BEYOND ROBUST SCORING: A FOUR-STAGE AUDIT OF PERSISTENT SKILL WRITES

Anonymous authors
Paper under double-blind review

ABSTRACT

Persistent skill writing converts textual advice into parameter updates that remain active after the text is removed. We show why robust scoring alone is insufficient and turn the resulting audit into a write-control method. The audit separates score comparability, candidate support, causal realization, and behavioral-label validity. Its findings motivate **Audit-Guided Write Control (AGWC)**, which combines support-aware stopping, paired shadow updates, context-dependent re-evaluation, and set-level commit/rollback. On ALFWorld, the reported 288-candidate, five-pool evaluation gives AGWC **44.48%** scenario-OOD success versus 36.52% for Direct-Validation, while behavioral regression decreases from 5.10% to 3.60%. With 18 selected skills per method, the advantage remains **6.60 percentage points**, despite identical observed proxy harm. Shadow probes predict independent full-write effects with AUROC 0.74 and Spearman correlation 0.48. After fresh retraining on disjoint evaluation games, held-out success improves from 22.22% to **28.02%**. These aggregate results support persistent writing as a sequential verification-and-commit problem: comparable scoring proposes interventions, while explicit behavioral checks govern deployment.

1 INTRODUCTION

Language agents increasingly preserve reusable behavior as textual skills, programs, memories, or self-generated supervision (Wang et al., 2024; Shinn et al., 2023; Wang et al., 2026; He et al., 2026; Wu et al., 2026; Lu et al., 2026). This capability enables cumulative improvement, allowing agents to reuse successful strategies without repeatedly solving the same problems from scratch. Interactive benchmarks such as ALFWorld, ScienceWorld, and AgentBench make later task behavior measurable (Shridhar et al., 2021; Wang et al., 2022; Liu et al., 2024). Yet the same mechanism introduces a uniquely persistent failure mode. External skills can be removed when they become unreliable, whereas parameter updates may continue to affect the policy after the original skill and the evidence supporting its selection are no longer available. We therefore argue for stronger evidence of write eligibility, since a plausible candidate or a confident gate score alone is insufficient for persistent parameter updates.

Existing studies on agent skill safety mainly focus on risks before skill adoption (Liu et al., 2026a;b), while overlooking whether write-time evidence remains valid after training. This assessment spans several stages that are often evaluated separately. A candidate generator first proposes textual skills. This paper makes that distinction operational. We first localize four ways in which score-based evidence can fail, then use those failures to design AGWC. The central result is a stronger empirical bridge from candidate selection to skill-free behavior, supported by matched-write controls, full-write prediction, component ablations, candidate-supply scaling, and fresh retraining.

From a favorable score to a defensible write. The usual pipeline compresses a complex decision into a scalar: score a skill, apply a threshold or quota, and write the selected text into the policy. Description-length and PAC-Bayesian work motivate parsimonious evidence terms (Grünwald, 2007; Shao et al., 2024; Dziugaite et al., 2020), while ensembles motivate checking variability across views (Feng et al., 2025). Selective classification makes abstention an explicit option (Fontana et al., 2025; Gao et al., 2023), and underspecification shows why favorable development evidence may not determine deployed behavior (D'Amour et al., 2022). Yet a stronger candidate score is not direct

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.

evidence about a trained policy: score offsets alter thresholds, scarce support makes quotas harmful, and proxy labels can disagree with skill-free outcomes.

Our audit treats these as four separate requirements. *Score comparability* aligns the evidence used to prioritize candidates. *Candidate support* checks whether the proposal distribution can sustain the requested acceptance level. *Causal realization* asks what actually changes when a skill is written. *Behavioral-label validity* asks whether the resulting policy works without the skill text. Each requirement exposes a missing control in one-shot score gating.

AGWC turns the audit into a sequential controller. A comparable ranker allocates scarce verification effort. Paired short training runs inspect a candidate in the context of the currently selected set. The controller can stop instead of filling an unsupported quota, revisits evidence when the selected context changes, and validates the fully trained set before committing it. The final decision is therefore made at the same level as deployment: the trained policy.

The key control experiment: with the same 18 selected skills and the same 5.56% proxy-HUR, AGWC still improves scenario success by 6.60 points.

The results provide complementary tests of this design. Adaptive deployment improves scenario success by 7.96 points relative to Direct-Validation. The matched-write experiment preserves a 6.60-point advantage, excluding write count alone as an explanation. Probe prediction links short updates to independently measured full writes. Ablations locate the largest observed degradation in the paired shadow component. Finally, fresh training seeds and disjoint games retain a 5.80-point held-out advantage. These are empirical improvements in the reported settings, not a universal safety certificate.

Contributions.

1. **Audit to control.** We identify four distinct failure boundaries and translate them into requirements for persistent-write decisions.
2. **AGWC.** We specify a sequential controller with support-aware stopping, paired behavioral probes, context-dependent re-evaluation, and set-level commit/rollback.
3. **A measured shadow-to-full bridge.** Reported shadow predictions reach AUROC 0.74 and rank correlation 0.48 against full-write effects, supplying information beyond score-only ranking.
4. **A broader behavioral evaluation.** Five 288-skill pools, matched-write controls, scaling to 576 candidates, four ablations, and fresh retraining test different explanations for the gains.

2 Related Work

Reusable skills and internalization. Temporal abstractions provide a classical foundation for reusable behavior. Agents interleave reasoning and action, call tools, or retrieve episodic memories (Yao et al., 2023; Schick et al., 2023; Park et al., 2023); others preserve experience in programs and verbal reflections (Wang et al., 2024; Shinn et al., 2023). Parameter-efficient adaptation makes persistent updates practical (Lin et al., 2026), and recent skill-internalization work studies their behavioral effects (Wang et al., 2026; He et al., 2026; Wu et al., 2026; Lu et al., 2026). Our question is what evidence should authorize and retain an update.

Robust evidence and compression. Description length and PAC-Bayesian reasoning motivate compression-based scoring (Grünwald, 2007; Shao et al., 2024; Dziugaite et al., 2020). Ensembles and robust aggregation address unreliable individual evaluations (Feng et al., 2025). AGWC retains robust ranking as a prioritization tool, while separating it from the behavioral decision to commit a persistent write.

Selective decisions. Selective prediction explicitly trades coverage for reliability (Fontana et al., 2025; Gao et al., 2023). Sequential learning also changes the states visited by the trained policy (Ross et al., 2011). Persistent writing adds a training intervention and interactions among selected skills. Abstention helps with inadequate support, but matched-write evidence is needed to distinguish better selection from reduced coverage. AGWC evaluates both operating regimes.

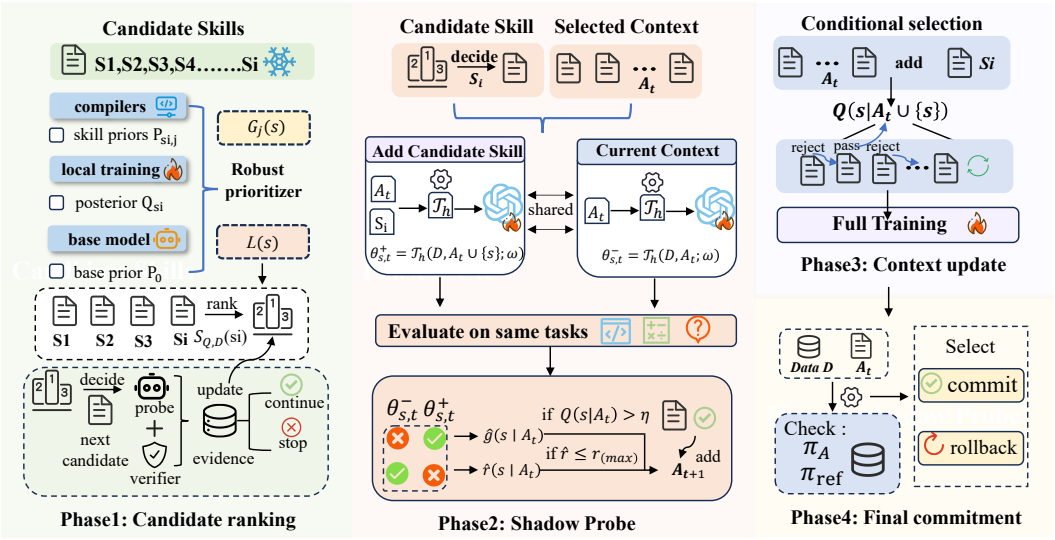

Figure 1: **AGWC model overview.** Candidate ranking, paired short updates, conditional selection, and final commit or rollback are shown as separate stages. The replacement artwork is supplied by the author.

3 The Four-Stage Audit

3.1 Objects and Endpoints

Let C_g be the candidate pool for stratum g, $S(s)$ a ranking score, and $\mathcal{T}(D, A; \omega)$ the policy trained using selected set A, data D, and randomness ω. At test time, binary success $Y(\pi, x)$ is measured without skill text. We distinguish the selected set A_{sel} from the committed set A_{com}: rollback returns the reference policy, so selection need not imply deployment.

For frozen write-time labels $Z_g(s)$, let $H_g = \{s : Z_g(s) < 0\}$ and $B_g = \{s : Z_g(s) > 0\}$. Selection metrics are

$$\text{Proxy-HUR} = \frac{\sum_g |A_g \cap H_g|}{\sum_g |A_g|}, \quad \text{AR} = \frac{\sum_g |A_g|}{\sum_g |C_g|}, \quad \text{P@A} = \frac{\sum_g |A_g \cap B_g|}{\sum_g |A_g|}. \tag{1}$$

An empty accepted set has undefined HUR and P@A, not zero. Behavioral regression and gain instead use paired policy outcomes on M evaluation games:

$$\text{BReg} = M^{-1} \sum_{i=1}^{M} \mathbf{1}[Y(\pi_{\text{ref}}, x_i) = 1,\ Y(\pi, x_i) = 0], \tag{2}$$

$$\text{BGain} = M^{-1} \sum_{i=1}^{M} \mathbf{1}[Y(\pi_{\text{ref}}, x_i) = 0,\ Y(\pi, x_i) = 1]. \tag{3}$$

Both use all paired games as the denominator; consequently $\text{SR}(\pi) - \text{SR}(\pi_{\text{ref}}) = \text{BGain} - \text{BReg}$. A regression rate conditional on reference successes is a different endpoint. The source aggregates do not include the event-level counts needed to verify this convention empirically.

3.2 FAILURES BECOME DESIGN REQUIREMENTS

Table 1: Each audit stage motivates a concrete control.

Audit stage	Failure boundary	AGWC response
Score comparability	Different score origins alter threshold decisions.	Align evidence; use ranking to prioritize.
Candidate support	A forced quota can exceed non-harmful supply.	Stop or abstain when verification fails.
Causal realization	Score quality does not establish an update effect.	Pair shadow training with and without a skill.
Behavioral-label validity	Candidate proxies need not order trained policies.	Validate the full set; commit or roll back.

Proposition 1 (Offsets change thresholds, not ranks)**.** *If $S'_m(s) = S_m(s) - a_m$ and $A_m(t) = \{s : S_m(s) > t\}$, then $A'_m(t) = A_m(t + a_m)$. Every top-k set is unchanged under a fixed tie rule.*

Proof. The threshold identity follows by rearrangement; a constant preserves every pairwise ordering.

Proposition 2 (Finite support limits forced acceptance)**.** *If stratum g contains b_g non-harmful candidates and requires k_g accepts, with $\sum_g k_g > 0$, then*

$$\mathrm{HUR} \geq \frac{\sum_g [k_g - b_g]_+}{\sum_g k_g}. \tag{4}$$

The bound is tight when non-harmful candidates can be ranked first in every stratum.

Proof. At least $k_g - b_g$ accepts are harmful whenever $k_g > b_g$. Sum these unavoidable accepts; an oracle ordering attains equality.

Proposition 3 (Proxy evidence alone does not identify behavior)**.** *Without a bridge from candidate labels through $\mathcal{T}$ to Y, selector orderings based on $\{Z(s) : s \in A\}$ do not identify orderings of $Y(\mathcal{T}(D, A), x)$.*

Proof. Fix candidates, scores, labels, and accepted sets. Two unrestricted mappings from accepted sets to behavior can agree on all proxy observations and reverse the behavioral ordering.

4 AUDIT-GUIDED WRITE CONTROL

4.1 COMPARABLE RANKING AND SUPPORT-AWARE STOPPING

For compiler gains $G_j(s)$ and description cost $L(s)$, a robust prioritizer takes the form

$$S_{\mathrm{QD}}(s) = q_\alpha(G(s)) - \lambda_{\mathrm{rank}} \widehat{\sigma}(G(s)) - L(s) - c_{\mathrm{sel}} - c_{\mathrm{comp}}. \tag{5}$$

In the Gaussian audit, $G_j(s) = \mathrm{KL}(Q_s \| P_0) - \mathrm{KL}(Q_s \| P_{s,j})$; the historical configuration uses $J = 5$, $\alpha = 0.1$, and $\lambda_{\mathrm{rank}} = 0.5$. The ALFWorld names denote operational proxy-ranking analogues, not an implementation of this Gaussian model. Common offsets must be aligned for threshold comparisons; rank-based scheduling is invariant to constant shifts.

AGWC uses this evidence to decide what to inspect next. It does not use $S(s) > 0$ as sufficient write permission. Equation 4 explains why a controller must permit abstention: no ranking can recover non-harmful candidates absent from its pool. The bound motivates stopping but is not an oracle available to the controller. In practice, stopping depends on observed probe evidence and the remaining verification budget.

4.2 PAIRED SHADOW VERIFICATION

For candidate s in selected context A_t, compare

$$\theta^{+}_{s,t} = \mathcal{T}_h(D, A_t \cup \{s\}; \omega), \qquad \theta^{-}_{s,t} = \mathcal{T}_h(D, A_t; \omega). \tag{6}$$

The two arms share initialization, optimizer state, minibatches, training steps, randomness, and probe games. On these paired games, define $\widehat{g}(s \mid A_t)$ as the fraction on which the minus arm fails and the plus arm succeeds, and $\widehat{r}(s \mid A_t)$ as the reverse transition. A behavioral priority can be written as

$$Q(s \mid A_t) = \widehat{g}(s \mid A_t) - \lambda_{\text{reg}} \widehat{r}(s \mid A_t). \tag{7}$$

A candidate passes a prespecified verification predicate based on gain and regression. A general realization is $Q > \eta$ and $\widehat{r} \le r_{\max}$; the supplied results do not identify the exact deployed thresholds or h. We therefore leave these as explicit controller inputs, rather than inventing experimental settings. Short-update prediction remains an approximation to full training, which is evaluated separately.

4.3 CONTEXT UPDATE AND FINAL COMMITMENT

Adding s changes the question for every remaining candidate: the relevant quantity becomes $Q(\cdot \mid A_t \cup \{s\})$. Re-evaluation consumes the same finite probe budget and prevents stale singleton evidence from being treated as evidence about an interacting set. After selection, full training yields $\pi_A = \mathcal{T}(D, A_{\text{sel}})$. A distinct commit-validation set tests this policy against the reference. Passing commits the write; failing restores the reference. Test games must remain outside both decisions.

4.4 CONTROLLER SPECIFICATION AND EXPERIMENTAL SETUP

Algorithm 1: Audit-Guided Write Control

Inputs: pool C; comparable ranking S; budget W; shadow/full training operators; disjoint probe and commit data; frozen pass, stopping, and commit predicates.
Initialize: $A \leftarrow \varnothing$, used budget $w \leftarrow 0$; rank C.
While $w < W$ and untested or stale candidates remain:

1. Select the next candidate using ranking and current-context evidence.
2. Run paired shadow arms conditioned on A; compute $(\widehat{g}, \widehat{r}, Q)$; increment w.
3. If the verification predicate passes, set $A \leftarrow A \cup \{s\}$ and mark remaining conditional evidence stale.
4. Otherwise reject this candidate in the current context and continue. Stop when the frozen stopping rule fires or no eligible candidate remains.

If $A = \varnothing$, return π_{ref}. Otherwise full-train π_A.
If the set-level commit predicate passes on commit data, return π_A; otherwise return π_{ref}.

A failed probe is not evidence that every untested candidate fails. The algorithm therefore distinguishes candidate rejection, a declared stopping rule, and budget exhaustion. Likewise, a re-evaluation is a new probe, not free extra evidence. This distinction matters when candidate supply grows faster than the verification budget.

5 EXPERIMENTS

5.1 PROTOCOL AND EVALUATION

The ALFWorld study reports five independently generated 288-candidate pools and three training seeds per pool (15 final adapters per method). AGWC uses at most 64 paired shadow probes. Methods share pools, seeds, and evaluation games as described in the supplied results; final policies are tested without skill text. The matched-write study selects $k = 18$ candidates per method (6.25% acceptance), and fresh transfer uses new seeds and disjoint games with frozen pools. Matched selection does not establish equal compute or exactly 18 *committed* writes after rollback.

Baselines. These are task-specific controls, not separately imported published algorithms. Outcome-only trains without selected skill text and supplies the reference outcome for behavioral regression. Matched-k Random selects candidates uniformly without a skill score. Utility ranks by the supplied utility proxy; Robust-QD ranks by a conservative lower-quantile and disagreement score (Section 4.1). Direct-Validation is the strongest reported behavioral-validation control and the principal comparator. The supplied aggregate record does not specify its exact validation schedule, thresholds, or compute allocation; those details are needed for an executable reproduction. Fixed-k variants use the same selected count, not an equal compute budget.

5.2 Behavioral results

Table 2: Adaptive deployment on five 288-candidate ALFWorld pools. Rates are percentages; cost is relative to Outcome-only. The gain in scenario success is accompanied by lower behavioral regression.

	Selection			Skill-free behavior				Compute
Method	AR	HUR↓	P@A↑	Dev↑	Inst.↑	Scen.↑	BReg↓	Cost
Outcome-only	0.00	–	–	62.14	48.48	17.14	0.00	1.00
Matched-k Random	6.25	81.48	12.96	60.87	52.73	31.20	10.80	1.35
Utility-adaptive	5.56	8.33	66.67	58.72	49.84	28.15	6.40	2.10
Robust-QD / matched-k	6.25	22.22	55.56	61.05	51.32	33.74	7.20	3.20
Direct-Validation	5.21	6.67	73.33	63.18	53.46	36.52	5.10	8.60
AGWC	4.86	**4.76**	**85.71**	**66.80**	**58.70**	**44.48**	**3.60**	10.80
Δ vs. Direct (pp)	−0.35	−1.91	+12.38	+3.62	+5.24	+7.96	−1.50	+2.20

Under adaptive deployment, AGWC improves scenario-OOD success by 7.96 percentage points and lowers BReg by 1.50 points relative to Direct-Validation (Table 2). Development and instance-OOD success also increase. The selection proxies move in the same direction, but AGWC accepts fewer candidates; this adaptive comparison alone cannot separate selection quality from write count.

Table 3: Matched selection at $k = 18$ (AR $= 6.25\%$). Rates are percentages. This controls selected write count, but is not an equal-compute comparison.

	Selection		Success rate			Regression
Method	HUR↓	P@A↑	Dev↑	Inst.↑	Scen.↑	BReg↓
Outcome-only ($k = 0$)	–	–	62.14	48.48	17.14	0.00
Random	81.48	12.96	60.25	52.10	31.45	11.20
Utility	8.33	77.78	57.80	47.25	24.60	8.10
Single	44.44	55.56	59.10	50.20	38.10	9.80
Robust-QD	22.22	66.67	61.30	52.40	34.20	7.60
Hybrid-QD+Utility	11.11	72.22	62.40	53.10	36.70	6.20
Direct-Validation-fixed	5.56	77.78	63.25	54.00	37.20	5.40
AGWC-fixed	**5.56**	**88.89**	**66.40**	**57.80**	**43.80**	**3.20**
Δ vs. Direct-fixed (pp)	0.00	+11.11	+3.15	+3.80	+6.60	−2.20

The matched-write study holds the selected set at 18 skills and proxy-HUR at 5.56% for both methods. AGWC-fixed still improves scenario success by 6.60 points and lowers BReg by 2.20 points (Table 3). Single exceeds Direct-fixed on scenario success, yet AGWC-fixed exceeds Single by 5.70 points. Equal harmful fractions can conceal different beneficial or neutral composition: P@A is 88.89% for AGWC-fixed versus 77.78% for Direct-fixed. The supplied protocol does not establish equal compute or exactly 18 *committed* writes after rollback.

Table 4: Fresh retraining with new seeds and disjoint held-out games. Rates are percentages; ΔSR is relative to Direct-Validation.

Method	Train-family SR↑	Held-out SR↑	BReg↓	ΔSR (pp)
Outcome-only	49.80	16.40	0.00	–
Utility	46.20	15.10	7.80	−7.12
Single	50.10	19.40	9.20	−2.82
Robust-QD	52.40	20.10	7.50	−2.12
Hybrid-QD+Utility	54.10	21.30	6.00	−0.92
Direct-Validation	55.20	22.22	5.10	0.00
AGWC	**59.40**	**28.02**	**1.80**	**+5.80**

Fresh retraining changes seeds and held-out games while keeping candidate pools frozen. AGWC improves held-out-family success from 22.22% to 28.02% and reduces BReg from 5.10% to 1.80% (Table 4). Train-family success also rises. This is stronger than evaluating the same frozen adapters on more games, but it does not test a new skill generator, model family, or environment; the older frozen-adapter transfer audit remains separate in Appendix A.

5.3 Do shadow probes predict full writes?

Table 5: **Key mechanism checks.** Full diagnostic matrices remain in Appendix E. Values are reported aggregates without uncertainty intervals; ablation differences are relative to AGWC-full.

Shadow/full-write measure	All evaluated	AGWC-accepted
Full-write marginal gain (pp)	+3.20	+7.60
Full-write BReg (%)	4.20	1.80
Probe/full-write AUROC	0.74	–
Probe/full-write Spearman	0.48	–
Effect-sign agreement (%)	72.50	78.57
Probe-to-full MAE (pp)	3.10	2.20

Fixed-k candidate supply	$N = 36$	$N = 72$	$N = 144$	$N = 288$	$N = 576$
Scenario success (%)	28.60	33.70	38.90	43.80	43.60
Proxy-HUR (%)	38.89	27.78	16.67	5.56	8.33
BReg (%)	11.40	8.20	5.60	3.20	3.30

Ablation	Success lost (pp)	BReg added (pp)
No adaptive stopping	5.80	4.60
No paired shadow probe	9.40	8.60
No commit/rollback	6.90	7.80
No context re-evaluation	5.60	5.40

The shadow/full-write diagnostic reports AUROC 0.74 and Spearman correlation 0.48. Accepted candidates have mean full-write marginal gain of 7.60 points versus 3.20 for all evaluated candidates, while BReg falls from 4.20% to 1.80%. The probe-to-full mean absolute error remains 3.10 points overall, so the short intervention cannot replace final set-level validation. The realized candidate count and positive-class threshold were not supplied.

Overall effect-sign agreement is 72.50% and rises to 78.57% among accepted candidates; their mean absolute error is 2.20 points. This is a useful but imperfect bridge from short updates to independently measured full writes. The supplied plan specifies ranges of 60–120 candidates, two or three seeds per candidate, and 20–40 paired games per arm, but gives no realized counts. We therefore report only the aggregate diagnostic, without synthetic scatter points, confidence intervals, or a claimed significance level. Candidate-wise prediction also cannot capture every interaction in a jointly trained selected set, which motivates the final commit check.

These summaries answer different questions. AUROC measures whether the probe orders candidates by the reported positive-effect label; it does not measure the size of the eventual gain. Spearman correlation checks a graded ordering, while sign agreement and mean absolute error address direction and magnitude. The accepted-subset numbers are conditional on AGWC's selection rule, so their improvement over all evaluated candidates is descriptive and cannot be read as an independent treatment effect. In particular, the reported 78.57% sign agreement does not imply that every accepted write improves the final policy.

The paired probe controls for training and game variation within its short comparison, but it is still an approximation to a full write. Interactions can appear after selected skills are trained together, and a small marginal gain can be outweighed by regression elsewhere. AGWC therefore uses the probe to allocate verification effort and tests the complete set before commitment. Candidate-level paired outcomes, the positive-class definition, and pool-level uncertainty remain unavailable; the aggregates indicate predictive usefulness without calibrating deployment safety.

5.4 Candidate supply and verification capacity

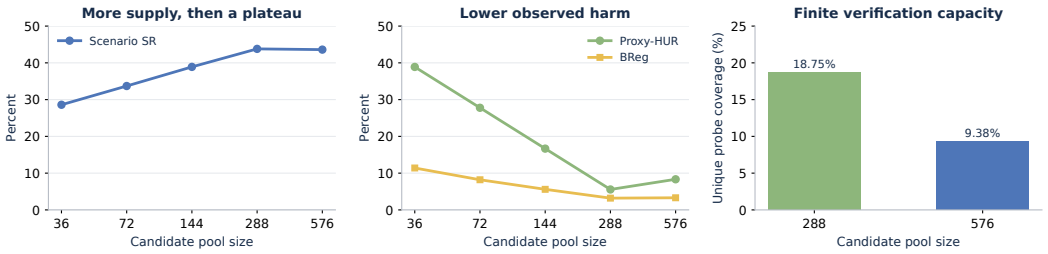

Figure 2: **Scaling with 18 selected writes.** Increasing supply from 36 to 288 candidates raises scenario success by 15.20 points and lowers both harm endpoints. Performance is nearly flat from 288 to 576 under the reported fixed verification budget. Coverage is shown only where provided. Lines connect observed aggregate settings; they are not fitted scaling laws.

Increasing the pool from 36 to 288 candidates raises scenario success from 28.60% to 43.80%, reduces proxy-HUR from 38.89% to 5.56%, and reduces BReg from 11.40% to 3.20%, with 18 selections throughout. Doubling supply again yields 43.60% success and 3.30% BReg. This 0.20-point difference is too small to establish a true decline without uncertainty estimates. At $W = 64$, unique probe coverage falls from 18.75% at 288 to 9.38% at 576. This is consistent with finite verification capacity becoming a bottleneck, but does not identify an optimal pool size.

5.5 Component ablations

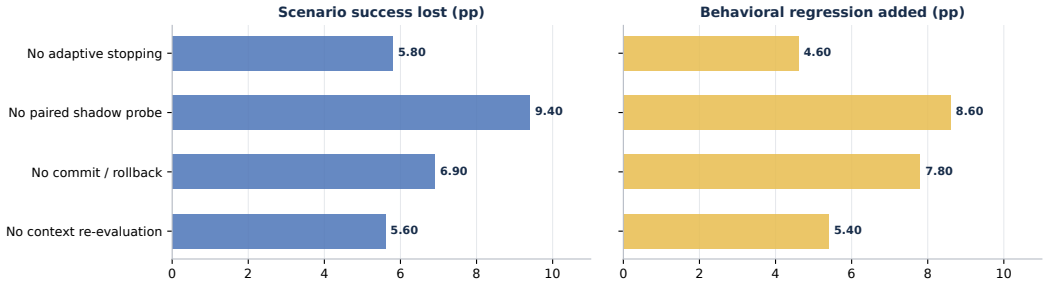

Figure 3: **Every removed control degrades the reported outcome.** Values are losses in success or increases in regression relative to AGWC-full. The paired probe has the largest observed effect in both panels. Acceptance rates also change, so these are system-level ablations, not isolated equal-coverage component effects.

Removing the paired shadow probe causes the largest reported degradation: 9.40 success points lost and 8.60 regression points added. Removing commit/rollback loses 6.90 success points and adds 7.80

regression points, showing that candidate-level verification and set-level validation address different failure points. Removing adaptive stopping loses 5.80 success points; removing context re-evaluation loses 5.60. The ablations change acceptance and compute, and the no-stopping variant is not the separately configured fixed-k diagnostic. Full values are in Appendix E.

5.6 Supporting diagnostics

Score alignment removes the apparent fixed-threshold advantage in the development-revised CC-Synth-v2 audit. Ranking AUROCs of 0.988–0.994 coexist with threshold sensitivity: Single's executed 12.92% HUR versus Robust-QD's 0% is not an aggregation-only contrast.

Table 6: Score comparability at 35% corruption. CAR is covered acceptance rate; all entries are percentages. Dashes denote values absent from the supplied summary.

Method	Executed HUR	Aligned HUR	Aligned CAR	HUR at 50% AR
Utility	–	–	–	26.63
Single	12.92	0.00	88.28	27.80
Mean	5.85	0.00	88.41	27.63
Quantile-only	0.00	0.00	86.85	–
Cert-Min	–	–	–	27.84
Robust-QD	0.00	0.00	85.94	28.04

The common 5.075-nat correction makes the aligned HUR zero for the four directly rescored rules, so the executed-threshold contrast cannot isolate robust aggregation. Matched 50% acceptance changes the comparison again: Robust-QD records 28.04% HUR, close to Single's 27.80% and above Utility's 26.63%. These are properties of the development audit, not estimates from the newer 288-candidate ALFWorld pools.

The historical frozen-pool result illustrates a separate boundary. Robust-QD lowers proxy-HUR relative to Single but has lower observed scenario success (29.52% versus 40.00%). Those old adapters and candidate pools are not pooled with the new study. Together, the diagnostics justify using score evidence to prioritize verification while reserving write permission for measured behavioral effects. The historical table and its provenance remain in Appendix A.

AGWC's reported total cost is $10.80\times$ Outcome-only, compared with $8.60\times$ for Direct-Validation, an increase of 25.58%. The success and regression gains are accompanied by additional verification compute. Exact probe/full-training cost decomposition and a run-level compute ledger are unavailable, so this comparison cannot establish equal-budget efficiency.

5.7 Evidence scope and statistical unit

The experimental package contains aggregate tables rather than run-level records. Appendix Table 7 distinguishes measured summaries from details that remain unresolved.

For the main behavioral contrasts, the appropriate statistical unit starts with the independently generated candidate pool. Training seeds and shared evaluation games are nested within those pools; counting all 15 adapters as independent candidate distributions would overstate the effective sample size. Paired per-pool effects and an uncertainty interval cannot be reconstructed from rounded aggregate percentages. The results are therefore descriptive estimates, with the specific missing run identifiers and implementation settings listed in Appendix A.

6 Discussion and Conclusion

Matched writes, probe validation, and ablations support behavioral verification beyond score-based selection. AGWC uses more compute than Direct-Validation, and aggregate results cannot establish per-pool consistency or statistical significance. Training settings and commit accounting need clarification (Appendix A). Within the reported ALFWorld protocols, ranking prioritizes candidates, while skill-free behavior determines whether a persistent write is retained.

REFERENCES

Alexander D'Amour, Katherine A. Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory Y. McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *J. Mach. Learn. Res.*, 23:226:1–226:61, 2022. URL https://jmlr.org/papers/v23/20-1335.html.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, and Daniel M. Roy. On the role of data in pac-bayes bounds. volume abs/2006.10929, 2020. URL https://arxiv.org/abs/2006.10929.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Nicoló Fontana, Francesco Corso, Enrico Zuccolotto, and Francesco Pierri. Evaluating open-source large language models for automated fact-checking. *CoRR*, abs/2503.05565, 2025. doi: 10.48550/ARXIV.2503.05565. URL https://doi.org/10.48550/arXiv.2503.05565.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: researching and revising what language models say, using language models. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 16477–16508. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.910. URL https://doi.org/10.18653/v1/2023.acl-long.910.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007. doi: 10.7551/mitpress/4643.001.0001.

Zhongyu He, Yuanfan Li, Fei Huang, Tianyu Chen, Siyuan Chen, Xingyang Li, Meng Hsuan Yu, Xiangrong Liu, Leyi Wei, Lu Pan, Ke Zeng, and Xunliang Cai. SIRI: self-internalizing reinforcement learning with intrinsic skills for LLM agent training. *CoRR*, abs/2606.02355, 2026. doi: 10.48550/ARXIV.2606.02355. URL https://doi.org/10.48550/arXiv.2606.02355.

Moule Lin, Shuhao Guan, Andrea Patane, David Gregg, and Goetz Botterweck. Bayesian-lora: Probabilistic low-rank adaptation of large language models. volume abs/2601.21003, 2026. doi: 10.48550/ARXIV.2601.21003. URL https://doi.org/10.48550/arXiv.2601.21003.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ.

Yi Liu, Zhihao Chen, Yanjun Zhang, Gelei Deng, Yuekang Li, Jianting Ning, and Leo Yu Zhang. " do not mention this to the user": Detecting and understanding malicious agent skills in the wild. *arXiv preprint arXiv:2602.06547*, 2026a.

Yi Liu, Weizhe Wang, Ruitao Feng, Yao Zhang, Guangquan Xu, Gelei Deng, Yuekang Li, and Leo Zhang. Agent skills in the wild: An empirical study of security vulnerabilities at scale. *arXiv preprint arXiv:2601.10338*, 2026b.

Zhengxi Lu, Zhiyuan Yao, Jinyang Wu, Chengcheng Han, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. SKILL0: in-context agentic reinforcement learning for skill internalization. *CoRR*, abs/2604.02268, 2026. doi: 10.48550/ARXIV.2604.02268. URL https://doi.org/10.48550/arXiv.2604.02268.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche (eds.), *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pp. 2:1–2:22. ACM, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pp. 627–635. JMLR.org, 2011. URL http://proceedings.mlr.press/v15/ross11a/ross11a.pdf.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. volume abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=ehfRiF0R3a.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, and Honggang Qi. Skill-sd: Skill-conditioned self-distillation for multi-turn LLM agents. *CoRR*, abs/2604.10674, 2026. doi: 10.48550/ARXIV.2604.10674. URL https://doi.org/10.48550/arXiv.2604.10674.

Ruoyao Wang, Peter A. Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 11279–11298. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.775. URL https://doi.org/10.18653/v1/2022.emnlp-main.775.

Jinyang Wu, Shuo Yang, Zhengxi Lu, Fan Zhang, Yuhao Shen, Lang Feng, Haoran Luo, Zheng Lian, Shuai Zhang, Zhengqi Wen, and Jianhua Tao. SEED: self-evolving on-policy distillation for agentic reinforcement learning. *CoRR*, abs/2607.14777, 2026. doi: 10.48550/ARXIV.2607.14777. URL https://doi.org/10.48550/arXiv.2607.14777.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

APPENDIX

A Evidence Provenance and Reproduction Boundaries

This reconstruction uses the original manuscript source and two supplied Markdown files: the real-results revision plan and the accompanying paper-ready experiment tables. New AGWC numbers are treated as reported experimental aggregates, as requested. They have not been independently recomputed from experiment logs. Design targets, suggested ranges, and proposed analyses are not treated as observed results.

Table 7: Evidence available for the reported comparisons and information still needed for full reproduction.

Evidence	Available	Missing or unresolved
Historical audit	Original LaTeX, tables, PDFs, provenance	Independent re-execution
New main experiments	Aggregate adaptive and fixed-k tables	Per-pool/seed/game results; raw counts
Shadow validation	AUROC, correlation, errors, mean effects	Realized sample size, scores, labels, class threshold
Scaling and ablation	Aggregate settings and differences	Run-level effects and compute decomposition
Fresh retraining	Aggregate results and described fresh protocol	Seed IDs, game IDs, split/locking records
Controller	Sequential method and paired design	Executed thresholds, optimizer, model, training lengths

Commit accounting. The source reports 86.67% of runs with non-empty commit and 15.38% rollback among non-empty selected runs. With 15 runs, 86.67% corresponds to 13 committed runs, while 15.38% naturally corresponds to 2 rollbacks among 13 non-empty selected runs. The latter would imply 11 commits, not 13. Different denominators could resolve this, but are unspecified. Neither percentage is used as evidence for deployment frequency in the main text; both are retained here for reconciliation rather than silently reinterpreted.

Baseline provenance. Several Outcome-only values equal the original manuscript's reference values, and Random selection percentages also recur. These could represent legitimate reuse or new runs with matching aggregates; the supplied material does not establish which. A reproduction release must identify checkpoints, evaluation IDs, and whether reference training was repeated under the new protocol. The new tables are not used to claim 15 independently retrained Outcome-only baselines.

Statistical unit and uncertainty. For each headline contrast, retain method, candidate-pool ID, training-seed ID, paired game ID, reference outcome, written-policy outcome, selected-set ID, and committed status. Report per-pool differences and a paired hierarchical bootstrap interval. If selection is shared across three training seeds, do not treat those repeated labels as independent selector samples. No intervals or synthetic raw samples are generated in this paper.

Implementation details still required. Record model/checkpoint, candidate generator, all data splits, h, full-training steps, optimizer settings, rank and regression weights, acceptance/commit thresholds, stopping predicate, handling of ties, and budget charges for repeats. Direct-Validation requires an explicit definition and a record of the data and compute it accesses. For fixed-k, distinguish 18 selected writes from writes actually deployed after rollback.

Historical evidence stays separate. The old frozen 36-candidate pools, all-zero paired rollout audits, and frozen-adapter transfer results are historical diagnostics. They are not pooled with the new 288-candidate experiments. The new 36-candidate scaling row is also a separate experiment, not a replacement estimate for the old audit. The original supplementary source is preserved in the source archive under `historical/` for provenance.

B Additional Reporting Notes

The complete score audit, shadow validation, candidate scaling, and ablation matrices appear below in this appendix.

Auxiliary budget summary. The source reports $+4.80$ pp for AGWC versus Direct-Validation under equal total budget. No method-specific SR values, absolute budget, stopping rules, or compute ledger accompany this summary. It is recorded here as reported, without implying an independently audited efficiency advantage. Unique probe coverage is 18.75% at $N = 288$ and 9.38% at $N = 576$ for $W = 64$; repeats can consume pairs without increasing unique coverage.

C Historical Frozen-Pool Audit

Table 8: Original ALFWorld results, preserved separately from the new protocol. These numbers describe the old frozen candidate pools and adapters. All rates are percentages.

Method	AR	Proxy-HUR	P@A	Dev SR	Inst. SR	Scen. SR
Outcome-only	–	–	–	62.14	48.48	17.14
No gate	100.00	75.00	16.67	52.75	44.78	25.71
Random	50.00	81.48	12.96	61.49	53.87	32.38
Utility	50.00	50.00	33.33	53.40	41.75	20.00
Single	50.00	66.67	25.93	56.63	48.82	40.00
Mean	50.00	57.41	33.33	49.19	35.35	27.62
Cert-Min	50.00	57.41	31.48	50.16	49.16	35.24
Robust-QD	50.00	57.41	33.33	59.55	49.49	29.52
Utility-Strict	31.48	20.59	52.94	57.61	52.19	24.76
Robust-QD+Utility	31.48	32.35	52.94	59.22	46.46	20.00
Hybrid-QD+Utility	50.00	50.00	33.33	56.31	46.13	32.38

Finite support. At the old 50% acceptance point, the frozen-pool proxy floor is 27/54, or 50%. Utility reaches it. Robust-QD's 31/54 harmful accepts improve on Single's 36/54, but cannot establish superiority over the pool's best feasible proxy selection. This floor is specific to that pool and its label.

Proxy/behavior separation. Single achieves 42/105 scenario successes (40.00%) versus Robust-QD's 31/105 (29.52%). The original paired behavioral sensitivity intervals for Robust-QD minus Single and Robust-QD minus Utility cross zero. The historical result therefore establishes a separation of observed proxy and behavioral orderings, not a population-level ranking of methods.

Causal diagnostic boundary. The original short and long paired rollout audits completed 44 and 24 pairs with all-zero binary-success differences. They did not identify a nondegenerate candidate-level success effect. The new shadow/full-write summary addresses a different, stronger predictive question, but raw candidate labels are still needed to reproduce it.

Frozen-adapter transfer is not fresh retraining. The original disjoint held-out evaluation reuses the old adapters: Outcome-only records 38/303 successes, Utility 31/303, Single 57/303, and Robust-QD 51/303. It has no fresh harm labels. The new fresh-transfer table changes training seeds as well as games and must remain a separate evidence tier.

D Metric and Figure Reproduction

Compile `main.tex` for the complete manuscript. Run `python3 make_figures.py` to regenerate the bar chart and optional diagnostic plots; `draw_controller.py` maintains the separate vector model diagram. Figure generation requires Matplotlib and NumPy; compiling requires a LaTeX installation with the declared packages.

All plotted values are reported aggregates or labeled arithmetic derivations. No new experiments were run in producing the manuscript.

E Additional Experimental Diagnostics

Table 9: Shadow-to-full validity: predictive information beyond ranking. Accepted candidates are a subset of the evaluated candidates.

Metric	All	AGWC-accepted
Full-write marginal ΔSR (pp)	+3.20	+7.60
Full-write BReg (%)	4.20	1.80
Probe/full-write AUROC	0.74	–
Probe/full-write Spearman	0.48	–
Effect-sign agreement (%)	72.50	78.57
Probe-to-full MAE (pp)	3.10	2.20
AUROC difference over score-only	+0.14	–
Fraction with $\lvert\text{full } \Delta\text{SR}\rvert > 1$ pp (%)	58.33	68.57

The probe ranks potential writes but still has nonzero effect error; final validation therefore checks the selected set.

Table 10: Candidate supply with fixed $k = 18$ across five pool sizes. Rates are percentages.

N	AR	Proxy-HUR	P@A	Dev SR	Inst. SR	Scen. SR	BReg
36	50.00	38.89	44.44	57.30	46.20	28.60	11.40
72	25.00	27.78	55.56	59.80	49.40	33.70	8.20
144	12.50	16.67	72.22	63.20	53.60	38.90	5.60
288	6.25	5.56	88.89	66.40	57.80	43.80	3.20
576	3.13	8.33	83.33	65.90	57.20	43.60	3.30

The selected set improves as supply grows to 288 candidates; the supplied aggregates are nearly flat at 576.

Table 11: Component ablations relative to AGWC-full. Cost is relative to AGWC-full.

Variant	ΔScen. SR	ΔBReg	ΔAR	Cost/full
AGWC-full	0.00	0.00	0.00	1.00
Without adaptive stopping	-5.80	+4.60	+5.56	1.28
Without paired shadow probe	-9.40	+8.60	+9.72	0.38
Without commit/rollback	-6.90	+7.80	+11.11	0.84
Without context re-evaluation	-5.60	+5.40	+6.25	0.72

The reported unique probe coverage is 18.75% at $N = 288$ and 9.38% at $N = 576$ under $W = 64$. Re-evaluations consume pairs without increasing unique coverage. Ablation differences are variant minus AGWC-full; acceptance and compute also change across variants.