↓	Len ↓
SFT init.	$5.61_{\pm 0.93}$	$36.10_{\pm 0.22}$	$13.26_{\pm 0.23}$	$13.37_{\pm 0.07}$
PPO	$33.76_{\pm 0.94}$	$58.79_{\pm 0.06}$	$5.91_{\pm 0.01}$	$9.94_{\pm 0.04}$
GRPO	$31.89_{\pm 0.40}$	$57.32_{\pm 0.24}$	$7.38_{\pm 0.11}$	$10.88_{\pm 0.01}$
GiGPO	$35.53_{\pm 1.29}$	$59.16_{\pm 0.72}$	$4.54_{\pm 0.18}$	$8.44_{\pm 0.04}$
Flow-GRPO	$33.15_{\pm 0.37}$	$54.65_{\pm 0.03}$	$12.28_{\pm 0.21}$	$12.35_{\pm 0.10}$
FlowRL	$3.86_{\pm 0.58}$	$11.32_{\pm 0.61}$	$12.34_{\pm 0.52}$	$9.33_{\pm 0.19}$
TB	$15.63_{\pm 0.73}$	$37.00_{\pm 0.24}$	$15.04_{\pm 0.39}$	$14.24_{\pm 0.17}$
SubTB (Hu et al.)	$11.22_{\pm 1.35}$	$37.57_{\pm 0.74}$	$13.80_{\pm 0.12}$	$13.65_{\pm 0.01}$
G^{2}ARL (ours)	$\mathbf{48.04}_{\pm \mathbf{0.91}}$	$\mathbf{66.79}_{\pm \mathbf{0.48}}$	$\mathbf{2.20}_{\pm \mathbf{0.05}}$	$\mathbf{7.94}_{\pm \mathbf{0.01}}$

Subscripts are standard deviations over decoding seeds. Rep is the repeat-action rate over successful rollouts (both bases: Appendix B.3).

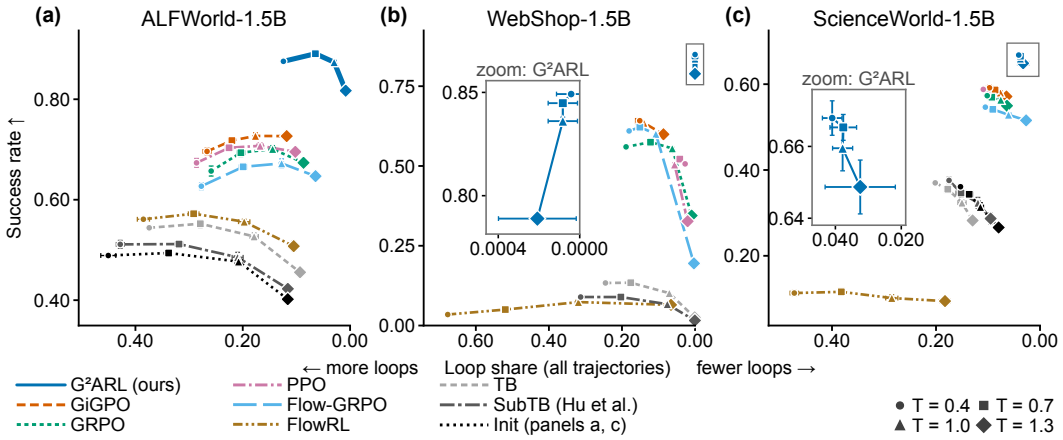

Figure 2: **Success and loop share across decoding temperatures at 1.5B.** Panels show ALFWorld, WebShop, and ScienceWorld. The reversed horizontal axis places fewer loops to the right. Markers indicate temperatures 0.4, 0.7, 1.0, and 1.3. Insets enlarge G^{2}ARL's results. Bars show decoding-seed standard deviations, except the ALFWorld G^{2}ARL and GiGPO points, whose bars are over their training runs (Table 8). Single-evaluation points have no error bar.

4.3 Main Results

Task success. G^{2}ARL achieves the highest success in all five benchmark and model-size settings (Tables 1 and 2). At 1.5B, gains over GiGPO, the strongest retrained comparator, are 17.9, 20.7, and 7.6 percentage points on ALFWorld, WebShop, and ScienceWorld, respectively. The corresponding 95% task-bootstrap intervals are $[12.8, 23.1]$, $[17.8, 23.7]$, and $[5.3, 10.0]$ points. At 7B, gains over the strongest comparator are 4.2 points on ALFWorld and 10.7 on WebShop, with intervals $[-1.2, 9.8]$ and $[8.2, 13.2]$ points. The WebShop-7B FlowRL cell, which solves no task, is outside the paired comparisons (Appendix D.2).

Training controls. Using TB and SubTB's reward scale, $\beta = 8$, G^{2}ARL reaches 80.18% at 1.5B and 92.14% at 7B (Table 4). These and the full-training-set control of Section 4.1 exceed the corresponding GiGPO results. On the full ALFWorld training set, on which every policy-gradient row trains, the paired difference against GiGPO is $+17.72$ points, $[+12.65, +22.91]$ over tasks (Appendix B.5).

4.4 Behavior

Higher success is accompanied by shorter successful trajectories and lower repeat-action rates on both successful and all rollouts in every main setting (Tables 1 and 2, Appendix B.3). Loop share on all rollouts is also lowest in each setting (Appendix B.3). On WebShop at 1.5B, loop share is below 0.01%, compared with 15.13% for GiGPO.

Across four decoding temperatures, G^{2}ARL has the highest success in all 12 benchmark and temperature combinations and the lowest loop share in 11 (Figure 2). The exception is ScienceWorld at

Table 3: **Ablations on ALFWorld-1.5B**. Each row is one configuration of G^{2}ARL. Row (1) is the Table 1 row (two training runs), rows (2) to (7) are single training runs at β=8, read against row (2). Row (3) replaces the loss by its equal-weight degenerate case, Var-TB (Lemma 1(i)), and keeps the gate and buffer. Row (6) replaces the coverage rule by reward-prioritized replay. Each row changes only the setting its name gives (Appendix B.5). Subscripts are standard deviations over decoding seeds.

Row	Configuration	Succ (%) ↑	UV ↑
(1)	Full recipe, β=64 (Table 1)	**87.55**$_{\pm 0.17}$	7.037$_{\pm 0.094}$
(2)	Full recipe, β=8	80.18$_{\pm 0.50}$	**7.414**$_{\pm 0.142}$
(3)	Var-TB in place of the flow reallocation loss	66.31$_{\pm 0.69}$	6.680$_{\pm 0.315}$
(4)	Failure-failure pair weight 0.3 → 1.0	67.74$_{\pm 0.35}$	6.433$_{\pm 0.112}$
(5)	Replay consumption off	65.71$_{\pm 0.71}$	5.318$_{\pm 0.169}$
(6)	Reward-prioritized replay instead	60.48$_{\pm 0.49}$	5.106$_{\pm 0.067}$
(7)	Replay off, Var-TB in place of the loss	55.47$_{\pm 0.73}$	5.625$_{\pm 0.216}$

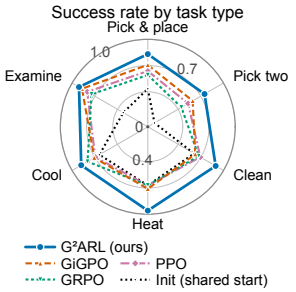

Figure 3: **ALFWorld success by task family at 1.5B.** Radius shows success rate on tasks in unseen rooms, at decoding temperature 0.4.

temperature 1.3, where Flow-GRPO has fewer loops. On ALFWorld at 1.5B, the success advantage extends to all six task families (Figure 3). UV is above GiGPO's in three of the four blocks of Table 1 and below it on WebShop-1.5B, where every policy-gradient row is higher. The highest UV in each block belongs to another row. For binary scores, the target's conditional distribution over successful trajectories depends on the starting policy and the invalid-action cost (Equations (1) and (13)). Repetition within an episode and variety across solutions are separate measurements (Figure 8).

4.5 Ablations and Training Dynamics

Rows (3) to (6) of Table 3 each alter one design choice of the β=8 recipe, row (2). Row (7) alters two. *Flow reallocation loss.* Var-TB, its equal-weight case, row (3), reads 13.87 points of success and 0.73 UV below row (2). Raising only the failure-failure weight to 1, row (4), reads 12.44 points and 0.98 UV below (Appendix A.5). *Replay.* Turning replay consumption off, row (5), reads 14.47 points and 2.10 UV below row (2). Reward-prioritized replay (Shen et al., 2023), row (6), reads 19.70 points and 2.31 UV below. Row (6) restores the contrast yet reads 5.23 points below row (5). Rows (2), (3), (5), and (7) form a 2×2 ablation of loss and replay: the flow reallocation loss adds 10.24 points without replay and 13.87 points with replay, while replay adds 10.84 points under Var-TB and 14.47 points under the flow reallocation loss. These are conditional rather than additive effects. *Reward scale.* The β=64 configuration, row (1), reads 7.37 points of success higher and 0.38 lower in UV than the β=8 configuration, row (2).

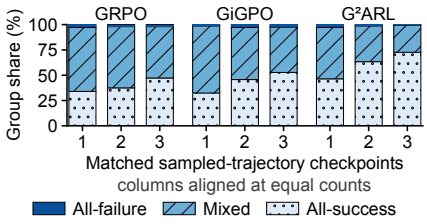

Figure 4: **ALFWorld-1.5B training.** Group composition at three matched points (Appendix B.5).

Groups in which all trajectories succeed or all fail become more frequent as training progresses (Figure 4). Offline sampling from the GiGPO checkpoint at the last plotted point of Figure 4 produces 55% homogeneous groups (all-success or all-failure), on which a trajectory-level group-normalized advantage carries no success-failure contrast. Among freshly sampled groups in our run, row (2) of Table 3, the all-success share reaches 72.6% at the last plotted point (Appendix B.5). On ALFWorld, G^{2}ARL drops these all-success groups before the pairwise update. The two components act on the same contrast: the weights set the share of the success-failure contrast within a mixed group, while replay determines whether such a contrast is present in the group and the selection rule determines which stored successes supply it (row (6) against row (5)). The same buffer and injection rule added to GiGPO, with injected rows entering the group normalization as sampled rows, reads -2.30 points $[-4.79, +0.12]$ relative to GiGPO over two training runs each (Appendix B.5), showing no reliable change there. Under the pairwise loss, the same replay rule reads 14.47 points higher (row (2) against row (5)).

5 OBSERVATIONS AND DIAGNOSIS ON GFLOWNET METHODS

On WebShop-1.5B from the instruction-tuned base, the two reference GFlowNet objectives, TB and SubTB (Hu et al.), and FlowRL finish below every policy-gradient row of Table 1. The best of the three reaches 13.34% success. In each objective the terminal score enters only part of the loss terms (Appendix C).

(i) All-success and all-failure groups under TB. With binary scores, the score enters the TB gradient on a group in two places: a calibration term that applies one scalar coefficient to every trajectory's score vector, and a contrast term that separates its successes from its failures (Proposition 5). On a group whose trajectories all succeed or all fail, the contrast term is zero, and the remaining update only fits the group to the shape of the cost-adjusted reference at a level set by the partition parameter (Proposition 6). A group-normalized advantage is zero on such a group, so the trajectory-level advantage term of a policy gradient is zero there, whereas the TB update can remain nonzero (Proposition 7, Appendix C.3). Replacing the replay rule of the WebShop-1.5B TB training by injection of stored successes into its all-failure groups, and dropping the all-failure groups that no stored success can refill, with the objective otherwise unchanged, raises success from 13.34% to 16.73%, a task-paired difference of $+3.39$ points (Appendix C.6). In the first third of that training, 65.0% of the sampled groups are all-failure before injection.

(ii) The substituted prefix flow under SubTB. The modified objective of Hu et al. (2024) expresses the flow of a prefix through the reward of stopping there, which assumes that a trajectory can stop at any state, and needs no separate flow network. An agent episode ends only when the environment ends it, so the port, which keeps that form, gives each interior prefix a flow that ignores what the prefix can still lead to (Appendix C.4). The loss then reaches zero only when every valid interior action keeps its reference probability and the final action alone is reweighted by the score (Proposition 8). Wherever a final action can succeed under the reference policy at a prefix whose available actions are all terminal and valid, that assignment is not a normalized policy, so the loss has no zero (Proposition 9, Appendix C.4). On ALFWorld-1.5B the trained policy stays close to its warm start: success rises by 2.22 points (Table 1) and the per-turn log-ratio to the reference stays near zero (Appendix C.4).

(iii) Zero-advantage groups under FlowRL. FlowRL assigns zero advantage to a group whose trajectories all have the same score. The reward then leaves that group's loss, and the remaining residual fits each length-normalized log-ratio to a learned offset (Propositions 10 and 11). DAPO excludes such groups by resampling (Yu et al., 2026). In the WebShop-7B run these groups carry a mean per-step loss share of 0.7846, and training success falls to zero (Appendix C.5).

6 LIMITATION

The objective focuses on tasks with a terminal reward (Section 3.1). Experiments cover three text environments with a verifier terminal score: ALFWorld, WebShop, and ScienceWorld. All methods train LoRA adapters from a shared starting checkpoint at 1.5B and 7B. Full fine-tuning and larger models are not studied.

7 CONCLUSION

We presented G^{2}ARL, a distribution-matching post-training method for multi-turn language-model agents. Group flow reallocation cancels the partition function in pairwise residuals and weights success-failure comparisons highest. Coverage-guided dynamic rebalancing combines group gating, submodular replay, and conditional injection of stored same-task successes. G^{2}ARL improves success over GiGPO at 1.5B by 17.93, 20.73 and 7.63 points on ALFWorld, WebShop and ScienceWorld (25.8%, 32.3% and 12.9% relative).

REPRODUCIBILITY STATEMENT

Section 3 states the objective, the group-level gate and the replay selection rule. Appendix A gives the standing assumptions and complete proofs. Appendix D.1 fixes the evaluation protocol (task splits, k, decoding seeds, resampling units), Tables 20 and 21 list the adapter, rollout, reward and replay settings of the main G^2ARL rows and of every retrained baseline. The training stack, the evaluation entry point, the frozen analysis scripts behind the metrics of Appendix D.1, and the training and evaluation code will be released upon acceptance.

AI USE STATEMENT

In this work, we used generative AI tools for polishing and shortening prose drafted by the authors, and checking the manuscript for internal consistency. We also used AI coding assistants to implement and debug parts of the training and evaluation code. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. *Advances in neural information processing systems*, 30, 2017.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025.

Brian Bartoldson, Siddarth Venkatraman, James Diffenderfer, Moksh Jain, Tal Ben-Nun, Seanie Lee, Minsu Kim, Johan Obando Ceron, Yoshua Bengio, and Bhavya Kailkhura. Trajectory balance with asynchrony: Decoupling exploration and learning for fast, scalable llm post-training. *Advances in Neural Information Processing Systems*, 38:113901–113931, 2026.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in neural information processing systems*, 34:27381–27394, 2021.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023.

Xiwen Chen, Wenhui Zhu, Peijie Qiu, Xuanzhao Dong, Hao Wang, Haiyu Wu, Huayu Li, Aris Sotiras, Yalin Wang, and Abolfazl Razi. Dra-grpo: Your grpo needs to know diverse reasoning paths for mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13995–14019, 2026.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Tiago Da Silva, Luiz Max Carvalho, Amauri Souza, Samuel Kaski, and Diego Mesquita. Embarrassingly parallel gflownets. *arXiv preprint arXiv:2406.03288*, 2024.

Tristan Deleu, Padideh Nouri, Nikolay Malkin, Doina Precup, and Yoshua Bengio. Discrete probabilistic inference as control in multi-path environments. *arXiv preprint arXiv:2402.10309*, 2024.

Tristan Deleu, Padideh Nouri, Yoshua Bengio, and Doina Precup. Relative trajectory balance is equivalent to trust-pcl. *arXiv preprint arXiv:2509.01632*, 2025.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026.

Luca Gioacchini, Giuseppe Siracusano, Davide Sanvito, Kiril Gashteovski, David Friede, Roberto Bifulco, and Carolin Lawrence. Agentquest: A modular benchmark framework to measure

progress and improve llm agents. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pp. 185–193, 2024.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

Edward Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio, and Nikolay Malkin. Amortizing intractable inference in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 43053–43083, 2024.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Haoqiang Kang, Enna Sachdeva, Piyush Gupta, Sangjae Bae, and Kwonjoon Lee. Gflowvlm: Enhancing multi-step reasoning in vision-language models with generative flow networks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3815–3825. IEEE, 2025.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. In *International Conference on Learning Representations*, volume 2024, pp. 20620–20653, 2024.

Minchan Kwon, Sunghyun Baek, Minseo Kim, Jaemyung Yu, Dongyoon Han, and Junmo Kim. Stable-gflownet: Toward diverse and robust llm red-teaming via contrastive trajectory balance. *arXiv preprint arXiv:2605.00553*, 2026.

Seanie Lee, Minsu Kim, Lynn Cherif, David Dobre, Juho Lee, Sung Ju Hwang, Kenji Kawaguchi, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, et al. Learning diverse attacks on large language models for robust red-teaming and safety tuning. In *International Conference on Learning Representations*, volume 2025, pp. 63141–63163, 2025.

Siheng Li, Zhanhui Zhou, Wai Lam, Chao Yang, and Chaochao Lu. Repo: Replay-enhanced policy optimization. *arXiv preprint arXiv:2506.09340*, 2025.

Zhuofeng Li, Haoxiang Zhang, Seungju Han, Sheng Liu, Jianwen Xie, Yu Zhang, Yejin Choi, James Y Zou, and Pan Lu. In-the-flow agentic system optimization for effective planning and tool use. In *International Conference on Learning Representations*, volume 2026, pp. 50524–50570, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024.

Xiaodong Liu, Michael Xu, Jack W Stokes, Paul Smolensky, Doug Burger, and Jianfeng Gao. Gflowrl: Scaling distribution-matching rl to large language models. *arXiv preprint arXiv:2607.13394*, 2026a.

Xiaoqian Liu, Ke Wang, Yuchuan Wu, Fei Huang, Yongbin Li, Jianbin Jiao, and Junge Zhang. Agentic reinforcement learning with implicit step rewards. In *International Conference on Learning Representations*, volume 2026, pp. 129271–129291, 2026b.

Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K Qiu, and Yuqing Yang. Agent lightning: Train any ai agents with reinforcement learning. *arXiv preprint arXiv:2508.03680*, 2025.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. Agentboard: An analytical evaluation board of multi-turn llm agents. *Advances in neural information processing systems*, 37:74325–74362, 2024.

Kanika Madan, Jarrid Rector-Brooks, Maksym Korablyov, Emmanuel Bengio, Moksh Jain, Andrei Cristian Nica, Tom Bosc, Yoshua Bengio, and Nikolay Malkin. Learning gflownets from partial episodes for improved convergence and stability. In *International Conference on Machine Learning*, pp. 23467–23483. PMLR, 2023.

Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in gflownets. *Advances in Neural Information Processing Systems*, 35:5955–5967, 2022a.

Nikolay Malkin, Salem Lahlou, Tristan Deleu, Xu Ji, Edward Hu, Katie Everett, Dinghuai Zhang, and Yoshua Bengio. Gflownets and variational inference. *arXiv preprint arXiv:2210.00580*, 2022b.

Sobhan Mohammadpour, Emmanuel Bengio, Emma Frejinger, and Pierre-Luc Bacon. Maximum entropy gflownets with soft q-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2593–2601. PMLR, 2024.

Junhyuk Oh, Yijie Guo, Satinder Singh, and Honglak Lee. Self-imitation learning. In *International conference on machine learning*, pp. 3878–3887. PMLR, 2018.

Hyun Bin Park and Du-Seong Chang. Headroom-drift replay: A primitive for principled replay control in grpo. *arXiv preprint arXiv:2609.03941*, 2026.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, et al. Webrl: Training llm web agents via self-evolving online curriculum reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 79791–79821, 2025.

Lorenz Richter, Ayman Boustati, Nikolas Nüsken, Francisco Ruiz, and Omer Deniz Akyildiz. Vargrad: a low-variance gradient estimator for variational inference. *Advances in Neural Information Processing Systems*, 33:13481–13492, 2020.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Max W Shen, Emmanuel Bengio, Ehsan Hajiramezanali, Andreas Loukas, Kyunghyun Cho, and Tommaso Biancalani. Towards understanding and improving gflownet training. In *International conference on machine learning*, pp. 30956–30975. PMLR, 2023.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Ryoichi Takase, Masaya Tsunokake, Yuta Tsuchiya, and Shota Inuzuka. Gflownet fine-tuning for diverse correct solutions in mathematical reasoning tasks. *arXiv preprint arXiv:2410.20147*, 2024.

Daniil Tiapkin, Nikita Morozov, Alexey Naumov, and Dmitry P Vetrov. Generative flow networks as entropy-regularized rl. In *International Conference on Artificial Intelligence and Statistics*, pp. 4213–4221. PMLR, 2024.

Nikhil Vemgal, Elaine Lau, and Doina Precup. An empirical study of the effectiveness of using a replay buffer on mode discovery in gflownets. *arXiv preprint arXiv:2307.07674*, 2023.

Siddarth Venkatraman, Moksh Jain, Luca Scimeca, Minsu Kim, Marcin Sendera, Mohsin Hasan, Luke Rowe, Sarthak Mittal, Pablo Lemos, Emmanuel Bengio, et al. Amortizing intractable inference in diffusion models for vision, language, and control. *Advances in neural information processing systems*, 37:76080–76114, 2024.

Xu Wan, Yansheng Wang, Wenqi Huang, and Mingyang Sun. Buffer matters: Unleashing the power of off-policy reinforcement learning in large language model reasoning. *arXiv preprint arXiv:2602.20722*, 2026.

Chen Wang, Lai Wei, Yanzhi Zhang, Chenyang Shao, Zedong Dan, Weiran Huang, Yuzhi Zhang, and Yue Wang. Eframe: Deeper reasoning via exploration-filter-replay reinforcement learning framework. *arXiv preprint arXiv:2506.22200*, 2025a.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022.

Xi Wang, Wenbo Lu, and Shengjie Wang. Rooted absorbed prefix trajectory balance with submodular replay for gflownet training. *arXiv preprint arXiv:2603.00454*, 2026.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025b.

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Xin Guo, Dingwen Yang, Chenyang Liao, Wei He, et al. Agentgym: Evaluating and training large language model-based agents across diverse environments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27914–27961, 2025a.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Honglin Guo, Jiaqi Liu, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, et al. Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning. *arXiv preprint arXiv:2509.08755*, 2025b.

Zhikun Xu, Yu Feng, Jacob Dineen, Taiwei Shi, Jieyu Zhao, and Ben Zhou. Skill reuse as compression in agentic rl. *arXiv preprint arXiv:2605.31509*, 2026.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *Advances in Neural Information Processing Systems*, 38:117157–117186, 2026.

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. Diversity-aware policy optimization for large language model reasoning. *Advances in Neural Information Processing Systems*, 38: 94801–94826, 2026.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022b.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Fangxu Yu, Lai Jiang, Haoqiang Kang, Shibo Hao, and Lianhui Qin. Flow of reasoning: Training llms for divergent reasoning with minimal examples. *arXiv preprint arXiv:2406.05673*, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Runzhe Zhan, Yafu Li, Zhi Wang, Xiaoye Qu, Dongrui Liu, Jing Shao, Derek Wong, and Yu Cheng. Exgrpo: Learning to reason from experience. In *International Conference on Learning Representations*, volume 2026, pp. 12810–12846, 2026.

David W Zhang, Corrado Rainone, Markus Peschl, and Roberto Bondesan. Robust scheduling with gflownets. *arXiv preprint arXiv:2302.05446*, 2023.

Hongzhi Zhang, Jia Fu, Jingyuan Zhang, Kai Fu, Qi Wang, Fuzheng Zhang, and Guorui Zhou. Rlep: Reinforcement learning with experience replay for llm reasoning. *arXiv preprint arXiv:2507.07451*, 2025.

Kaichen Zhang, Yuzhong Hong, Junwei Bao, Hongfei Jiang, Yang Song, Hong Dingqian, and Hui Xiong. Gvpo: Group variance policy optimization for large language model post-training. *Advances in Neural Information Processing Systems*, 38:165798–165820, 2026a.

Mingda Zhang, Tiesunlong Shen, Haoran Luo, Wenjin Liu, Zikai Xiao, Erik Cambria, and Xiaoying Tang. Skillflow: Flow-driven recursive skill evolution for agentic orchestration. *arXiv preprint arXiv:2605.14089*, 2026b.

Zijing Zhang, Ziyang Chen, Mingxiao Li, Zhaopeng Tu, and Xiaolong Li. Rlvmr: Reinforcement learning with verifiable meta-reasoning rewards for robust long-horizon agents. In *International Conference on Learning Representations*, volume 2026, pp. 114676–114700, 2026c.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, volume 2024, pp. 15585–15606, 2024.

Xuekai Zhu, Daixuan Cheng, Dinghuai Zhang, Henry Li, Kaiyan Zhang, Che Jiang, Youbang Sun, Ermo Hua, Yuxin Zuo, Xingtai Lv, et al. Flowrl: Matching reward distributions for llm reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 151758–151780, 2026.

A METHOD: TARGET, LOSS AND PROCEDURE

We use q_i, W, L, and G from Section 3.2. All results concern a fixed task x. Let $\mathcal{T}_x$ be the finite set of trajectories that end at an environment terminal state or at the turn limit. Finiteness follows from a finite token vocabulary, per-turn token caps, and a finite episode horizon. Both $R_x(\tau)$ and the policies considered below are positive on this common support, and $0 < Z_x = \sum_{\tau \in \mathcal{T}_x} R_x(\tau) < \infty$.

Assumptions and scoring. Environment transitions are deterministic given the task and action sequence. Each action is a generated token string that ends at the first end-of-turn token or at the per-turn token cap. These outcomes define a normalized action distribution. Together with the episode turn limit, they give $\sum_{\tau \in \mathcal{T}_x} \pi_\theta(\tau \mid x) = 1$. An action includes generated reasoning and the environment command. Its log-probability sums all generated-token log-probabilities, but no prompt or environment-observation tokens. Consequently $\log \pi_\theta(\tau \mid x) = \sum_{t<T} \log \pi_\theta(a_t \mid s_t)$, and $q(\tau) = \log \pi_\theta(\tau \mid x) - \log R_x(\tau)$ is the residual of Equation (2). The policy is trained through LoRA, and π_{ref} is the same starting checkpoint with the trainable adapter disabled.

On ALFWorld and WebShop, y is the binary success bit (Section 3.1). The sets W and L always use the separate binary success flag. Any statement that requires binary y is restricted to ALFWorld and WebShop. Pair weights, reference probabilities, scores, and invalid-turn counts are constants with respect to differentiation. We assume $G > 1$ and $\sum_{i<j} w_{ij} > 0$ for every evaluated pairwise loss. For an empty class, its residual mean and within-class sum of squares are set to zero.

A.1 THE TARGET DISTRIBUTION AND ITS VARIATIONAL CHARACTERIZATION

The target solves a regularized control problem. For the deployed choice $\rho = 1$, let $r(\tau) = \beta y(\tau) - \lambda n(\tau)$, so that $\log \pi^*(\tau \mid x) = \log \pi_{\text{ref}}(\tau \mid x) + r(\tau) - \log Z_x$. Substituting this into the divergence gives, for any normalized policy π on $\mathcal{T}_x$, $\mathrm{KL}(\pi \| \pi^*) = \mathrm{KL}(\pi \| \pi_{\text{ref}}) - \mathbb{E}_\pi[r] + \log Z_x$, which rearranges to Equation (6). Non-negativity of KL divergence, with equality only for identical distributions, makes π^* the unique maximizer of $\mathcal{J}$ over all trajectory distributions, the variational characterization of the reference-tilted target (Venkatraman et al., 2024; Deleu et al., 2025). It identifies the solution, not an equality between the gradients of $\mathcal{J}$ and the weighted pairwise loss.

Class marginal and conditionals. For binary scores, let $B_1 = \sum_{\tau:y(\tau)=1} \pi_{\mathrm{ref}}(\tau \mid x)^\rho e^{-\lambda n(\tau)}$ and let B_0 be the same sum over failures. When both are positive,

$$\frac{\pi^*(y=1 \mid x)}{\pi^*(y=0 \mid x)} = e^\beta \frac{B_1}{B_0}, \tag{12}$$

and among successful trajectories

$$\pi^*(\tau \mid x, y=1) \propto \pi_{\mathrm{ref}}(\tau \mid x)^\rho e^{-\lambda n(\tau)}, \tag{13}$$

with the same form for the conditional over failed trajectories. Hence β enters the target only through the class marginal, and the two conditionals depend on the reference and on λ alone.

A.2 Trajectory probabilities as flows

Consider the prefix tree of complete trajectories, retaining the complete history rather than merging different histories with the same environment observation. For a prefix h, let $\mathcal{T}_x(h)$ be its complete extensions and set $F_\theta(h) = Z_x \sum_{\tau \in \mathcal{T}_x(h)} \pi_\theta(\tau \mid x)$, so that $F_\theta(\emptyset) = Z_x$ and $F_\theta(\tau) = Z_x \pi_\theta(\tau \mid x)$. At every non-terminal prefix, the possible next actions partition its complete extensions, so

$$F_\theta(h) = \sum_a F_\theta(ha), \tag{14}$$

the flow-conservation condition on a tree (Bengio et al., 2023), where the flow into ha is the flow on its unique incoming edge. Thus $\pi_\theta(\tau \mid x) = F_\theta(\tau)/F_\theta(\emptyset)$ is the trajectory's fraction of total flow, and matching probability ratios matches relative terminal flows, while Z_x only exhibits a valid flow and is never evaluated or learned. At $\pi_\theta = \pi^*$, $F^*(h) = \sum_{\tau \in \mathcal{T}_x(h)} R_x(\tau)$, so terminal flows satisfy $F^*(\tau) = R_x(\tau)$ and the induced action policy

$$\pi^*(a \mid h) = \frac{F^*(ha)}{F^*(h)} = \frac{\sum_{\tau \in \mathcal{T}_x(ha)} R_x(\tau)}{\sum_{\tau \in \mathcal{T}_x(h)} R_x(\tau)} \tag{15}$$

is normalized, because the numerators over the available actions sum to the denominator. No prefix reward or learned state-flow head is introduced.

A.3 Closed form of the pairwise loss

Define the class means $\bar{q}_W$, $\bar{q}_L$, their difference $\Delta = \bar{q}_W - \bar{q}_L$, and the within-class sums of squares $V_W = \sum_{i \in W}(q_i - \bar{q}_W)^2$ and $V_L = \sum_{i \in L}(q_i - \bar{q}_L)^2$.

Proposition 2 (Closed form)**.** *For all $q \in \mathbb{R}^G$,*

$$\begin{aligned}\sum_{i<j} w_{ij}(q_i - q_j)^2 = (w_{WW} n_W + w_{WL} n_L) V_W \\ + (w_{LL} n_L + w_{WL} n_W) V_L + w_{WL} n_W n_L \Delta^2.\end{aligned} \tag{16}$$

Differentiating either side gives the same gradient.

Proof. For numbers $z_1, \ldots, z_n$ with mean $\bar{z}$, expanding the squares gives

$$\sum_{i<j}(z_i - z_j)^2 = n \sum_i (z_i - \bar{z})^2, \tag{17}$$

so the within-class pair sums are $n_W V_W$ and $n_L V_L$. For cross-class pairs, write $q_i - q_j = (q_i - \bar{q}_W) - (q_j - \bar{q}_L) + \Delta$. The centered terms sum to zero, giving

$$\sum_{i \in W} \sum_{j \in L} (q_i - q_j)^2 = n_L V_W + n_W V_L + n_W n_L \Delta^2. \tag{18}$$

Multiplying the three sums by their class weights yields Equation (16), an identity in q whose two sides therefore have the same derivatives. □

The implementation evaluates the right-hand side, whose class means and sums are accumulated in linear time in G and remain differentiable, instead of constructing all pairs.

Uniform weights and the score term. Let $\bar{q}_g = G^{-1}\sum_i q_i$. Optimizing TB over its shared additive constant gives

$$\min_c \frac{1}{G}\sum_i (q_i + c)^2 = \frac{1}{G}\sum_i (q_i - \bar{q}_g)^2 = \mathcal{L}_{\mathrm{c}}, \qquad c^* = -\bar{q}_g. \tag{19}$$

This is the group-centered log-variance objective (Richter et al., 2020; Zhang et al., 2023; Malkin et al., 2022b).

Lemma 1 (Uniform weights and score differences)**.** *(i) If $w_{WW} = w_{LL} = w_{WL} > 0$, then $\mathcal{L}_{\mathrm{pairs}} = \mathcal{L}_{\mathrm{c}}$ in value and gradient. (ii) For binary y, the terminal-score difference in $q_i - q_j$ vanishes within a class and equals $\pm\beta$ across classes. Only successful-failed pairs therefore contain this score difference.*

Proof. For equal weights $w > 0$, Equation (17) gives $\sum_{i<j} w(q_i - q_j)^2 = wG\sum_i (q_i - \bar{q}_g)^2$, and $\mathcal{N}_g = \frac{2G}{G-1} w\binom{G}{2} = wG^2$. Their ratio, $\mathcal{L}_{\mathrm{pairs}}$, equals $\mathcal{L}_{\mathrm{c}}$ identically in q, hence also in gradient. Detaching $\bar{q}_g$ gives the same gradient because $\sum_i (q_i - \bar{q}_g) = 0$. For part (ii), Equation (2) gives $q_i - q_j = (D_i - D_j) - \beta(y_i - y_j) + \lambda(n_i - n_j)$. The score term cancels when $y_i = y_j$, and $y_i - y_j = \pm 1$ for binary outcomes in different classes. □

A.4 Target preservation under any pair weighting

For a global statement, enumerate the distinct trajectories in $\mathcal{T}_x$ as $\tau_1, \ldots, \tau_M$. Let $a_{uv} \geq 0$ be fixed aggregate comparison weights and define

$$\mathcal{E}(\pi) = \sum_{u<v} a_{uv}\left[\log\frac{\pi(\tau_u \mid x)}{R_x(\tau_u)} - \log\frac{\pi(\tau_v \mid x)}{R_x(\tau_v)}\right]^2. \tag{20}$$

Each a_{uv} aggregates the pair weights, normalizers, group weights and sampling frequencies of every group in which the pair appears. The comparison graph has one vertex per trajectory and an edge whenever $a_{uv} > 0$.

Proposition 3 (Target preservation under connected comparisons)**.** *Assume the graph of a_{uv} is connected and the weights are fixed. (i) $\mathcal{E}(\pi) = 0$ if and only if $\pi = \pi^*$. (ii) Parameterize an arbitrary positive trajectory distribution by $\pi_z(\tau_u \mid x) = e^{z_u}/\sum_v e^{z_v}$. Then $\mathcal{E}(\pi_z)$ is a quadratic form in z whose Hessian is positive definite on the complement of the constant-logit direction, and every stationary point of $\mathcal{E}(\pi_z)$ represents π^*. (iii) Changing non-negative comparison weights while preserving connected support changes neither this stationary distribution nor the set of zero-loss distributions.*

Proof. Let $r = (r_1, \ldots, r_M)^\top$ with $r_u = \log R_x(\tau_u)$, and let H be the weighted graph Laplacian, $H_{uv} = -a_{uv}$ for $u \neq v$ and $H_{uu} = \sum_{v\neq u} a_{uv}$, so that $H\mathbf{1} = 0$ and, for every $b \in \mathbb{R}^M$, $b^\top H b = \sum_{u<v} a_{uv}(b_u - b_v)^2 \geq 0$. By connectivity, $\ker H = \mathrm{span}\{\mathbf{1}\}$, since a zero quadratic form forces equality along every edge. Every positive π is π_z with $z_u = \log\pi(\tau_u \mid x)$, and the shared normalizer $\log\sum_v e^{z_v}$ cancels from every residual difference, so $\mathcal{E}(\pi_z) = (z - r)^\top H(z - r)$, with gradient $2H(z-r)$ and Hessian $2H$, which is positive definite on the complement of $\mathbf{1}$. Hence $\mathcal{E}(\pi_z) = 0$ and $\nabla_z\mathcal{E} = 0$ are each equivalent to $z - r = c\mathbf{1}$, and then $\pi_z(\tau_u \mid x) = e^{r_u+c}/\sum_v e^{r_v+c} = R_x(\tau_u)/Z_x$, which proves (i) and (ii). Conversely, $z = r$ gives $\pi_z = \pi^*$ with $\mathcal{E} = 0$ and $\nabla_z\mathcal{E} = 0$. For (iii), (i) and (ii) hold for every connected weighting and identify the same distribution π^*. □

Finite groups. On one group, $\mathcal{L}_{\mathrm{pairs}}$ is the Laplacian quadratic form of the group graph divided by $\mathcal{N}_g$, and a mixed group with $w_{WL} > 0$ is connected even when the within-class weights are zero. The edge-equality argument of part (i) then gives the finite-group statement of Proposition 1. If the graph is disconnected, each connected component may instead have its own residual constant, leaving the relative mass between components unconstrained. In particular, when both classes are nonempty, $w_{WL} = 0$ separates successes from failures (each class can split further if its within-class weight is zero). Rescaling the two classes separately and renormalizing preserves every within-class ratio and leaves the success probability free, so the implementation requires $w_{WL} > 0$.

A.5 PAIR WEIGHTS FROM THE QUADRATIC FORM

Let $\Sigma_w = w_{WW}\binom{n_W}{2} + w_{LL}\binom{n_L}{2} + w_{WL} n_W n_L = \sum_{i<j} w_{ij}$, so that the normalizer of Equation (3) is $\mathcal{N}_g = \frac{2G}{G-1}\Sigma_w$. The successful-failed pairs receive coefficient share

$$s_{WL} = \frac{w_{WL} n_W n_L}{\Sigma_w}. \tag{21}$$

By Proposition 3, the weights move neither the zero set nor the stationary set of the loss. They set the curvature of the quadratic form along its error directions ($\mathbf{1}_C$ denotes the indicator vector of a set C of group members).

Proposition 4 (Spectrum of the group loss)**.** *For a mixed group with weights (w_{WW}, w_{LL}, w_{WL}), let H be the Laplacian of the group graph, $H_{ij} = -w_{ij}$ for $i \neq j$ and $H_{ii} = \sum_{j\neq i} w_{ij}$, so that $\sum_{i<j} w_{ij}(q_i - q_j)^2 = q^\top H q$, and set $A = w_{WW} n_W + w_{WL} n_L$ and $B = w_{LL} n_L + w_{WL} n_W$. (i) The eigenvalues of H are 0 on $\mathbf{1}$, A on the within-success deviations (the vectors supported on W with zero sum, $n_W - 1$ directions), B on the within-failure deviations ($n_L - 1$ directions), and $w_{WL} G$ on the contrast direction $v = n_L \mathbf{1}_W - n_W \mathbf{1}_L$. (ii) On these three eigenspaces, the Hessian of $\mathcal{L}_{\text{pairs}}$ in residual coordinates has the curvatures*

$$\kappa_W = \frac{(G-1)A}{G\,\Sigma_w}, \qquad \kappa_L = \frac{(G-1)B}{G\,\Sigma_w}, \qquad \kappa_{WL} = \frac{(G-1)\,w_{WL}}{\Sigma_w}. \tag{22}$$

(iii) $\kappa_{WL} - \kappa_W = (G-1)(w_{WL} - w_{WW})\,n_W/(G\,\Sigma_w)$ and $\kappa_{WL} - \kappa_L = (G-1)(w_{WL} - w_{LL})\,n_L/(G\,\Sigma_w)$, so, for $G \geq 3$, the contrast direction is strictly the stiffest direction of the loss at every mixed composition if and only if $w_{WL} > \max(w_{WW}, w_{LL})$. With equal weights, all three curvatures equal $2/G$ at every composition.

Proof. For b supported on W with $\sum_{i\in W} b_i = 0$, $(Hb)_u = w_{WW}\sum_{v\in W, v\neq u}(b_u - b_v) + w_{WL} n_L b_u = A b_u$ for $u \in W$ and $(Hb)_u = -w_{WL}\sum_{v\in W} b_v = 0$ for $u \in L$. The failure class is symmetric, and $(Hv)_u = w_{WL} n_L (n_L + n_W) = w_{WL} G\, v_u$ for $u \in W$, likewise on L. These G directions are mutually orthogonal, which proves (i). The Hessian of $q^\top H q/\mathcal{N}_g$ is $2H/\mathcal{N}_g$ with $\mathcal{N}_g = 2G\Sigma_w/(G-1)$, which gives Equation (22). The gaps in (iii) are $(G-1)/(G\Sigma_w)$ times $w_{WL}G - A = (w_{WL} - w_{WW})n_W$ and $w_{WL}G - B = (w_{WL} - w_{LL})n_L$. Both are positive at every mixed composition when $w_{WL} > \max(w_{WW}, w_{LL})$, and for $G \geq 3$ the mixed compositions with $n_W = 2$ and with $n_L = 2$ show that both inequalities are necessary. Equal weights w give $A = B = w_{WL}G = wG$ and $\Sigma_w = w\binom{G}{2}$. □

The deployed weights. We use $w_{WW} = 1$, $w_{LL} = 0.3$ and $w_{WL} = 3$ in every main configuration (Table 20). These values satisfy two requirements. First, the contrast direction is the only direction that carries β (Lemma 1(ii)), and by Proposition 4(iii) it is the stiffest direction of the loss at every mixed composition exactly when w_{WL} exceeds both within-class weights. Uniform weights lie on the boundary of this condition: all directions are equally stiff, and the successful-failed pairs carry the share $2p(1-p)$ of the pair weight in the large-group limit (Equation (21) with $n_W = pG$), which vanishes as the success fraction p tends to one. Second, in an all-failure group ($n_W = 0$) the failure-failure pairs are the only pairs. For $i, j \in L$ both ratios in e_{ij} are unchanged by conditioning on L, so these pairs constrain only the target's conditional over failures, a class whose odds against success the target scales by $e^{-\beta}$ relative to the cost-adjusted reference (Equation (12)), and $w_{LL} > 0$ keeps this constraint, because that conditional still carries the invalid-action cost λ. In a mixed group the cross-class pairs also constrain both conditionals, through the coefficients of Equation (16), so $w_{LL} < w_{WW}$ lowers only the direct failure-failure coefficient. It is a design choice, tested by row (4) of Table 3. The two requirements fix $w_{WL} > \max(w_{WW}, w_{LL})$ and $w_{LL} > 0$ and leave the ratios free. The deployed ratios are one fixed choice within these constraints, used unchanged across benchmarks and scales. With the deployed values at $G = 16$, κ_{WL}/κ_W rises from 1.04 to 2.67 and κ_{WL}/κ_L falls from 6.40 to 1.06 as n_W runs from 1 to 15, and the successful-failed pairs carry at least half of the total pair weight for every composition with $n_W \leq 13$, against $6 \leq n_W \leq 10$ under uniform weights. Coefficient share is not realized loss share, which also depends on the squared residual differences within each pair class.

A.6 REPLAY SELECTION

We use the selection rule of Wang et al. (2026). If a task's candidate bucket exceeds capacity C, initialize $S = \emptyset$ and repeatedly add the candidate with largest positive gain. Stop at $|S| = C$ or when no remaining candidate has positive gain. The gain in Equation (7) uses $\phi(v, d) = w_{\mathrm{div}} v d$ on ALFWorld and $\phi(v, d) = v + w_{\mathrm{div}} d$ on WebShop and ScienceWorld, with $v(c) = y_{\mathrm{ver}}(c) + \mathbf{1}\{c \text{ succeeds}\}$, where y_{ver} is the verifier's raw terminal score, not the reference-tilted log-reward or the normalized ScienceWorld score. We use $w_{\mathrm{div}} = 1$ and $d(c \mid \emptyset) = 1$.

Trajectory similarity. The fixed similarity $\mathrm{sim}(c, s) \in [0, 1]$ combines four components, with weights 0.4, 0.3, 0.2, and 0.1. They are action-type bigram Jaccard similarity, visited-state Jaccard similarity, normalized edit similarity of tool-type sequences, and exact command-signature agreement. For tool sequences u, u', edit similarity is $1 - \mathrm{lev}(u, u')/\max(|u|, |u'|)$. On ALFWorld and WebShop, the signature keeps object-manipulation commands, removes instance identifiers, and collapses consecutive repeats (level 3). On ScienceWorld, it uses the verbatim command sequence (level 0).

Coverage. For each family f, let $b_f(c)$ be the category assigned to trajectory c, and let $n_f(b, S)$ count selected trajectories in category b. Define

$$F_{\mathrm{cov}}(S) = \sum_f \sum_b w_f \alpha_f(b) \log\big(1 + n_f(b, S)\big), \tag{23}$$

where $w_f \geq 0$ and $\alpha_f(b) \in (0, 1]$ are fixed during each selection pass. The coverage increment in Equation (7) is $\Delta F_{\mathrm{cov}}(c \mid S) = F_{\mathrm{cov}}(S \cup \{c\}) - F_{\mathrm{cov}}(S) = \sum_f w_f \alpha_f\big(b_f(c)\big) \log\big(1 + 1/(1 + n_f(b_f(c), S))\big)$, because adding c raises the count of its category by one in each family. Each term is non-negative and non-increasing in that count, and counts do not decrease when S grows, so F_{cov} is monotone and submodular: for $S \subseteq T$ and $c \notin T$, $\Delta F_{\mathrm{cov}}(c \mid S) \geq \Delta F_{\mathrm{cov}}(c \mid T) \geq 0$. The active families are solution length and tool-use patterns. Length categories use the turn count ℓ and coefficient $(\ell + 1)/(\ell_{\max} + 1)$, where $\ell_{\max}$ is fixed from the current candidate pool. The tool family uses the per-turn tool-type sequence, or the set of tool types on ScienceWorld, with coefficient one. The length-family weight is 0.1. The tool-family weight is 0.5 on ALFWorld and WebShop, and 1.0 on ScienceWorld. The visited-state coverage-family weight is zero in all reported configurations. Visited states still contribute to the trajectory similarity.

A.7 GATING, REFILL, AND MIXED-GROUP INJECTION

For each task, execute the following steps in order.

Sample and store. Sample G trajectories, score their terminal outcomes, and offer verified successes to the task's replay bucket before applying the gate. ALFWorld admits successful trajectories without signature-based duplicate replacement. WebShop and ScienceWorld replace a trajectory with the same signature only when the new verifier score is higher. The duplicate-checking signatures are level 3 on WebShop and level 0 on ScienceWorld. On ScienceWorld, a task without stored successes may also admit positive-score failed trajectories from the task's top decile as partial-score examples. This admission channel closes once a success is stored. An over-capacity bucket is rebuilt using Appendix A.6.

All-success groups. After offering their successes to the buffer, assign weight α and do not inject trajectories. We use $\alpha = 0$ on ALFWorld and WebShop-7B, and $\alpha = 0.25$ on WebShop-1.5B and ScienceWorld. Thus, a zero-weight all-success group can still contribute trajectories to later updates.

All-failure groups. Draw up to $\lfloor G/2 \rfloor$ trajectories uniformly without replacement from the same task's stored successes. On ScienceWorld only, when none are available, use stored partial-score examples. Replay each candidate's action sequence in the environment and re-score its generated text under the current policy by teacher forcing. Accept a candidate only if execution ends at its stored boundary and any stored success is verified again. Otherwise, mark it stale and exclude it from later draws. Each accepted candidate replaces one group member, leaving G unchanged.

Assign weight one if a verified success enters the group, 0.25 if only partial-score examples enter, and zero if no replacement is accepted.

Mixed groups. At training progress $u \in [0, 1]$, inject with probability $p_{\text{replay}}(u) = 0.75 - 0.50u$. Conditional on injection, draw the requested replacement count uniformly from $\{\lceil 0.25G \rceil, \ldots, \lfloor 0.5G \rfloor\}$. Select group members uniformly for replacement, subject to retaining at least one failure. Draw stored successes uniformly without replacement and apply the same checks as above. When fewer candidates are available, use the available eligible trajectories. The group retains weight one and size G.

Loss and parameters. Apply Equation (9) to the resulting group, using the current-policy log-probability of each trajectory. A zero-weight group is skipped rather than requiring a loss evaluation. The pairwise numerator is evaluated by Equation (16). No importance weight is applied to a replayed residual. The penalty term $\widehat{\mathrm{KL}}_x$ of Equation (9) averages, over the turns of each row with stored sampling-time log-probabilities,

$$k_3(\delta_t) = e^{-\delta_t} - 1 + \delta_t, \qquad \delta_t = \log \pi_{\text{old}}(a_t \mid s_t) - \log \pi_\theta(a_t \mid s_t) \tag{24}$$

where π_{old} is the policy that generated the trajectory. On a fresh trajectory scored identically at sampling and training, $\delta_t = 0$. On a replayed trajectory the penalty measures displacement on stored actions rather than an unbiased divergence estimate. We use $\eta = 0.01$, capacity $C = 64$ on ALFWorld and WebShop, and $C = 16$ on ScienceWorld. Full training configurations are in Appendix D.3.

A.8 Scope

Propositions 1 and 3 identify the zero-loss and stationary distributions of the pairwise objective. They do not establish convergence of the optimizer, or distribution matching from zero loss on a finite collection of groups. The global statement requires residual agreement over the full target support, or the connected-comparison condition of Appendix A.4. Buffer coverage concerns behavioral categories and does not establish that trajectory-level condition, and the submodularity of F_{cov} (Appendix A.6) concerns that function alone, not the combined gain with its order-dependent dissimilarity term. Proposition 3 and the curvatures of Appendix A.5 treat trajectory logits as independent coordinates with fixed weights. A language-model or LoRA parameterization ties them together, its gradient can vanish without every pair error vanishing, and the target need not be exactly representable, so we do not claim that neural-parameter stationary points, restricted-model optima or the optimizer's fixed points are unchanged by the choice of weights, nor that a weighting changes the speed of training in parameter space. At an exactly represented target every pair gradient is zero whatever the weights. The gate and the replay buffer determine which trajectories receive constraints and how often, not the target ratio of an evaluated pair. A stored success can be replayed only after the task has produced one, and ScienceWorld's partial-score fallback supplies graded-score information without a successful-failed comparison. Replay changes a trajectory's frequency in the training distribution, not its residual equation, and the absence of importance weighting is not a claim that replay gives an unbiased on-policy policy-gradient estimate. Positive target mass on every successful trajectory (Equation (13)) is a support statement, not a finite-sample mode-discovery guarantee. The penalty of Equation (24) vanishes with its derivative on fresh trajectories, but on replayed trajectories its gradient need not vanish where the pair errors do, so the preservation results apply to $\mathcal{L}_{\text{pairs}}$ and not to the full loss of Equation (9).

B Supplementary results

This appendix collects the measurements behind the figures and tables of Section 4: the full temperature sweeps behind Figures 2 and 5, the paired intervals of the printed rows, the behavior panel, the controls behind the ablation and the main rows, and the additional ScienceWorld readings. Metric definitions, seed conventions, table notes and training settings are in Appendix D.

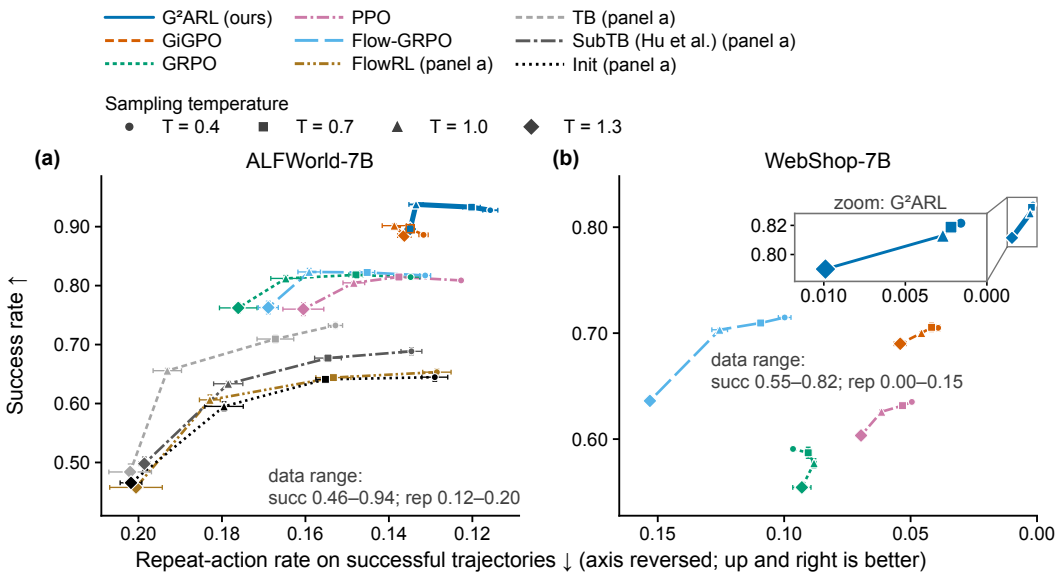

Figure 5: **Success against repeat-action rate at 7B**, ALFWorld (left) and WebShop (right), four decoding temperatures (marker shape and size), reversed horizontal axis: the successful-trajectory repeat-action rate of the 7B temperature sweeps. The line beginning `data range` gives the smallest and the largest value reached by any point drawn in that panel, on each axis. The inset magnifies our four WebShop points, which span 0.0083. The WebShop panel omits the reference objectives, absent from the 7B sweep (Table 5), and FlowRL, whose successful-trajectory repeat-action rate is undefined there (Table 6). Error bars are decoding-seed standard deviations, except the WebShop-7B G^{2}ARL point at t=0.4, whose bar is taken over its six evaluation runs (Table 5).

B.1 Full temperature sweeps

Tables 4 and 5 give the full sweeps behind the ALFWorld and WebShop panels of Figure 2 and behind Figure 5. Table 7 gives the ScienceWorld sweep. At each of the four temperatures, on ALFWorld and on WebShop at both scales, the G^{2}ARL row that Table 1 prints has the highest success rate in its block, in all sixteen cells. The loop-share coordinates of Figure 2 are listed in Table 8, and its success coordinates are those of the three sweep tables. In all twelve 1.5B cells our point has the highest success, and in eleven of the twelve also the lowest loop share. The exception is ScienceWorld at t=1.3, where Flow-GRPO's loop share is 0.0267 against our 0.0325.

Task families in Figure 3. The figure splits the t=0.4 ALFWorld-1.5B cells of Table 1 by the six task families (Pick & place, Examine, Pick two, Cool, Clean and Heat), drawing GiGPO, PPO, GRPO and the shared starting point beside our row. Averaging a method's six axes weighted by the number of tasks in each family reproduces its ALFWorld success entry. Our printed row leads the strongest of the seven retrained baselines and the starting point on every family, the widest gap against GiGPO being +0.2694 on Clean. Pick two is the hardest family for every method.

The 7B sweeps. Figure 5 repeats Figure 2 at 7B on the successful-trajectory repeat-action rate (Tables 4 and 6), loop share being registered for the 1.5B sweeps.

B.2 Paired intervals for the printed rows

Table 9 gives the paired difference and its 95% interval between each printed G^{2}ARL row and the baselines listed in its block, under the task resampling unit. Table 10 does the same for ScienceWorld under both of its units. Differences are ours minus the baseline throughout. The WebShop-7B FlowRL cell is excluded (Appendix D.2). Under the coarser ALFWorld unit, its 47 task configurations, every ALFWorld success comparison of Table 9 keeps its sign and, where it excludes zero under the task unit, its exclusion of zero. Our printed row produces shorter successful trajectories than every baseline in all four blocks, by 0.53 to 6.50 turns, each side pooled over its own solved set.

Table 4: **Temperature sweep on ALFWorld** (`valid_unseen`, 134 tasks, k=16). Panel (a) is the success rate and panel (b) the successful-trajectory repeat-action rate. Cells are means over four decoding seeds ± sample standard deviation, with three exceptions: the 1.5B G^{2}ARL (β=64) row and the 1.5B GiGPO row, the 1.5B G^{2}ARL and GiGPO rows of Table 1, are each a mean over its training runs with ± over those runs (Appendix B.5), and the t=0.4 cell of the 1.5B G^{2}ARL (β=8) row pools repeated evaluations of the same seeds (Appendix D.1). The 7B G^{2}ARL (β=128) row is the 7B G^{2}ARL row of Table 1.

(a) Success rate ↑

Method	t=0.4	t=0.7	t=1.0	t=1.3
Qwen2.5-1.5B-Instruct				
G^{2}ARL (β=8)	0.8018 ± 0.0050	0.8469 ± 0.0054	0.8426 ± 0.0014	0.7966 ± 0.0070
G^{2}ARL (β=64)	0.8755 ± 0.0248	0.8907 ± 0.0276	0.8738 ± 0.0215	0.8171 ± 0.0136
GiGPO	0.6962 ± 0.0026	0.7178 ± 0.0012	0.7270 ± 0.0072	0.7266 ± 0.0062
GRPO	0.6571 ± 0.0100	0.6936 ± 0.0076	0.7009 ± 0.0068	0.6736 ± 0.0080
PPO	0.6734 ± 0.0087	0.7034 ± 0.0026	0.7073 ± 0.0061	0.6954 ± 0.0051
Flow-GRPO (released recipe)	0.6269 ± 0.0068	0.6655 ± 0.0056	0.6723 ± 0.0090	0.6469 ± 0.0067
FlowRL	0.5611 ± 0.0026	0.5721 ± 0.0023	0.5562 ± 0.0049	0.5075 ± 0.0101
TB	0.5441 ± 0.0021	0.5522 ± 0.0080	0.5269 ± 0.0071	0.4555 ± 0.0069
SubTB (Hu et al.)	0.5111 ± 0.0079	0.5117 ± 0.0020	0.4854 ± 0.0098	0.4226 ± 0.0030
Init.	0.4889 ± 0.0052	0.4939 ± 0.0030	0.4768 ± 0.0012	0.4019 ± 0.0076
Qwen2.5-7B-Instruct				
G^{2}ARL (β=128)	0.9279 ± 0.0022	0.9331 ± 0.0021	0.9377 ± 0.0018	0.8959 ± 0.0033
G^{2}ARL (β=8)	0.9214 ± 0.0015	0.9326 ± 0.0022	0.9338 ± 0.0048	0.8924 ± 0.0046
GiGPO	0.8861 ± 0.0021	0.8965 ± 0.0024	0.9017 ± 0.0027	0.8849 ± 0.0073
GRPO	0.8144 ± 0.0013	0.8182 ± 0.0031	0.8122 ± 0.0049	0.7622 ± 0.0084
PPO	0.8088 ± 0.0031	0.8146 ± 0.0043	0.8047 ± 0.0021	0.7601 ± 0.0108
Flow-GRPO (released recipe)	0.8172 ± 0.0022	0.8223 ± 0.0050	0.8232 ± 0.0059	0.7628 ± 0.0107
FlowRL	0.6532 ± 0.0050	0.6440 ± 0.0020	0.6062 ± 0.0083	0.4577 ± 0.0068
TB	0.7324 ± 0.0050	0.7094 ± 0.0062	0.6557 ± 0.0039	0.4839 ± 0.0135
SubTB (Hu et al.)	0.6885 ± 0.0058	0.6771 ± 0.0018	0.6335 ± 0.0040	0.4979 ± 0.0112
Init.	0.6447 ± 0.0069	0.6410 ± 0.0021	0.5951 ± 0.0077	0.4655 ± 0.0044

(b) Repeat-action rate on successful trajectories ↓

Method	t=0.4	t=0.7	t=1.0	t=1.3
Qwen2.5-1.5B-Instruct				
G^{2}ARL (β=8)	0.1992 ± 0.0032	0.2050 ± 0.0008	0.2184 ± 0.0025	0.2363 ± 0.0021
G^{2}ARL (β=64)	0.1571 ± 0.0069	0.1693 ± 0.0088	0.1851 ± 0.0075	0.2044 ± 0.0048
GiGPO	0.2064 ± 0.0139	0.2126 ± 0.0148	0.2178 ± 0.0110	0.2244 ± 0.0107
GRPO	0.2157 ± 0.0055	0.2289 ± 0.0058	0.2328 ± 0.0022	0.2398 ± 0.0030
PPO	0.2244 ± 0.0040	0.2365 ± 0.0046	0.2406 ± 0.0031	0.2412 ± 0.0014
Flow-GRPO (released recipe)	0.2038 ± 0.0016	0.2251 ± 0.0030	0.2344 ± 0.0050	0.2407 ± 0.0037
FlowRL	0.2209 ± 0.0057	0.2444 ± 0.0050	0.2558 ± 0.0060	0.2633 ± 0.0045
TB	0.2296 ± 0.0048	0.2546 ± 0.0066	0.2756 ± 0.0042	0.2908 ± 0.0032
SubTB (Hu et al.)	0.2307 ± 0.0008	0.2646 ± 0.0055	0.2833 ± 0.0011	0.2879 ± 0.0065
Init.	0.2352 ± 0.0055	0.2705 ± 0.0087	0.2871 ± 0.0047	0.2961 ± 0.0081
Qwen2.5-7B-Instruct				
G^{2}ARL (β=128)	0.1157 ± 0.0018	0.1202 ± 0.0021	0.1335 ± 0.0017	0.1349 ± 0.0003
G^{2}ARL (β=8)	0.1190 ± 0.0009	0.1275 ± 0.0012	0.1377 ± 0.0018	0.1505 ± 0.0035
GiGPO	0.1317 ± 0.0011	0.1349 ± 0.0009	0.1387 ± 0.0029	0.1363 ± 0.0016
GRPO	0.1348 ± 0.0022	0.1479 ± 0.0017	0.1647 ± 0.0036	0.1761 ± 0.0045
PPO	0.1227 ± 0.0004	0.1376 ± 0.0031	0.1484 ± 0.0026	0.1605 ± 0.0049
Flow-GRPO (released recipe)	0.1313 ± 0.0013	0.1452 ± 0.0018	0.1591 ± 0.0027	0.1689 ± 0.0024
FlowRL	0.1285 ± 0.0034	0.1533 ± 0.0042	0.1829 ± 0.0025	0.2006 ± 0.0063
TB	0.1528 ± 0.0017	0.1672 ± 0.0044	0.1931 ± 0.0034	0.2020 ± 0.0051
SubTB (Hu et al.)	0.1346 ± 0.0025	0.1546 ± 0.0032	0.1785 ± 0.0035	0.1986 ± 0.0013
Init.	0.1290 ± 0.0031	0.1552 ± 0.0023	0.1795 ± 0.0046	0.2018 ± 0.0026

B.3 BEHAVIOR PANEL

Tables 11 and 12 give every behavioral quantity we compute, on both bases, for the rows listed from the four main blocks and the ScienceWorld block, from the frozen analysis scripts behind the metrics of Appendix D.1 applied to the same evaluation products (the stored output of one evaluation run) as Tables 1 and 2.

Loop share on WebShop under restricted scopes. On the all-trajectory basis of Figure 2, the WebShop-1.5B G^{2}ARL loop share is the lowest of the compared methods, and it stays lowest among

Table 5: **Temperature sweep on WebShop** (test split, goal indices $[0, 500)$, 500 tasks, k=16). Cells are means over decoding seeds $\pm$ sample standard deviation (divisor $n-1$). Entries without $\pm$ come from a single seed. Seed counts differ by column, so dispersions are not comparable across columns. The counts are three at t=0.4, two at t=0.7 and t=1.3 (seeds 0 and 1), and three, two or one at t=1.0 depending on the row. The 7B G^{2}ARL row at t=0.4 averages six evaluation runs over three decoding seeds, its $\pm$ taken over those runs. The FlowRL row in the 7B block aggregates its decoding seeds at the pre-registered reporting point, and Success and Score are exactly zero in every seed at every temperature: all trajectories exhaust the turn limit without issuing a purchase action, so its success-trajectory repeat-rate cells are n/a. The 7B SubTB (Hu et al.) and printed 7B TB configurations are absent from this sweep, as is the 1.5B Init. row. Their t=0.4 results appear in Table 1. The 7B GRPO cell at t=0.4 and the 59.08% of Table 1 are one mean lying at a rounding boundary, rounded as a fraction here and as a percentage there.

(a) Success rate ↑

Method	t=0.4	t=0.7	t=1.0	t=1.3
Qwen2.5-1.5B-Instruct				
G^{2}ARL (printed)	0.8492 ± 0.0003	0.8448 ± 0.0008	0.8361 ± 0.0019	0.7889 ± 0.0013
GRPO	0.5599 ± 0.0010	0.5746 ± 0.0036	0.5551 ± 0.0030	0.3446 ± 0.0006
GiGPO	0.6419 ± 0.0034	0.6401 ± 0.0040	0.6351	0.5992 ± 0.0001
PPO	0.5066 ± 0.0023	0.5216 ± 0.0041	0.5043 ± 0.0031	0.3267 ± 0.0033
Flow-GRPO (released recipe)	0.6098 ± 0.0019	0.6216 ± 0.0021	0.6002	0.1947 ± 0.0042
FlowRL	0.0345 ± 0.0014	0.0501 ± 0.0016	0.0735 ± 0.0040	0.0648 ± 0.0022
TB (λ=0.1)	0.1334 ± 0.0037	0.1342 ± 0.0020	0.1015	0.0254 ± 0.0027
SubTB (Hu et al.) (λ=0.1)	0.0894 ± 0.0016	0.0893 ± 0.0024	0.0669	0.0158 ± 0.0001
Qwen2.5-7B-Instruct				
G^{2}ARL (printed)	0.8215 ± 0.0005	0.8188 ± 0.0004	0.8130	0.7900 ± 0.0023
GRPO	0.5907 ± 0.0020	0.5871 ± 0.0052	0.5772 ± 0.0043	0.5544 ± 0.0021
GiGPO	0.7050 ± 0.0026	0.7055 ± 0.0039	0.7001 ± 0.0010	0.6900 ± 0.0041
PPO	0.6351 ± 0.0016	0.6316 ± 0.0003	0.6258 ± 0.0011	0.6034 ± 0.0039
Flow-GRPO (released recipe)	0.7148 ± 0.0016	0.7096 ± 0.0013	0.7032 ± 0.0012	0.6361 ± 0.0027
FlowRL	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000

(b) Score ↑

Method	t=0.4	t=0.7	t=1.0	t=1.3
Qwen2.5-1.5B-Instruct				
G^{2}ARL (printed)	91.79 ± 0.07	91.59 ± 0.12	91.07 ± 0.14	87.82 ± 0.07
GRPO	71.49 ± 0.09	75.14 ± 0.39	75.77 ± 0.09	58.73 ± 0.77
GiGPO	77.39 ± 0.30	77.61 ± 0.50	78.01	76.63 ± 0.04
PPO	75.88 ± 0.26	76.01 ± 0.10	74.09 ± 0.31	57.20 ± 0.39
Flow-GRPO (released recipe)	73.22 ± 0.21	75.20 ± 0.39	75.25	34.37 ± 0.38
FlowRL	14.61 ± 0.29	21.30 ± 0.10	27.39 ± 0.15	27.09 ± 0.25
TB (λ=0.1)	49.04 ± 0.35	45.57 ± 0.77	39.23	15.94 ± 0.21
SubTB (Hu et al.) (λ=0.1)	40.81 ± 0.45	37.31 ± 0.20	32.68	12.52 ± 0.09
Qwen2.5-7B-Instruct				
G^{2}ARL (printed)	89.97 ± 0.07	89.84 ± 0.06	89.71	88.72 ± 0.04
GRPO	71.23 ± 0.29	71.14 ± 0.66	70.69 ± 0.07	68.54 ± 0.35
GiGPO	81.32 ± 0.14	81.43 ± 0.38	81.24 ± 0.17	80.38 ± 0.18
PPO	78.29 ± 0.07	78.17 ± 0.08	77.82 ± 0.07	75.39 ± 0.15
Flow-GRPO (released recipe)	81.14 ± 0.15	80.75 ± 0.00	80.18 ± 0.04	73.09 ± 0.28
FlowRL	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Table 6: WebShop temperature sweep, continued: panel (c), the *success-trajectory* repeat-action rate, averaged over the successful trajectories of the solved tasks only, which is why it is undefined (n/a) for a method that solves nothing. Protocol, seed counts and all qualifications are those of Table 5.

(c) Repeat-action rate (success trajectories) ↓

Method	t=0.4	t=0.7	t=1.0	t=1.3
Qwen2.5-1.5B-Instruct				
G^{2}ARL (printed)	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0005 ± 0.0001	0.0039 ± 0.0002
GRPO	0.1605 ± 0.0019	0.1576 ± 0.0026	0.1483 ± 0.0020	0.1191 ± 0.0013
GiGPO	0.1361 ± 0.0012	0.1437 ± 0.0017	0.1502	0.1418 ± 0.0026
PPO	0.1216 ± 0.0022	0.1403 ± 0.0036	0.1738 ± 0.0030	0.1602 ± 0.0012
Flow-GRPO (released recipe)	0.1321 ± 0.0004	0.1354 ± 0.0016	0.1390	0.0966 ± 0.0015
FlowRL	0.3359 ± 0.0048	0.3371 ± 0.0022	0.3295 ± 0.0080	0.2264 ± 0.0084
TB (λ=0.1)	0.2907 ± 0.0035	0.2922 ± 0.0058	0.2395	0.0600 ± 0.0089
SubTB (Hu et al.) (λ=0.1)	0.2898 ± 0.0057	0.2739 ± 0.0099	0.2197	0.0377 ± 0.0003
Qwen2.5-7B-Instruct				
G^{2}ARL (printed)	0.0016 ± 0.0002	0.0022 ± 0.0002	0.0027	0.0099 ± 0.0003
GRPO	0.0965 ± 0.0005	0.0905 ± 0.0004	0.0882 ± 0.0008	0.0930 ± 0.0036
GiGPO	0.0391 ± 0.0002	0.0417 ± 0.0000	0.0457 ± 0.0002	0.0541 ± 0.0024
PPO	0.0495 ± 0.0006	0.0532 ± 0.0010	0.0615 ± 0.0007	0.0696 ± 0.0010
Flow-GRPO (released recipe)	0.0998 ± 0.0025	0.1093 ± 0.0010	0.1255 ± 0.0031	0.1531 ± 0.0013
FlowRL	n/a	n/a	n/a	n/a

Table 7: **ScienceWorld, four decoding temperatures on the full L1 test pool** (1,684 tasks, one trajectory per task, the protocol of Table 2, whose cells the t=0.4 column reproduces). Each cell is success (%) ↑ / Rep-succ (%) ↓, mean over three decoding seeds with the sample standard deviation in brackets in units of the last printed digit. The supervised cold start has four decoding seeds at t=0.4 and two at t=1.3. UV is identically 1 at one trajectory per task and is omitted.

Method	t=0.4	t=0.7	t=1.0	t=1.3
SFT init.	36.10(22) / 13.26(23)	34.34(3) / 13.11(17)	31.37(100) / 13.35(53)	26.60(42) / 13.61(36)
PPO	58.79(6) / 5.91(1)	58.61(31) / 5.95(5)	58.21(18) / 6.20(5)	57.82(14) / 6.26(33)
GRPO	57.32(24) / 7.38(11)	56.97(12) / 7.40(8)	56.35(57) / 7.26(18)	54.97(66) / 7.14(29)
GiGPO	59.16(72) / 4.54(18)	58.65(51) / 4.70(19)	58.04(33) / 5.03(17)	57.13(60) / 5.21(8)
Flow-GRPO	54.65(3) / 12.28(21)	54.14(40) / 12.07(11)	52.87(15) / 11.98(28)	51.56(55) / 11.88(56)
FlowRL	11.32(61) / 12.34(52)	11.60(43) / 13.38(132)	10.15(78) / 13.02(117)	9.44(45) / 13.31(97)
TB	37.00(24) / 15.04(39)	35.53(15) / 15.52(11)	32.38(65) / 15.87(54)	28.23(111) / 15.77(20)
SubTB (Hu et al.)	37.57(74) / 13.80(12)	34.66(97) / 13.76(9)	32.34(126) / 14.18(27)	28.68(129) / 14.53(33)
G^{2}ARL (ours)	66.79(48) / 2.20(5)	66.53(36) / 2.72(8)	65.95(63) / 3.81(9)	64.87(75) / 5.02(20)

Table 8: **Loop-share coordinates of Figure 2**: all-trajectory basis, mean ± sample standard deviation over a row's decoding seeds, except the two ALFWorld rows G^{2}ARL and GiGPO, the rows Table 1 prints, whose cells are the mean over their training runs (Appendix B.5) with ± over those runs. Cells without ± come from a single decoding seed. Rows as in Table 1, except that the WebShop-1.5B block omits the Init. row. WebShop's reference rows use the same invalid-action penalty as our row (Appendix D.1). Success coordinates of the same cells are in Tables 4, 5 and 7.

Method	t=0.4	t=0.7	t=1.0	t=1.3
ALFWorld-1.5B				
G^{2}ARL (ours)	0.1242 ± 0.0321	0.0646 ± 0.0290	0.0297 ± 0.0182	0.0082 ± 0.0026
GiGPO	0.2666 ± 0.0058	0.2202 ± 0.0128	0.1763 ± 0.0089	0.1181 ± 0.0096
GRPO	0.2589 ± 0.0016	0.2035 ± 0.0076	0.1444 ± 0.0045	0.0868 ± 0.0061
PPO	0.2861 ± 0.0036	0.2255 ± 0.0068	0.1677 ± 0.0058	0.1020 ± 0.0064
Flow-GRPO	0.2772 ± 0.0040	0.1992 ± 0.0075	0.1276 ± 0.0033	0.0644 ± 0.0049
FlowRL	0.3851 ± 0.0113	0.2915 ± 0.0098	0.1973 ± 0.0105	0.1051 ± 0.0075
TB	0.3743 ± 0.0032	0.2790 ± 0.0044	0.1788 ± 0.0077	0.0933 ± 0.0084
SubTB (Hu et al.)	0.4283 ± 0.0073	0.3188 ± 0.0071	0.2095 ± 0.0102	0.1160 ± 0.0067
Init.	0.4506 ± 0.0140	0.3382 ± 0.0058	0.2078 ± 0.0067	0.1161 ± 0.0067
WebShop-1.5B				
G^{2}ARL (ours)	0.0000 ± 0.0001	0.0001 ± 0.0001	0.0001 ± 0.0001	0.0002 ± 0.0002
GiGPO	0.1513 ± 0.0017	0.1556 ± 0.0029	0.1401	0.0865 ± 0.0014
GRPO	0.1888 ± 0.0034	0.1219 ± 0.0031	0.0623 ± 0.0005	0.0094 ± 0.0016
PPO	0.0272 ± 0.0004	0.0418 ± 0.0001	0.0560 ± 0.0052	0.0204 ± 0.0009
Flow-GRPO	0.1805 ± 0.0024	0.1511 ± 0.0002	0.1072	0.0031 ± 0.0006
FlowRL	0.6769 ± 0.0045	0.5193 ± 0.0024	0.3193 ± 0.0068	0.0659 ± 0.0042
TB	0.2452 ± 0.0049	0.1760 ± 0.0002	0.0704	0.0010 ± 0.0002
SubTB (Hu et al.)	0.3131 ± 0.0010	0.2031 ± 0.0033	0.0759	0.0006 ± 0.0003
ScienceWorld-1.5B				
G^{2}ARL (ours)	0.0410 ± 0.0030	0.0376 ± 0.0040	0.0378 ± 0.0030	0.0325 ± 0.0107
GiGPO	0.0972 ± 0.0059	0.0865 ± 0.0042	0.0762 ± 0.0030	0.0645 ± 0.0035
GRPO	0.1015 ± 0.0031	0.0928 ± 0.0035	0.0754 ± 0.0041	0.0635 ± 0.0039
PPO	0.1091 ± 0.0040	0.1021 ± 0.0039	0.1037 ± 0.0046	0.0909 ± 0.0062
Flow-GRPO	0.1051 ± 0.0047	0.0909 ± 0.0048	0.0606 ± 0.0046	0.0267 ± 0.0031
FlowRL	0.4725 ± 0.0081	0.3820 ± 0.0036	0.2850 ± 0.0117	0.1827 ± 0.0099
TB	0.2015 ± 0.0030	0.1774 ± 0.0027	0.1500 ± 0.0078	0.1297 ± 0.0061
SubTB (Hu et al.)	0.1750 ± 0.0033	0.1516 ± 0.0097	0.1194 ± 0.0046	0.0950 ± 0.0089
Init.	0.1529 ± 0.0040	0.1362 ± 0.0077	0.1146 ± 0.0058	0.0796 ± 0.0042

trajectories of six or more turns (Table 13), on the tasks both compared methods solved at least once under the same seed, and under both restrictions together (Table 14).

B.4 SAMPLED TRAJECTORIES

Each figure in this subsection prints, verbatim, the command sequences an evaluation product stores for one task, one box per command. A sequence runs down a column in Figure 6 and along a row, wrapping onto the line below, in Figures 7 and 8, where a WebShop search string takes a line of its own, and "$\times n$" is the number of the task's 16 rollouts that produced that sequence. A box is filled orange when its command is identical to an earlier command of the same rollout, the event the repeat-action rate of Table 1 counts, and a block a rollout emits several times back to back is drawn

Table 9: **Paired differences and 95% intervals, ours minus each listed baseline**, cluster bootstrap over tasks, $10{,}000$ resamples. Success and Rep-succ as in Appendix D.1, differences in fractions of $[0, 1]$. The ALFWorld-1.5B G^{2}ARL and GiGPO rows are each the mean over that row's training runs, whose evaluation chains are pooled. The intervals resample tasks with each task's rollouts and decoding-seed average held fixed, and condition on those training runs.

Baseline	Success	Rep-succ
ALFWorld-1.5B, against G^{2}ARL (mean over the training runs on the 64-task subset)		
GiGPO	$+0.1793$ [$+0.1282$, $+0.2312$]	-0.0493 [-0.0704, -0.0289]
GRPO	$+0.2185$ [$+0.1631$, $+0.2752$]	-0.0586 [-0.0828, -0.0352]
PPO	$+0.2021$ [$+0.1485$, $+0.2573$]	-0.0673 [-0.0921, -0.0431]
Flow-GRPO	$+0.2487$ [$+0.1930$, $+0.3046$]	-0.0467 [-0.0688, -0.0254]
FlowRL	$+0.3144$ [$+0.2408$, $+0.3847$]	-0.0638 [-0.0896, -0.0385]
TB	$+0.3314$ [$+0.2591$, $+0.4009$]	-0.0725 [-0.0982, -0.0473]
SubTB (Hu et al.)	$+0.3644$ [$+0.2903$, $+0.4358$]	-0.0736 [-0.0997, -0.0484]
Init.	$+0.3866$ [$+0.3128$, $+0.4578$]	-0.0781 [-0.1045, -0.0522]
ALFWorld-7B, against G^{2}ARL (β=128)		
GiGPO	$+0.0419$ [-0.0120, $+0.0976$]	-0.0160 [-0.0302, -0.0022]
GRPO	$+0.1136$ [$+0.0518$, $+0.1774$]	-0.0191 [-0.0341, -0.0034]
PPO	$+0.1192$ [$+0.0576$, $+0.1855$]	-0.0070 [-0.0221, $+0.0089$]
Flow-GRPO	$+0.1108$ [$+0.0507$, $+0.1734$]	-0.0157 [-0.0308, $+0.0001$]
FlowRL	$+0.2747$ [$+0.2056$, $+0.3467$]	-0.0129 [-0.0321, $+0.0069$]
TB	$+0.1955$ [$+0.1320$, $+0.2627$]	-0.0371 [-0.0557, -0.0180]
SubTB (Hu et al.)	$+0.2394$ [$+0.1739$, $+0.3083$]	-0.0189 [-0.0375, -0.0003]
Init.	$+0.2832$ [$+0.2144$, $+0.3552$]	-0.0134 [-0.0324, $+0.0063$]
WebShop-1.5B, against the printed G^{2}ARL row		
GiGPO	$+0.2073$ [$+0.1775$, $+0.2370$]	-0.1361 [-0.1510, -0.1223]
GRPO	$+0.2893$ [$+0.2580$, $+0.3210$]	-0.1605 [-0.1748, -0.1468]
PPO	$+0.3426$ [$+0.3137$, $+0.3724$]	-0.1216 [-0.1321, -0.1115]
Flow-GRPO	$+0.2394$ [$+0.2097$, $+0.2701$]	-0.1321 [-0.1441, -0.1204]
FlowRL	$+0.8147$ [$+0.7823$, $+0.8457$]	-0.3359 [-0.3825, -0.2907]
TB	$+0.7158$ [$+0.6851$, $+0.7455$]	-0.2907 [-0.3142, -0.2677]
SubTB (Hu et al.)	$+0.7598$ [$+0.7290$, $+0.7892$]	-0.2898 [-0.3153, -0.2650]
Supervised WebShop ckpt	$+0.5012$ [$+0.4627$, $+0.5386$]	-0.0589 [-0.0669, -0.0515]
WebShop-7B, against the printed G^{2}ARL row		
GiGPO	$+0.1165$ [$+0.0897$, $+0.1442$]	-0.0375 [-0.0430, -0.0324]
GRPO	$+0.2308$ [$+0.2020$, $+0.2603$]	-0.0949 [-0.1032, -0.0870]
PPO	$+0.1864$ [$+0.1595$, $+0.2143$]	-0.0479 [-0.0540, -0.0421]
Flow-GRPO	$+0.1067$ [$+0.0823$, $+0.1318$]	-0.0983 [-0.1066, -0.0899]
TB	$+0.8111$ [$+0.7772$, $+0.8435$]	-0.2141 [-0.2337, -0.1922]
SubTB (Hu et al.)	$+0.7980$ [$+0.7641$, $+0.8307$]	-0.1538 [-0.1748, -0.1362]

Table 10: **ScienceWorld paired differences under two resampling units**, ours minus each baseline, $10{,}000$ resamples. Task unit: $1{,}684$ tasks. Family unit: 29 task types. Success is in fractions of $[0, 1]$ and Score in the benchmark's native points. Metrics as in Appendix D.1. The Len column is restricted to the tasks both methods solve, while Rep-succ is taken over each side's own solved set and is reported under the task unit only. The supervised cold-start comparisons pair the three decoding seeds that row shares with ours, so they differ from the subtraction of the two cells of Table 2, whose cold-start row averages its four seeds.

Baseline	Succ, tasks	Succ, fam.	Score, tasks	Score, fam.
Against G^{2}ARL				
GiGPO	$+0.0762$ [$+0.0529$, $+0.0996$]	$+0.0762$ [$+0.0186$, $+0.1516$]	$+12.50$ [$+8.06$, $+16.94$]	$+12.50$ [$+0.97$, $+28.64$]
PPO	$+0.0800$ [$+0.0604$, $+0.1000$]	$+0.0800$ [$+0.0256$, $+0.1392$]	$+14.27$ [$+10.49$, $+18.14$]	$+14.27$ [$+6.67$, $+23.68$]
GRPO	$+0.0946$ [$+0.0740$, $+0.1156$]	$+0.0946$ [$+0.0281$, $+0.1681$]	$+16.15$ [$+12.09$, $+20.24$]	$+16.15$ [$+5.74$, $+30.04$]
Flow-GRPO	$+0.1213$ [$+0.0980$, $+0.1441$]	$+0.1213$ [$+0.0490$, $+0.2128$]	$+14.89$ [$+10.60$, $+19.09$]	$+14.89$ [$+7.00$, $+24.57$]
FlowRL	$+0.5546$ [$+0.5319$, $+0.5782$]	$+0.5546$ [$+0.3706$, $+0.7116$]	$+44.18$ [$+40.18$, $+48.11$]	$+44.18$ [$+13.67$, $+74.66$]
TB	$+0.2979$ [$+0.2753$, $+0.3205$]	$+0.2979$ [$+0.1776$, $+0.3945$]	$+32.41$ [$+28.03$, $+36.93$]	$+32.41$ [$+5.07$, $+61.20$]
SubTB (Hu et al.)	$+0.2922$ [$+0.2692$, $+0.3143$]	$+0.2922$ [$+0.1770$, $+0.3812$]	$+36.82$ [$+32.40$, $+41.18$]	$+36.82$ [$+11.15$, $+62.09$]
SFT init.	$+0.3060$ [$+0.2831$, $+0.3284$]	$+0.3060$ [$+0.1846$, $+0.4021$]	$+42.05$ [$+37.55$, $+46.55$]	$+42.05$ [$+15.41$, $+67.90$]
Baseline	**Rep-succ, tasks**	**Rep-all, tasks**	**Rep-all, fam.**	**Len, tasks**
Against G^{2}ARL				
GiGPO	-0.0234 [-0.0288, -0.0182]	-0.0298 [-0.0372, -0.0226]	-0.0298 [-0.0609, $+0.0060$]	-0.84 [-1.01, -0.67]
PPO	-0.0371 [-0.0430, -0.0316]	-0.0328 [-0.0396, -0.0259]	-0.0328 [-0.0705, $+0.0087$]	-2.04 [-2.25, -1.84]
GRPO	-0.0532 [-0.0592, -0.0472]	-0.0411 [-0.0483, -0.0337]	-0.0411 [-0.0845, $+0.0037$]	-2.92 [-3.18, -2.68]
Flow-GRPO	-0.1008 [-0.1076, -0.0939]	-0.1057 [-0.1141, -0.0972]	-0.1057 [-0.1539, -0.0537]	-4.37 [-4.67, -4.07]
FlowRL	-0.1012 [-0.1222, -0.0815]	-0.3555 [-0.3690, -0.3417]	-0.3555 [-0.4043, -0.2814]	-3.97 [-4.66, -3.31]
TB	-0.1291 [-0.1378, -0.1205]	-0.1665 [-0.1768, -0.1562]	-0.1665 [-0.2188, -0.0993]	-6.06 [-6.50, -5.64]
SubTB (Hu et al.)	-0.1168 [-0.1244, -0.1092]	-0.1380 [-0.1473, -0.1288]	-0.1380 [-0.1742, -0.0795]	-5.55 [-5.95, -5.15]
SFT init.	-0.1105 [-0.1178, -0.1033]	-0.1220 [-0.1310, -0.1131]	-0.1220 [-0.1529, -0.0657]	-5.36 [-5.74, -4.97]

Table 11: **Behavior panel, ALFWorld**, `valid_unseen`, k=16, temperature 0.4, four decoding seeds. The ALFWorld-1.5B G^{2}ARL and GiGPO rows are each the mean over that row's training runs (Appendix B.5). Rep and Loop are the repeat-action rate and the loop share of Appendix D.1, each on the successful-trajectory and on the all-trajectory basis. Loop-all is the loop record stored in the evaluation product and Loop-succ is recomputed from the stored solution signatures of successful trajectories, so the two are never subtracted. LRB is the mean over all trajectories of the total span of the longest back-to-back verbatim repetition (0 when there is none). Len is the mean turns of a successful trajectory. The mean length of a failed trajectory is 50.00 turns for every row, the turn limit. All rates are percentages. Subscripts are sample standard deviations over decoding seeds. For the two two-run rows they are taken over decoding seeds of the two-run mean.

Method	Succ ↑	UV ↑	Rep-succ ↓	Rep-all ↓	Loop-succ ↓	Loop-all ↓	LRB-all ↓	Len
Qwen2.5-1.5B-Instruct								
Init.	48.89±0.52	7.321±0.129	23.52±0.55	46.36±0.53	11.19±1.13	45.06±1.40	8.975±0.382	19.91±0.27
PPO	67.34±0.87	5.749±0.295	22.44±0.40	38.52±0.20	6.51±0.66	28.61±0.36	5.733±0.048	16.59±0.19
GRPO	65.71±1.00	6.289±0.164	21.57±0.55	38.03±0.10	5.10±0.81	25.89±0.16	4.986±0.053	18.16±0.35
GiGPO	69.62±0.66	5.316±0.082	20.64±0.26	36.77±0.47	5.45±0.37	26.66±0.74	5.602±0.104	15.62±0.17
Flow-GRPO	62.69±0.68	6.682±0.136	20.38±0.16	37.94±0.26	5.89±0.27	27.72±0.40	5.111±0.092	18.18±0.16
FlowRL	56.11±0.26	6.936±0.032	22.09±0.57	42.07±0.43	9.93±1.04	38.51±1.13	7.817±0.120	19.04±0.29
TB	54.41±0.21	7.296±0.143	22.96±0.48	43.44±0.17	9.34±1.20	37.43±0.32	7.122±0.104	19.28±0.27
SubTB (Hu et al.)	51.11±0.79	7.253±0.113	23.07±0.08	45.19±0.40	11.34±0.33	42.83±0.73	8.506±0.166	19.53±0.05
G^{2}ARL (printed)	87.55±0.17	7.037±0.094	15.71±0.13	22.85±0.05	3.90±0.08	12.42±0.13	2.166±0.041	13.41±0.09
Qwen2.5-7B-Instruct								
Init.	64.47±0.69	4.426±0.168	12.90±0.31	28.59±0.27	4.38±0.22	16.39±0.69	3.077±0.106	19.61±0.16
PPO	80.88±0.31	3.614±0.078	12.27±0.04	21.47±0.06	0.75±0.25	5.98±0.47	0.853±0.095	16.91±0.04
GRPO	81.44±0.13	3.930±0.100	13.48±0.22	22.46±0.11	1.09±0.14	6.52±0.36	0.900±0.038	16.91±0.08
GiGPO	88.61±0.21	3.448±0.081	13.17±0.11	19.58±0.07	3.03±0.50	10.59±0.54	2.239±0.099	14.50±0.04
Flow-GRPO	81.72±0.22	3.982±0.065	13.13±0.13	21.87±0.10	1.63±0.20	7.88±0.32	1.124±0.021	16.63±0.06
FlowRL	65.32±0.50	4.367±0.067	12.85±0.34	28.00±0.42	4.39±0.54	16.17±0.23	2.952±0.106	19.47±0.15
TB	73.24±0.50	5.294±0.070	15.28±0.17	26.19±0.09	4.73±0.48	12.14±0.48	2.172±0.077	19.58±0.13
SubTB (Hu et al.)	68.85±0.58	4.472±0.072	13.46±0.25	26.75±0.27	3.99±0.45	13.58±0.43	2.552±0.112	19.49±0.15
G^{2}ARL (β=128, printed)	92.79±0.22	5.217±0.071	11.57±0.18	15.14±0.29	1.32±0.05	4.92±0.16	0.754±0.037	13.86±0.07

once and closed by a return arrow carrying that count. All three figures read the products behind the G^{2}ARL and GiGPO rows of Table 1 at decoding seed 0 and temperature 0.4, and name ALFWorld tasks by their entry in the pinned `valid_unseen` order. The tasks are selected by the following rules. Figure 6 takes the ALFWorld task with the largest difference between the two methods in the number of distinct successful command sequences, among the 35 tasks both methods solve on which every G^{2}ARL success takes at most 13 commands. Figure 7 takes the one ALFWorld task, of the 10 on which no G^{2}ARL success repeats a command, on which every GiGPO success repeats one. Figure 8 takes, among the 204 WebShop tasks on which both methods succeed in all 16 rollouts and no G^{2}ARL success repeats a command, the task on which GiGPO's longest successful rollout is longest. The behavior it shows is the one behind the WebShop repeat-action column of Table 1.

B.5 Controls for the ablation and the main rows

Ablation rows. Every row of Table 3 is a configuration of G^{2}ARL, run on ALFWorld from the warm start on the ALFWorld-1.5B configuration of Table 20. Row (1) is the row Table 1 prints, at reward scale β=64, and rows (2) to (7) train at β=8. Row (3) replaces the group flow reallocation loss by its equal-weight degenerate case, Var-TB (Lemma 1(i)), and keeps the group-level gate and the coverage-selected buffer of Section 3.3 in place. Row (4) raises the failure-failure pair weight. Row (5) switches replay consumption off, so that refill, mixed-group replacement and the container that exists only for the replay path move together, and all-success groups keep the weight 0 the gate would otherwise give them. Row (6) replaces coverage selection by reward-prioritized selection, which scores stored trajectories by the mixed log-reward of Equation (1), keeps the highest-scoring ones within capacity, and draws for injection by a softmax over those scores, the rule of Shen et al. (2023) in the form used by GFlowNet language-model fine-tuning (Hu et al., 2024). Row (7) combines (5) with (3).

Pair weights experiments. Table 15 reports three further training runs of the row (2) configuration of Table 3. The evaluation protocol is that of Table 3. The first run is the full recipe and serves as the control. The second run sets the pair weights (w_{WW}, w_{LL}, w_{WL}) to $(0, 0, 3)$ instead of $(1, 0.3, 3)$, so that only successful-failed pairs have nonzero weight. The third run sets the three pair weights equal, the Var-TB form of row (3), and reads 11.82 points of success and 1.01 UV below the control.

Table 12: **Behavior panel, WebShop and ScienceWorld.** Init. is the bare instruction-tuned model of the WebShop-1.5B block of Table 1. Columns are as in Table 11, with Score added and the mean length of a failed trajectory printed, since it is not pinned to the turn limit here. WebShop: 500 held-out tasks, k=16, temperature 0.4, three decoding seeds. ScienceWorld: 1,684 L1 test tasks, one trajectory per task, temperature 0.4, three decoding seeds and four for the supervised start. Loop-succ is recomputed from the stored solution signatures on WebShop and read from the product's per-trajectory loop record on ScienceWorld. Rates are percentages. Subscripts are sample standard deviations over decoding seeds. `n/a` marks a quantity undefined on an empty successful-trajectory set. A dash marks UV on ScienceWorld, identically 1 at one trajectory per task.

Method	Succ ↑	Score ↑	UV ↑	Rep-succ ↓	Rep-all ↓	Loop-succ ↓	Loop-all ↓	LRB-all ↓	Len	Len (fail)
WebShop, Qwen2.5-1.5B-Instruct										
Init. (Table 1)	5.26±0.15	22.27±0.07	2.524±0.078	34.08±2.51	47.84±0.16	16.18±2.70	46.05±0.19	5.077±0.015	8.89±0.35	12.77±0.04
Supervised WebShop ckpt	34.80±0.60	41.97±0.44	2.326±0.009	5.89±0.25	43.37±0.26	3.40±0.39	47.20±0.36	5.470±0.031	6.33±0.03	12.31±0.03
PPO	50.66±0.23	75.88±0.26	9.323±0.009	12.16±0.22	10.51±0.14	2.81±0.17	2.72±0.04	0.342±0.005	6.25±0.04	6.17±0.04
GRPO	55.99±0.10	71.49±0.09	9.503±0.147	16.05±0.19	20.57±0.07	11.16±0.73	18.88±0.34	1.859±0.020	6.51±0.03	8.89±0.02
GiGPO	64.19±0.34	77.39±0.30	7.332±0.057	13.61±0.12	16.68±0.10	10.57±0.13	15.13±0.17	1.644±0.023	6.37±0.01	8.12±0.06
Flow-GRPO	60.98±0.19	73.22±0.21	11.310±0.132	13.21±0.04	18.52±0.06	9.72±0.16	18.05±0.24	1.765±0.016	6.33±0.03	9.10±0.02
FlowRL	3.45±0.14	14.61±0.29	3.965±0.284	33.59±0.48	56.17±0.35	13.30±0.22	67.69±0.45	7.510±0.044	6.35±0.11	12.69±0.06
TB (λ=0.1)	13.34±0.37	49.04±0.35	3.647±0.180	29.07±0.35	32.00±0.18	12.58±0.99	24.52±0.49	2.779±0.053	7.99±0.02	8.94±0.02
SubTB (Hu et al.) (λ=0.1)	8.94±0.16	40.81±0.45	3.159±0.101	28.98±0.57	36.83±0.10	12.49±1.27	31.31±0.10	3.602±0.002	8.02±0.13	10.02±0.01
TB (λ=6.4)	5.71±0.18	58.14±0.11	3.031±0.147	12.08±0.67	4.79±0.06	3.78±0.58	2.28±0.11	0.278±0.006	5.23±0.07	4.05±0.02
SubTB (Hu et al.) (λ=6.4)	2.41±0.11	54.19±0.20	5.335±0.157	0.13±0.17	0.14±0.02	0.00±0.00	0.08±0.02	0.008±0.003	3.17±0.04	3.11±0.00
G^{2}ARL (printed)	84.92±0.03	91.79±0.07	6.300±0.084	0.00±0.00	0.00±0.01	0.00±0.00	0.00±0.01	0.001±0.001	5.06±0.00	4.54±0.00
WebShop, Qwen2.5-7B-Instruct										
PPO	63.51±0.16	78.29±0.07	7.967±0.104	4.95±0.06	10.65±0.19	2.94±0.18	5.99±0.39	0.502±0.034	5.94±0.02	9.32±0.07
GRPO	59.08±0.20	71.23±0.29	8.168±0.067	9.65±0.05	17.61±0.03	4.65±0.17	10.46±0.16	0.949±0.015	6.31±0.01	10.92±0.01
GiGPO	70.50±0.26	81.32±0.14	6.409±0.017	3.91±0.02	9.40±0.12	1.02±0.10	4.23±0.10	0.370±0.021	5.58±0.01	9.29±0.06
Flow-GRPO	71.48±0.16	81.14±0.15	10.104±0.052	9.98±0.25	14.34±0.17	8.95±0.11	12.71±0.20	1.030±0.022	6.17±0.03	9.58±0.05
FlowRL	0.00±0.00	0.00±0.00	n/a	n/a	38.48±0.10	n/a	15.51±0.40	2.611±0.015	n/a	15.00±0.00
TB	1.04±0.04	4.48±0.23	2.917±0.402	21.57±1.13	53.26±0.16	13.60±1.79	36.67±0.49	3.813±0.012	7.55±0.39	14.57±0.02
SubTB (Hu et al.)	2.35±0.08	11.71±0.09	2.357±0.184	15.54±0.39	40.91±0.13	2.31±0.39	20.34±0.58	2.518±0.027	7.87±0.30	13.87±0.03
G^{2}ARL (printed)	82.15±0.05	89.97±0.07	9.199±0.076	0.16±0.02	0.18±0.02	0.12±0.04	0.12±0.04	0.007±0.002	5.05±0.00	4.56±0.02
ScienceWorld, Qwen2.5-1.5B-Instruct										
SFT init.	36.10±0.22	5.61±0.93	—	13.26±0.23	21.42±0.37	6.79±0.61	15.29±0.40	1.820±0.073	13.37±0.07	19.75±0.18
PPO	58.79±0.06	33.76±0.94	—	5.91±0.01	12.55±0.23	4.75±0.10	10.91±0.40	1.115±0.082	9.94±0.04	18.13±0.16
GRPO	57.32±0.24	31.89±0.40	—	7.38±0.11	13.37±0.07	2.21±0.31	10.15±0.31	1.022±0.030	10.88±0.01	19.31±0.09
GiGPO	59.16±0.72	35.53±1.29	—	4.54±0.18	12.25±0.27	0.43±0.21	9.72±0.59	1.152±0.021	8.44±0.04	16.62±0.18
Flow-GRPO	54.65±0.03	33.15±0.37	—	12.28±0.21	19.83±0.11	2.28±0.19	10.51±0.47	1.292±0.017	12.35±0.10	21.99±0.02
FlowRL	11.32±0.61	3.86±0.58	—	12.34±0.52	44.82±0.18	3.86±1.15	47.25±0.81	5.785±0.065	9.33±0.19	24.65±0.08
TB	37.00±0.24	15.63±0.73	—	15.04±0.39	25.91±0.17	8.46±0.69	20.15±0.30	2.459±0.080	14.24±0.17	22.36±0.24
SubTB (Hu et al.)	37.57±0.74	11.22±1.35	—	13.80±0.12	23.07±0.08	7.74±0.12	17.50±0.33	2.070±0.013	13.65±0.01	20.91±0.10
G^{2}ARL (printed)	66.79±0.48	48.04±0.91	—	2.20±0.05	9.27±0.07	0.09±0.09	4.10±0.30	0.522±0.016	7.94±0.01	16.35±0.15

Table 13: **WebShop-1.5B loop share under restrictions applied to each method separately**, all-trajectory basis, temperature 0.4, three decoding seeds. Each cell is the mean ± sample standard deviation over seeds of a per-seed fraction. Scope (a) is the loop share of Figure 2. Scope (b) keeps trajectories of six or more turns, the shortest length at which a loop can occur. Scope (e) splits (a) by outcome, keeping the per-trajectory loop record stored in the evaluation product, so its success column is a different computation from the signature-recomputed Loop-succ of Table 12. Rows as in Table 1.

Method	(a) all trajectories	(b) length ≥ 6	(e) successes	(e) failures
G^{2}ARL (ours)	0.0000 ± 0.0001	0.0002 ± 0.0004	0.0000 ± 0.0000	0.0003 ± 0.0005
GiGPO	0.1513 ± 0.0017	0.2825 ± 0.0035	0.1056 ± 0.0019	0.2334 ± 0.0038
GRPO	0.1888 ± 0.0034	0.3488 ± 0.0076	0.1113 ± 0.0080	0.2875 ± 0.0027
PPO	0.0272 ± 0.0004	0.0637 ± 0.0004	0.0281 ± 0.0017	0.0262 ± 0.0009
Flow-GRPO	0.1805 ± 0.0024	0.3452 ± 0.0084	0.0972 ± 0.0016	0.3108 ± 0.0054
FlowRL	0.6769 ± 0.0045	0.8358 ± 0.0010	0.1294 ± 0.0022	0.6965 ± 0.0052
TB	0.2452 ± 0.0049	0.3724 ± 0.0085	0.1226 ± 0.0094	0.2641 ± 0.0054
SubTB (Hu et al.)	0.3131 ± 0.0010	0.4272 ± 0.0035	0.1203 ± 0.0121	0.3321 ± 0.0004

Table 14: **WebShop-1.5B loop share under paired restrictions**, all-trajectory basis, temperature 0.4, three decoding seeds. Each cell is the mean ± sample standard deviation over seeds of a per-seed fraction. Each block pairs the printed G^{2}ARL row with one baseline: (c) keeps the tasks both methods solved at least once under the same seed, and (b)+(c) adds the length restriction of Table 13. *: two of three seeds enter. In the third, one side of the pair has no trajectory in this scope and the seed is dropped from both sides.

Side	(c) jointly solved	(b)+(c)
Paired with GiGPO		
G^{2}ARL (ours)	0.0001 ± 0.0001	0.0003 ± 0.0005
GiGPO	0.1547 ± 0.0044	0.2862 ± 0.0083
Paired with GRPO		
G^{2}ARL (ours)	0.0001 ± 0.0001	0.0003 ± 0.0005
GRPO	0.1965 ± 0.0101	0.3606 ± 0.0163
Paired with PPO		
G^{2}ARL (ours)	0.0001 ± 0.0001	0.0003 ± 0.0005
PPO	0.0314 ± 0.0006	0.0711 ± 0.0017
Paired with Flow-GRPO		
G^{2}ARL (ours)	0.0001 ± 0.0001	0.0003 ± 0.0006
Flow-GRPO	0.1757 ± 0.0056	0.3421 ± 0.0119
Paired with FlowRL		
G^{2}ARL (ours)	0.0000 ± 0.0000	$0.0000 \pm 0.0000^*$
FlowRL	0.5394 ± 0.0172	$0.7768 \pm 0.0043^*$
Paired with TB		
G^{2}ARL (ours)	0.0001 ± 0.0001	0.0006 ± 0.0010
TB	0.2375 ± 0.0136	0.3565 ± 0.0184
Paired with SubTB (Hu et al.)		
G^{2}ARL (ours)	0.0001 ± 0.0002	0.0010 ± 0.0017
SubTB (Hu et al.)	0.2931 ± 0.0046	0.3997 ± 0.0050

Table 15: **Pair weights on ALFWorld-1.5B.** Three training runs evaluated on `valid_unseen` (134 tasks, k=16, temperature 0.4) with four decoding seeds. Rep-succ and Rep-all are the repeat-action rate over successful and over all rollouts. Len is the mean number of turns of successful rollouts. Subscripts are standard deviations over decoding seeds.

Configuration	Succ (%) ↑	UV ↑	Rep-succ (%) ↓	Rep-all (%) ↓	Len ↓
Control, full recipe	$76.80_{\pm 0.38}$	$6.608_{\pm 0.096}$	$17.77_{\pm 0.19}$	$30.33_{\pm 0.35}$	$15.67_{\pm 0.11}$
Successful-failed pairs only	$77.34_{\pm 0.88}$	$6.900_{\pm 0.078}$	$15.35_{\pm 0.39}$	$29.03_{\pm 0.19}$	$13.95_{\pm 0.19}$
Equal pair weights (Var-TB)	$64.98_{\pm 0.32}$	$5.601_{\pm 0.171}$	$19.72_{\pm 0.29}$	$38.53_{\pm 0.12}$	$16.89_{\pm 0.15}$
Difference against the control					
Successful-failed pairs only	$+0.54$	$+0.292$	-2.42	-1.30	-1.72
Equal pair weights (Var-TB)	-11.82	-1.007	$+1.95$	$+8.20$	$+1.22$

Replay inside a group-normalized policy gradient. To test the replay mechanism within GiGPO, we added our replay buffer and injection rule while keeping every other setting of its printed ALFWorld-1.5B recipe unchanged: the toolchain, full training split, starting checkpoint and training length. Stored successes are selected by the submodular coverage rule of Section 3 and injected by the same rule as in G^{2}ARL. They enter GiGPO's group normalization as if they had been sampled. We trained this twice and compare it against the two GiGPO training runs pooled into the ALFWorld-1.5B row of Table 1, every run evaluated on `valid_unseen` at temperature 0.4 with four decoding seeds. Averaged over its two runs it reads 67.33% success (66.05% and 68.61%), 5.52 UV and a successful-trajectory repeat-action rate of 21.1%, against 69.62%, 5.32 and 20.6% for GiGPO. The paired per-task success difference, each side averaged over its two runs, is -2.30 points, $[-4.79, +0.12]$. Solution length reads 15.95 turns against 15.62 and loop share 28.7% against 26.7% of rollouts. The interval resamples tasks with each task's rollouts and its average over decoding seeds and training runs held fixed.

Reward scale. Table 16 gives the development-split readings behind the 7B reward scale. The ALFWorld-1.5B reward scale was fixed on the same development split (`valid_seen`, 140 tasks, k=16, temperature 0.4): the training run on which β was selected read 0.9046 ± 0.0042 success and 6.178 ± 0.177 UV at β=64, versus 0.8398 ± 0.0046 and 5.873 ± 0.074 at β=8. Table 22 prints the

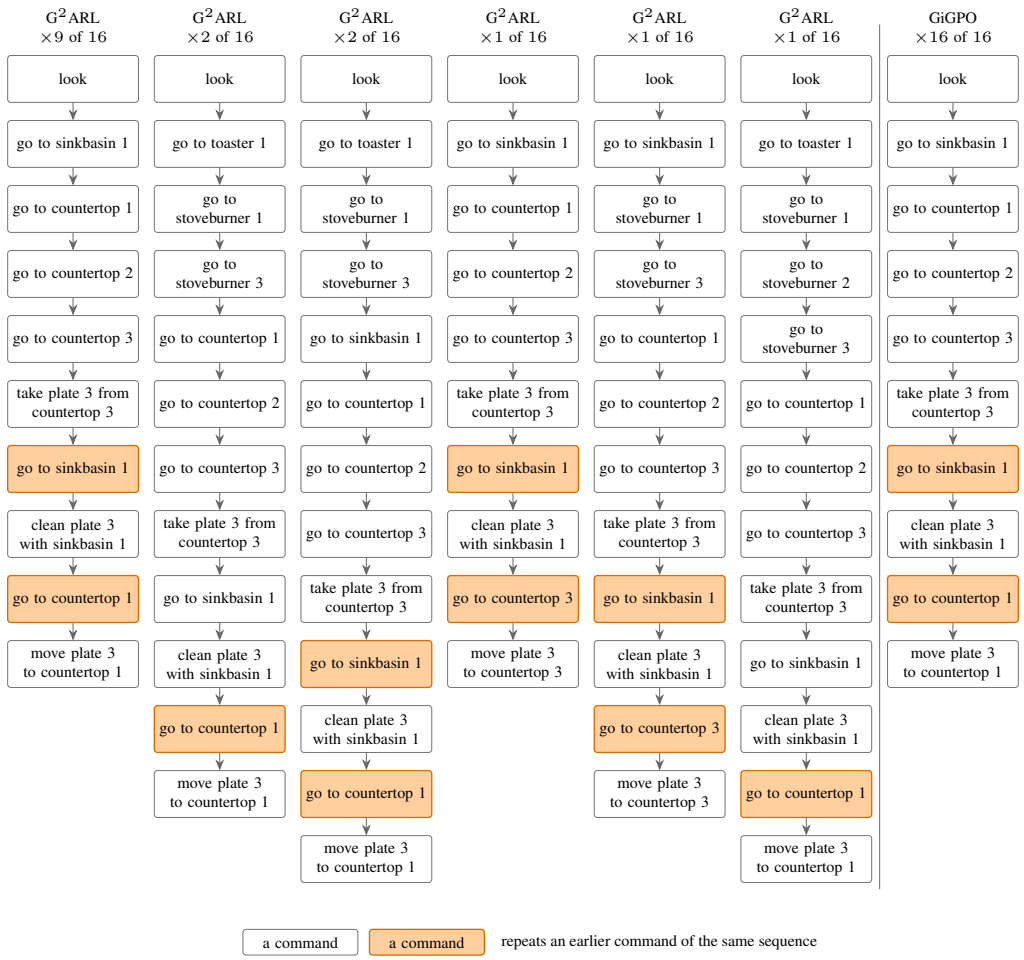

Figure 6: **Every way the two policies solved one ALFWorld task.** Task pick_clean_then_place_in_recep-Plate-None-CounterTop-10, k=16 rollouts at temperature 0.4, decoding seed 0, from the products behind the ALFWorld-1.5B rows of Table 1. One column per verbatim-distinct successful command sequence, read top to bottom, $\times n$ rollouts each. G^{2}ARL solved the task in all 16 rollouts along 6 sequences. GiGPO also solved it in all 16, along the one right of the rule. Orange marks a command identical to an earlier command of the same rollout.

Table 16: **Reward scale on ALFWorld-7B.** Read on the development split (`valid_seen`, 140 tasks, k=16, temperature 0.4, four decoding seeds), not the split Table 1 reports from. The split selects β=128 on success and on UV, the row Table 1 prints.

Configuration	Success ↑	UV ↑
β=8	0.9604 ± 0.0020	3.005 ± 0.091
β=64	0.9624 ± 0.0013	3.192 ± 0.060
β=128	0.9756 ± 0.0022	4.156 ± 0.077

mean over that configuration's training runs. TB trained at β=64, the reward scale of the printed ALFWorld-1.5B G^{2}ARL row, reads $56.49(83)\%$ success on `valid_unseen`, against $54.41(21)\%$ at β=8 (Table 1). At a shared reward scale the G^{2}ARL row leads TB by 31.06 points at β=64 (the two readings above) and by 25.77 points at β=8 (Table 4).

Training task pool on ALFWorld. Two training pools appear on ALFWorld: the full official training split (3,553 tasks), on which the four policy-gradient baselines and FlowRL train, and a fixed subset of 64 tasks drawn from that split, separately within each of ALFWorld's task families, on which the G^{2}ARL rows of Table 1 at both scales, the ablations of Table 3 and the two reference rows train. The subset is disjoint from both evaluation splits. G^{2}ARL was developed on the subset,

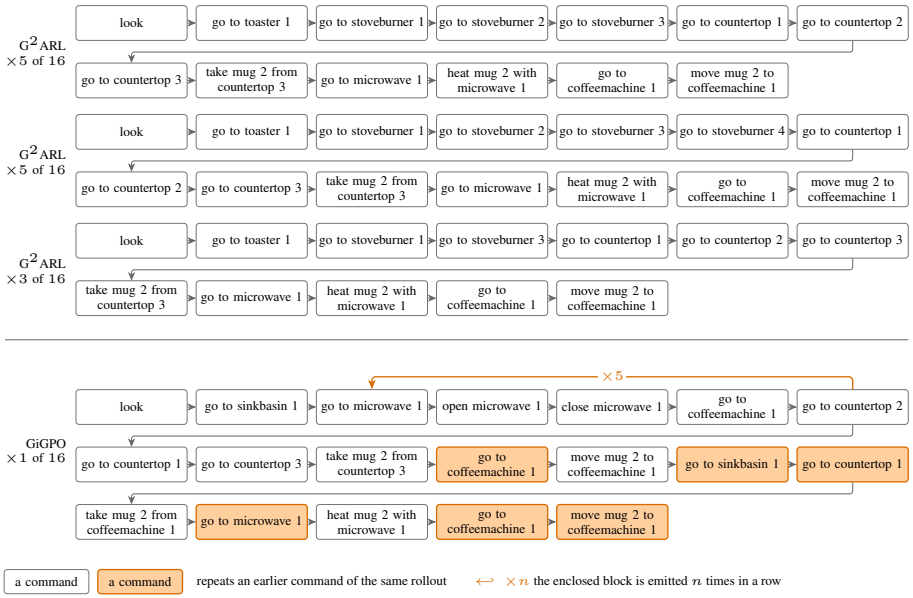

Figure 7: **A repeat-free set of solutions against a baseline that loops.** ALFWorld task pick_heat_then_place_in_recep-Mug-None-CoffeeMachine-10, k=16 rollouts at temperature 0.4, seed 0, from the G^{2}ARL and GiGPO products behind Table 1. G^{2}ARL solved it in all 16 rollouts along 6 repeat-free sequences of 11 to 16 commands, 3 of them drawn above the rule and 3 omitted. GiGPO solved it in 3 of 16, all repeating commands. The drawn success emits the 5-command block closed by the return arrow 5 times in a row, commands 3 to 27 of 39. The other two, 34 commands with 15 repeats and 45 commands with 30 repeats, are omitted.

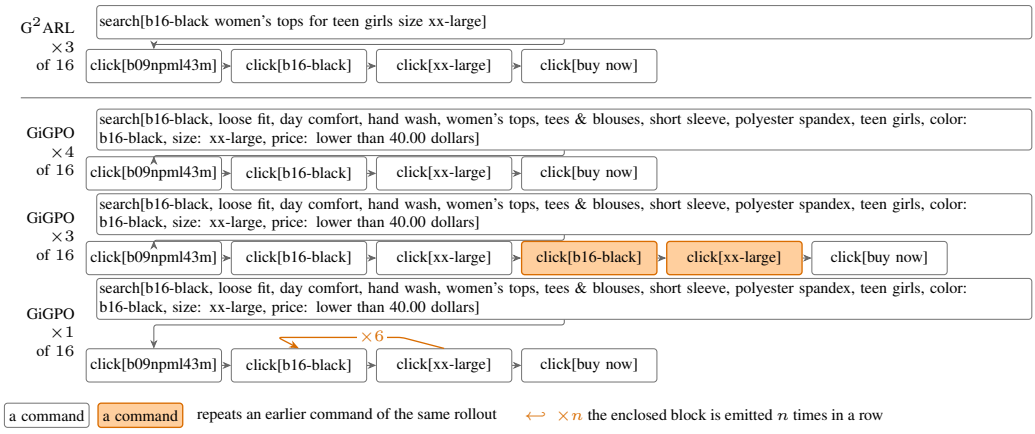

Figure 8: **Where the WebShop repeat-action rate comes from.** Test-split goal 475, k=16 rollouts at temperature 0.4, seed 0, from the G^{2}ARL and GiGPO products behind Table 1. Both buy the target item in all 16 rollouts. Above the rule the most frequent G^{2}ARL sequence, below it GiGPO's two most frequent and its longest, $\times n$ rollouts each. No G^{2}ARL rollout repeats a command, while 12 of GiGPO's 16 successes click an option button again, 10 back to back. Its longest runs 15 commands, the return arrow enclosing a block emitted 6 times in a row, commands 3 to 14.

Table 17: **Training pool and method on ALFWorld**, `valid_unseen`, temperature 0.4, k=16, four decoding seeds. The G^{2}ARL subset cells are the rows Table 1 prints. Each G^{2}ARL cell at 1.5B, and the 1.5B GiGPO full-split cell, is a mean over two training runs on its pool with the standard deviation over those runs in $\pm$. In every other cell the $\pm$ gives the decoding-seed standard deviation.

Scale	Method	Training pool	Succ (%)
1.5B	G^{2}ARL	64-task subset	87.55 ± 2.48
1.5B	G^{2}ARL	full split (3,553)	87.34 ± 2.73
1.5B	GiGPO	64-task subset	63.54 ± 1.10
1.5B	GiGPO	full split (3,553)	69.62 ± 0.26
7B	G^{2}ARL	64-task subset	92.79 ± 0.22
7B	G^{2}ARL	full split (3,553)	91.93 ± 0.28
7B	GiGPO	full split (3,553)	88.61 ± 0.21

Table 18: **Training-run variation on ALFWorld-1.5B for G^{2}ARL and GiGPO.** Each method was trained twice with its printed configuration. Each run was evaluated on `valid_unseen` (134 tasks, k=16, temperature 0.4) with four decoding seeds. Run rows give the mean over the four decoding seeds $\pm$ the decoding-seed standard deviation. The mean rows give the mean over the two runs, the value Table 1 prints, $\pm$ the standard deviation over the two runs. Rep is reported on successful and on all rollouts.

Method	Training run	Succ (%) ↑	Rep-succ (%) ↓	Rep-all (%) ↓	Len ↓	UV ↑
G^{2}ARL	run 1	89.31 ± 0.17	15.22 ± 0.24	21.18 ± 0.19	13.21 ± 0.09	7.616 ± 0.108
G^{2}ARL	run 2	85.80 ± 0.30	16.20 ± 0.13	24.53 ± 0.14	13.61 ± 0.14	6.458 ± 0.142
G^{2}ARL	mean over runs	87.55 ± 2.48	15.71 ± 0.69	22.85 ± 2.37	13.41 ± 0.28	7.037 ± 0.819
GiGPO	run 1	69.44 ± 0.78	21.63 ± 0.38	37.36 ± 0.46	16.26 ± 0.23	5.534 ± 0.112
GiGPO	run 2	69.81 ± 0.70	19.66 ± 0.18	36.18 ± 0.49	14.97 ± 0.12	5.099 ± 0.088
GiGPO	mean over runs	69.62 ± 0.26	20.64 ± 1.39	36.77 ± 0.83	15.62 ± 0.91	5.316 ± 0.307

and the G^{2}ARL rows of Table 1 and the ablations of Table 3 report runs trained there. Table 17 places the full-split runs alongside, with GiGPO on both pools. With the evaluation held fixed (`valid_unseen`, temperature 0.4, k=16, four decoding seeds), the two pools agree within the dispersion over training runs at 1.5B (87.55 ± 2.48 against 87.34 ± 2.73, each the mean over two training runs) and within one point at 7B (92.79 against 91.93). GiGPO reads 63.54% on the subset against 69.62% on the full split, the subset result coming from one training run of the full-split row's recipe with the task pool as the sole change. On the full split, on which every policy-gradient row trains, the paired difference between G^{2}ARL and GiGPO is $+17.72$ points, $[+12.65, +22.91]$ over tasks. TB, whose Table 1 row trains on the subset, retrained on the full split under its own recipe (β=8, its own update rule) reads $55.56 \pm 0.48\%$ at 1.5B, against the 54.41% of Table 1.

Group composition during training (Figure 4). Our bars are telemetry from the training run of row (2) of Table 3, averaged over a short window of consecutive optimizer steps around the center. The baseline bars are reconstructed offline from that method's saved training checkpoints: the adapter saved at the matched point is loaded onto the shared base, a frozen list of training tasks is re-sampled at the group size of the figure and temperature 1.0 under one decoding seed, and the three group classes are counted again. Each numbered column holds one saved checkpoint per method, the three chosen so that the cumulative number of sampled trajectories is equal across the three runs. At the last column, all-success and all-failure groups together are 55% of the GiGPO sample and the all-success share of our row-(2) run is 72.6%. The group-normalized trajectory-level advantage is identically zero on such a group under a binary terminal score. GiGPO's step-level term and GRPO's invalid-action penalty term can remain non-zero there, and neither carries a success-failure contrast.

All-failure groups before injection. In the per-step training logs of G^{2}ARL, all-failure groups are 3.6% of the sampled groups in the first third of the ALFWorld-1.5B run (an earlier training of the recipe on the task subset of Appendix B.5), 25.5% on WebShop-1.5B and 20.4% on ScienceWorld-1.5B, and over the whole run 17.6% on WebShop-1.5B and 14.9% on ScienceWorld-1.5B. A third is one of three consecutive, equal-length parts of the logged step sequence.

Table 19: **Three trainings of SubTB (Hu et al.) on ALFWorld-1.5B.** Variant (1) is the row Table 1 prints. Readings are on `valid_unseen` (134 tasks, k=16), means over four decoding seeds $\pm$ the sample standard deviation over those seeds. Pre-clip norm is the range of the gradient norm before clipping over the run's logged optimizer steps, and Steps clipped is the share of logged steps whose norm exceeded the threshold. The telemetry of variant (2) covers 76% of its optimizer steps. Init. is the warm start and is not trained. Table 11 gives its UV and repeat-action rate.

(a) Training telemetry and the reading at t=0.4

Row	Pre-clip norm	Steps clipped	Succ ↑	UV ↑	Rep-succ ↓
Init.	—	—	0.4889 ± 0.0052	—	—
(1) printed form	$[3535, 8244]$	100%	0.5111 ± 0.0079	7.253 ± 0.113	0.2307 ± 0.0008
(2) released reduction	$[5.59, 12.13]$	100%	0.5028 ± 0.0054	7.453 ± 0.191	0.2326 ± 0.0083
(3) threshold 1→3535	$[3764, 9835]$	100%	0.4959 ± 0.0074	7.159 ± 0.122	0.2341 ± 0.0030

(b) Four temperatures

Row	t=0.4	t=0.7	t=1.0	t=1.3
Success rate ↑				
(1) printed form	0.5111 ± 0.0079	0.5117 ± 0.0020	0.4854 ± 0.0098	0.4226 ± 0.0030
(2) released reduction	0.5028 ± 0.0054	0.5111 ± 0.0095	0.4838 ± 0.0012	0.4159 ± 0.0031
(3) threshold 1→3535	0.4959 ± 0.0074	0.5058 ± 0.0043	0.4865 ± 0.0047	0.4172 ± 0.0038
Repeat-action rate on successful trajectories ↓				
(1) printed form	0.2307 ± 0.0008	0.2646 ± 0.0055	0.2833 ± 0.0011	0.2879 ± 0.0065
(2) released reduction	0.2326 ± 0.0083	0.2633 ± 0.0052	0.2872 ± 0.0056	0.2940 ± 0.0042
(3) threshold 1→3535	0.2341 ± 0.0030	0.2697 ± 0.0082	0.2851 ± 0.0073	0.2955 ± 0.0022

Injection after the buffer fills. On WebShop-7B from the instruction-tuned base, the mean number of retrievable stored trajectories per buffer query rises from 0.05 early in training to about 6 later: the opening updates come from the naturally mixed groups, and injection contributes once those early successes have filled the buffer.

B.6 SubTB (Hu et al.): three trainings of the ALFWorld-1.5B row

The ALFWorld-1.5B SubTB (Hu et al.) row of Table 1 is one of three trainings of that objective. All three keep the form of Hu et al. (2024), which uses prefix potentials in place of a learned state flow. Variant (1) is the printed row: the turn-level port of Appendix C.4, whose per-trajectory window sum is that of Equation (3) of Hu et al. (2024), with the gradient norm clipped at 1. Variant (2) replaces that reduction by the one in the released implementation, which sums the squared window residuals of the whole micro-batch and divides by the number of windows in the batch, $N_{\text{win}} = \sum_i T_i(T_i+1)/2$, rather than by G. The reward configuration, the prefix potentials and the clipping threshold are unchanged. Variant (3) keeps the reduction of variant (1) and raises the clipping threshold from 1 to 3535, the rounded minimum of the printed run's pre-clip gradient norm. Each variant is one training run and reports its final checkpoint (Table 19).

B.7 ScienceWorld: benchmark and evaluation setup

ScienceWorld (Wang et al., 2022) is a text simulator for school-level science experiments, such as measuring a melting point or growing a plant, performed by issuing natural-language commands such as `open door` or `activate stove`. It has 29 task types with several instances each. We use the L1 task split of the RLVMR code base (Zhang et al., 2026c) and its test segment of 1,684 tasks (L0 and L1 are split names of that code base, unrelated to the signature level of UV), a turn limit of 30 and that code base's prompt template. All rows share one supervised cold-start checkpoint, which solves 36.10% of its evaluated episodes.

C Diagnosis: statements, proofs and telemetry

Notation is that of Sections 3.1 and 3.2: $q_i = D_i - \beta y_i + \lambda n_i$ (Equation (2)), a_x the learned partition parameter of TB, in the place of $\log Z_x$, $\bar{q} = \frac{1}{G}\sum_i q_i$, and $g_i = \nabla_\theta \log \pi_\theta(\tau_i) = \nabla_\theta q_i$, because y_i, n_i and the reference terms carry no gradient. The class means $\bar{q}_W$, $\bar{q}_L$, their difference $\Delta = \bar{q}_W - \bar{q}_L$ and the within-class sums V_W, V_L are those of Appendix A.3, and $\bar{g}$, $\bar{g}_W$ and $\bar{g}_L$

are the corresponding means of g_i, zero for an empty class. Telemetry that matches a proposition is reported as consistent with it.

C.1 Setting and provenance of the telemetry

Training-side numbers are step-level values logged by the training named where they are used. Interval statements are over all its recorded steps. On ALFWorld-1.5B from the warm start, TB (one learned partition parameter per task, initialized at 0), SubTB (Hu et al.) (no partition parameter) and G^{2}ARL share one task per micro-batch, gradient-norm clip 1, $\beta = 8$ and $\lambda = 0.1$. FlowRL runs from its released recipe. On WebShop, both reference objectives are trained at 1.5B with $\lambda = 6.4$ and $\lambda = 0.1$, and at 7B with $\lambda = 6.4$. Evaluation-side numbers are those of Table 1.

C.2 The gradient split and the success channel

Proposition 5 (Where the success signal lives)**.** *For binary y, the TB gradient on a group splits into a calibration term, within-class terms and a between-class contrast term:*

$$\nabla_\theta \mathcal{L}_{\mathrm{TB}} = 2(\bar{q} + a_x)\,\bar{g} + \frac{2}{G}\Big[\sum_{i\in W}(q_i - \bar{q}_W)\,g_i + \sum_{i\in L}(q_i - \bar{q}_L)\,g_i\Big] + \frac{2n_W n_L}{G^2}\,\Delta\,(\bar{g}_W - \bar{g}_L). \quad (25)$$

The within-class terms contain no β. The score enters the calibration coefficient, through $\bar{q} = \bar{D} - \beta n_W/G + \lambda\bar{n}$, one scalar on the common direction $\bar{g}$, and, for mixed groups, the contrast coefficient, through $\Delta = (\bar{D}_W - \bar{D}_L) - \beta + \lambda(\bar{n}_W - \bar{n}_L)$, the only term of this decomposition with an explicit factor $\bar{g}_W - \bar{g}_L$. If $n_W n_L = 0$ the contrast term is zero. On ScienceWorld y varies inside a class and the within-class terms carry β as well. The statement is used on ALFWorld and WebShop only.

Proof. $\nabla_\theta \mathcal{L}_{\mathrm{TB}} = \frac{2}{G}\sum_i (a_x + q_i) g_i$. For i in class c write $a_x + q_i = (a_x + \bar{q}) + (q_i - \bar{q}_c) + (\bar{q}_c - \bar{q})$, with $\bar{q}_W - \bar{q} = \frac{n_L}{G}\Delta$ and $\bar{q}_L - \bar{q} = -\frac{n_W}{G}\Delta$. The first two parts give the first two terms of Equation (25). The class-mean parts give $\frac{2\Delta}{G}\big[\frac{n_L}{G} n_W \bar{g}_W - \frac{n_W}{G} n_L \bar{g}_L\big] = \frac{2n_W n_L}{G^2}\Delta(\bar{g}_W - \bar{g}_L)$. Within a class y_i is constant, so $q_i - \bar{q}_c = (D_i - \bar{D}_c) + \lambda(n_i - \bar{n}_c)$ contains no β, and $\bar{q}$ and Δ are read off Equation (2). □

Proposition 6 (All-success and all-failure groups)**.** *If $y_i = 0$ for every trajectory of a group, then $\nabla_\theta \mathcal{L}_{\mathrm{TB}} = \frac{2}{G}\sum_i (a_x + D_i + \lambda n_i)\, g_i$, which carries no explicit current-batch success term. For fixed a_x the loss on that group is zero exactly when $D_i = -a_x - \lambda n_i$ for every i. If $y_i = 1$ for every trajectory, then $\nabla_\theta \mathcal{L}_{\mathrm{TB}} = \frac{2}{G}\sum_i (a_x - \beta + D_i + \lambda n_i)\, g_i$: the score enters only as the shift $-\beta$ of the calibration coefficient, common to every trajectory, and the loss is zero exactly when $D_i = \beta - a_x - \lambda n_i$ for every i. In both cases the zero set is the cost-adjusted reference shape $\pi_\theta(\tau_i \mid x) \propto \pi_{\mathrm{ref}}(\tau_i \mid x)^\rho e^{-\lambda n_i}$ across the group, at a level set by a_x.*

Proof. With $y_i \equiv 0$ the residual is $a_x + D_i + \lambda n_i$, and with $y_i \equiv 1$ it is $a_x - \beta + D_i + \lambda n_i$. Differentiate, and a sum of squares vanishes exactly when every term does. □

Where the two modifications act. By Proposition 2, $\nabla_\theta \mathcal{L}_{\mathrm{pairs}}$ has within-class terms of the form in Equation (25), the between-class term $\frac{2w_{WL} n_W n_L}{\mathcal{N}_g}\Delta\,(\bar{g}_W - \bar{g}_L)$, and no calibration term. With uniform weights it is the last two terms of Equation (25) (Lemma 1). The pair weights set the between-class coefficient share (Equation (21)). The gate and injection set whether the term exists (Section 3.3).

C.3 The group-normalized advantage on all-success and all-failure groups

Proposition 7 (The group-normalized advantage is a pure contrast)**.** *Let $\hat{A}_i = (y_i - \bar{y})/\sigma$, where $\bar{y} = n_W/G$ and σ is the population standard deviation of the group's binary scores, with $\hat{A}_i = 0$ when $\sigma = 0$, and let $u = \sum_i \hat{A}_i\, g_i$ be the trajectory-level update. Then $u = \sqrt{n_W n_L}\,(\bar{g}_W - \bar{g}_L)$: it is zero on an all-success or all-failure group, it contains no common-mode term, and its coefficient depends on the group's composition only.*

Proof. For $\sigma > 0$, $\hat{A}_i = \sqrt{n_L/n_W}$ on W and $-\sqrt{n_W/n_L}$ on L, so $u = n_W\sqrt{n_L/n_W}\,\bar{g}_W - n_L\sqrt{n_W/n_L}\,\bar{g}_L$. Since $\sum_i \hat{A}_i = 0$, $u = \sum_i \hat{A}_i(g_i - \bar{g})$, so no multiple of $\bar{g}$ enters. With the sample standard deviation the coefficient is multiplied by $\sqrt{(G-1)/G}$. □

Telemetry: what the reference objectives optimize from the instruction-tuned base. On WebShop a large share of the sampled groups is all-failure early in training (Appendices B.5 and C.6), and on those groups the residual is $a_x + D_i + \lambda n_i$ (Proposition 6). The 1.5B trainings follow this residual in a form set by λ: at $\lambda = 6.4$ the cost dominates and the episodes shorten, at $\lambda = 0.1$ the log-ratio term D_i dominates and grows.

Training	λ	Loss	Valid-action rate	Mean episode length	Training success
TB	6.4	$5280 \to 148$	$0.243 \to 0.959$	$14.69 \to 4.56$	3.78% (single-step max. 93.75%)
SubTB (Hu et al.)	6.4	–	$\to 1.0$	$\to 3.0$	2.27%
TB	0.1	$3.92 \to 241$	$\to 0.272$	–	7.08%
SubTB (Hu et al.)	0.1	$133 \to 1301$	$\to 0.217$	–	5.17%

Arrows give the first and the last recorded value. A single arrowed value is the last. Training success is the mean over recorded steps. The TB training at $\lambda = 6.4$ also raises its logged mean task score, a training quantity that carries the invalid-turn cost and differs from the evaluation Score of Appendix D.1, from -71.20 to -1.20. The evaluation rollouts show the two forms. At $\lambda = 6.4$ both checkpoints emit well-formed actions, repeat queries to the turn limit and rarely succeed (Table 12). The 7B checkpoint's evaluation-averaged invalid-action rate is 0.0011, and one 7B rollout and one 1.5B rollout have repeat-action rates of 0.867 and 0.933. At $\lambda = 0.1$ the 1.5B checkpoints emit unparseable text, with evaluation-averaged invalid-action rates of 0.82 to 0.85.

C.4 SubTB (Hu et al.): the substituted flow and terminal concentration

Throughout this subsection, $\rho = 1$, $\beta > 0$, $\lambda \geq 0$, and $y \in \{0, 1\}$. Fix a task x on the finite prefix tree of Appendix A. The root is nonterminal. Transitions are deterministic given the history. Every trajectory ends at an environment terminal state or at the turn limit. The policies π_{ref} and π_θ are positive on the same support. Population losses use a fixed sampling distribution μ that is positive on every trajectory. They are minimized over normalized tabular action policies. The infimum over any parameterized policy family is at least the tabular minimum. All gradients hold trajectories, reference probabilities, scores, and invalid marks fixed. An edge (h, a) is *terminal* if ha is complete and *interior* otherwise. Its invalid mark is $\iota(h, a) \in \{0, 1\}$. For $h = s_t$, define $n(h) = \sum_{k<t} \iota_k$.

The deployed form and Equation (3) of Hu et al. (2024). The loss of Appendix D.4 averages over the rows of a micro-batch the per-trajectory sum of squared window residuals. For a trajectory of T turns with node potentials $A_0 = 0, A_1, \ldots, A_T$,

$$\mathcal{L}_{\mathrm{SubTB}}(\tau) = \sum_{0 \leq u < v \leq T} (A_v - A_u)^2, \quad A_t = \sum_{k<t} \big[\log \pi_\theta(a_k \mid s_k) - \log \pi_{\mathrm{ref}}(a_k \mid s_k)\big] - \beta y_t + \lambda n^{(t)}, \tag{26}$$

where y_t is the score of the prefix s_t scored as a finished episode ($y_t = 0$ for $t < T$ and $y_T = y$ on ALFWorld and WebShop) and $n^{(t)} = n(s_t)$. Write $\varepsilon_k = \log \pi_\theta(a_k \mid s_k) - \log \pi_{\mathrm{ref}}(a_k \mid s_k) + \lambda \iota_k$ for the cost-adjusted log-ratio of turn k and $\delta_k = A_{k+1} - A_k$ for its increment, so that $\delta_k = \varepsilon_k$ for $k < T - 1$, $\delta_{T-1} = \varepsilon_{T-1} - \beta y$, and $A_T = q$ (Equation (2)).

Write q_{GFN} for the policy in the formulation of Hu et al. (2024) and $\top$ for its stop symbol, so that $s\top$ is the sequence terminated at s. Hu et al. (2024) replace the state flow in the SubTB residual of Madan et al. (2023) by $F(s) = R(s\top)/q_{\mathrm{GFN}}(\top \mid s)$, the reward of stopping at s divided by the learned stop probability, "to account for trajectories being terminable at all states". Their Appendix A.2 derives the substitution from $R(s\top) = F(s)\, q_{\mathrm{GFN}}(\top \mid s)$ at convergence. Agent trajectories in our formulation have no separate stop action, so the port sets both stop-probability factors of a window to one and fixes the flow of every prefix to the target's unnormalized density of the prefix scored as an unfinished task,

$$\Phi(s_t) = \pi_{\mathrm{ref}}(s_t \mid x)\, e^{\beta y_t - \lambda n^{(t)}}, \qquad \Phi(s_t) = \pi_{\mathrm{ref}}(s_t \mid x)\, e^{-\lambda n^{(t)}}\ (t < T), \qquad \Phi(\tau) = R_x(\tau). \tag{27}$$

The residual of Equation (3) for the window (u, v) then reads $\log \Phi(s_u) + \sum_{k=u}^{v-1} \log \pi_\theta(a_k \mid s_k) - \log \Phi(s_v) = A_v - A_u$, so Equation (26) is Equation (3) with the flow of every prefix fixed to Φ and the stop probabilities set to one.

The prefix flow of the target and the dropped factor. For a prefix h let $n_{\mathrm{rest}}(\tau) = n(\tau) - n(h)$ count the invalid turns after h on a completion $\tau \supseteq h$, and let $\mathbb{E}_{\mathrm{ref}}[\cdot \mid h]$ be the expectation over completions under $\pi_{\mathrm{ref}}(\tau \mid h) = \pi_{\mathrm{ref}}(\tau \mid x)/\pi_{\mathrm{ref}}(h \mid x)$. Define the continuation value

$$\Gamma(h) = \mathbb{E}_{\mathrm{ref}}\big[e^{\beta y - \lambda n_{\mathrm{rest}}} \mid h\big] > 0, \qquad \Gamma(\tau) = e^{\beta y(\tau)} \text{ at a terminal node.} \tag{28}$$

By Equation (15), $F^*(h) = \sum_{\tau \in \mathcal{T}_x(h)} \pi_{\mathrm{ref}}(\tau \mid x) e^{\beta y(\tau) - \lambda n(\tau)} = \pi_{\mathrm{ref}}(h \mid x) e^{-\lambda n(h)} \Gamma(h)$, so at an interior prefix $F^*(h) = \Phi(h)\Gamma(h)$ and at a terminal node $F^*(\tau) = \Phi(\tau)$. Flow conservation, Equation (14), divided by $\Phi(h)$ reads $\Gamma(h) = \sum_a \pi_{\mathrm{ref}}(a \mid h) e^{-\lambda \iota(h,a)} \Gamma(ha)$, and the target's action policy is

$$\pi^*(a \mid h) = \frac{F^*(ha)}{F^*(h)} = \pi_{\mathrm{ref}}(a \mid h)\, e^{-\lambda \iota(h,a)}\, \frac{\Gamma(ha)}{\Gamma(h)}. \tag{29}$$

At the root $\Phi(\emptyset) = 1$ and $\Gamma(\emptyset) = Z_x$. The deployed form substitutes $\Phi(h)$ for the flow $\Phi(h)\Gamma(h)$ of every interior prefix and drops the factor $\Gamma(h)$, the reference expectation of the tilt over the remainder of the episode. At a terminal node nothing is dropped.

A trajectory of T turns contributes $T(T+1)/2$ window residuals, each the sum $\delta_u + \cdots + \delta_{v-1}$ of the increments it spans, so $\mathcal{L}_{\mathrm{SubTB}}(\tau)$ is a positive definite quadratic form in $(\delta_0, \ldots, \delta_{T-1})$: the one-step windows alone contribute $\sum_k \delta_k^2$, and every other window a non-negative term. For binary scores the score βy enters only δ_{T-1}, hence only the T windows (u, T) that end at the terminal node of a successful trajectory (50 of the 1275 windows at the ALFWorld turn limit $T = 50$) and no window of a failed one. Since the windows penalize sums of increments, this count is not a share of the loss or of its gradient.

Let $\Gamma_0(ha) = e^{\beta y(ha)}$ if the edge (h, a) is terminal and $\Gamma_0(ha) = 1$ if it is interior, and define the edge assignment

$$p^\dagger(a \mid h) = \frac{\Phi(ha)}{\Phi(h)} = \pi_{\mathrm{ref}}(a \mid h)\, e^{-\lambda \iota(h,a)}\, \Gamma_0(ha). \tag{30}$$

It is a positive function on the edges of the tree. It is a policy only at the prefixes where it sums to one.

Proposition 8 (Zero-loss assignment and terminal concentration)**.** *(i) $\mathcal{L}_{\mathrm{SubTB}}(\tau) = 0$ if and only if $\delta_k = 0$ for every k, that is $\varepsilon_k = 0$ for $k < T-1$ and $\varepsilon_{T-1} = \beta y$. Equivalently $\pi_\theta(a_k \mid s_k) = p^\dagger(a_k \mid s_k)$ for every turn. On ALFWorld and WebShop this keeps every valid action before the final turn at its reference log-probability and places the whole tilt βy on the final action. The population loss under μ is zero if and only if $\pi_\theta = p^\dagger$ on every edge of the tree. (ii) For every edge, $\pi^*(a \mid h) = p^\dagger(a \mid h)\, \Gamma(ha)/\big(\Gamma_0(ha)\Gamma(h)\big)$. Hence $p^\dagger(\cdot \mid h) = \pi^*(\cdot \mid h)$ at a prefix h if and only if $\Gamma(ha)/\Gamma_0(ha) = \Gamma(h)$ for every available action a. If every available action at h is interior and valid, this holds if and only if $\Gamma(ha)$ takes the same value for every available a. Two such actions with different values of $\mathbb{E}_{\mathrm{ref}}[e^{\beta y - \lambda n} \mid ha]$ make the assignment and the target differ at h. (iii) The following are equivalent: $p^\dagger$ sums to one at every non-terminal prefix, $p^\dagger = \pi^*$ on every edge, $\Phi = F^*$ at every prefix, and $\Gamma(h) = 1$ at every non-terminal prefix. A normalized zero-loss policy exists if and only if $\Gamma(h) = 1$ at every nonterminal prefix. It then equals π^* on the full tree, and $Z_x = 1$.*

Proof. (i) The one-step windows contribute $\sum_k \delta_k^2$ and every window residual is a sum of increments, so the loss is zero if and only if every $\delta_k = 0$. By the definitions of ε_k and δ_k, $\delta_k = 0$ reads $\pi_\theta(a_k \mid s_k) = \pi_{\mathrm{ref}}(a_k \mid s_k) e^{-\lambda \iota_k} \Gamma_0(s_k a_k)$, which is Equation (30). Since every $\mu(\tau)$ is positive and every edge lies on some trajectory, the population loss is zero exactly when $\pi_\theta = p^\dagger$ on every edge. (ii) Divide Equation (29) by Equation (30). If all available actions are interior and valid, $\Gamma_0 = 1$ and $\iota = 0$ at h, so the condition is $\Gamma(ha) = \Gamma(h)$ for all a. Conversely a common value c gives $\Gamma(h) = c \sum_a \pi_{\mathrm{ref}}(a \mid h) = c$ by flow conservation. For a valid edge $n(ha) = n(h)$, so $\mathbb{E}_{\mathrm{ref}}[e^{\beta y - \lambda n} \mid ha] = e^{-\lambda n(h)} \Gamma(ha)$. (iii) If $p^\dagger$ sums to one at every non-terminal prefix, then $\Phi(h) = \sum_a \Phi(ha)$ there and $\Phi(\tau) = R_x(\tau)$ at every leaf, so induction from the leaves gives $\Phi = F^*$ and $p^\dagger = F^*(ha)/F^*(h) = \pi^*$. Conversely $p^\dagger = \pi^*$ sums to one. Since

$F^*(h) = \Phi(h)\Gamma(h)$ at interior prefixes and $F^* = \Phi$ at terminal nodes, $\Phi = F^*$ is $\Gamma(h) = 1$ at every non-terminal prefix, and at the root it is $Z_x = \Gamma(\emptyset) = 1$. □

Let $\Sigma^\dagger(h) = \sum_a p^\dagger(a \mid h) = \mathbb{E}_{\pi_{\mathrm{ref}}(\cdot|h)}\left[e^{-\lambda\iota(h,a)}\Gamma_0(ha)\right]$ be the mass the assignment places at h.

Proposition 9 (Normalization obstruction)**.** *Let y be binary and $\beta > 0$. (i) If every available action at h is terminal and valid and the successful ones have reference mass $\pi_{\mathrm{ref}}(y = 1 \mid h) > 0$, then $\Sigma^\dagger(h) = 1 + (e^\beta - 1)\,\pi_{\mathrm{ref}}(y = 1 \mid h) > 1$. For every normalized policy and every successful action $a^\star$ at h with $p_\star = \pi_{\mathrm{ref}}(a^\star \mid h)$, the final increment of the trajectory $ha^\star$ satisfies $\delta_{T-1} = \log \pi_\theta(a^\star \mid h) - \log p_\star - \beta \le \log(1/p_\star) - \beta$, which is negative whenever $p_\star > e^{-\beta}$. At $\beta = 8$, $e^{-\beta} \approx 3.4 \times 10^{-4}$. (ii) If every available action at h is interior, $\lambda > 0$, and the invalid actions have reference mass $p_{\mathrm{inv}} > 0$, then $\Sigma^\dagger(h) = 1 - (1 - e^{-\lambda})\,p_{\mathrm{inv}} < 1$. (iii) If $\Sigma^\dagger(h) \neq 1$ at some prefix h of the tree, the population loss is positive for every normalized policy, and its minimum over normalized tabular policies is attained and positive.*

Proof. (i) With every action terminal and valid, $\Sigma^\dagger(h) = \sum_a \pi_{\mathrm{ref}}(a \mid h)e^{\beta y(ha)} = \pi_{\mathrm{ref}}(y = 0 \mid h) + e^\beta \pi_{\mathrm{ref}}(y = 1 \mid h)$. The increment bound follows from $\varepsilon_{T-1} = \log \pi_\theta(a^\star \mid h) - \log p_\star$ and $\pi_\theta(a^\star \mid h) \le 1$. (ii) $\Gamma_0 \equiv 1$, so $\Sigma^\dagger(h) = (1 - p_{\mathrm{inv}}) + e^{-\lambda}p_{\mathrm{inv}}$. (iii) A normalized policy has $\sum_a \pi_\theta(a \mid h) = 1 \neq \Sigma^\dagger(h)$, so it differs from $p^\dagger$ on some edge, and a trajectory through that edge has positive mass under μ and a nonzero one-step residual there. For attainment, let B be the finite loss of the reference and fix for every edge (h, a) a trajectory τ_{ha} through it. A policy with loss at most B satisfies $\mu(\tau_{ha})\big(\log \pi_\theta(a \mid h) - \log p^\dagger(a \mid h)\big)^2 \le B$, which bounds every $\pi_\theta(a \mid h)$ away from zero. The sublevel set is thus a closed subset of a compact part of the product of action simplices on which the loss is continuous, so the minimum is attained, and it is positive by the first sentence. □

The batch reduction and the loss scale. The released reduction divides the micro-batch sum of squared window residuals by the number of windows $\sum_i T_i(T_i + 1)/2$ instead of by the number of rows G, which multiplies the printed loss and its gradient by one positive scalar per micro-batch. At fixed rows the gradient direction, the zero set and the normalization obstruction (Propositions 8 and 9) are unchanged, and norm clipping likewise preserves the gradient direction. Appendix B.6 reports the trainings with the released reduction and with the clipping threshold raised.

Telemetry. For the printed ALFWorld-1.5B run, variant (1) of Table 19, the per-turn log-ratio $\log \pi_\theta(a_t \mid s_t) - \log \pi_{\mathrm{ref}}(a_t \mid s_t)$, averaged over all nonpadding turns of the training batch, has recorded minimum and maximum that round to -0.008 and 0.014, and mean 0.0013 over the first third of the run and 0.0036 over the last third. For TB, the maximum rounds to 0.061 and the mean over the last third is 0.0263. For G^{2}ARL run 1 of Table 18, trained at the score weight of Table 20 rather than at the diagnostic setting of Appendix C.1, the maximum rounds to 2.42 and the mean over the last third is 1.13.

C.5 FlowRL: deployed form, assumptions, propositions and proofs

FlowRL (Zhu et al., 2026) is the externally released member of the trajectory-balance form. Its deployed loss, matching the released implementation, is $\frac{1}{B}\sum_i w_i\delta_i^2$, where B is the number of trajectory rows in the loss batch, with

$$\delta_i = \log Z_\phi(x_i) + \frac{S_i - S_i^{\mathrm{ref}}}{N_i} - \beta_F \hat{A}_i, \qquad w_i = \mathrm{clamp}\big(e^{\,S_i - S_i^{\mathrm{old}}},\ \max = 10\big)\ \text{(detached)}, \tag{31}$$

where S_i, S_i^{ref} and S_i^{old} are trajectory-level sums of response-token log-likelihoods under the current, reference and behind-by-one-step policies, N_i is the response-token count, $\hat{A}_i$ is the group-standardized advantage (score minus group mean, divided by the group standard deviation), $\beta_F = 15$ is the released reward scale, distinct from the β of Equation (1), and $\log Z_\phi$ is a learned projection head conditioned on the initial prompt, and $u_i = (S_i - S_i^{\mathrm{ref}})/N_i$ denotes the length-normalized log-ratio. The implementation assigns zero advantage to a group whose scores have zero variance and tags a group as degenerate when $\max_i |\hat{A}_i| < 10^{-8}$. The statement below assumes **(F1)** deterministic transitions, as in Appendix A, and **(F2)** that under on-policy sampling only

the support of the current policy is constrained, so every global statement is read on the empirical objective.

Proposition 10 (Groups with a single score value carry no reward)**.** *Let a sampled group for task x have all trajectories at the same score, so that its group-standardized advantage is identically zero. Then the group's contribution to the loss is $\frac{1}{B}\sum_{i\in\text{group}} w_i\big(\log Z_\phi(x)+u_i\big)^2$, independent of the task reward. Its zero set constrains the length-normalized log-ratio u_i to one value across the group, a per-task constant absorbed by the head.*

Proof. Substitute $\hat{A}_i \equiv 0$ into Equation (31). □

Proposition 11 (Length-normalized fixed point)**.** *If $\delta_i = 0$ for every trajectory of a group, then $\log\pi_\theta(\tau_i \mid x) - \log\pi_{\text{ref}}(\tau_i \mid x) = N_i\big(\beta_F\hat{A}_i - \log Z_\phi(x)\big)$. Two trajectories of the group with equal advantage and token counts $N_i \neq N_j$ differ in log-ratio by $(N_i - N_j)\big(\beta_F\hat{A}_i - \log Z_\phi(x)\big)$, and on a group with a single score value the zero set is $\pi_\theta(\tau \mid x) = \pi_{\text{ref}}(\tau \mid x)\, e^{-N_\tau \log Z_\phi(x)}$ on the sampled trajectories, the reference tilted by token count. The trajectory-level target of Equation* (1) *carries no such factor.*

Proof. Multiply $\delta_i = 0$ by N_i. Set $\hat{A}_i \equiv 0$ for the uniform case. □

An all-failure policy whose rollouts for a task are copies of one trajectory, a single `search[...]` action repeated for 15 turns, is a zero-loss point of the empirical objective: every sampled group has a single score value, so the reward is absent from its loss (Proposition 10), and $\log Z_\phi(x) = -u(x)$, with $u(x)$ the group-constant log-ratio of that trajectory, makes every δ_i zero. Every w_i of Equation (31) is strictly positive, so $\sum_i w_i\delta_i^2 = 0$ if and only if every $\delta_i = 0$: the importance truncation changes the gradient field away from the zero set and leaves the zero set unchanged.

Telemetry. On WebShop-7B from the instruction-tuned base, the per-step share of loss mass carried by zero-advantage groups averages 0.7846 over the recorded steps and the degenerate-group fraction averages 0.8383. Early in training the minimum of the logged group scores is exactly -0.1 on every recorded step, the value of the invalid-action penalty. This source of within-group score variance recedes as the valid-action ratio rises, and a group whose trajectories then share one score carries no reward (Proposition 10). The recorded steps are divided into six equal segments. The rows below are the first, the fifth and the last, each entry the mean over the steps of the segment:

Training segment	valid-action ratio	degenerate groups	degenerate loss mass	training success
Early	0.7805	0.4800	0.4011	1.38%
Late	0.9844	0.9650	0.9612	0.13%
Final	0.9569	0.9837	0.9757	0.00%

These readings are consistent with Proposition 10 and the zero-loss construction above. The evaluation row of that training solves no task at any of the four evaluated decoding temperatures (Table 5), and its run settings are in Appendix D.2. On ALFWorld-7B from the warm start the same loss-mass key averages 0.0014 while the degenerate-group fraction is 0.4188, consistent with the mixed groups of a warm-started run carrying the residual $\beta_F\hat{A}_i$ while the reference-fitting residual $\log Z_\phi(x) + u_i$ of a uniform group is closed by the head.

C.6 Controls on the TB rows: injection, clipping and tempering

We take the TB training on WebShop-1.5B at $\lambda = 0.1$ and replace its replay rule, which supplies every task a fixed number of stored successes regardless of the group's outcome, with the intervention of Section 3.3: stored successes are injected into all-failure groups, and an all-failure group whose task has no stored success enters the loss with weight zero. The objective, the base checkpoint, the prompt template and the low-rank adaptation are unchanged. The injected training reaches $16.73 \pm 0.30\%$ test success against $13.34 \pm 0.37\%$ for the unmodified training, three decoding seeds each. The paired difference is $+3.39$ points, $[+2.80, +4.02]$ over tasks. Its evaluation-averaged invalid-action rate is 0.79 against the 0.82 to 0.85 of Appendix C.3, and its training valid-action rate

reads 0.344 against 0.272. Before injection, 50.4% of its sampled groups are all-failure (65.0% over the first third of training).

Clipping and tempering controls for the ALFWorld-1.5B TB row. Two further TB trainings with the gradient-norm clipping threshold raised from 1 to 96 read $55.85 \pm 0.76\%$ success with a constant sampling temperature and $54.82 \pm 0.37\%$ with a tempered-sampling schedule specified for this comparison, against the printed $54.41 \pm 0.21\%$ (Table 1).

D EXPERIMENTAL DETAILS

D.1 SETUP DETAILS

Metrics. The metrics are those of Section 4.2. *Success* is the fraction of the $|\text{tasks}| \times k$ rollouts that the environment's own symbolic verifier accepts (on ALFWorld a check of the goal predicates). No model judges its own output anywhere in this paper. *Score* is on each benchmark's native scale: 0 to 100 on WebShop, and on ScienceWorld partial credit for a partly completed experiment, with maximum 100 and possibly negative. ALFWorld has no continuous score. *UV* is Equation (11). Level 0 is the finest command signature, and the mean is at least 1. *Rep-succ* averages the repeat-action rate over the trajectories the verifier marks solved and *Rep-all* over every trajectory of that evaluation. The two have different denominators and are never subtracted from one another. *Loop share* is reported on the all-trajectory basis. The panel of Appendix B.3 prints all of these and the mean longest repeated block. Rep-succ is computed on ALFWorld and WebShop from the stored command signatures of successful trajectories, the computation the paired intervals of those two benchmarks use. The WebShop evaluator also writes a per-trajectory field from the executed action text, and on our printed rows the two agree to four decimals. On ScienceWorld the printed cells and the paired intervals read that per-trajectory field over the solved trajectories, the convention Table 12 states for the loop columns.

Resampling units. Paired intervals come from a cluster bootstrap over tasks with $10{,}000$ resamples: a drawn task carries all of its rollouts as one block, its decoding-seed average unchanged, so the intervals describe task-sampling uncertainty. The primary unit is the task (134 on ALFWorld, 500 on WebShop, $1{,}684$ on ScienceWorld). A second, coarser unit is used where the benchmark registers one, ALFWorld's 47 task configurations and ScienceWorld's 29 task families, while WebShop's test segment (goal indices $[0, 500)$, one shopping instruction per task) has the task as its only unit. The ALFWorld-1.5B G^{2}ARL intervals pool the evaluation chains of its training runs (Appendix B.5), each task carrying the mean of those chains, and condition on those training runs as fixed panels.

Seeds and dispersion. ALFWorld cells average four decoding seeds $(0, 1, 2, 1000)$, WebShop three $(0, 1, 2)$, ScienceWorld three, with four for the supervised cold start at $t{=}0.4$ and two at $t{=}1.3$ (Table 7). Every $\pm$ is a sample standard deviation over decoding seeds with divisor $n - 1$. Row (2) of Table 3 and the WebShop-7B G^{2}ARL row evaluate each of their seeds more than once and take that dispersion over evaluation runs instead, eight and six of them respectively. The ALFWorld-1.5B G^{2}ARL and GiGPO rows of Table 1 are means over two training runs on their pools (Appendix B.5), with $\pm$ the standard deviation over decoding seeds of that mean. Table 18 gives the per-run readings and the standard deviation over runs, 2.48 points of success for G^{2}ARL.

Terms. A *block* is one benchmark-and-scale sub-table of Tables 1 and 2, and the *printed* row of a method is the one configuration of that method reported in a block. A replay *container* is the buffer's de-duplication rule: a candidate whose command-sequence signature, at the level given in Table 20, matches a stored trajectory replaces it only if its verifier score is higher, and the setting *none* admits every candidate (Appendix A.7).

Statistical reporting. A difference quoted from two printed cells is the subtraction of the cells as printed. A paired difference is computed from the stored evaluation outputs before rounding, under the pairing stated in its caption. The two can differ where both are printed.

Reward parameters across rows. Tables 20 and 21 give β and λ per row. Within each printed WebShop block our row and both reference rows use the same λ (0.1 at 1.5B, 6.4 at 7B), and at 1.5B

each reference objective was trained at both values. The policy-gradient rows carry their toolchain's invalid-action penalty. On ALFWorld the two reference rows train at β=8 and the printed G^{2}ARL rows at β=64 (1.5B) and β=128 (7B). Table 4 prints G^{2}ARL at β=8 as well, above both reference rows at every temperature and at both model sizes. The ALFWorld diagnostic runs in Appendix C.1 share β=8 and λ=0.1.

Evaluation checkpoint, serving and task ordering. Each printed row is evaluated at the adapter of its planned end of training. Every post-trained row of Tables 1, 2 and 3 is served as the shared base model plus that method's adapter. Task enumeration order is pinned to sorted order for every comparison, because filesystem order selects a different task subset.

D.2 TABLE NOTES

Table 1: dashes, and `n/a` in Tables 6 and 12. ALFWorld is the `valid_unseen` split (134 tasks, k=16) and WebShop the held-out test split (500 tasks, k=16). Table 1 prints a dash in two situations. The WebShop columns of the 7B *Init.* row were not evaluated: the 7B bare instruction-tuned base has no evaluation on the test split this table uses. The WebShop cells of the 7B FlowRL row belong to a training that solves no task in any of its three decoding seeds (8,000 trajectories per seed): its Success and Score are zero and its Rep-succ, Len and UV, which are computed over successful trajectories, are undefined. Tables 5 and 12 print those zeros, and Tables 6 and 12 mark the undefined cells `n/a`. The row's all-trajectory repeat rate, 38.48%, is in Table 12.

The starting point of the WebShop-1.5B block. Every row of the WebShop-1.5B block starts from the bare instruction-tuned Qwen2.5-1.5B model, and its *Init.* row is that model measured on the same test split at k=16, temperature 0.4 and three decoding seeds. The behavior-cloned checkpoint named *Supervised WebShop ckpt* in Appendices B.2 and B.3 is not the starting point of any printed row.

Dispersions. Every numeric cell of Tables 1, 2 and 3 and of the behavior panels (Tables 11 and 12) prints its sample standard deviation over decoding seeds as a subscript, except row (2) of Table 3 and the WebShop-7B G^{2}ARL row, whose dispersion is over repeated evaluation runs (Appendix D.1). For the two ALFWorld-1.5B rows that average two training runs, the subscript is the standard deviation over decoding seeds of the two-run mean. Their dispersion over training runs is in Table 18.

The WebShop-7B FlowRL row and the ALFWorld-7B G^{2}ARL row. The FlowRL row at WebShop-7B aggregates three decoding seeds at the pre-registered reporting point, and success and Score are exactly zero in every seed: every trajectory exhausts the turn limit without issuing a purchase action. The run sets the coefficient on the penalty that keeps the policy close to its reference model to 0.0, the literal value in the method's official script, and its partition head is trained at full rank while sharing the adapter's gradient-clipping threshold (Appendix D.4). This cell is excluded from the paired-bootstrap analyses because its success is zero in every seed, so that its paired difference would equal our success rate and its Rep-succ and Len are undefined. The ALFWorld-7B G^{2}ARL row is the β=128 configuration (Appendix B.5). Table 4 prints G^{2}ARL at β=8 at each scale as well.

Table 2: dispersion convention. Every $\pm$ is a sample standard deviation over decoding seeds with divisor $n-1$, computed from the per-seed products without intermediate rounding (the results record carries the repeat-action and length columns with divisor n, and the point estimates are identical). The supervised cold-start row uses four decoding seeds while every other row uses three.

D.3 HYPERPARAMETERS

Table 20 lists the configuration of every G^{2}ARL row of Tables 1 and 2. Rows (2) to (7) of Table 3 use its ALFWorld-1.5B configuration at reward scale β=8 and vary the settings named in each row, and row (1) is the Table 1 row (Appendix B.5). Table 21 lists the baseline settings, read from the launch scripts and configuration files that produced the printed rows. Every baseline is retrained under the same low-rank adaptation as our rows, and its other settings follow its released recipe (Appendix D.4). The trajectory similarity of Equation (7) is the four-term function of Appendix A.6, and

Table 20: **Training configuration of the main G^{2}ARL rows.** The signature container's level 0 is the verbatim command sequence. Level 3 keeps only the object-manipulation commands with their instance identifiers removed and consecutive repeats collapsed. Baseline settings are in Table 21.

	ALFWorld		WebShop		ScienceWorld
	1.5B	7B	1.5B	7B	1.5B
Starting checkpoint	behavior-cloned		instruction-tuned base		supervised cold start
Low-rank dimension / scaling	8 / 16	8 / 16	32 / 64	32 / 64	32 / 64
Adapted modules	attention projections q, k, v, o				
Turn limit	50	50	15	15	30
Group size G	16	16	16	16	32
Training sampling temperature	1.0				
Reward scale β	64	128	64	64	64
Invalid-action penalty λ	0.1	0.1	0.1	6.4	6.4
Reference exponent ρ	1.0				
Pair weights w_{WW}, w_{LL}, w_{WL}	$1, 0.3, 3$				
All-success group weight α	0	0	0.25	0	0.25
Log-probability penalty coefficient	0.01				
Mixed-group replacement probability	annealed $0.75 \rightarrow 0.25$ over the run				
Replay capacity per task C	64	64	64	64	16
Signature container	none	none	level 3	level 3	level 0
Selection gain form	quality-gated		additive		
Selection weights (dissimilarity, length, state, tool)	$1.0, 0.1, 0, 0.5$				$1.0, 0.1, 0, 1.0$
Partial-credit admission	off	off	off	off	on

Table 21: **Training settings of the retrained baselines**, read from the launch scripts and configuration files that produced the printed rows. Upper block: the baselines trained with the verl-agent toolchain (PPO, GRPO, GiGPO, Flow-GRPO and FlowRL). Lower block: the two reference GFlowNet objectives, trained with our stack. Two values a launch script does not set are read from the toolchain's own configuration: the training sampling temperature on ALFWorld and ScienceWorld, and the ratio-clipping bound. The invalid-action penalty is off for Flow-GRPO on every benchmark (its released recipe) and for GRPO and FlowRL on ScienceWorld (the upstream scripts). Method-specific settings: GiGPO discounts its step-level returns at γ=0.95. FlowRL uses β=15, an importance-weight cap of 10 and weight decay 0.01. PPO uses a low-rank critic and generalized advantage estimation with discount and trace decay both 1. For three rows (ALFWorld-1.5B PPO, ScienceWorld PPO and Flow-GRPO) the launch script is not in our repository. Of their settings our records establish those the shared script template fixes. The G^{2}ARL settings are in Table 20.

	ALFWorld		WebShop		ScienceWorld
	1.5B	7B	1.5B	7B	1.5B
Baselines using the verl-agent toolchain					
Low-rank dimension / scaling (q, k, v, o)	8 / 16	8 / 16	32 / 64	32 / 64	32 / 64
Turn limit	50	50	15	15	30
Group size (rollouts per task)	8	8	8	8	8
Sampling temperature (training)	1.0				
KL coefficient: PPO, GRPO, GiGPO	off	off	0.01	0.01	off
KL coefficient: Flow-GRPO / FlowRL	0.001 / 0				
Ratio clipping	0.2 (Flow-GRPO upper clip 0.3)				
Invalid-action penalty	0.1	0.1	0.1	0.1	0.1 (GiGPO)
Reference GFlowNet objectives: TB and SubTB (Hu et al.) (our stack)					
Low-rank dimension / scaling (q, k, v, o)	8 / 16	8 / 16	32 / 64	32 / 64	32 / 64
Reward scale β, invalid penalty λ	8, 0.1	8, 0.1	64, 0.1	64, 6.4	64, 6.4
Sampling temperature, turn limit	1.0, 50	1.0, 50	1.0, 15	1.0, 15	1.0, 30

its command-signature term uses level 3 on ALFWorld and WebShop and level 0 on ScienceWorld, independently of the container row. Each injected stored trajectory is teacher-forced through the current policy and re-executed in the environment (Appendix A.7). Over training, the stored trajectories the replay mechanism injected into training groups amount to 5.9% of the trajectories sampled in the second ALFWorld-1.5B training run, 3.0% on ALFWorld-7B, 6.7% on WebShop-7B and 10.4% on ScienceWorld-1.5B. The telemetry of the first ALFWorld-1.5B run and of the WebShop-1.5B run predates this counter.

D.4 BASELINES: PROVENANCE AND KNOWN DEVIATIONS

Implementations. PPO, GRPO, GiGPO and Flow-GRPO run in `langfengQ/verl-agent` at commit `20bd331`, an agent-oriented derivative of verl (Sheng et al., 2025) that is the reference implementation released with GiGPO. Flow-GRPO is the method of Li et al. (2026), taken from `lupantech/AgentFlow` at `b940064`, in its released-recipe configuration. FlowRL is Zhu et al. (2026), its objective taken from `Xuekai-Zhu/FlowRL` and run as a patch inside the same verl-agent checkout, at the recommended settings of its released recipe. Under our unified low-rank setting it trains its partition head at full rank while its policy is low-rank, the two sharing one gradient-clipping threshold, whereas its released recipe fine-tunes all parameters. On ScienceWorld these baselines run in the same framework at the same commit, with rank-32 low-rank adaptation, a turn limit of 30 and the prompt template of the RLVMR code base (FlowRL: Appendix D.2). The two reference GFlowNet objectives run in our own stack, whose in-process vLLM sampling engine, one environment actor per concurrent trajectory, in-memory adapter hot-swap and fixed-length history prompt are shared with the stack the policy-gradient baselines run in (Feng et al., 2026; Sheng et al., 2025). The GRPO and GiGPO recipes we retrain register no reference model on ALFWorld and ScienceWorld, and use one on WebShop as a KL penalty. **Every baseline cell in our tables is that method retrained by us.**

Deviations from the released language-model TB implementation. Our TB baseline learns one partition parameter per training task. The released language-model TB implementation learns a single scalar. Tempered sampling with reward-temperature annealing, which the released implementation uses to begin training from a weak starting point, is disabled on both sides of our comparison. The two reference rows retain the released implementation's replay interface, which retrieves a fixed number of stored trajectories per task or none. G^{2}ARL's gate injects whichever stored successes are available (Section 3.3). The invalid-action penalty takes two values on WebShop (Appendix D.1).

The deployed SubTB form of Hu et al. (2024). The baseline is our turn-level port of the modified sub-trajectory balance of Hu et al. (2024), Equation (3) of that paper: a node is a turn boundary (an action is one turn of generated text, and the observation the environment returns is not a policy step), a fixed prefix potential replaces the learned state flow, and the form carries no partition parameter. Writing y_t and $n^{(t)}_{\text{invalid}}$ for the verifier score and the invalid-turn count of the prefix terminated at turn t, that potential is $\Phi(s_t) = \pi_{\text{ref}}(s_t \mid x)^{\rho} e^{\beta y_t - \lambda n^{(t)}_{\text{invalid}}}$, the node potential is $A_t = \sum_{k<t} \big[\log \pi_\theta(a_k \mid s_k) - \rho \log \pi_{\text{ref}}(a_k \mid s_k)\big] - \beta\, y_t + \lambda\, n^{(t)}_{\text{invalid}}$, the residual of Equation (2) for that prefix, and the loss is $\mathcal{L}_{\text{SubTB}} = \frac{1}{G} \sum_{i=1}^{G} \sum_{0 \le u < v \le T_i} \big(A^{(i)}_v - A^{(i)}_u\big)^2$ over the G trajectories of a micro-batch. The released implementation divides the batch sum by the number of windows in the batch, $N_{\text{win}} = \sum_i T_i(T_i+1)/2$, instead of by G. Equation (3) of Hu et al. (2024) carries a stop probability at each end of a window, since there the policy itself emits the end token. The turn-level port has no separate stop action and sets both stop-log-probability terms of the residual to zero. Appendix C.4 derives the zero-loss edge assignment of this form and its normalization obstruction. Appendix B.6 reports two further trainings of this row.

Placement against the published split. The ALFWorld tables of the framework we retrain in are measured on `valid_seen`, an in-distribution split (the released environment wrapper passes `eval_in_distribution` to the ALFWorld task loader, and no released script overrides it), whereas the main tables of this paper are measured on the held-out `valid_unseen` split, whose room layouts do not appear in training. The published rows also train all parameters from the instruction-tuned base, whereas every ALFWorld row in our tables trains a rank-8 adapter from the shared behavior-cloned warm start. Table 22 places our runs on the published split, which removes the first difference and leaves the other two. Under this change of training regime the retrained values move in both directions: PPO at 1.5B reads 17.4 points above its published value, GiGPO and GRPO at 7B read 6.5 and 14.9 points above, GRPO at 1.5B lies within the published standard deviation, and GiGPO at 1.5B reads 12.4 points below. On the published split G^{2}ARL reads 89.85% at 1.5B for the row Table 1 prints, 93.61% for its full-split counterpart (Appendix B.5), and 97.56% at 7B. GiGPO retrained under the same regime reads 74.35% at 1.5B.

Table 22: **ALFWorld success rate on the split used by the published tables** (valid_seen, 140 tasks, temperature 0.4, k=16), with the valid_unseen column giving the held-out split of Table 1 (134 tasks). The G^{2}ARL rows marked * are trained on the 64-task subset and are the rows Table 1 prints. The unmarked 1.5B row is trained on the full training split (Appendix B.5). Published values are the "All" column of the GiGPO paper for the standardized variant, which is the variant our retraining runs. Each 1.5B G^{2}ARL cell, and the 1.5B GiGPO valid_unseen cell, is a mean over training runs with $\pm$ over those runs (Appendix B.5). The number of decoding seeds behind a training run's valid_seen reading varies. In every other cell of ours the $\pm$ is over decoding seeds, the valid_seen readings of the GiGPO and GRPO 7B rows being a single seed. The published $\pm$ is over training seeds. valid_seen is the split on which our reward scale was selected (Appendix B.5). Dashes: cells the published table does not contain. Bold marks the best valid_seen value in each block.

Method	Ours, valid_seen	Ours, valid_unseen	Published, valid_seen
Qwen2.5-1.5B-Instruct			
G^{2}ARL*	0.8985 ± 0.0086	0.8755 ± 0.0248	—
G^{2}ARL, full split	$\mathbf{0.9361 \pm 0.0058}$	0.8734 ± 0.0273	—
GiGPO	0.7435 ± 0.0039	0.6962 ± 0.0026	0.867 ± 0.017
PPO	0.7183 ± 0.0032	0.6734 ± 0.0087	0.544 ± 0.031
GRPO	0.7009 ± 0.0012	0.6571 ± 0.0100	0.728 ± 0.036
Qwen2.5-7B-Instruct			
G^{2}ARL*	$\mathbf{0.9756 \pm 0.0022}$	0.9279 ± 0.0022	—
GiGPO	0.9732	0.8861 ± 0.0021	0.908 ± 0.013
GRPO	0.9254	0.8144 ± 0.0013	0.776 ± 0.052

Supervised starting points. The ALFWorld warm start is supervised on the successful one-step-expert replays of the reinforcement-learning baselines' own code base, rendered through the same environment manager those baselines use, with rank-32 low-rank adaptation over all linear layers. Its supervision target is a fixed template <think>...</think><action>...</action> whose reasoning text is generated from the target action itself, shared by every method that warm-starts from it. The supervised adapter is merged into the starting-point weights, and every ALFWorld row trains a fresh adapter on top of that single merged model. The ScienceWorld cold start is a supervised fine-tune of the instruction-tuned 1.5B model on expert trajectories replayed from the simulator's reference solutions and rendered through the same prompt template. Every ScienceWorld row trains a fresh adapter on top of it.