REWARD REASONING FOR ITS GAIN OVER A SNAP JUDGMENT: INVERSE-RL COLD STARTS FOR GUI AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cold-starting GUI agents requires learning from expert demonstrations before sparse task-completion rewards can support effective online reinforcement learning. Supervised fine-tuning (SFT) provides this initialization, but fitting expert traces can come at the cost of general capability. Learning a reward from demonstrations offers another route: the agent trains on its own decisions, guided by a critic that reasons about whether a decision came from the expert or the policy. The challenge is that source labels supervise the critic's final answer, not whether its reasoning helped. Obvious errors and stylistic cues can make a pair easy to classify, allowing uninformative reasoning to receive the same reward as reasoning that improves the judgment. We introduce ReGain, which trains the critic on the improvement its reasoning provides over a direct, no-reasoning judgment. For each policy-expert pair, we score the verdict distributions with and without chain-of-thought reasoning using a detached Brier score and reward their difference. The continuous score captures improvements even when the selected verdict is unchanged, while centring rewards across distinct pairs gives greater relative credit to reasoning that improves the comparison. Confidently correct direct judgments leave little room for positive gain, reducing the reward for reasoning that merely repeats obvious source cues. On a 27B GUI agent trained on 20,480 demonstrations, ReGain outperforms the adversarial baseline RARO and SFT on the same data and delays the critic's reliance on stylistic cues. It matches SFT trained on nine times as many demonstrations in pass@1 and pass@4, without the 14% loss of general capability incurred by either SFT baseline.

1 INTRODUCTION

Graphical user interface (GUI) agents that operate applications from screenshots and natural-language instructions are a practical route to general digital assistance (Zhang et al., 2025a; Nguyen et al., 2025), and online reinforcement learning (RL) lets them learn to complete tasks end to end (Bai et al., 2024; Qi et al., 2025). The online reward is sparse, however. It arrives once per episode, identifies neither the failed step nor the reasoning behind a decision, and is rarely positive while the agent cannot yet complete tasks. Many GUI applications also have a richer resource: annotators record expert demonstrations step by step, each pairing a screen and an instruction with the expert's decision, its reasoning and its action (Wu et al., 2025; Xu et al., 2025; Qin et al., 2025). These pairs supervise individual decisions, which the episode reward cannot, and raise the question of how to use this richer feedback before online training.

The standard answer is supervised fine-tuning (SFT), which maximizes the likelihood of the demonstrated tokens. Fitting fixed targets can reduce output diversity (Li et al., 2025) and out-of-distribution performance, with later RL recovering only part of the loss (Jin et al., 2025), and prolonged post-training can reduce plasticity (Hernandez-Garcia et al., 2026). These findings motivate learning from the policy's own decisions rather than repeatedly fitting expert text. Existing on-policy methods, however, each rely on something a set of demonstrations does not provide. On-policy distillation scores every token of the student's rollouts with a teacher (Agarwal et al., 2024; Song & Zheng, 2026), is sensitive to the choice of teacher and loss (Zhu et al., 2026; Fu et al., 2026), and helps only when the teacher is stronger and reasons compatibly (Li et al., 2026), yet public models do not

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.

know the in-house action protocols of a proprietary GUI domain. On-policy self-distillation lets the model, conditioned on the demonstration, serve as its own teacher (Zhao et al., 2026), but it still matches per-token distributions, and the student can learn to behave as if it still saw the reference, a privilege illusion that destabilizes training (Wu et al., 2026b; Yang et al., 2026). Filtered self-training (Zelikman et al., 2022; Yuan et al., 2023; Singh et al., 2024) and rule-based step rewards (Lu et al., 2026; Luo et al., 2025) need a correctness check, and matching one recorded action rejects the valid alternatives a screen often admits.

Inverse RL offers a promising alternative: it learns a reward from demonstrations, without requiring an external teacher or task verifier (Ng & Russell, 2000; Ho & Ermon, 2016). RARO (Cai et al., 2026) brings this approach to language-model reasoning. One model acts as policy and critic, the critic sees a policy and an expert decision for the same state in random order and names the expert's side (A, B, or tie), and the policy is rewarded for being judged expert, so it learns from its own samples and is judged on whole decisions. When applied to a 27B GUI agent, however, this recipe degraded within about 100 steps, with three symptoms. **S1, verdict collapse:** ties crowd out A/B verdicts while critic accuracy stays near chance, on up to two thirds of pairs at a tie value of 0.75. **S2, the CoT adds little:** the critic's gain from reasoning over committing without it grows uniform across pairs, and at a tie value of 0.55 the critic stops writing a chain of thought (CoT). **S3, authorship cues:** late in training the critic increasingly picks the shorter decision, the expert's being short, while the policy it trains gets worse (§5.4).

All three arise from pairs whose source the critic can identify before it reasons, namely malformed policy decisions, easy pairs, and pairs whose authorship shows in their style. A verdict-only reward pays these pairs in full, so they dominate the critic update. We therefore ask how a CoT should be rewarded when the critic could already make a useful judgment without it, and answer by rewarding the gain from reasoning. ReGain scores the critic twice on every pair, once after a sampled CoT and once when it is forced to commit without reasoning, its *snap judgment*, and rewards the difference (§4.3). A pair the snap judgment already decides correctly can gain at most the margin it leaves open, so group centring moves the credit to pairs where reasoning decided the verdict. The gain is measured with a detached Brier score read from the verdict distribution before sampling, over a critic group that spans distinct pairs, as RARO's grouping over the policy rollouts of a task does. Because style-decided pairs are among those withheld, S3 is delayed rather than removed (§5.4).

Contributions. *(1)* A diagnosis of RARO at single-step granularity that separates verdict collapse from healthy abstention and traces S1–S3 to pairs decided at a glance (§4.1). *(2)* The gain-from-reasoning reward, which withholds credit from pairs decided at a glance, with its expectation under calibration and the proof that it needs a heterogeneous critic group (§4.3). *(3)* A detached Brier score on the verdict distribution that makes the gain measurable on a single sample, of which the discrete reward with tie value 0.75 is the three-point restriction (§4.2). On a 27B GUI agent, ReGain outperforms RARO and data-matched SFT and matches full-data SFT in pass@1 and pass@4 without its 14% loss of general capability (§5).

2 Related Work

Training GUI agents. Trained GUI agents are typically initialized by SFT on demonstration trajectories, often after grounding-oriented pre-training (Wu et al., 2025; Xu et al., 2025; Qin et al., 2025). Online RL then optimizes task outcome against an environment signal or a learned evaluator (Bai et al., 2024; Qi et al., 2025). R1-style recipes instead score each step by rule-based agreement with a reference action (Lu et al., 2026; Luo et al., 2025); like the exact-match filter of §1, this gives zero reward to a valid decision whenever it differs from the reference, although several actions are often admissible at a state. Our stage precedes these pipelines as a decision-level cold start from the same demonstrations that assumes no environment reward, no learned outcome evaluator, and no exact-match rule.

Cold starts for RL. The standard cold start for RL with verifiable rewards (DeepSeek-AI, 2025; Shao et al., 2024b; Yu et al., 2025; Zheng et al., 2025) is SFT on curated reasoning traces (DeepSeek-AI, 2025), but extended fixed-target training can reduce output diversity (Li et al., 2025), interfere with later post-training stages (Fernando et al., 2024), reduce out-of-distribution performance (Jin et al., 2025), and reduce plasticity (Hernandez-Garcia et al., 2026). These findings motivate judging

an initialization by what later RL can reach from it rather than by its offline score alone, a distinction also drawn in offline-to-online RL (Nakamoto et al., 2023); we approximate it by pass@4, general-capability retention, and trajectory-level success before online training (§5.2). Filtered self-training (Zelikman et al., 2022; Yuan et al., 2023; Singh et al., 2024) and on-policy distillation (Zhao et al., 2026; Song & Zheng, 2026; Li et al., 2026) keep the samples on-policy but train against token-level targets selected by a filter or a teacher. A different line removes the verifier from RL itself. VeriFree (Zhou et al., 2026) and JEPO (Tang et al., 2025) score reasoning by the likelihood of a reference answer, while General-Reasoner (Ma et al., 2025) and Writing-Zero (Jia et al., 2025) use a trained judge or pairwise reward model. Applied to single-step decisions, the first family commits to the recorded action as a soft target and inherits the exact-match limitation above; the second requires a separately trained judge, an external evaluator that a cold start from demonstrations alone does not have.

Inverse reinforcement learning. Learning a reward from demonstrations is the inverse-RL problem (Ng & Russell, 2000), and adversarial imitation solves it with a discriminator that separates expert from policy behavior (Ho & Ermon, 2016). Related language-model work learns alignment rewards from demonstrations (Sun & van der Schaar, 2024), trains generative verifiers (Zhang et al., 2025b), and pits a reasoner against an adversarial critic (Liu et al., 2025a; Cai et al., 2026); RARO is the recipe we adapt. Two properties of the objective carry over from the imitation literature. First, a discriminator uses whatever separates the sources, and shortcut features are documented in adversarial imitation (Zolna et al., 2021) and in learned classifiers generally (Karras et al., 2020; Geirhos et al., 2020), and text carries detectable authorship signatures (Huang et al., 2024; Kumarage & Liu, 2023). Second, demonstrations do not determine the reward, since multiple rewards fit the same data even in the infinite-data limit (Skalse et al., 2023), and identification requires information that a source label does not provide (Cao et al., 2021). The residual failure we report is therefore reward ambiguity, not reward hacking against a fixed misspecified reward (Skalse et al., 2022; Gao et al., 2023), and we document it without claiming to remove it (§5.4). The critic in these methods is rewarded through a discrete verdict. Strictly proper scoring rules instead score the reported probability and are maximized by honest reports (Gneiting & Raftery, 2007); they have served as generation losses (Shao et al., 2024a) and, in RLCR (Damani et al., 2025), as an RL reward for a verbalized confidence. RLCR is the closest prior work to our critic return; §4.2 states how the scored event and gradient path differ.

3 PRELIMINARIES

We adapt RARO (Cai et al., 2026) to single-step supervision. A shared model θ plays two roles. As a *policy*, it produces a decision $y \sim \pi_\theta(\cdot \mid x)$ for a context x, the current GUI screen and task instruction (Zhang et al., 2025a; Nguyen et al., 2025); a decision contains free-form reasoning, a statement of intent, and an action from a restricted grammar. As a *critic*, it receives a side-randomized pair $(y, y^\star)$ of one policy and one expert decision for the same context and generates a CoT followed by a verdict distribution (p_A, p_B, p_T) over side A, side B, and tie. We keep RARO's name for this role, although the critic is the discriminator of adversarial imitation, not the value function of actor–critic methods.

The critic is supervised by source identity, earning 1 when its verdict selects the expert side, 0 when it selects the policy side, and $c = \tau_{\text{crit}}$ for a tie; the policy receives the opposing reward, with τ_{pol} for a tie. The tie lets an uncertain critic abstain. In the original study, it is emitted on about 70% of pairs, removing it reduces final accuracy, and training is reported stable for $\tau_{\text{crit}} \in \{0.50, 0.55\}$ (Cai et al., 2026).

Both roles are optimized with Group Relative Policy Optimization (GRPO) (Shao et al., 2024b). A policy group consists of G rollouts for the same context. Each rollout is paired with the expert decision, with an independently randomized side order, and becomes one critic prompt, so the critic group for that context consists of G different prompts, one per policy rollout. Within a group the advantage is the mean-centred reward $A_i = r_i - \bar{r}$, without the standard-deviation scaling whose bias Liu et al. (2025b) document, and the same scalar is applied to every token in a rollout.

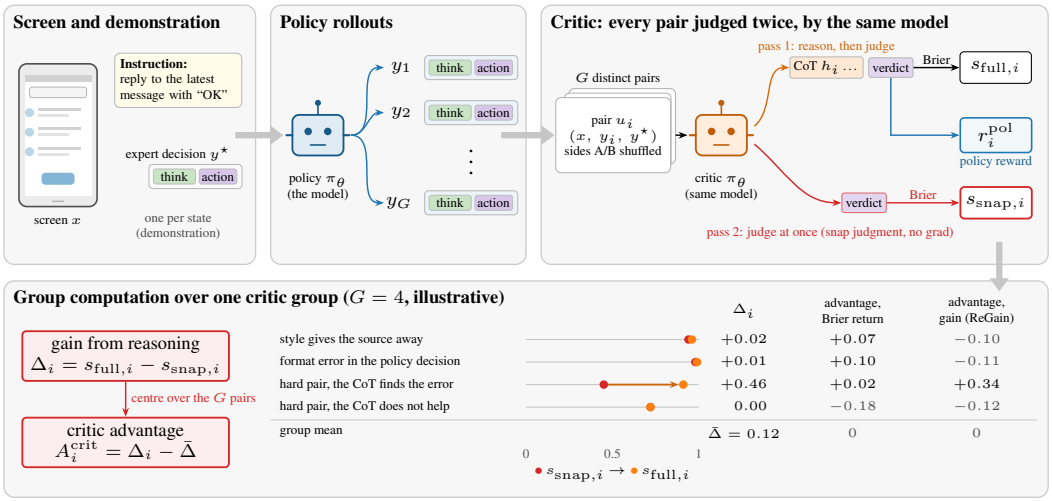

Figure 1: Overview of ReGain. The critic is trained on the gain of its post-CoT verdict over its snap judgment, centred over a group of distinct pairs, and the post-CoT verdict also rewards the policy. In the illustrative group at the bottom, pairs decided at a glance (rows 1–2) earn positive advantage under the Brier return but negative advantage under the gain, and the pair whose CoT found the error (row 3) receives most of the credit only under the gain.

4 METHOD

ReGain keeps the objective, the two roles, and the shared backbone of §3 and credits a critic CoT only for what it adds to the verdict (Figure 1). A detached Brier score on the verdict distribution makes this measurable on a single sample (§4.2), and the reward is its gain over the critic's snap judgment (§4.3).

4.1 TWO ORIGINS OF THE OBSERVED FAILURES

Table 2 (Appendix A) separates the symptoms into two origins. Origin A is reward design. The three-valued return discards the confidence behind each sampled outcome, since every correct commitment earns one, every tie c, and every error zero, and credits the CoT for the verdict it leads to whether or not the reasoning was needed, which gives S1 and S2; the score makes both measurable, and the gain acts on both (§5.4). Origin B is the adversarial objective. The critic is rewarded for classifying source identity, so authorship cues serve it without improving judgments of quality (S3), and demonstrations alone do not identify which source differences matter; neither component changes this objective (Appendix G).

4.2 DETACHED PROPER-SCORING REWARD

Construction. Let (p_A, p_B, p_T) be the critic's three-way verdict distribution and $Z \in \{A, B\}$ the expert-side label. Redistributing tie mass gives a binary source report $\tilde{p}_A = p_A + \frac{1}{2}p_T$ and $\tilde{p}_B = p_B + \frac{1}{2}p_T$, so that $\tilde{p}_{\text{exp}} = \tilde{p}_Z$ and $\tilde{p}_{\text{pol}} = 1 - \tilde{p}_{\text{exp}}$ are the masses on the expert and policy sides. The critic's source score is the Brier score (Gneiting & Raftery, 2007; Shao et al., 2024a), written so that larger is better, and the policy receives the opposing continuous reward in place of the discrete $0/\tau_{\text{pol}}/1$ return of §3:

$$s = 1 - (1 - \tilde{p}_{\text{exp}})^2 \in [0, 1], \qquad r_{\text{pol}} = 1 - (1 - \tilde{p}_{\text{pol}})^2. \tag{1}$$

Both scalars are read from the verdict distribution before the verdict token is sampled and are detached. The Brier-only ablation uses s as the critic reward, ReGain builds the reward from two such scores (§4.3), and Appendix A relates s to the two-class Brier loss.

Proposition 4.1 (Strict properness of the terminal score). *Let $\mathcal{I} = (x, y_A, y_B, h)$ be the information available after a sampled critic CoT h, let $q = \Pr(Z = A \mid \mathcal{I})$, and let $\tilde{p}_A$ be the terminal report.*

Under Eq. (1), $\mathbb{E}[s \mid \mathcal{I}] = 1 - q + 2q\tilde{p}_A - \tilde{p}_A^2$*, which has the unique maximizer* $\tilde{p}_A^\star = q$ *and optimal value* $G(q) = 1 - q(1-q)$.

Unlike RLCR (Damani et al., 2025), where a proper rule rewards a confidence the model emits, the verdict probabilities here are detached, so the verdict position receives zero expected gradient and the score acts only through the CoTs it selects (Appendix A); the proposition characterizes the scoring rule, not the optimizer.

The discrete reward at $c = 0.75$ is a three-point restriction. At the three hard reports the expected score is q for A, $1 - q$ for B, and 0.75 for a tie at every q, so

$$\text{discrete reward with } c = 0.75 \;=\; \text{the Brier rule restricted to } \tilde{p}_A \in \{0, \tfrac{1}{2}, 1\}, \tag{2}$$

which Figure 3 (Appendix A) plots; at the values $c \in \{0.50, 0.55\}$ that Cai et al. (2026) report, the tie lies below the curve. The same coefficient sets the abstention band $q \in [1 - c, c]$ (Appendices E and F), and §5.4 shows that a narrow and a wide band fail in different ways.

4.3 Rewarding the gain from reasoning

Two scores per pair. We score the critic twice on every pair. The first score s_{full} follows a sampled CoT; for the second we append the verdict delimiter directly to the prompt, with the same side assignment, so the model commits without reasoning; we call this report the critic's *snap judgment*, and s_{snap} is its Brier score. Their difference,

$$\Delta_i = s_{\text{full},i} - s_{\text{snap},i} \in [-1, 1], \tag{3}$$

is the gain from reasoning, measuring how much the sampled CoT improved the source report over the snap judgment. It is also the measurement behind S2. Because $s_{\text{snap},i}$ conditions only on the pair, one no-grad forward pass per pair supplies it.

The gain as the critic reward. For one task x the critic group holds one prompt per policy rollout, the pair $(x, y_i, y^\star)$ with an independently randomized side order, so the G prompts differ and each has its own reference $s_{\text{snap},i}$. The gain is the critic reward, centred by the group mean:

$$r_i^{\text{crit}} = \Delta_i, \qquad A_i^{\text{crit}} = \Delta_i - \bar{\Delta} = \big(s_{\text{full},i} - \bar{s}_{\text{full}}\big) - \big(s_{\text{snap},i} - \bar{s}_{\text{snap}}\big), \tag{4}$$

and the critic role is trained on its CoT and verdict tokens with this advantage (Eq. (9)). Under the Brier return alone the advantage is the first bracket, so a CoT on an easy pair earns a positive advantage whether or not the reasoning contributed; the second bracket is the per-pair baseline that a group mean shared across pairs of varying difficulty does not supply.

What the gain measures. The CoT is computation, not evidence. Given the pair, the ideal posterior over Z is fixed, and the critic's reports approximate it with different amounts of computation. The gain is therefore a statement about reports. Let r_0 be the snap judgment of $\tilde{p}_A$ and r_1 the report after a sampled CoT.

Proposition 4.2 (Expected gain equals report movement)**.** *Suppose both reports are calibrated,* $\Pr(Z = A \mid r_0) = r_0$ *and* $\Pr(Z = A \mid r_0, r_1) = r_1$*. Then*

$$\mathbb{E}[\Delta \mid r_0] = \mathrm{Var}(r_1 \mid r_0) \;\le\; r_0(1 - r_0), \tag{5}$$

with equality on the right if and only if $r_1 \in \{0, 1\}$ *almost surely.*

Over pairs the critic initially judges alike, the expected gain is how far reasoning moves the report, capped by the uncertainty the snap judgment leaves. Appendix B gives the proof and the expected gain without calibration, and §5.3 shows that the trained critic is not calibrated.

Pairs decided at a glance. Whatever the calibration, the cap $\Delta_i \le 1 - s_{\text{snap},i}$ withholds credit from pairs the snap judgment already decides correctly, by a format error, an obvious mistake, or style, and after mean centring such a pair receives advantage $-\bar{\Delta}$, negative whenever the rest of the group gained, so positive credit goes to the pairs where reasoning decided the verdict (Figure 1, bottom). Pairs whose report does not move cannot be helped by reasoning, since their visible content carries no source information the critic can extract. No preprocessing could apply this filter, since which pairs are decided at a glance changes as both roles train; RARO's tie is its discrete precursor on the hard end (Appendix B). For style-decided pairs this is the extent of the mitigation, because the snap judgment still uses style and the policy is still scored by a critic that reads it (§5.4).

Why a scored probability is required. A sampled verdict cannot supply this gain, because a difference of two sampled returns is dominated by verdict noise and a discrete return does not record a change of belief that crosses no decision boundary. Reading the distribution before sampling and scoring it with a strictly proper rule removes both problems, and the Brier rule makes the expected gain equal to the variance (Appendix B).

The grouping is necessary. If all G critic rollouts were sampled on one pair, s_{snap} would be constant across the group and

$$\Delta_i - \bar{\Delta} = s_{\mathrm{full},i} - \bar{s}_{\mathrm{full}}, \tag{6}$$

so centring would cancel the reference exactly. Building the critic group from the G policy rollouts of a task is therefore a design constraint that standard RL with verifiable rewards, where a group shares one prompt, does not meet. Because $s_{\mathrm{snap},i}$ does not depend on the sampled CoT, it is an admissible baseline, so the expected update is that of GRPO on the Brier return and the reference changes only which pairs carry each finite-group update (Appendices B and C, Algorithm 1).

5 Experiments

5.1 Setup

Data and model. The demonstration set contains 190k single-step GUI-operation demonstrations from more than 200 applications of a mobile OS; each pairs a screen and an instruction with an expert decision, typically a short CoT, a statement of intent, and an action. The backbone is Qwen3.8-27B (Qwen Team, 2026), a public vision-language model used from its released checkpoint; one set of parameters serves as both policy and critic, with thinking mode on at the `xhigh` reasoning-effort level in training and evaluation. The data-matched comparison holds the number of demonstrations at 20,480. Adversarial runs draw tasks from the full set until they have consumed that many, and SFT trains on a random subset of that size; one SFT run on the full 190k set serves as a reference. Demonstrations supply no trajectory-level outcome.

Benchmarks. *Single-step.* An in-house benchmark of 250 held-out states with one reference action each, in five sub-suites of 50: grounding, instruction following, abnormal states, keyword extraction, and safety (Appendix K). A grounding prediction is correct if its click falls inside the reference element's box, a safety prediction if it refuses, and any other if it matches the reference action type and arguments. We sample four decisions per state at temperature 1 and report the mean per-sample accuracy (pass@1) and the fraction of states with a correct sample (pass@4). Because several actions can be valid and the reference follows the conventions that SFT imitates, pass@1 favours SFT. *Trajectory-level.* MobileGym (Wu et al., 2026a) is a browser-hosted simulator of 28 applications whose structured state lets deterministic rules judge success. An in-house harness maps each decision's action onto MobileGym's action space (Appendix J), and the test set is 100 in-house multi-step tasks over its applications that cover the business scenarios the online stage targets, each with a rule-based success check and a budget of 60 steps; the task set's training split is reserved for the online stage. Each cold-started policy is rolled out with no trajectory-level training, the state in which the online stage would receive it; each task is attempted four times, and the success rate averages the attempts. *General capability.* Eleven public multimodal benchmarks: MMMU (Yue et al., 2024), MMMU-Pro (Yue et al., 2025), MMStar (Chen et al., 2024), MathVista (Lu et al., 2024), MathVision (Wang et al., 2024), DynaMath (Zou et al., 2025), RealWorldQA (xAI, 2024), OCRBench v2 (Fu et al., 2025), SimpleVQA (Cheng et al., 2025), MMLongBench-Doc (Ma et al., 2024), and HallusionBench (Guan et al., 2024). Because their scales differ, we report the mean relative change from the backbone; Appendix M gives every score.

Baselines and budget. We compare the untrained backbone, SFT for one epoch on the 20,480 subset and on the full 190k set, RARO with the discrete reward at $c = 0.75$, this implementation's value, and at $c = 0.55$, the value Cai et al. (2026) report stable, the Brier-only ablation, and ReGain, all from the same checkpoint. Adversarial runs use 128 tasks per step with $G = 4$ policy rollouts each and are evaluated at step 160, when they have consumed 20,480 demonstrations, each seen once; the step is fixed by this budget, and §5.4 follows every run beyond it. Each demonstration costs G policy rollouts, G critic rollouts, and G snap-judgment passes, so the method is demonstration-efficient

rather than compute-efficient. The ReGain run takes 29.6 hours on 64 A800 GPUs to reach step 160 (Appendix J).

Metrics and seeds. The training-dynamics quantities, taken from the training logs and defined in Appendix J, are a validation score on a held-out split of the demonstrations, the tie rate, critic accuracy with a tie counted as one half, the critic advantage variance, and CoT lengths; curves show ten-step trailing means. ReGain and its Brier-only ablation use two random seeds, whose single-step samples are pooled and general-capability scores averaged, with trajectory results and curves from one; every other run uses one seed. Entries are given $\pm$ one standard error. The equilibrium-constrained curriculum of Appendix H is disabled in every run.

5.2 Main results

Table 1 reports the data-matched comparison on all three benchmarks, with the full-data SFT run below the rule as a reference.

Against RARO. At the same 160 steps and the same 20,480 demonstrations, ReGain reaches 70.0 pass@1 against 65.6 for the discrete reward at $c = 0.75$ and 63.3 at $c = 0.55$. The two discrete runs fail in different ways, the first by settling into ties and the second by dropping its reasoning (§5.4). The Brier-only row isolates the continuous score from the gain reward. At 65.9 it matches the discrete reward at $c = 0.75$, whose training it also follows, so the continuous score alone does not explain the improvement; §5.3 examines what the gain adds.

Against SFT at matched demonstrations. SFT on the same 20,480 demonstrations reaches 57.4 pass@1, 12.6 points below ReGain and 3.8 below the untrained backbone, so at this data scale imitating the demonstration tokens makes the agent worse. This carries the finding of Cai et al. (2026), that an adversarial critic extracts more from fixed demonstrations than token-level imitation, to single-step granularity.

The full-data reference. ReGain matches SFT on all 190k demonstrations with one ninth of them, 70.0 against 66.1 pass@1 with standard errors of 2 to 4 points, and reaches the same pass@4, 86.0 against 86.8, where every other data-matched cold start stays at 79.6 to 81.2, so ReGain solves more of the same set on the first sample. The difference is the cost, as full-data SFT loses 14% of the backbone's general capability and ReGain 0.9%.

Trajectory-level transfer. Without trajectory-level training, ReGain succeeds on 32.7% of the MobileGym tasks, against 24.0% to 28.8% for the other cold starts; the backbone, which has not learned the harness's action protocol, succeeds on none. With standard errors of about 4 points the ranking broadly agrees with the single-step benchmark without resolving any single gap: the lead over data-matched SFT narrows from 12.6 points to 5.1, and the lead over full-data SFT is 3.9 on both benchmarks. By difficulty (Appendix L), ReGain leads on the intermediate levels, and no cold start solves more than 15% of the hardest tasks.

General capability. Both SFT runs lose about 14% of the backbone's general capability, falling below it on ten and eleven of the eleven benchmarks, whereas every adversarial run stays within about 2% (Appendix M).

5.3 The gain reward is what helps

The Brier-only row of Table 1 holds the score fixed and removes the snap-judgment reference, so its difference from ReGain is the effect of the gain reward.

The reference is what helps. Replacing the Brier return with the gain Δ raises single-step pass@1 by 4.1 points and trajectory-level success by 4.6 points, neither above 1.4 standard errors of the difference, so the training dynamics carry most of the evidence. The two rewards have the same expected update (§4.3) but reinforce different pairs. Recomputed on the same rollouts at step 160, pairs decided at a glance, 35% of all pairs, receive 40% of the positive credit under the Brier return

Table 1: Data-matched cold-start comparison at step 160 (20,480 demonstrations), with full-data SFT as a reference. Trajectory success is measured on the 100 in-house MobileGym tasks without trajectory-level training, a task that a run did not finish counting as a failure, and general capability is the mean relative change (%) from the backbone over eleven benchmarks (Appendix M).

		Single-step		**Trajectory**	**General capability**
Cold start	Demos	pass@1	pass@4	Success	Mean Δ vs. backbone
Backbone (no cold start)	0	61.2 ± 3.1	79.2 ± 2.9	–	–
SFT	20,480	57.4 ± 4.0	81.2 ± 3.4	27.6 ± 3.7	-14.2
RARO, discrete, $c = 0.75$	20,480	65.6 ± 3.2	79.6 ± 2.9	25.9 ± 3.5	-2.1
RARO, discrete, $c = 0.55$	20,480	63.3 ± 2.8	80.4 ± 2.9	24.0 ± 3.8	$+0.8$
ReGain, Brier return only	20,480	65.9 ± 2.2	79.8 ± 2.0	28.1 ± 3.9	$+0.2$
ReGain	20,480	$\mathbf{70.0} \pm 2.1$	86.0 ± 1.8	$\mathbf{32.7} \pm 3.7$	-0.9
SFT (full-data reference)	190,000	66.1 ± 3.9	$\mathbf{86.8} \pm 2.7$	28.8 ± 3.8	-14.1

and 20% under the gain, and pairs whose CoT overturned the snap judgment, 18% of all pairs, receive 37% and 56% (Appendix P).

The critic is not calibrated. Over the critic rollouts of ReGain, s_{snap} averages 0.586 at the first step and 0.722 at step 160, and s_{full} 0.491 and 0.620. A calibrated report scores at least 0.75 in expectation ($G(q) \geq \frac{3}{4}$ in Eq. (7)), so neither report is calibrated, and the mean gain, about -0.1 at both points, is negative, which Proposition 4.2 excludes for calibrated reports. The method does not rest on it: the expected update equals that of the Brier return, and by the cap $\Delta_i \leq 1 - s_{\text{snap},i}$ a pair the snap judgment already decides correctly, by style included, earns at most the margin it leaves open, whatever the calibration. Nor does the reference collapse, as s_{snap} rises by 0.14 over the first 160 steps, about as much as s_{full} (Appendix C).

5.4 Training dynamics: two failure modes and a delayed shortcut

Figure 2 follows the four adversarial runs past the data-matched checkpoint, to step 561 for ReGain and to steps 240 to 278 for the others.

Two failure modes of the discrete reward. Both discrete runs first fill with ties while their critics stay near chance: over the first 120 steps critic accuracy stays between 0.43 and 0.51 while the tie rate climbs past 0.46. They then part in line with the premium a tie earns over a random commitment, $c - \frac{1}{2}$ (Appendix A). At $c = 0.75$ the premium is 0.25 and the critic keeps tying, on up to two thirds of pairs, with accuracy no higher than 0.51 through step 160: the verdict collapse S1. Its remaining decisive verdicts drift toward the shorter decision, the first sign of S3 (Figure 6). At $c = 0.55$ the premium is 0.05, and the critic leaves the tie by giving up its reasoning: nearly every verdict is emitted with an empty CoT by step 200, its tie rate falls from 0.64 to 0.24, and its accuracy rises to 0.63 while the validation score falls from 0.38 to 0.24. The verdict no longer depends on reasoning, the extreme of S2. Its decisive verdicts show no preference for brevity, so its gain in accuracy fits a degrading policy that is easier to spot at a glance (Appendix Q).

The continuous score alone keeps the tie attractor. The Brier-only run follows the run at $c = 0.75$ and ties even more, on 64% of pairs at step 160 and 75% at step 240, while its validation score peaks at 0.37 at step 80 and falls to 0.23. Eq. (2) predicts the resemblance: the even report earns 0.75 under the Brier score, the same premium of 0.25 over a random commitment as the tie at $c = 0.75$. Under both rewards the policy's CoT lengthens with the tie rate, to four to five times the 107-token mean of the demonstrations, so the lengthening is not imitation (Appendix N).

The gain delays the shortcut. Under the gain reward, a CoT that retreats to the even report on a pair whose snap judgment already favoured the expert earns $\Delta_i < 0$, so the retreat is no longer safe. ReGain ties on 35% of pairs at step 160 and never on more than half, keeps the highest critic advantage variance of the four runs (0.11 at step 160 against 0.05 at $c = 0.75$ and for Brier-only), and never loses its critic CoT. Its validation score rises to 0.42 at step 160 and peaks at 0.44 at step

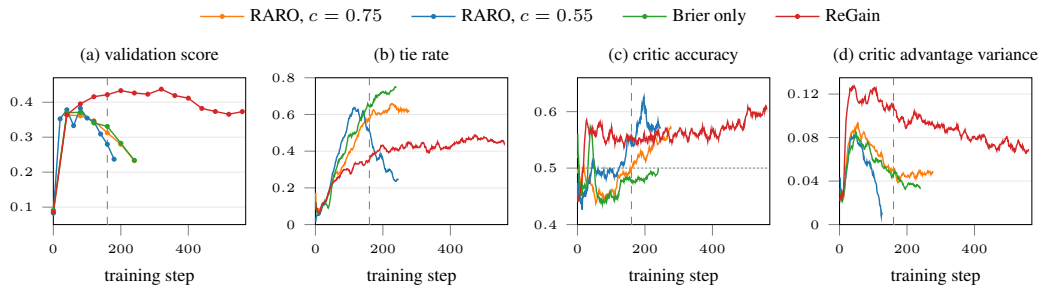

Figure 2: Training dynamics of the four adversarial runs, with the data-matched checkpoint at step 160 dashed and ten-step trailing means in (b) to (d). (a) Validation score. (b) Tie rate. (c) Critic accuracy with a tie counted as $\frac{1}{2}$, chance dotted. (d) Critic advantage variance. Each curve is a single training run.

320; beyond that, critic accuracy climbs from 0.57 to 0.60 while the validation score falls to 0.37, and its decisive verdicts come to pick the shorter decision three times in four, with accuracy up to 0.18 higher when the expert's is the shorter (Figure 6): the authorship shortcut S3, arriving later than under the discrete reward. Withdrawing credit from style-decided pairs (§4.3) delays the shortcut but does not prevent it, because the snap judgment still uses style and the policy is still scored by a critic that reads it.

The reported checkpoint. Step 160 is fixed by the data budget, not selected on validation. It falls after the validation peaks of the other three runs, at steps 40 to 80, and before that of ReGain; even at their peaks, 0.36 to 0.38, those runs stay below ReGain at step 160, 0.42. The recipe still needs a stopping rule.

6 Conclusion and Limitations

Adapting adversarial reasoning to single-step GUI decisions exposes verdict collapse, weak dependence on the critic CoT, and authorship cues, all arising from pairs the critic decides at a glance. ReGain rewards the critic CoT for its gain over its snap judgment, so such pairs earn at most the margin the snap judgment leaves open and credit moves to pairs where reasoning worked; this needs a proper score on the verdict distribution and a critic group spanning distinct pairs, which standard RLVR does not supply. On the same rollouts, the gain halves the positive credit that pairs decided at a glance receive under the Brier return. On a 27B GUI agent it matches SFT trained on nine times as many demonstrations without its 14% loss of general capability. The authorship shortcut is delayed rather than fixed; whether outcome feedback in a later online stage separates quality from source identity is left open.

Limitations. *Authorship*: demonstrations do not identify which source differences are quality-relevant (Skalse et al., 2023; Cao et al., 2021), so the validation score of ReGain declines after step 320 while its critic comes to favour the shorter decision, and the recipe needs a stopping rule; brevity is the one cue we measure, and the action-only diagnostic of Appendix G is untested. *Handoff*: general-capability retention and pass@4, which prior work links to RL headroom (Li et al., 2025; Jin et al., 2025), and trajectory-level success are proxies, not the handoff itself, since no online RL stage is run. *Evaluation*: the single-step benchmark scores one reference action and favours SFT; the trajectory-level set has no such bias but only 100 tasks, and the gains over RARO and the Brier-only ablation are within 1.4 standard errors. *Cost and scope*: each demonstration costs G policy rollouts, G critic rollouts, and G snap-judgment passes, and the study uses one 27B backbone, in-house test sets, and two seeds for ReGain and its ablation but one for every other run. *Reference and calibration*: Δ is self-generated and can respond to surface form as well as to valid inference (Appendix C), although the snap score rises over the first 160 steps rather than degrading, and the critic is not calibrated, so Proposition 4.2 explains the design rather than describing the trained critic (§5.3).

AI USE STATEMENT

We used generative AI tools (large language model assistants) for the following tasks. *Writing:* drafting and editing the text of every section, proposing the title and keywords, and suggesting the structure of the paper. *Literature:* searching for related work and formatting references; every citation was checked against the cited source by the authors. *Methodology and experiments:* discussing and giving feedback on the experimental design, including which baselines, ablations, and diagnostic curves to report. *Mathematical claims:* restating Proposition 4.2 in terms of calibrated reports and deriving the decomposition in Appendix B; the authors re-derived and verified every proof independently. *Implementation:* writing and editing the training and evaluation code; the code was reviewed and tested by the authors. We did not use generative AI to generate data or to produce experimental results. All AI-assisted content was reviewed by the authors, who take full responsibility for the final text, claims, and artifacts of this work.

REPRODUCIBILITY STATEMENT

All runs start from the public Qwen3.8-27B checkpoint (Qwen Team, 2026). Algorithm 1 gives the full training step, Appendix J the hyperparameters and the trajectory-level harness, and §5.1 the compute and the seeds. The single-step benchmark and the trajectory-level task set are described in §5.1 and Appendices K and L; the trajectory-level simulator is public (Wu et al., 2026a). The demonstrations and both test sets are internal and will not be released. The training code will be released as a recipe for verl (Sheng et al., 2025).

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Hao Bai, Yifei Zhou, Mert Cemri, Jiayi Pan, Alane Suhr, Sergey Levine, and Aviral Kumar. DigiRL: Training in-the-wild device-control agents with autonomous reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

David Berthelot, Thomas Schumm, and Luke Metz. BEGAN: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

Alex Bie, Gautam Kamath, and Guojun Zhang. Private GANs, revisited. *Transactions on Machine Learning Research*, 2023.

Locke Cai, Max Ryabinin, and Ivan Provilkov. Escaping the verifier: Learning to reason via demonstrations. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. arXiv:2511.21667.

Haoyang Cao, Samuel N. Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, Yutao Zeng, Zhoufutu Wen, Ke Jin, Baorui Wang, Weixiao Zhou, Yunhong Lu, Hangyuan Ji, Tongliang Li, Wenhao Huang, and Zhoujun Li. SimpleVQA: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4637–4646, 2025.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training LMs to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*, 2025.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Heshan Fernando, Han Shen, Parikshit Ram, Yi Zhou, Horst Samulowitz, Nathalie Baracaldo, and Tianyi Chen. Understanding forgetting in LLM supervised fine-tuning and preference learning – a convex optimization perspective. *arXiv preprint arXiv:2410.15483*, 2024.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, 2025.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. *arXiv preprint arXiv:2210.10760*, 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

J. Fernando Hernandez-Garcia, Tomas Figliolia, and Beren Millidge. Can scale save us from plasticity loss in large language models? *arXiv preprint arXiv:2606.24752*, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, 2016.

Baixiang Huang, Canyu Chen, and Kai Shu. Authorship attribution in the era of LLMs: Problems, methodologies, and challenges. *arXiv preprint arXiv:2408.08946*, 2024.

R. Jia, Y. Yang, Y. Gai, et al. Writing-Zero: Bridge the gap between non-verifiable tasks and verifiable rewards. *arXiv preprint arXiv:2506.00103*, 2025.

Hangzhan Jin, Sitao Luan, Sicheng Lyu, Guillaume Rabusseau, Reihaneh Rabbany, Doina Precup, and Mohammad Hamdaqa. RL fine-tuning heals OOD forgetting in SFT. *arXiv preprint arXiv:2509.12235*, 2025.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems*, 2020.

Tharindu Kumarage and Huan Liu. Neural authorship attribution: Stylometric analysis on large language models. *arXiv preprint arXiv:2308.07305*, 2023.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving diversity in supervised fine-tuning of large language models. In *International Conference on Learning Representations*, 2025. arXiv:2408.16673.

Qihao Liu, Luoxin Ye, Wufei Ma, Yu-Cheng Chou, and Alan Yuille. Generative adversarial reasoner: Enhancing LLM reasoning with adversarial reinforcement learning. *arXiv preprint arXiv:2512.16917*, 2025a.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-zero-like training: A critical perspective. In *Conference on Language Modeling*, 2025b. arXiv:2503.20783.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, et al. UI-R1: Enhancing efficient action prediction of GUI agents by reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2503.21620.

Run Luo, Lu Wang, Wanwei He, Longze Chen, Jiaming Li, and Xiaobo Xia. GUI-R1: A generalist R1-style vision-language action model for GUI agents. *arXiv preprint arXiv:2504.10458*, 2025.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. General-Reasoner: Advancing LLM reasoning across all domains. In *NeurIPS*, 2025.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. MMLongBench-Doc: Benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, 2024.

Mitsuhiko Nakamoto, Simon Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. Cal-QL: Calibrated offline RL pre-training for efficient online fine-tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.

Dang Nguyen, Jian Chen, Yu Wang, Gang Wu, Namyong Park, Zhengmian Hu, Hanjia Lyu, Junda Wu, Ryan Aponte, Yu Xia, Xintong Li, Jing Shi, Hongjie Chen, Viet Dac Lai, Zhouhang Xie, Sungchul Kim, Ruiyi Zhang, Tong Yu, Mehrab Tanjim, Nesreen K. Ahmed, Puneet Mathur, Seunghyun Yoon, Lina Yao, Branislav Kveton, Jihyung Kil, Thien Huu Nguyen, Trung Bui, Tianyi Zhou, Ryan A. Rossi, and Franck Dernoncourt. GUI agents: A survey. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22522–22538, 2025.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Wenyi Zhao, Yu Yang, Xinyue Yang, Jiadai Sun, Shuntian Yao, Tianjie Zhang, Wei Xu, Jie Tang, and Yuxiao Dong. WebRL: Training LLM web agents via self-evolving online curriculum reinforcement learning. In *International Conference on Learning Representations*, 2025. arXiv:2411.02337.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. UI-TARS: Pioneering automated GUI interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. `https://qwen.ai/blog?id=qwen3.8`, 2026. Model card: `https://huggingface.co/Qwen/Qwen3.8-27B`.

Chenze Shao, Fandong Meng, Yijin Liu, and Jie Zhou. Language generation with strictly proper scoring rules. In *Proceedings of the 41st International Conference on Machine Learning*, 2024a. arXiv:2405.18906.

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

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*, 2025.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, 2024.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *arXiv preprint arXiv:2209.13085*, 2022.

Joar Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Hao Sun and Mihaela van der Schaar. Inverse-RLignment: Large language model alignment from demonstrations through inverse reinforcement learning. *arXiv preprint arXiv:2405.15624*, 2024.

Y. Tang, S. Wang, L. Madaan, and R. Munos. Beyond verifiable rewards: Scaling reinforcement learning in language models to unverifiable data. In *NeurIPS*, 2025.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, 2024.

Dingbang Wu, Rui Hao, Haiyang Wang, Shuzhe Wu, Han Xiao, Zhenghong Li, Bojiang Zhou, Zheng Ju, Zichen Liu, Lue Fan, and Zhaoxiang Zhang. MobileGym: A verifiable and highly parallel simulation platform for mobile GUI agent research. *arXiv preprint arXiv:2605.26114*, 2026a.

Jianyu Wu, Yizhou Wang, Encheng Su, Chen Tang, and Shixiang Tang. DAPD: Dual-anchored policy distillation. *arXiv preprint arXiv:2608.01735*, 2026b.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. OS-ATLAS: A foundation action model for generalist GUI agents. In *International Conference on Learning Representations*, 2025. arXiv:2410.23218.

xAI. Grok-1.5 vision preview. `https://x.ai/news/grok-1.5v`, 2024. Introduces the RealWorldQA benchmark.

Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, and Caiming Xiong. Aguvis: Unified pure vision agents for autonomous GUI interaction. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. arXiv:2412.04454.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. *arXiv preprint arXiv:2604.03128*, 2026.

Q. Yu, Z. Zhang, R. Zhu, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *NeurIPS*, 2025.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, 2022.

Chaoyun Zhang, Shilin He, Jiaxu Qian, Bowen Li, Liqun Li, Si Qin, Yu Kang, Minghua Ma, Guyue Liu, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. Large language model-brained GUI agents: A survey. *Transactions on Machine Learning Research*, 2025a. arXiv:2411.18279.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *International Conference on Learning Representations*, 2025b. arXiv:2408.15240.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

C. Zheng, S. Liu, M. Li, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. Reinforcing general reasoning without verifiers. In *ICLR*, 2026.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *arXiv preprint arXiv:2605.11182*, 2026.

Konrad Zolna, Scott Reed, Alexander Novikov, Sergio Gómez Colmenarejo, David Budden, Serkan Cabi, Misha Denil, Nando de Freitas, and Ziyu Wang. Task-relevant adversarial imitation learning. In *Proceedings of the 2020 Conference on Robot Learning*, volume 155, pp. 247–263, 2021.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. DynaMath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, 2025.

A Full Derivation of the Proper Scoring Reward

Proof of Proposition 4.1. Given the information $\mathcal{I}$, $Z = A$ with probability q. With $\tilde{p}_B = 1 - \tilde{p}_A$, the score is $1 - (1 - \tilde{p}_A)^2$ if $Z = A$ and $1 - \tilde{p}_A^2$ if $Z = B$, so

$$\mathbb{E}[s \mid \mathcal{I}] = q\big(1 - (1 - \tilde{p}_A)^2\big) + (1 - q)\big(1 - \tilde{p}_A^2\big) = 1 - q + 2q\tilde{p}_A - \tilde{p}_A^2 = G(q) - (\tilde{p}_A - q)^2, \quad (7)$$

with $G(q) = 1 - q(1 - q)$. The last form shows that the expected score is a concave quadratic in the report with its unique maximum at the honest report $\tilde{p}_A = q$, which is strict properness; it is also the form used in Appendix B. As a numerical check at $q = 0.8$: the honest report earns $0.8 \cdot 0.96 + 0.2 \cdot 0.36 = 0.840$, the overconfident report $\tilde{p}_A = 1$ earns 0.800, and the hedge $\tilde{p}_A = \frac{1}{2}$ earns 0.750.

Table 2: The two origins of the observed failures. S1 and S2 arise from verdict-only reward design; S3 follows from the source-classification objective and admits no fix from demonstrations alone.

	Origin A (reward design)	**Origin B** (objective)
Statement	Three-valued rewards ignore confidence; CoT credit ignores what the CoT added	The critic is trained to discriminate source identity
Source	This implementation of RARO	Adversarial imitation from demonstrations
Symptoms	S1, S2	S3
Status	Targeted by the score and the gain	Not identified from demonstrations alone
Response	§4.2, §4.3	§5.4: delayed, not fixed

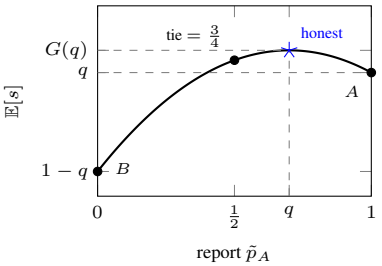

Figure 3: Expected Brier score of a report $\tilde{p}_A$ at posterior $q = 0.7$. The hard reports of the discrete reward with $c = 0.75$ lie on the curve (B earns $1-q$, A earns q, a tie $\frac{3}{4}$), so the tie beats the correct commitment, and the honest report $G(q)$ beats all three.

Relation to the two-class Brier loss. The Brier score is classically defined as a loss over all classes, lower being better. For the binary report $(\tilde{p}_A, \tilde{p}_B)$ with $\tilde{p}_B = 1 - \tilde{p}_A$ and outcome $Z = A$, this loss is $(1-\tilde{p}_A)^2 + (0-\tilde{p}_B)^2 = 2(1-\tilde{p}_A)^2$, and symmetrically $2(1-\tilde{p}_B)^2$ for $Z = B$. The score of Eq. (1), $s = 1-(1-\tilde{p}_{\text{exp}})^2$, keeps only the expert-side term and flips the sign so that larger is better; it is therefore an affine transform of the two-class loss, $s = 1 - \frac{1}{2}\,\text{Brier}$, with the same optimizer and the same properness.

Zero expected gradient at the verdict position. Let v be the sampled verdict token and h the preceding CoT. The score is computed from the verdict distribution before v is sampled and detached before the update, so conditional on h the advantage $A(h)$ is independent of v and

$$\mathbb{E}_{v|h}[A(h)\nabla_\theta \log \pi_\theta(v \mid h)] = A(h)\nabla_\theta \sum_v \pi_\theta(v \mid h) = 0. \tag{8}$$

The non-zero expected update acts on h alone, through $\mathbb{E}_h[A(h)\nabla_\theta \log \pi_\theta(h \mid u)]$ for critic prompt u. Including the verdict token in the token sum of Eq. (9) therefore adds variance but no expected signal; masking it is a variance reduction.

What the tie redistribution does. The score depends on the three-way report only through $\tilde{p}_A = p_A + \frac{1}{2}p_T$, so every report on the line $p_A + \frac{1}{2}p_T = q$ is optimal: $(0.6, 0.4, 0)$, $(0.5, 0.3, 0.2)$, and $(0.3, 0.1, 0.6)$ receive identical scores. The rule is strictly proper for the binary source report and places no constraint on p_T given $\tilde{p}_A$; this is the sense in which §4.2 says the score calibrates the binary report and not the three-way distribution. The redistribution demotes the tie from a third action to the midpoint of the belief axis. At $q = \frac{1}{2}$ the hedge $\tilde{p}_A = \frac{1}{2}$ earns 0.75 with certainty, whereas committing to a side at random earns 1 or 0 with equal probability, 0.5 in expectation: by concavity of Eq. (7), the score prices overconfidence, and the difference $G(q) - 0.75 = (q - \frac{1}{2})^2$ is exactly the premium for knowing more than nothing.

Why the redistribution is needed. Two alternatives fail. If the tie mass is ignored and the raw side probability is scored, $s' = 1-(1-p_Z)^2$, then $\mathbb{E}[s'] = q\,(1-(1-p_A)^2) + (1-q)(1-(1-p_B)^2)$ increases in both p_A and p_B, so the optimum sets $p_T = 0$: the score is proper for the side probabilities but leaves the tie no role. If instead the tie is treated as a third label, the one-sided quadratic score is not proper. For a belief (q_A, q_B, q_T) over three labels, maximizing $\sum_k q_k(2p_k - p_k^2)$ subject to $\sum_k p_k = 1$ gives $p_k = 1 - \lambda/(2q_k)$ with $\lambda = 4/\sum_k q_k^{-1}$; at $(0.5, 0.3, 0.2)$ the optimal report is $(0.613, 0.355, 0.032)$, which pushes low-probability labels toward zero. The one-sided score is proper only for two labels, where it is affine to the full Brier loss (above); the redistribution is what reduces the three-way report to that case.

The discrete reward as a three-point restriction. Under the discrete reward the critic emits a hard verdict, and its expected reward is q for A, $1-q$ for B, and c for a tie. Eq. (7) at $\tilde{p}_A \in \{1, 0, \frac{1}{2}\}$

gives q, $1-q$, and $\frac{3}{4}$, so for $c = \frac{3}{4}$ the two schemes agree in expectation at the three reports the discrete scheme can express, which is Eq. (2). The realized rewards differ: the discrete reward is a sampled $0/1$ outcome, the Brier score a deterministic function of the report and Z. The optimal hard report is $\arg\max\{q, 1-q, c\}$, so a tie is chosen for $q \in (1-c, c)$, a band of width $2c-1$: $c < \frac{1}{2}$ forces a side, $c \geq 1$ forces a tie, and $c = \frac{3}{4}$ covers half of the belief axis. Inside the band the discrete optimum earns c while the honest continuous report earns $G(q)$, so at $c = \frac{3}{4}$ the reward forgone is $(q - \frac{1}{2})^2 \leq \frac{1}{16}$; what is lost is not reward but information, since $q = 0.55$ and $q = 0.70$ produce the same verdict and the same reward. Tuning c moves the band; only a larger report space can carry the belief. Appendices E and F discuss how adversarial training pushes q into the band.

Scope. Two limits follow. Properness holds for the binary source event: the redistribution merges "the two decisions are equivalent" with "the critic cannot tell", and the score cannot separate them. And Proposition 4.1 characterizes the scoring rule, not the optimizer: the detached score never trains the report directly, so calibration of the reports is an empirical property (§5.3).

B Properties and Costs of the Gain Reward

This appendix collects properties of the gain reward of §4.3 that the main text states without discussion.

Critic objective. With the advantage of Eq. (4), the critic role minimizes

$$\mathcal{L}(\theta) = -\frac{1}{G}\sum_{i=1}^{G} \mathrm{detach}(A_i^{\mathrm{crit}}) \sum_t \log \pi_\theta(z_{i,t} \mid z_{i,<t}, u_i), \tag{9}$$

where u_i is the critic prompt for pair i and z_i the critic's CoT and verdict.

Proof of Proposition 4.2. By Proposition 4.1 and calibration, $\mathbb{E}[s_{\mathrm{full}} \mid r_0, r_1] = G(r_1)$ and $\mathbb{E}[s_{\mathrm{snap}} \mid r_0] = G(r_0)$. The tower property gives $\mathbb{E}[r_1 \mid r_0] = \Pr(Z = A \mid r_0) = r_0$, so $\mathbb{E}[G(r_1) \mid r_0] - G(r_0) = \mathbb{E}[r_1^2 \mid r_0] - r_0^2 = \mathrm{Var}(r_1 \mid r_0)$, and $r_1^2 \leq r_1$ gives the bound $r_0(1-r_0)$.

Why a sampled verdict is not enough. A difference of two sampled returns is dominated by verdict noise, and the expected discrete reward of the best hard report, $\max\{q, 1-q, c\}$, is flat on the tie band and linear elsewhere, so a movement of the critic's belief that crosses no decision boundary is never recorded. Reading the distribution before sampling removes the first problem and a strictly proper score the second.

RARO's tie as a discrete filter. A pair the critic cannot separate receives a fixed c under the discrete reward instead of a sampled $0/1$ verdict, which keeps its noise from dominating the update, and Eq. (2) identifies $c = 0.75$ with the Brier score of an uninformed report. The tie therefore acts as a filter on the hard end of the difficulty range, through one fixed value. The gain reward filters both ends, continuously, against a reference specific to each pair: pairs that reasoning cannot move and pairs that need no reasoning both receive $\Delta_i \approx 0$ and, after centring, an advantage of $-\bar{\Delta}$.

Invariances. Subtracting $s_{\mathrm{snap},i}$, a quantity independent of the report, does not move the argmax of Proposition 4.1: the optimal report after a CoT is still $\tilde{p}_A^\star = q$. What changes is the credit assigned to a computation, not the target of the terminal score. Because the reference is a valid baseline, the expected update of Eq. (9) coincides with that of GRPO on the Brier return, before ratio clipping and the KL term; the gain reward changes the variance of the update and its finite-sample allocation across pairs, which is what a curriculum does and what a baseline is permitted to do.

What mean centring does to accurate CoTs. Mean centring guarantees that part of every group receives negative advantage, and under Eq. (4) that part includes accurate CoTs whose gain is below the group average: on a pair the model already classifies at zero thought, a correct CoT is discouraged relative to CoTs that changed a verdict. This is intended, since the reward is for reasoning that adds information, but it means the critic is not trained to reproduce reasoning it does not need.

Expected gain without calibration. Let $q_0 = \Pr(Z = A \mid r_0)$ and $q_1 = \Pr(Z = A \mid r_0, r_1)$ be the calibrated values of the two reports of Proposition 4.2. By Proposition 4.1 the expected score of a report r when the calibrated value is q is $G(q) - (r - q)^2$, so

$$\mathbb{E}[\Delta \mid r_0] = \mathrm{Var}(q_1 \mid r_0) \; - \; \mathbb{E}\big[(r_1 - q_1)^2 \mid r_0\big] \; + \; (r_0 - q_0)^2. \tag{10}$$

The first term is Proposition 4.2. The second says that miscalibration after the CoT lowers the gain: a CoT cannot earn credit by inflating confidence without information, because the terminal score is proper. The third says that miscalibration of the snap judgment raises the gain, which is the channel examined in Appendix C.

Why any strictly proper score would do. Proper scoring rules are exactly those whose expected value under honest reporting is a convex function of the reported probability (Gneiting & Raftery, 2007), so for any strictly proper rule the tower property gives a non-negative expected gain that is positive whenever reasoning moves the report. The Brier rule is the choice for which the gain equals the variance of the report movement exactly (Proposition 4.2) and for which the discrete reward is a three-point restriction; other rules would change the weighting of report movements, not the existence of the curriculum.

Cost. The reference costs one no-grad forward pass per pair, on top of the G policy and G critic rollouts an adversarial step already generates. Both roles use mean-centred GRPO advantages, all scores are detached before the update, and Appendix I gives the full step.

C Can the Critic Game Its Own Baseline?

Equation (4) makes the critic reward depend on s_{snap}, a quantity the same parameters produce. This appendix records the resulting incentive question.

The channel. Since $r_i^{\mathrm{crit}} = s_{\mathrm{full},i} - s_{\mathrm{snap},i}$, lowering $s_{\mathrm{snap},i}$ raises the reward of the rollout on pair i. A model that degraded its own snap judgment would therefore inflate its credit. Nothing in Eq. (9) pays for a good reference.

Why the direct path is blocked. s_{snap} is computed under no-grad and detached, so it contributes no gradient term. The update at step t treats it as a constant, and a parameter change that lowers s_{snap} receives no immediate credit for doing so. Degradation could only arise as a slow selection effect across steps, in which configurations with weaker snap judgments survive because they generate larger rewards.

Why uniform degradation cancels. Group centring removes any shift shared by the group: lowering every $s_{\mathrm{snap},i}$ in a group by the same δ leaves $\Delta_i - \bar{\Delta}$ unchanged. Degradation could pay only if it were selective across pairs, lowering the reference on some pairs relative to others in the same group, and only through the slow selection effect above. The concern is therefore a pair-selective drift, not a uniform collapse of the snap judgment.

What we measure. The account above is an argument, not a guarantee: s_{snap} and s_{full} share parameters, so we cannot rule out a drift that lowers both. We therefore report s_{snap} and s_{full} at the first step and at step 160 (§5.3). A run in which s_{snap} declines while s_{full} stays flat, so that Δ grows without any gain in source accuracy, is the signature of reference degradation, and we treat it as a falsifier for the design (§5.3). Anchoring s_{snap} to a frozen copy of the initial model would remove the channel entirely at the cost of an extra set of weights; we do not use it and monitor s_{snap} instead.

D Routing on the Reference-Match Check

A possible alternative is to route exact action matches to a reference-match check. Under the assumption that each recorded expert action is valid, a match gives a high-precision positive signal without applying token-level cross-entropy to the reasoning. A mismatch remains inconclusive because multiple actions may be valid. A GRPO group in which every rollout receives the same

match reward has zero advantage and contributes no gradient, but selective routing can still change the distribution of states handled by the critic. This possible selection effect would need to be measured. The reported runs consult the critic on all pairs.

E Sensitivity of the Tie Rate to the Tie Coefficient

The optimal discrete report under Eq. (2) is $\arg\max\{q, 1-q, c\}$, so tie is selected exactly on $q \in [1-c, c]$. The band is empty for $c < \frac{1}{2}$ and has width $2c-1$ otherwise. Let the continuous part of the source-posterior distribution have density f. Then

$$P_{\text{tie}}(c) = \int_{1-c}^{c} f(q)\, dq, \qquad \frac{dP_{\text{tie}}}{dc} = f(c) + f(1-c). \tag{11}$$

A large density near $1/2$ therefore makes the tie rate respond steeply to c: under a diffuse posterior distribution, moving c from 0.5 to 0.75 changes the tie rate little, whereas under mass concentrated near $1/2$ the same move can carry most of the distribution into the abstention band. Which regime holds is an empirical property of the policy–expert pair distribution, so Eq. (11) is a testable account of S1 rather than a proof that adversarial training must produce a step. Such concentration is plausible as policy and expert content distributions approach one another (Li et al., 2026), but authorship cues can keep the posterior away from $1/2$.

Why single-step granularity may matter. Exact action matches are more common for decisions at the same GUI state than for two complete long-form answers, and a match removes one source of content-level evidence. This may enrich the source posterior near $1/2$, but it does not force $q = 1/2$: reasoning, intent, and coordinate choices can still differ. The action match rate is therefore a stratification variable, not an estimate of a point mass, and a comparison across model sizes and task granularities is needed before attributing the difference from Cai et al. (2026) to either factor.

F On the Reported Tie Coefficients

For the critic, $c = 0.75$ matches the Brier score at report $1/2$ under our chosen normalization, but it also creates an abstention band of width 0.5. Taking $c \to \frac{1}{2}^{+}$ narrows the band while changing the value assigned to the same report. One scalar therefore controls both the tie payoff and the set of source posteriors mapped to tie. This trade-off provides one possible explanation for reported values $\tau_{\text{crit}} \in \{0.50, 0.55\}$; it is not evidence that those values were selected for this reason.

For the policy, group centring can reduce sensitivity to τ_{pol}. If a fraction p of a group receives the tie reward, then

$$\tau_{\text{pol}} - \bar{r} = (1-p)(\tau_{\text{pol}} - \bar{r}_{\neg\text{tie}}). \tag{12}$$

The tie advantage is therefore damped by the non-tie fraction. Changes in the sampled group modify this relation, so downstream insensitivity is an empirical question rather than a guarantee.

G An Action-Only Diagnostic for the Authorship Confound

Structured decisions separate free-form rationale fields from an action field with a restricted grammar. Let the observation window $\rho \in [0, 1]$ control how much free-form decision content the critic sees: at $\rho = 1$ the full decision, at $\rho = 0$ only the canonical action call with its structured name and arguments. An action is *open-ended* when an argument itself contains free-form text. Setting $\rho = 0$ removes the rationale and intent fields and their surface-form cues from the critic's input. Let $\text{acc}_{\rho=0}$ be critic source accuracy on this action-only view. It measures the discriminative information that survives removal of the free-form fields and needs no learned invariance model; the gap $a_{\text{full}} - \text{acc}_{\rho=0}$ is a proxy for the critic's sensitivity to those fields, not an additive style share, since accuracy differences absorb feature interactions.

Limits. The view is not style-invariant: coordinate precision, choices among equivalent actions, and open-ended arguments containing free-form text can still reveal source identity, so we do not assume chance-level accuracy at $\rho = 0$. Several actions may be valid at a state, and a critic restricted

to actions cannot evaluate the preceding reasoning. Because both views use the same source-labelled demonstrations, the diagnostic resolves neither reward non-identifiability nor the absence of an outcome verifier; it bounds the problem rather than removing it.

Planned measurements. Three uses of the view follow from the main text. First, tracking a_{full} and $\text{acc}_{\rho=0}$ over training tests the reading of the plateau in §5.4: if a_{full} keeps rising past step 160 while $\text{acc}_{\rho=0}$ stays flat, the critic's remaining gains come from the free-form fields, and the widening gap is a candidate stopping signal. Second, the correlation of s_{snap} with $\text{acc}_{\rho=0}$ across pairs tests the claim of §4.3 that pairs solvable at zero thought are enriched for source cues, together with the share of critic gradient the gain reward allocates to the top $\text{acc}_{\rho=0}$ quartile relative to the Brier return. Third, $\text{acc}_{\rho=0}$ stratified by structured and open-ended actions bounds how much source information the action field itself carries.

H Equilibrium-Constrained Curriculum (Not Enabled)

This appendix records a curriculum-control heuristic that we implemented but did not enable in any reported run. It limits critic–policy imbalance and does not remove authorship confounding, so it is not claimed as a contribution. We describe it because the action-only view of Appendix G was originally designed to anchor it, and because the balance problem it targets is real: source cues can let the critic improve faster than the policy, in which case policy rewards lose useful variation and learning stalls, an imbalance also reported for discriminators in vision-based imitation (Zolna et al., 2021).

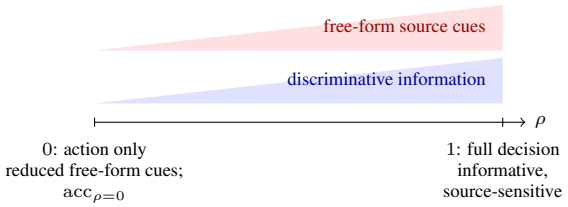

Figure 4: Increasing ρ adds task information and free-form source cues. The action-only endpoint removes rationale style but remains source-sensitive (Appendix G).

Controlling the relative strength of adversarial players is standard in GAN training (Berthelot et al., 2017; Heusel et al., 2017; Bie et al., 2023). We act on three parts of the training schedule.

Observation window. The masking ratio ρ controls how much free-form content the critic sees. Lowering ρ removes some rationale-level source cues but also removes task information. The action-only endpoint remains source-sensitive on open-ended actions and through structured behavioral choices.

Reference-anchored critic-loss gating. Let a_{full} be source-classification accuracy under the full view and $\text{acc}_{\rho=0}$ the corresponding accuracy under the action-only view. Both are estimated periodically on the same fixed, held-out set of source-labelled pairs; the evaluation period is a reported hyperparameter. We include critic-role updates only while

$$a_{\text{full}} \leq g(\text{acc}_{\rho=0}), \tag{13}$$

where g is a monotone mapping specified with the experimental hyperparameters. This anchors the gate to an observed reference but does not make it parameter-free. Since the policy and critic share model parameters, closing the gate removes the direct critic loss; policy-role updates can still change critic behavior. We therefore avoid describing the operation as freezing the critic itself.

Policy-side ladder. If the policy cannot sample useful behavior, critic-loss gating alone can deadlock. We allow two escalations: a larger sampling budget (L0) and prefix scaffolding (L1). The source and schedule of prefixes must be reported because scaffolding can introduce its own imitation and style effects. We do not use stronger token-level supervision in this stage.

I Overall Algorithm

Let G be the group size; the policy group of G decisions induces the critic group of G pairs. As in RARO, half of the critic groups are replayed: a replayed group pairs the G stored rollouts of an earlier task with that task's expert decision, receives snap judgments from the current model, and is centred in the same way, so the grouping of §4.3 holds for it. The action-only diagnostics of Appendix G are not computed in the reported runs, and the gate of Appendix H is disabled throughout.

Algorithm 1 ReGain training step.

Sample policy rollouts $\{y_i\}_{i=1}^{G}$ for task x
for each i: pair y_i with $y^\star$, randomize sides, and form the critic prompt u_i; for a replayed group, take $\{y_i\}$ and $y^\star$ from the buffer
with no gradient **do** // once per pair, same side assignment as u_i
  $s_{\mathrm{snap},i} \leftarrow$ Brier score of the snap judgment, obtained by appending the verdict delimiter to u_i
Sample one critic rollout z_i on each u_i and read pre-sampling $(p_A, p_B, p_T)_i$ at the verdict position
$\tilde{p}_{\mathrm{exp},i},\ \tilde{p}_{\mathrm{pol},i} \leftarrow$ redistributed verdict mass on the expert and policy sides
$r_i^{\mathrm{pol}} \leftarrow \mathrm{detach}\big(1-(1-\tilde{p}_{\mathrm{pol},i})^2\big)$; $A_i^{\mathrm{pol}} \leftarrow r_i^{\mathrm{pol}} - \bar{r}^{\mathrm{pol}}$
$s_{\mathrm{full},i} \leftarrow \mathrm{detach}\big(1-(1-\tilde{p}_{\mathrm{exp},i})^2\big)$; $\Delta_i \leftarrow s_{\mathrm{full},i} - s_{\mathrm{snap},i}$
$A_i^{\mathrm{crit}} \leftarrow \Delta_i - \bar{\Delta}$ // the G pairs differ, so s_{snap} does not cancel; see Eq. (6)
Accumulate the policy-role loss with $\{A_i^{\mathrm{pol}}\}$ on $\{y_i\}$ and the critic-role loss with $\{A_i^{\mathrm{crit}}\}$ on $\{z_i\}$
Update the shared model with the accumulated role losses

J Experimental Details and Hyperparameters

Trajectory-level harness. At every step the policy writes its reasoning in a think block and then one GUI action as a function call, `<think>...</think>` followed by `<action>name(arguments)</action>`. The in-house harness parses the action and maps it onto the action space of MobileGym (Wu et al., 2026a), which covers taps, double taps, long presses, text entry, swipes, drags, the back, home, recent-apps, and enter keys, waiting, launching an application, submitting an answer, and ending the task. Each action of the policy becomes one MobileGym action or a short sequence of them, actions without a counterpart in the simulator are handled inside the harness, and coordinates pass through on a normalized 0–1000 frame. The 100 test tasks are defined over MobileGym's 28 applications.

Training hyperparameters. Table 3 lists the settings of the adversarial runs and of SFT. All runs start from the released Qwen3.8-27B checkpoint with thinking mode on at the `xhigh` reasoning-effort level.

Evaluation settings. Single-step: four samples per state at temperature 1, at most 16,384 generated tokens. Validation: every 20 to 40 steps, greedy decoding with one sample per state. Trajectory level: a budget of 60 steps, temperature 1, four rollouts per task, and a prompt that keeps the screenshots of the last three steps and the full text history. General capability: rule-based answer matching together with GPT-5.6 as a judge.

Training-dynamics metrics. The validation score is the mean score of the policy's decisions on a held-out split of the demonstration set, not used for training, each scored against its human-annotated ground truth by a verifier used only for evaluation, every 20 to 40 steps. The tie rate is the fraction of sampled critic verdicts that are T, and critic accuracy the fraction that select the expert side, with a tie counted as one half. The critic advantage variance is the variance of the gain advantage $\Delta_i - \bar{\Delta}$ over the pairs on which the snap judgment is probed, computed with the same formula in every run; it measures how differently the CoT moves the report across pairs. CoT lengths are mean token counts of policy decisions and critic verdicts. A value quoted at a step is the ten-step trailing mean at that step. The ReGain run takes 29.6 hours to reach step 160, about 1,900 GPU-hours.

Table 3: Training hyperparameters. The adversarial runs differ only in the critic reward (§5.1).

	Adversarial runs	SFT
Vision encoder	trained	frozen
Optimizer	Adam, $\beta = (0.9, 0.999)$, weight decay 0.01, gradient clip 1.0	AdamW
Learning rate	1×10^{-6}, constant, no warmup	1×10^{-5}
Batch	128 tasks $\times$ $G = 4$ rollouts, one update per step	global batch 64
Maximum sequence length	prompt $16{,}384$, response $2{,}048$ tokens	$120{,}000$ tokens
Image size	at most 640 image tokens	at most 640 image tokens
Loss tokens	generated tokens, token mean	target tokens only
Training length	160 steps ($20{,}480$ demonstrations, each once)	1 epoch
Hardware	A800; $4{+}4$ nodes of 8 GPUs	A800
Critic group	one prompt per policy rollout, random side order	–
Tie value c (discrete)	0.75 or 0.55	–
Policy tie reward τ_{pol} (discrete)	0.55	–
Advantage	group mean-centred, not divided by the standard deviation	–
Role loss weights	policy $\frac{1}{3}$, critic $\frac{2}{3}$	–
Clip ratio	0.2 low, 0.28 high, dual clip 3.0	–
KL regularization	low-variance KL loss to the initial model, 0.001; no entropy bonus	–
Rollout temperature / top-p	1.0 / 1.0	–
Maximum response length (policy / critic)	$2{,}048$ / $2{,}048$	–
Critic replay	half of the critic groups rebuilt from a buffer of the last $1{,}024$ policy rollouts, one earlier task per group, as in RARO	–

K Single-Step Benchmark by Sub-Suite

Table 4 breaks the single-step pass@1 of Table 1 down by sub-suite. Each sub-suite holds 50 held-out states, so a sub-suite entry carries a standard error of 5 to 10 points and only the larger differences are informative. The largest gains of ReGain over the backbone are on abnormal-state handling ($+18$ points), keyword extraction ($+9$), and safety ($+7.5$); data-matched SFT falls below the backbone on instruction following, safety, and grounding.

Table 4: Single-step pass@1 (%) by sub-suite. The overall row is the score reported in Table 1.

			SFT		RARO, discrete		ReGain	
Sub-suite	N	Backbone	20,480	190,000	c=0.75	c=0.55	Brier only	full
Grounding	50	77.0	70.0	77.5	78.0	78.0	80.5	83.3
Instruction following	50	73.5	62.0	76.5	74.5	61.0	70.8	76.3
Abnormal-state handling	50	42.5	47.0	59.5	49.0	50.0	46.8	60.8
Keyword extraction	50	58.0	62.0	70.0	68.0	66.0	67.5	67.3
Safety	50	55.0	46.0	47.0	58.5	61.5	63.8	62.5
Overall (pooled)	250	61.2	57.4	66.1	65.6	63.3	65.9	70.0

L Trajectory-Level Success by Difficulty

Table 5 breaks the trajectory-level success of Table 1 down by the difficulty level of the MobileGym tasks. The levels hold 8, 26, 35, and 31 tasks, so the L1 entries are the least reliable. ReGain leads on L2 and L3; on L4, where every cold start stays at or below 15%, data-matched SFT and Brier-only lead.

Table 5: Trajectory-level success (%) by task difficulty. N is the number of tasks per level; a task that a run did not finish counts as a failure, and no run finished one L4 task. The last row is the success rate reported in Table 1. The backbone succeeds on no task (§5.2).

		SFT		RARO, discrete		ReGain	
Level	N	20,480	190,000	c=0.75	c=0.55	Brier only	full
L1	8	78.1	90.6	68.8	93.8	90.6	93.8
L2	26	41.4	51.0	44.9	44.2	45.2	56.8
L3	35	17.4	16.4	15.0	11.4	13.6	22.2
L4	31	14.5	8.1	11.3	3.2	14.0	8.5
All	100	27.6	28.8	25.9	24.0	28.1	32.7

M General-Capability Scores

Table 6 gives the per-benchmark scores behind the "General capability" column of Table 1. Each benchmark is scored with its standard metric, by rule-based answer matching together with GPT-5.6 as a judge, and the aggregate is the unweighted mean over benchmarks of the relative change from the backbone, $(\text{score} - \text{score}_{\text{base}})/\text{score}_{\text{base}}$, so that benchmarks with different scales contribute equally.

Table 6: General capability after each cold start on eleven multimodal benchmarks (accuracy, %). The mean relative change is the aggregate reported in Table 1. Both SFT runs lose about 14% of the backbone's general capability and do so on almost every benchmark, whereas every adversarial run stays within about 2% of the backbone.

		SFT		RARO, discrete		ReGain	
Benchmark	Backbone	20,480	190,000	c=0.75	c=0.55	Brier only	full
MMMU	81.0	65.0	70.0	77.0	76.0	78.5	76.5
MMMU-Pro	77.0	63.0	60.5	78.5	80.5	79.5	78.5
MMStar	71.6	64.9	68.3	69.5	73.3	73.1	70.2
MathVista	80.0	70.0	68.0	76.0	78.0	77.0	78.0
MathVision	84.0	60.0	60.0	76.0	84.0	81.0	79.0
DynaMath	80.0	78.0	82.0	80.0	82.0	82.0	81.0
RealWorldQA	87.0	77.0	66.0	82.0	84.0	85.0	82.0
OCRBench v2	74.2	64.6	65.2	73.8	75.5	74.2	74.2
SimpleVQA	42.0	37.0	36.0	44.0	46.0	44.5	48.0
MMLongBench-Doc	64.2	50.8	55.5	68.0	66.9	66.8	64.5
HallusionBench	90.0	83.0	81.0	84.0	86.0	87.5	85.0
Unweighted mean	75.6	64.9	64.8	73.5	75.7	75.4	74.3
Mean relative change vs. backbone (%)	–	−14.2	−14.1	−2.1	+0.8	+0.2	−0.9
Benchmarks below the backbone	–	11/11	10/11	7/11	4/11	5/11	7/11

N Chain-of-Thought Lengths During Training

Figure 5 shows the mean CoT length of policy decisions and of critic verdicts for the four adversarial runs of Figure 2.

In every run the policy's CoT first shrinks from over 500 tokens to between 140 and 200. It then lengthens with the critic's tie rate, to 410 tokens at $c = 0.75$ and 565 for Brier-only, four to five times the 107-token mean of the demonstrations, but stays between 140 and 260 under ReGain. The lengthening is therefore not imitation of the demonstrations. One reading, which we do not test, is that a decision the critic cannot separate earns the policy the tie payoff, so under a critic that ties readily a longer decision pays.

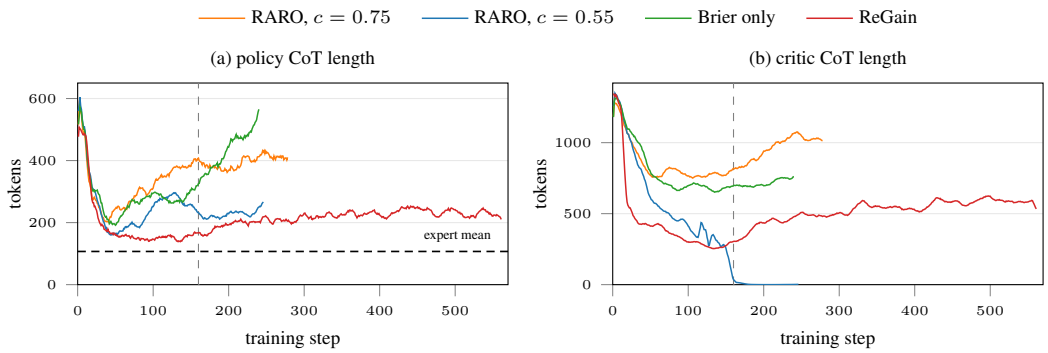

Figure 5: Mean CoT length of (a) policy decisions and (b) critic verdicts, as ten-step trailing means, for the four adversarial runs (§5.4); the vertical dashed line marks step 160, and the horizontal one in (a) the mean CoT length of the expert demonstrations, 107 tokens. At $c = 0.55$ the critic stops writing a CoT after step 200. Each curve is a single training run.

O CRITIC SHORTCUT DIAGNOSTICS

Figure 6 tracks two shortcuts in the logged critic outputs of the four adversarial runs. Panel (a) is the share of critic outputs whose CoT is empty. Panels (b) and (c) use the length of the think of each decision, which is short for the expert, a mean of 107 tokens. Panel (b) is the share of decisive (A or B) verdicts that pick the decision with the shorter think. Because the expert's think is usually the shorter one, panel (b) partly tracks accuracy, and panel (c) controls for this: critic accuracy, with a tie counted as one half, on pairs whose shorter think is the expert's minus that on pairs whose shorter think is the policy's. A critic that ignores length scores zero, and one that identifies the expert by brevity scores above zero. Over the first 100 steps the discrete and Brier-only critics prefer the longer think, picking the shorter one in 0.42 of decisive verdicts, and the ReGain critic has no preference (0.49). The empty-CoT share of the critic at $c = 0.55$ passes 0.05 at step 117 and reaches 0.99 by step 200, while its gap never exceeds 0.02; the critic at $c = 0.75$ drifts toward brevity after step 160, its shorter-think share rising from 0.58 to 0.72 and its gap to 0.08; Brier-only never does; and ReGain does so mainly after step 400, its shorter-think share rising from 0.57 to 0.77 and its gap reaching 0.18.

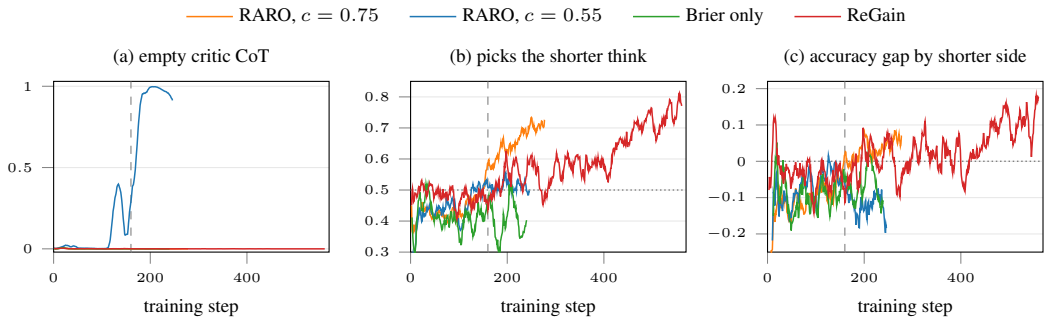

Figure 6: Critic shortcut diagnostics for the four adversarial runs, as ten-step trailing means; the dashed line marks step 160. (a) Share of critic outputs with an empty CoT. (b) Share of decisive verdicts that pick the decision with the shorter think; the dotted line is one half. (c) Critic accuracy, with a tie counted as $\frac{1}{2}$, on pairs whose shorter think is the expert's minus that on pairs whose shorter think is the policy's.

P WHERE THE CRITIC CREDIT GOES

Figure 7 recomputes the critic advantages of the ReGain run on its logged rollouts under both rewards, the gain advantage it was trained with and the Brier return. A pair is decided at a glance when $s_{\text{snap}} \geq 0.9$, and its CoT overturned the snap judgment when the snap judgment favoured the policy decision and the post-CoT report favoured the expert's. Panels (a) and (b) give the share of the positive credit, the sum of positive advantage times response length, that each type of pair receives, and panel (c) the share of the whole update, the sum of absolute advantage times length, on pairs decided at a glance. At step 160 pairs decided at a glance are 35% of all pairs and receive 40% of the positive credit under the Brier return but 20% under the gain, and pairs whose CoT overturned the snap judgment are 18% of all pairs and receive 37% and 56%. The whole update on pairs decided at a glance is about the same under both rewards, 33% and 31%: the gain does not silence these pairs but turns part of their credit negative, as the implicit curriculum of §4.3 predicts.

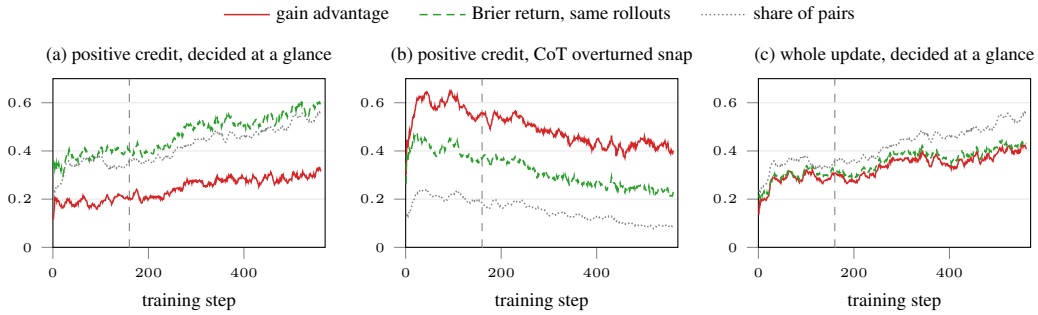

Figure 7: Where the critic credit of the ReGain run goes, under the gain advantage it was trained with and under the Brier return recomputed on the same rollouts, as ten-step trailing means; the dashed line marks step 160, and the dotted curve is the share of pairs of each type. (a) Positive credit on pairs decided at a glance ($s_{\text{snap}} \geq 0.9$). (b) Positive credit on pairs whose CoT overturned the snap judgment. (c) Whole update, positive and negative, on pairs decided at a glance.

Q QUALITATIVE CASES

Each case gives the instruction, the two decisions of the pair, and the critic's output. Transcripts are translated from Chinese and abridged, with omissions marked [...]; screens are described in words, and personal details are replaced by ⟨name⟩.

Case 1: a tie that overrides the reasoning ($c = 0.75$). The critic's CoT ends by preferring A, the expert, but the sampled verdict is a tie: the verdict distribution places 0.56 on A and 0.44 on T. At $c = 0.75$ a tie pays three quarters of a correct verdict with certainty, so the verdict can drift away from the conclusion of the reasoning at little cost to the critic (§5.4).

Screen. Home page of the grocery channel of a delivery app, with a search box at the top.
Instruction. "I want something different tonight. Please use the grocery channel to make a dish like braised pork, but with beef, and get the dish ready."

A (expert). `<think>` The user wants a dish like braised pork made with beef, so the core need is beef as the substitute main ingredient. No cut or brand is specified, so only the category word is kept; the recipe and the taste belong to cooking, not to the search query. `</think>`
`set_text([319, 116], "beef", clear=True, enter=True)`

B (policy). `<think>` [...] A beef dish like braised pork is usually called "braised beef", or one only needs to buy beef and seasoning. [...] Search "braised beef" first to see whether there is a ready-made ingredient kit; if not, buy the items separately. The search box is at the top, at about [499, 111]. `</think>`
`set_text([499, 111], "braised beef", clear=True, enter=True)`

Critic. [...] Since they are about the same, tentatively a tie. [...] Slightly leaning to A: searching "beef" corresponds more directly to "make it with beef". [...] Conclusion: A is slightly better, since its keyword strategy focuses on the ingredient. `</think>` [...] `<verdict>T</verdict>`

Case 2: a verdict without reasoning ($c = 0.55$, after the CoT collapse). Both decisions enter the same name into the same field, so the actions are equivalent, and the critic writes no reasoning at all before picking B, the expert. The verdict is emitted without the comparison the critic was trained to write, the collapse of §5.4.

Screen. A form for adding a frequent passenger, with fields for the name, the English surname and given name, and the ID number.
Instruction. "Fill in ⟨name⟩'s information."

A (policy). <think> The task is to fill in ⟨name⟩'s information in the app. According to the working memory, the information includes: * Chinese surname [...] * English given name [...] * Phone number [...] [about fifty more lines that re-check which ID number belongs to which family member and list the fields to fill] [...] Fill in the name first. </think>
`set_text([599, 227], "⟨name⟩")`

B (expert). <think> The current page adds a frequent passenger and asks for the passenger's details. By the instruction, the passenger is ⟨name⟩. The name comes first and should match the ID document, so enter ⟨name⟩ in the name field. </think>
`set_text([424, 238], "⟨name⟩")`

Critic. </think> [...] <verdict>B</verdict>

Case 3: the shortcut stated in words. In a preliminary run started from an SFT checkpoint, not among the runs of Table 1, the critic names the cue it uses. Both decisions tap the "Me" tab in the bottom bar of a video app; the expert's think is one sentence, and the policy's adds a reference to its plan.

Instruction. "Add a new purchaser's information in the video app."

A (expert). <think> To open the profile page and continue, tap the "Me" button at the bottom. </think>
`click([897, 968])`

B (policy). <think> According to the plan, I should now tap "Me" on the home page to open the profile page. The current page is the home page, and the bottom navigation bar has a "Me" button that can be tapped directly. </think> `click([867, 960])`

Critic (complete reasoning). Both answers correctly recognize the home page and tap "Me" as planned; the logic is correct. But A's think is concise and states the purpose directly, while B's think adds a description of the plan and is slightly redundant. Both follow the format, but A is more refined and closer to the expert's style. </think> [...] <verdict>A</verdict>

Case 4: reasoning about content under ReGain. The two thinks read alike, so their style does not separate the decisions, and the critic checks the landing points against the screen. The snap judgment already favoured the expert, with $s_{\text{snap}} = 0.969$, so the CoT, although it finds the decisive error, earns only $\Delta_i = 0.031$, the margin the snap judgment left open (§4.3).

Screen. Discovery feed of a social app, with content cards in the middle and a bottom navigation bar whose rightmost tab is "Me".
Instruction. "I want better privacy protection; please turn on the app's one-tap protection."

A (expert). <think> Following the plan, tap "Me" on the home page to open the profile page; the "Me" tab at the bottom is not selected, so tap it to go on. </think> `click([898, 972])`

B (policy). <think> According to the task plan I need to open the profile page; I am on the discovery page and need to tap "Me", at about 895, 708. </think> `click([895, 708])`

Critic. [...] Answer A: `click([898, 972])` [...] is exactly the "Me" button at the bottom right; the landing point is correct and the action advances the task. Answer B: `click([895, 708])` [...] $y = 708$ corresponds to the apple image in the right middle of the screenshot, not the "Me" button of the bottom bar; the landing point is wrong and would tap a content card. [...] A clearly advances the task; B has the right direction but the wrong landing point. [...] Not a tie. </think> [...] <verdict>A</verdict>