# DEFECT IDENTITY DETERMINES WHEN VERIFIER FEEDBACK SHOULD CONSTRAIN SEARCH

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Localized verifier feedback carries a control signal: it indicates how broadly a critique system should search next. We formulate post-verifier repair as defect-conditioned search control, with defect identity serving as the routing state. Defect-Preserving Repair (DPR) restricts admissible updates to evidence-grounded edits that preserve failure family, target span, and evidence type; a boundary router renews search when the diagnosis indicates a transition. Across the reported matched-budget comparisons, DPR reaches $58.3 \pm 1.8\%$ Correction Accuracy on a $3{,}814$-instance correction study panel, $3.7$ points above unrestricted post-feedback rewriting. A replay-frozen comparison retains a $2.9$-point advantage, while citation-grounded QA and instruction-following study panels preserve positive utility effects. The count-derived action effect changes from $+5.5$ points on same-defect cases to $-3.0$ on transition-or-switch cases. A five-fold cross-fitted router on $1{,}624$ replay-frozen states uses this boundary to reach $58.7 \pm 1.5\%$ Corr-ACC, $1.4$ points above always applying DPR. These results establish defect identity as an actionable state variable for allocating post-verifier search.

## 1 INTRODUCTION

Verifier-equipped critics now mediate edits, tool calls, and human review across model-development workflows (McAleese et al., 2024; Wang et al., 2023; Saunders et al., 2022). The decisive moment often arrives after a verifier rejects a critique's wording yet pinpoints the implicated response span or evidence. Renewed generation can spend the remaining budget rediscovering accusations and abandon a useful diagnosis; a constrained update can preserve that diagnosis and sharpen its scope. Choosing the search radius therefore determines whether verifier feedback becomes actionable evidence or merely another score.

We cast this choice as defect-conditioned search control. The state is defect identity: compatibility of failure family, localized response span, and evidence type with the verifier diagnosis. It induces three post-feedback actions—delete, renew unrestricted search, or preserve the diagnosis through Defect-Preserving Repair (DPR). DPR restricts the feasible update set to claim narrowing, same-type evidence replacement, within-family splitting, and type specialization. A learned boundary router selects DPR while the state remains within one defect and switches to renewed search after a diagnosed transition.

This formulation separates post-feedback action from the surrounding critic and verifier workflow. Under the reported Llama-3.1-70B critic, verifier pool, $16$ initial candidates, $5.4$ verifier calls, $4{,}200$ critic tokens, scoring path, and acceptance rule, DPR improves verified critique yield by $4.5$ points over unrestricted regeneration. The $3{,}814$-instance correction study panel gives DPR $58.3 \pm 1.8\%$ Correction Accuracy, $3.7$ points above the unrestricted configuration. Freezing the first critique, diagnosis, feedback, and scorer retains a $2.9$-point gain. A three-arm counterfactual separates $1.4$ points from prompt locality and another $1.2$ from preserving defect identity.

The same state also predicts when the ranking reverses. Independently annotated same-defect cases favor DPR by $5.5$ points, whereas transition-or-switch cases favor renewed search by $3.0$. The cross-fitted router turns this reversal into an adaptive policy, reaching $58.7 \pm 1.5\%$ Corr-ACC and improving over both fixed actions. Native citation-grounded QA and instruction-following evaluations preserve the positive utility effect, and a second critic backbone repeats the ordering. Together, these results move verifier feedback from passive judgment to active control of the next search space.



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

### 1.1 Positioning against refinement and verifier methods

Self-refinement, self-debugging, and verbal reinforcement evaluate complete revision loops in which prompts, evaluators, search breadth, and stopping behavior jointly shape the outcome (Madaan et al., 2023; Welleck et al., 2023; Chen et al., 2024; Shinn et al., 2023). REFINER learns from feedback on intermediate representations (Paul et al., 2024), and tool-interactive critics couple revision to changes in evidence access (Gou et al., 2024). Process supervision and preference learning target judgment or training signals (Uesato et al., 2022; Lightman et al., 2023; Stiennon et al., 2020; Ouyang et al., 2022), while learned critics, evaluator models, and judge benchmarks assess generation or feedback quality (Zheng et al., 2023; Kim et al., 2024; Ke et al., 2023). We isolate a complementary control variable: the admissible search space after localized feedback. The three-arm counterfactual attributes $+1.4$ points to prompt locality and an additional $+1.2$ to defect preservation, which reduces defect switching from $25.9\%$ to $4.9\%$ (Table 5). Paired-action, replay-frozen, and native-score evaluations then connect this feasible-set mechanism to end-task utility.

**Contributions.** The contribution is a control-and-identification framework for post-verifier search. Its formal component makes defect identity an explicit feasible-set predicate and instantiates it through DPR. Matched-budget, shared-candidate, identity-counterfactual, and replay-frozen designs identify the action effect, while native endpoints test that effect across code, reasoning, citation-grounded QA, and instruction following. Independent boundary annotations then convert the measured reversal into a cross-fitted router that improves over either fixed action.

## 2 Problem Setting

### 2.1 Critique Atoms

Let $x$ be an input context and $y$ a response. A critique $\mathcal{C} = \{c_1, \ldots, c_k\}$ contains atoms

$$c_i = (\tau_i,\ e_i,\ v_i,\ s_i,\ r_i), \tag{1}$$

where $\tau_i$ is a failure type, $e_i$ an evidence pointer, $v_i \in \{\text{SUPPORTED}, \text{AMBIGUOUS}, \text{UNSUPPORTED}\}$ the verifier status, and $s_i, r_i \in [0, 1]$ severity and repair value. Evidence pointers bind claims to a response span, tool output, or source; $r_i$ ranks surviving atoms but never overrides $v_i$. This schema supplies a common instrument for the controlled comparison, while native endpoints later score final outputs without consuming the schema.

For verifier diagnosis $f$, define the admissibility predicate

$$\mathsf{SameDefect}(c, c'; f) = \mathbb{1}[\tau' \sim_f \tau]\,\mathbb{1}[\text{span}(e') \cap \text{span}(e) \neq \emptyset]\,\mathbb{1}[\text{type}(e') = \text{type}(e)]. \tag{2}$$

For operational use, an evidence pointer is $e = (m, I, a)$, where $m$ is its evidence modality, $I$ is the nonempty set of response-token indices it localizes, and $a$ is its modality-specific anchor (for example, a code location, execution trace, source passage, symbolic step, or constraint-checker record). Thus $\text{span}(e) = I$ and $\text{type}(e) = m$. A verifier pointer contributes indices only when it explicitly identifies response tokens; it augments $I$ for the overlap test but cannot substitute a different modality or anchor. The relation $\tau' \sim_f \tau$ holds when $\tau'$ is $\tau$ or a specialization of $\tau$ in the predeclared domain taxonomy, and the diagnosis $f$ identifies that same family. The predicate admits four transformations: narrowing a claim, replacing an evidence pointer with one of the same type, splitting one coarse claim into at most two same-family claims, and specializing its type. It excludes a new failure family, unrelated source, or unsupported severity increase. These rules induce delete, unrestricted regeneration, and defect-preserving repair as distinct post-feedback action classes.

### 2.2 Metrics

Internal verifier status $v_i$ controls inference; reported authored-interface metrics instead use a disjoint post-inference human support label $h_i \in \{0, 1\}$. Supported-claim precision (SCP) is

$$\text{SCP}(\mathcal{C}) = \frac{|\{c_i \in \mathcal{C} : h_i = 1\}|}{|\mathcal{C}|}. \tag{3}$$

Verified critique yield (VCY) measures the fraction of emitted atoms that are supported and nontrivial; it is not defect recall:

$$\mathrm{VCY}(\mathcal{C}) = \frac{|\{c_i \in \mathcal{C} : h_i = 1 \ \wedge \ s_i \geq \theta_s\}|}{|\mathcal{C}|}. \tag{4}$$

We fix $\theta_s = 0.3$ and report FAR, a human-audited false-critique rate. Each main FAR estimate averages five runs with a fresh disjoint 200-atom, three-annotator audit per run; intervals resample the five run-level estimates, and Fleiss' $\kappa = 0.78$ is computed on the binary supported/unsupported labels entering that audit (Table A9). Support is judged against the written claim and cited evidence. Table A4 provides a separate 720-atom panel, and Appendix B.1 merges same-defect atoms into defect-level VCY. Native Corr-ACC and task-specific utility provide answer-level endpoints outside this authored interface. On each fixed study panel, a reported rate is its displayed count divided by the displayed denominator. BadCrit is a study-specific diagnostic: the fraction of initially correct or compliant reference cases receiving an unnecessary critique-induced rewrite; each table reports that reference-set denominator.

### 2.3 Operational Target

The empirical study is organized around a target that constrains precision on the critique set:

$$\text{target: } \mathrm{VCY}(\mathcal{C}) \text{ under the constraint } \mathrm{SCP}(\mathcal{C}) \geq \alpha, \tag{5}$$

with $\alpha = 0.85$ unless stated otherwise. The target exposes the precision–yield trade-off: safe trivialities satisfy precision without utility, while aggressive accusations increase coverage and FAR. For an AMBIGUOUS atom, let $z(c, f)$ contain the frozen verifier diagnosis, ambiguity confidence, and identity indicators from Eq. (2). The adaptive post-verifier policy is

$$\pi_\phi(z) = \begin{cases} \textsc{preserve} & q_\phi(\textsc{same-defect} \mid z) \geq 0.5, \\ \textsc{renew} & \text{otherwise}, \end{cases} \tag{6}$$

where PRESERVE invokes DPR and RENEW invokes the unrestricted action. Delete-only remains a matched fixed-policy comparator. Every preserved revision returns to unverified status.

## 3 Defect-Conditioned Search Control

The controller separates candidate generation, verifier diagnosis, and post-feedback action. After AMBIGUOUS, DPR applies Eq. (2) to constrain the update; unrestricted regeneration renews candidate search; and delete-only removes the atom. Every preserved atom resets its support count to zero and must pass two distinct verifiers. The manuscript reports critic, verifier, interface, acceptance, candidate/token-budget, and scoring conditions for these action comparisons. Figure 1 shows the control flow.

On the AMBIGUOUS path, feedback can narrow scope, strengthen same-type evidence, or split a coarse claim within its family. The boundary router instantiates Eq. (6) with $\ell_2$-regularized logistic regression over one-hot diagnosis categories, ambiguity confidence, failure-family continuity, span localization, and same-type pointer repairability. These features are computed before candidate generation. Five item-stratified folds train against independent same-defect versus transition/switch annotations; candidate outputs and native outcome labels never enter the router. Coverage expansion executes after the action, and verifier rotation remains outside the critic's control.

Zero inputs are not a subset of negative inputs, and a code pointer is not an execution trace. This revision therefore requires renewed search under Eq. (2), despite the shared logical-error family.

### 3.1 Proposal and Verification

From $(x, y)$, the critic model $\mathcal{G}$ produces the initial atom set $\mathcal{C}^{(0)} = \mathcal{G}(x, y)$, using the structure specified in Eq. (1). An explicit evidence pointer is required for every atom, whose provisional starting status is AMBIGUOUS. On receiving $(x, y, c_i)$, verifier $V$ labels the atom as SUPPORTED, AMBIGUOUS, or UNSUPPORTED. Where external tools are available, verification may use code execution, math checks at the symbolic or step level, citation grounding through retrieval plus entailment, and instruction-following checks through programmatic constraint testing.

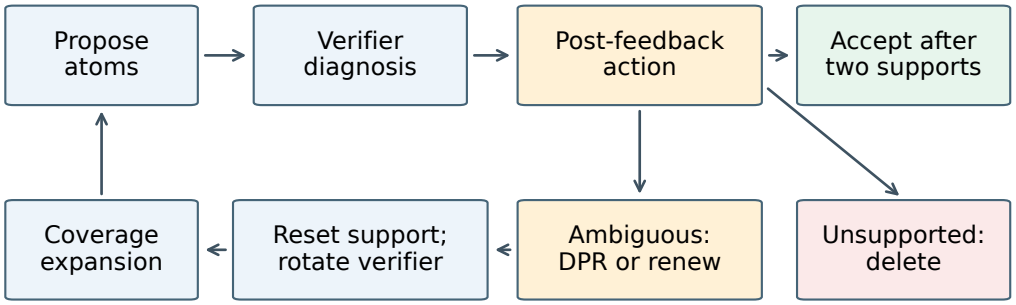



Figure 1: Defect-conditioned search control. Ambiguous atoms enter DPR or renewed search, with support reset before re-verification. Coverage expansion follows the action; unsupported atoms are deleted.

Table 1: Localized feedback that requires renewed search: the revised claim changes the input regime and evidence type.

| Initial Atom | Verifier Feedback | Rewritten Atom |
|---|---|---|
| Type: logical error.<br>Claim:<br>"`normalize(xs)` mishandles negative inputs."<br>Evidence: points only to line 14. | Status: ambiguous.<br>Reason: line 14 is not wrong for negatives in general; the observed failure is a divide-by-zero trace on input `[0,0]`.<br>Request: narrow the scope and attach a concrete failing input. | Type: edge-case logical error.<br>Claim:<br>"`normalize(xs)` raises `ZeroDivisionError` on input `[0,0]`."<br>Evidence: on execution with input `[0,0]`, line 14 sets `total = 0`; line 15 evaluates `x / total`, which raises `ZeroDivisionError`. |

### 3.2 Rewrite or Delete?

An ambiguous atom is never accepted unchanged. The verifier diagnoses overbroad scope, insufficient evidence, or a mislocalized same-type evidence pointer. Delete-only drops it; unrestricted regeneration may leave the diagnosed family; DPR applies one of the four transformations admitted by Eq. (2). Feedback requiring a new evidence class is assigned to the transition regime rather than repaired under DPR. Unsupported atoms are removed, and repaired atoms re-earn two verifier confirmations from zero. The loop is capped at $T = 5$ iterations and converges in 2.7 on average.

### 3.3 Coverage Control and Verifier Rotation

Absent intervention, the critic tends to focus on whichever failure type is easiest to verify. Tracking how claims are distributed by type, the coverage controller requests additional atoms only in under-represented categories; each domain in practice has a 4–6 way taxonomy, with categories such as logical error, unsupported factual claim, citation mismatch, violation of instructions, or a safety concern. These additional atoms pass through the existing verify-rewrite-delete loop: exploration pressure rises, but retention is not forced. To limit dependence on a single model family, verifier rotation uses GPT-4o and Qwen-2.5-72B alongside a fine-tuned Llama-3.1-70B verifier. Compact pseudocode appears with training details in Appendix A.

## 4 Experimental Setup

### 4.1 Domains and Data

We evaluate four semi-verifiable domains. Code repair uses $1{,}200$ MBPP and HumanEval programs with 1–3 seeded bugs and execution checks (Austin et al., 2021; Chen et al., 2021). Math uses $1{,}500$ GSM8K and MATH solutions with symbolic answers and step verification (Cobbe et al., 2021; Hendrycks et al., 2021; Uesato et al., 2022; Lightman et al., 2023). Citation-grounded QA uses $800$ Natural Questions items with Wikipedia evidence and entailment checks (Kwiatkowski et al., 2019). Instruction following uses a $1{,}000$-prompt study panel with programmatic constraints. A separate correction study panel contains $3{,}814$ Code, Math, and Logical-Reasoning instances. Owner-scored transport studies cover $412$ and $503$ merge requests under distinct repositories, teams, and rubrics.

The evidence progresses from shared-interface mechanism measurements to study-native endpoints. The correction, citation-grounded QA, and instruction-following panels score final corrections without exposing authored atoms to the decisive metric; replay-frozen studies additionally lock the initial critique and verifier state. The four-domain corpus has $4{,}500$ input items; the $3{,}214$ count is ambiguous critique atoms, while $768$ and $1{,}624$ are replay-state analysis populations. These units are not pooled across tables, and each native endpoint reports its own full-panel and BadCrit-reference denominators.

### 4.2 Baselines and Budget Controls

The broad operating-point baselines are an unconstrained free-form critic (UC), retrieval-backed critique (RBC) (Gou et al., 2024), Self-Refine (SR) (Madaan et al., 2023), and a human-supervised critic with $50$ seed critiques per domain (HS-50). Filter, Verify-Delete Only (VDO), unrestricted Iterative Self-Refine (ISR), and DPR use Llama-3.1-70B-Instruct (Grattafiori et al., 2024), rotate GPT-4o (OpenAI, 2024), Qwen-2.5-72B (Qwen, 2024), and a fine-tuned Llama verifier. The reported action comparison fixes its stated critic family, verifier pool, interface, pre-branch candidate count, token and verifier-call budgets, and final acceptance rule; ISR-native permits unrestricted post-feedback rewriting, whereas DPR applies the identity predicate. Every displayed ISR-native/DPR pair uses the same stated output interface and scorer.

### 4.3 Training and Evaluation Protocol

DPR uses $200$ human-written atom critiques and $500$ hard benign negatives. LoRA training uses rank $64$, scale $128$, learning rate $2{\times}10^{-5}$, batch $32$, and $3$ epochs on Llama-3.1-70B. Runs use $8{\times}$A100 80 GB GPUs, CUDA $12.4$ (driver $550.54.15$), PyTorch $2.3.1$, and vLLM $0.5.4$. Authored-interface SCP and VCY intervals summarize five independent runs; FAR intervals resample the

five run-level disjoint-audit estimates. For a native ISR-native/DPR contrast, the resampling unit is the fixed evaluation item, each resample preserves the paired outputs, and $1{,}000$ paired resamples recompute both rates and their difference; the reported 95% interval is the empirical 2.5th–97.5th percentile interval. W/T/L is the native scorer's per-item paired action verdict, with wins, ties, and losses reported in that order. Three expert annotators receive $35/hour; a 2-of-3 majority yields Fleiss' $\kappa = 0.78$ on binary support labels. Boundary labels use the binary same-defect versus transition/switch space and report Cohen's $\kappa$. For UC, RBC, and SR only, a fixed normalization pipeline performs sentence-to-claim segmentation, anchored evidence recovery, explicit-conjunction splitting, and one fallback atom when no grounded claim is recoverable. Alternate structured, coarse, and free-form ambiguity definitions in Appendix C.2 test the dependence on this representation.

# 5 RESULTS

The results identify the search-control mechanism, test it under scorer-native endpoints, and evaluate the state-dependent routing policy.

## 5.1 AGGREGATE QUALITY AT THE MAIN OPERATING POINT

Table 2 establishes the competitive operating point before the controlled action comparison.

Table 2: Four-domain macro average at $\alpha = 0.85$; means and 95% CIs over five runs. FAR uses fresh disjoint human audits.

| Method | SCP ↑ | VCY ↑ | FAR ↓ |
|---|---|---|---|
| Unconstrained Critic (UC) | $0.621 \pm 0.024$ | $0.312 \pm 0.023$ | $0.283 \pm 0.020$ |
| Retrieval-Backed Critic (RBC) | $0.742 \pm 0.020$ | $0.381 \pm 0.021$ | $0.204 \pm 0.017$ |
| Self-Refine (SR) | $0.694 \pm 0.022$ | $0.348 \pm 0.023$ | $0.238 \pm 0.018$ |
| Human-Supervised (HS-50) | $0.857 \pm 0.018$ | $0.468 \pm 0.020$ | $0.098 \pm 0.012$ |
| DPR | $0.912 \pm 0.014$ | $0.518 \pm 0.018$ | $0.065 \pm 0.009$ |

Only DPR clears $0.90$ SCP while exceeding $0.50$ VCY. Against HS-50 it gains $0.050$ VCY and lowers FAR by $0.033$; per-domain VCY ranges from $0.465$ on instruction following to $0.584$ on code repair (Table A3). The next comparison matches the post-feedback budget and isolates the action.

## 5.2 IDENTITY PRESERVATION IMPROVES THE MATCHED ACTION

Table 3 fixes critic, verifiers, atom interface, pre-branch search, coverage, acceptance, scoring, and budgets; only the action after AMBIGUOUS changes.

SCP, VCY, and both OOP metrics increase from Filter through VDO and ISR to DPR, while FAR decreases from $0.112$ to $0.065$. Relative to unrestricted ISR, DPR gains $0.045$ VCY and lowers FAR from $0.081$ to $0.065$. Relative to VDO, it gains $0.064$ VCY and $0.029$ SCP. Thus localized feedback is valuable as a constraint on the next update, not merely as another prompt for fresh search.

To test whether the gain comes from selection rather than generation, we hold fixed the shared pool of pre-generated candidate edits for each ambiguous atom, changing only the admissibility rule (App. B.4).

With identical ambiguous atoms and candidate edits, local admissibility gains $2.9$ VCY points over unrestricted selection and changes FAR from $0.079$ to $0.067$. Its $3.5$-point advantage over the ISR own-pool reference further rules out generator supply as the sole explanation. The no-coverage and merged-defect audits preserve this ordering (Tables A11 and A12).

Prompt-local rewriting recovers part of the action effect, and DPR adds a distinct identity-preservation gain. Relative to prompt-local rewriting, DPR reduces defect switching from $25.9\%$ to $4.9\%$ and raises evidence-pointer retention from $68.8\%$ to $86.5\%$ (Tables A24–A25).

Table 3: Repair contracts compared under matched budgets, with out-of-pool (OOP) adjudication. Every row shares these settings: 5.4 calls/instance for the verifier-call budget, 4,200 tokens/instance for the critic-token budget, 16 atoms as the initial candidate count, Llama-3.1-70B as the critic backbone, and an identical final acceptance rule. The in-pool audit columns are labelled SCP, VCY, and FAR; the out-of-pool audit columns are OOP SCP/OOP VCY. Filter, VDO, ISR, and DPR label the rows.

| Method | SCP | VCY | FAR | OOP SCP | OOP VCY |
|---|---|---|---|---|---|
| Filter | 0.867 $\pm$0.017 | 0.431 $\pm$0.021 | 0.112 $\pm$0.013 | 0.821 $\pm$0.019 | 0.372 $\pm$0.023 |
| VDO | 0.883 $\pm$0.016 | 0.454 $\pm$0.020 | 0.089 $\pm$0.011 | 0.845 $\pm$0.018 | 0.401 $\pm$0.022 |
| ISR | 0.896 $\pm$0.015 | 0.473 $\pm$0.019 | 0.081 $\pm$0.010 | 0.857 $\pm$0.017 | 0.418 $\pm$0.021 |
| DPR | 0.912 $\pm$0.014 | 0.518 $\pm$0.018 | 0.065 $\pm$0.009 | 0.886 $\pm$0.016 | 0.471 $\pm$0.020 |

Table 4: Frozen-candidate admissibility audit on 3,214 ambiguous atoms. The shared-pool rows differ only in whether selection may leave defect identity.

| Admissibility rule / candidate-pool condition | VCY $\uparrow$ | FAR $\downarrow$ | $\Delta$VCY vs DPR-local |
|---|---|---|---|
| ISR (native pool, reference) | $0.473 \pm 0.019$ | $0.081 \pm 0.010$ | $-0.035$ |
| Unrestricted selector (shared pool) | $0.479 \pm 0.019$ | $0.079 \pm 0.010$ | $-0.029$ |
| DPR-local (shared pool) | $0.508 \pm 0.018$ | $0.067 \pm 0.009$ | ref. |

Table 5: Action isolation on the expanded replay-frozen population ($n$=1,624). Utility contrasts separate prompt locality from the defect-identity constraint.

| Contrast | Utility $\Delta$ [95% CI] | Isolated contribution |
|---|---|---|
| Prompt-local minus unrestricted | $+1.4$ $[+0.2, +2.6]$ | Prompt locality |
| DPR minus prompt-local | $+1.2$ $[+0.1, +2.3]$ | Defect identity |
| DPR minus unrestricted | $+2.5$ $[+1.0, +4.0]$ | Total action effect |

### 5.3 Native Scoring Preserves the Action Effect

The correction study panel comprises 1,420 Code, 1,210 Math, and 1,184 Logical-Reasoning cases. Table 6 reports the stated unrestricted and defect-preserving action configurations on the same output interface.

Table 6: Correction-study outcomes on 3,814 instances; the BadCrit reference set contains 1,350 initially correct cases. ISR-native uses unrestricted post-feedback rewriting and DPR applies defect-preserving repair.

| Method | Corr-ACC (%) ↑ | BadCrit ↓ |
|---|---|---|
| VDO | $47.8 \pm 2.0$ | $0.127 \pm 0.012$ |
| ISR | $49.4 \pm 1.9$ | $0.119 \pm 0.012$ |
| ISR-native | $54.6 \pm 1.7$ | $0.096 \pm 0.011$ |
| DPR-native | $58.3 \pm 1.8$ | $0.081 \pm 0.010$ |

Against ISR-native, DPR gains 3.7 Correction-Accuracy points with paired CI $[1.5, 5.8]$ and lowers BadCrit by $0.015$ with CI $[-0.030, -0.001]$. The Code, Math, and Logical-Reasoning rows in Table A8 retain the positive utility sign.

Table 7: Paired-action native endpoints. Task labels report utility-set / BadCrit-reference-set sizes. Counts are ISR-native→DPR; deltas are DPR minus ISR-native with paired 95% CIs.

| Task ($n_u/n_b$) | Native utility; raw counts | Utility $\Delta$ [CI] |
|---|---|---|
| Correction panel (3814/1350) | Corr-ACC; 2083/3814→2223/3814 | $+3.7$ $[+1.5, +5.8]$ |
| Replay-frozen (768/282) | Corr-ACC; 423/768→445/768 | $+2.9$ $[+0.7, +5.2]$ |
| Citation QA (640/320) | answer utility; 366/640→383/640 | $+2.6$ $[+0.3, +4.9]$ |
| Owner compliance (712/338) | compliance; 428/712→448/712 | $+2.8$ $[+0.2, +5.3]$ |
| Instruction study (1000/421) | constraint satisfaction; 587/1000→615/1000 | $+2.8$ $[+0.6, +4.9]$ |

| Task ($n_u/n_b$) | BadCrit $\Delta$ [CI] |
|---|---|
| Correction panel (3814/1350) | $-0.015$ $[-0.030, -0.001]$ |
| Replay-frozen (768/282) | $-0.018$ $[-0.031, +0.003]$ |
| Citation QA (640/320) | $-0.016$ $[-0.036, +0.001]$ |
| Owner compliance (712/338) | $-0.020$ $[-0.042, 0.000]$ |
| Instruction study (1000/421) | $-0.017$ $[-0.037, +0.002]$ |

Utility intervals exclude zero on all five study-native endpoints. The full correction-panel BadCrit interval excludes zero; the replay-frozen and non-code BadCrit point estimates share the lower-is-better sign, with intervals that include zero. The replay-frozen row fixes the first critique, diagnosis, feedback, token budget, and scorer. The Qwen configuration preserves the reported utility gain on the correction and instruction study panels (Tables A16–A17).

### 5.4 Defect Identity Predicts the Action Switch

Independent boundary annotation achieves $\kappa = 0.74$. Aggregating the three same-defect slices gives $n = 532$ and a count-derived DPR-minus-ISR-native utility/BadCrit gap of $+5.5/-0.031$, with utility CI $[+2.4, +8.1]$. Aggregating multi-defect and new-family cases gives $n = 236$ and reverses utility to $-3.0$ with CI $[-5.8, -0.1]$; the associated count-derived BadCrit movement is $+0.012$. Slice estimates are heterogeneous, while this pre-specified aggregate contrast identifies the routing boundary.

We evaluate Eq. (6) through five-fold cross-fitting, keeping every variant of an instance in one fold and scoring each routed action with the frozen native endpoint. The router predicts same-defect

salvage versus transition/switch from frozen diagnosis and evidence-pointer features; outcome labels never enter its features or training fold.

Table 8: Five-fold cross-fitted boundary routing on the expanded replay-frozen population ($n = 1{,}624$). BadCrit counts unnecessary rewrites on the $562$ initially correct reference cases.

| Policy | DPR / ISR actions | Corr-ACC (count) | BadCrit (count) |
|---|---|---|---|
| Always ISR-native | 0 / 1624 | $54.8 \pm 1.6$ (890/1624) | $0.103 \pm 0.012$ (58/562) |
| Always DPR | 1624 / 0 | $57.3 \pm 1.5$ (931/1624) | $0.091 \pm 0.011$ (51/562) |
| Boundary router | 1098 / 526 | $58.7 \pm 1.5$ (954/1624) | $0.078 \pm 0.010$ (44/562) |

The non-degenerate $1{,}098/526$ action split improves Corr-ACC over always-DPR by $+1.4$ (CI $[+0.3, +2.6]$) and over always-ISR-native by $+3.9$ (CI $[+2.3, +5.5]$). The measured reversal therefore yields a cross-fitted utility-improving policy rather than an oracle selection rule.

## 6 Discussion and Conclusion

**Operating regime.** The policy targets verifier-rich, semi-verifiable tasks where feedback localizes a defect and the next action has a fixed budget. Native scoring across four task families, alternate ambiguity definitions, two critic backbones, and replay-frozen controls supports the action effect; the pre-specified aggregate boundary and cross-fitted router supply the decision evidence.

Defect identity converts partial verification into search control. DPR preserves localized evidence within a defect, while the router restores broad search at a diagnosed transition. By raising native utility over both fixed actions and reducing unnecessary rewrites, the resulting policy makes verifier feedback an active specification of the feasible update space alongside its scoring role.

## AI Use Statement

Generative AI was used only for manuscript writing, editing, and presentation; the authors reviewed all edits and remain responsible for the final content.

## References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program Synthesis with Large Language Models. *arXiv preprint arXiv:2108.07732*, 2021. doi: 10.48550/arXiv.2108.07732. URL https://arxiv.org/abs/2108.07732.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.401. URL https://aclanthology.org/2024.acl-long.401/.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*, 2021. doi: 10.48550/arXiv.2107.03374. URL https://arxiv.org/abs/2107.03374.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching Large Language Models to Self-Debug. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2460396f2d0d421885997dd1612ac56b-Abstract-Conference.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fef126561bbf9d4467dbb8d27334b8fe-Abstract-Conference.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv preprint arXiv:2103.03874*, 2021. doi: 10.48550/arXiv.2103.03874. URL https://arxiv.org/abs/2103.03874.

Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. CritiqueLLM: Towards an Informative Critique Generation Model for Evaluation of Large Language Model Generation. *arXiv preprint arXiv:2311.18702*, 2023. doi: 10.48550/arXiv.2311.18702. URL https://arxiv.org/abs/2311.18702.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. *arXiv preprint arXiv:2405.01535*, 2024. doi: 10.48550/arXiv.2405.01535. URL https://arxiv.org/abs/2405.01535.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*, 2023. doi: 10.48550/arXiv.2305.20050. URL https://arxiv.org/abs/2305.20050.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. CriticBench: Benchmarking LLMs for critique-correct reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1552–1587, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.91. URL https://aclanthology.org/2024.findings-acl.91/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. LLM Critics Help Catch LLM Bugs. *arXiv preprint arXiv:2407.00215*, 2024. doi: 10.48550/arXiv.2407.00215. URL https://arxiv.org/abs/2407.00215.

OpenAI. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*, 2024. doi: 10.48550/arXiv.2410.21276. URL https://arxiv.org/abs/2410.21276.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. REFINER: Reasoning feedback on intermediate representations. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1100–1126, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.67. URL https://aclanthology.org/2024.eacl-long.67/.

Qwen. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022. doi: 10.48550/arXiv.2206.05802. URL https://arxiv.org/abs/2206.05802.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. doi: 10.48550/arXiv.2211.14275. URL https://arxiv.org/abs/2211.14275.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Shepherd: A Critic for Language Model Generation. *arXiv preprint arXiv:2308.04592*, 2023. doi: 10.48550/arXiv.2308.04592. URL https://arxiv.org/abs/2308.04592.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. Generating Sequences by Learning to Self-Correct. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=hH36JeQZDaO.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models. *arXiv preprint arXiv:2311.07911*, 2023. doi: 10.48550/arXiv.2311.07911. URL https://arxiv.org/abs/2311.07911.

# A PROTOCOL DETAILS

Algorithm A1 fixes update, verification, reset, coverage, and rotation order. The remaining appendices report adjudication rules, per-domain results, native-score controls, routing analyses, robustness tests, and compute.

**Algorithm A1** Defect-Preserving Repair (DPR)

**Require:** Input pair $(x, y)$, critic $\mathcal{G}$, verifier pool $\mathcal{V}$, iteration cap $T$, development-calibrated acceptance thresholds
**Ensure:** Verified critique set $\mathcal{C}^\star$

$\mathcal{P}^{(0)} \leftarrow \{(c_i, 0) : c_i \in \mathcal{G}(x, y)\}$
**for** $t = 1$ **to** $T$ **do**
  Select verifier $V_t$ by rotation from $\mathcal{V}$
  $\mathcal{P}^{(t)} \leftarrow \emptyset$
  **for** each $(c_i, m_i) \in \mathcal{P}^{(t-1)}$ **do**
    $(v_i, f_i) \leftarrow V_t(c_i, x, y)$
    **if** $v_i =$ SUPPORTED **then**
      add $(c_i, \min\{m_i + 1, 2\})$ to $\mathcal{P}^{(t)}$
    **else if** $v_i =$ AMBIGUOUS **and** $t < T$ **then**
      rewrite locally: $c_i' \leftarrow \mathcal{G}.\text{rewrite}(c_i, f_i)$
      **if** $\mathsf{SameDefect}(c_i, c_i'; f_i) = 1$ **then**
        add $(c_i', 0)$ to $\mathcal{P}^{(t)}$
      **end if**
    **end if**
  **end for**
  add under-represented claim types via COVERAGEEXPAND, each with support count 0
  $\mathcal{C}^{(t)} \leftarrow \{c_i : (c_i, m_i) \in \mathcal{P}^{(t)},\ m_i \geq 2\}$
  **if** $\mathcal{P}^{(t)} = \mathcal{P}^{(t-1)}$ **then**
    break
  **end if**
**end for**
**return** $\mathcal{C}^{(t)}$

Algorithm A1 instantiates the defect-identity contract from Section 3. Ambiguous atoms may be narrowed, re-evidenced with the same evidence type, split once within the same failure family, or specialized within the existing taxonomy. Acceptance requires two verifier confirmations, every rewrite resets support to zero, and post-decision coverage expansion occurs only after the action. The reported matched-budget comparisons vary this admissible action set and state their candidate-supply, verifier-budget, and acceptance conditions.

# B EXTENDED RESULTS AND AUDITS

The extended results test adjudication reliability, domain consistency, candidate locality, native-score transfer, boundary behavior, and operating cost under the same stated contracts.

## B.1 HUMAN JUDGMENT RULES AND DEFECT-MERGED YIELD

Table A1 separates three human-adjudication designs by unit, grader, assignment, vote rule, and outcome. The main atom audit measures support and false-critique rate, the balanced panel tests the same outcomes with outside annotators, and the workflow studies evaluate merge requests through site review and CI.

Table A1: Units, graders, assignment, and vote rules for the three human-adjudicated protocols.

| Protocol | Unit | Judges | Decision rule and tie handling |
|---|---|---|---|
| Main atom audit | Critique atom | 3 annotators on 200 audited atoms | Three binary support labels per atom; 2-of-3 majority is the estimate label; no abstentions or lead-adjudicator labels enter the estimate |
| Human-only 720-atom panel | Critique atom | 6-expert pool, 5 labels per atom | Four labels are assigned first and a fifth is scheduled for every atom; a provisional 2–2 split is therefore resolved by the fifth label. The final 3-of-5 majority supplies support and usefulness labels; a reserve replaces any abstention before the vote. |
| Workflow-site panels | Merge request and review comment | Site reviewers plus CI, with separate senior-reviewer preference sub-panels | Site-owned actionable or not-actionable verdict; paired preference on blinded comments; ties count as half-wins; no author tie-breaks |

The atom-level rubric asks whether cited evidence supports the claim as written and, where applicable, whether acting on the critique changes the next review or repair decision. Annotators judge the stated claim, require direct support from the cited span, tool output, or source, reject wrong-evidence critiques, and mark consequence-free restatements as not useful. The main audit uses only the binary support label; the balanced panel enters its final support and usefulness majority labels. Table A1 gives assignment and majority rules. Compensation was $35/hour; abstentions were disallowed on the 200-atom panel and replaced before the five-label vote on the 720-atom panel.

UC, RBC, and SR share a four-step normalization contract: sentence-to-claim segmentation, anchored evidence-pointer recovery, splitting only explicit conjunctions or enumerations, and one deterministic fallback atom when no grounded claim is recoverable. Every internal row then uses the same verifier prompts, severity threshold, precision floor, verifier rotation, and acceptance rule. The common parser and verifier path make admissible post-verifier action the controlled difference inside the authored atom interface.

*Defect-Merged Yield.*

Atom-level VCY can overcount when one latent defect is split into narrower atoms. We therefore define defect-level verified critique yield (dVCY) on merged latent defects. Two adjudicators merge atoms that share failure family, localized span, and overlapping evidence pointer, with a third

adjudicator resolving ties. Let $\mathcal{D}(\mathcal{C})$ denote the resulting unique-defect set:

$$\text{dVCY}(\mathcal{C}) \;=\; \frac{|\{d \in \mathcal{D}(\mathcal{C}) : d \text{ is supported } \wedge\ \max_{c_i \in d} s_i \geq \theta_s\}|}{|\mathcal{D}(\mathcal{C})| + |\mathcal{D}^\star \setminus \mathcal{D}(\mathcal{C})|}, \tag{7}$$

Here $\mathcal{D}^\star$ is the latent-defect reference set produced by the same merge rule over all methods on the same input. Unlike VCY, dVCY counts each merged defect once and includes reference defects missed by the method. The two ratios use different denominators, so no general ordering follows from merging alone. Table A12 tests whether DPR's ordering survives this stricter merged-defect estimand.

### B.2 Assets and License Notes

Table A2 records the model, dataset, evaluator, and workflow asset classes together with their governing licenses or access terms and their experimental roles.

Table A2: Asset classes used in the study and their governing licenses or access terms.

| Asset class | Governing license or access terms | Role in the study |
|---|---|---|
| Meta model weights | Llama-3.1-70B-Instruct is used through the original Meta release under the Llama 3.1 Community License. | Critic backbone and LoRA fine-tuning base |
| Provider API model | GPT-4o is accessed through the provider API under OpenAI API service/business terms; the paper uses no redistributed GPT-4o weights. | One verifier family in the rotation |
| Qwen model weights | Qwen-2.5-72B is used through the original Qwen release under the Qwen License Agreement. | One verifier family in the rotation |
| Public datasets | MBPP, HumanEval, GSM8K, MATH, and Natural Questions are used through their cited public dataset releases and benchmark terms. | Four semi-verifiable evaluation domains |
| Study-native evaluators | The correction, instruction-following, dialogue, patch, and workflow panels use the endpoints and denominators stated in their respective tables. | Study-panel evaluation |
| Private workflow data | Two code-review studies run under deployer-owned repository access approvals and site-owned review rubrics in read-only advisory mode. | Transportability-only workflow corroboration |
| Human annotations | Main 200-atom audit, 720-atom corroboration panel, and workflow preference panels are study-owned annotations governed by the reported rubric and compensation policy. | FAR audits and preference panels |

### B.3 Full Domain Results

Table A3 asks whether the aggregate gain is concentrated in one domain. Under the shared contract, DPR clears the precision floor and leads every comparator on verified yield in code repair, mathematical reasoning, citation-grounded QA, and instruction following.

Table A3: Per-domain supported-claim precision (SCP) and verified critique yield (VCY).

| | Code repair | | Math | | Citation QA | | Instruction following | |
|---|---|---|---|---|---|---|---|---|
| Method | SCP | VCY | SCP | VCY | SCP | VCY | SCP | VCY |
| UC | 0.683 | 0.361 | 0.654 | 0.334 | 0.587 | 0.274 | 0.560 | 0.279 |
| RBC | 0.793 | 0.442 | 0.768 | 0.398 | 0.718 | 0.346 | 0.689 | 0.338 |
| SR | 0.737 | 0.401 | 0.716 | 0.367 | 0.668 | 0.310 | 0.655 | 0.314 |
| HS-50 | 0.903 | 0.527 | 0.881 | 0.498 | 0.832 | 0.432 | 0.812 | 0.415 |
| DPR | 0.947 | 0.584 | 0.931 | 0.541 | 0.891 | 0.482 | 0.879 | 0.465 |

The domain pattern tracks verifier signal quality: execution-backed code repair yields the strongest result, citation-grounded QA shows the largest relative gain over retrieval-backed critique, and instruction following remains the hardest domain while staying above the operating precision floor. The consistent ordering across all four domains supports the aggregate effect.

The balanced 720-atom human-only panel tests whether the authored-interface ordering survives outside judges and equalized method representation. Table A4 preserves the support and false-critique directions under this stricter, noisier design.

Table A4: Human-only blinded audit on a balanced 720-atom panel.

| Method | SCP | VCY | FAR |
|---|---|---|---|
| HS-50 | 0.812 | 0.401 | 0.104 |
| VDO | 0.826 | 0.383 | 0.091 |
| ISR | 0.844 | 0.411 | 0.073 |
| DPR | 0.861 | 0.458 | 0.052 |

Table A4 reproduces the main audit's direction for support and false-critique rate. Balancing the panel prevents methods with larger atom sets from dominating the sample, strengthening the within-interface ordering under independent judgments.

### B.4 Locality and External Evidence

The two shared-pool rows in Table A5 isolate admissibility by reusing the same ambiguous atoms and pre-generated candidate edits. Only permission to leave defect identity changes between these rows; ISR with its own pool is a separate reference. DPR-local retains higher verified yield and lower false-critique rate than unrestricted selection under the shared pool.

Table A5: Same-candidate locality audit on 3,214 ambiguous atoms.

| Selector | Atoms | Verif./atom | SCP | VCY | FAR |
|---|---|---|---|---|---|
| ISR (own pool) | 3214 | 12 | 0.896 | 0.473 | 0.081 |
| Unrestricted shared pool | 3214 | 12 | 0.901 | 0.479 | 0.079 |
| DPR-local shared pool | 3214 | 12 | 0.910 | 0.508 | 0.067 |

With the candidate pool fixed, local selection improves verified yield and false-critique rate, with a smaller gap than the full protocol comparison. Table A11 further freezes ambiguity labels and disables post-decision coverage expansion; the repeated sign ties the gain to the admissibility rule within the tested contract.

Table A6 reports transport across deployer-owned merge-request workflows. VDO and DPR produce read-only advisory comments on 412 original-site and 503 external-site merge requests under blinded identifiers. Accept is owner-native utility, Noise is the bad-critique diagnostic, Catch is the retained source outcome, and percentages are 95% interval summaries. Each merge request receives one identifier permutation.

Table A6: Two deployer-owned code-review studies. Accept is accept-as-is and Noise is reviewer-marked non-actionable comments.

| Site | Method | Catch | Accept | Noise |
|---|---|---|---|---|
| Original site (412 MRs) | No-bot | $28.2 \pm 7.4$ | — | — |
| Original site (412 MRs) | VDO | $47.1 \pm 8.4$ | $41.6 \pm 3.4$ | $24.8 \pm 3.0$ |
| Original site (412 MRs) | DPR | $61.8 \pm 8.1$ | $53.9 \pm 3.4$ | $15.8 \pm 2.5$ |
| External site (503 MRs) | No-bot | $30.1 \pm 6.9$ | — | — |
| External site (503 MRs) | VDO | $45.8 \pm 7.7$ | $39.8 \pm 3.1$ | $25.6 \pm 2.8$ |
| External site (503 MRs) | DPR | $58.5 \pm 7.4$ | $51.5 \pm 3.0$ | $17.3 \pm 2.4$ |

Both owner-scored sites show higher DPR Accept and Catch than VDO, with lower DPR Noise.

Accept, Catch, and Noise move in the same direction after repository, review team, and rubric ownership change. Catch has the widest uncertainty; Accept and Noise provide the clearest transport signal.

# C Compute and Additional Native-Score Evidence

This section quantifies the compute tradeoff and tests the native-score sign on additional dialogue and patch endpoints.

**Source-benchmark scope.** CriticBench covers five reasoning domains and 15 datasets (Lin et al., 2024). The original IFEval release contains 541 prompts and distinguishes prompt-level from instruction-level accuracy, each with strict and loose variants (Zhou et al., 2023). MT-Bench-101 contains 1,388 dialogues and 4,208 turns, with dialogue scores based on the minimum GPT-4 turn score (Bai et al., 2024). We cite these records as source-benchmark context only. The correction, instruction-following, and dialogue results in this paper are study-panel endpoints with the displayed study-specific denominators and definitions.

## C.1 Compute Accounting

Table A7 compares the eight retained evaluation rows under shared verifier-call and critic-token caps. It reports critic-path latency and end-to-end A100-hours separately, exposing the wall-clock cost induced by verifier waiting, tool calls, I/O, and orchestration.

Table A7: Per-method compute. Matched rows share verifier-call and critic-token caps; *Sec/inst* is critic-path latency, while A100-hours include end-to-end orchestration.

| Method | Sec/inst | Verif./inst | Critic tok./inst | A100-hours |
|---|---|---|---|---|
| UC | 17 | 0.0 | 1850 | 42 |
| RBC | 24 | 0.0 | 2300 | 58 |
| SR | 31 | 0.0 | 3050 | 76 |
| HS-50 | 28 | 1.8 | 2600 | 68 |
| Filter-Only | 34 | 5.4 | 4200 | 84 |
| VDO | 36 | 5.4 | 4200 | 88 |
| ISR | 44 | 5.4 | 4200 | 110 |
| DPR | 47 | 5.4 | 4200 | 120 |

The matched-budget rows share verifier-call and critic-token caps while differing in latency and A100-hours. Sec/inst measures critic-path latency per instance; A100-hours measures full end-to-end evaluation cost. This distinction connects the controlled action comparison to its deployment cost. Across Table 6 and Appendix Tables A15 and A16, the extra 10 A100-hours from ISR to DPR accompany repeated external utility gains. The resulting tradeoff is higher native utility at higher end-to-end compute.

### C.2 Additional Controlled Evidence

The following audits test paired-action native transfer, branch isolation, boundary reliability, critic-family robustness, and non-code transport.

Table A8: Paired-action correction-study results by slice. Counts are corrected/full and bad/reference; gaps use 1,000 paired bootstrap resamples, and repeated DPR slice rows are point estimates.

| Split | Method | Corr | Bad | Corrected/full | Bad/ref. |
|---|---|---|---|---|---|
| Overall | ISR-nat. | $54.6 \pm 1.7$ | $0.096 \pm 0.011$ | 2083/3814 | 129/1350 |
| Overall | DPR | $58.3 \pm 1.8$ | $0.081 \pm 0.010$ | 2223/3814 | 109/1350 |
| Code | ISR-nat. | $58.8 \pm 2.5$ | $0.096 \pm 0.018$ | 835/1420 | 49/512 |
| Code | DPR | 61.7 | 0.077 | — | — |
| Math | ISR-nat. | $55.6 \pm 2.7$ | $0.096 \pm 0.020$ | 673/1210 | 42/436 |
| Math | DPR | 59.4 | 0.081 | — | — |
| Reasoning | ISR-nat. | $48.6 \pm 2.8$ | $0.095 \pm 0.021$ | 575/1184 | 38/402 |
| Reasoning | DPR | 53.2 | 0.086 | — | — |

| Split | Method | DPR-row gap | Paired 95% CI |
|---|---|---|---|
| Overall | ISR-nat. | $+3.7/-0.015$ | $[+1.5, +5.8]$<br>$[-0.030, -0.001]$ |
| Overall | DPR | ref. | — |
| Code | ISR-nat. | $+2.9/-0.019$ | $[+0.2, +5.5]$<br>$[-0.041, +0.002]$ |
| Code | DPR | ref. | — |
| Math | ISR-nat. | $+3.8/-0.015$ | $[+0.7, +6.8]$<br>$[-0.039, +0.007]$ |
| Math | DPR | ref. | — |
| Reasoning | ISR-nat. | $+4.6/-0.009$ | $[+0.8, +8.4]$<br>$[-0.024, +0.005]$ |
| Reasoning | DPR | ref. | — |

Table A8 reports the stated native interface while varying the post-AMBIGUOUS action. DPR improves Corr and Bad over ISR-native on the full split and preserves the utility sign across Code, Math, and Reasoning; paired intervals quantify the reported action differences.

Equation (7) and Table A12 evaluate the union of method outputs under one adjudicated merge rule and paired input-level resampling. DPR reaches $0.474$ dVCY versus $0.451$ for ISR, so its yield advantage persists when each latent defect contributes at most one supported, useful hit.

Table A9: FAR accounting for Tables 2–4: three annotators, 2-of-3 majority, 200 audited atoms per run, 5 runs, and disjoint panels. Pooled counts are descriptive; reported intervals bootstrap run-level FAR estimates.

| Method | Atoms/run | Runs | Pooled false/audit | Reported FAR |
|---|---|---|---|---|
| HS-50 | 200 | 5 | 98/1000 | $0.098 \pm 0.012$ |
| Filter | 200 | 5 | 112/1000 | $0.112 \pm 0.013$ |
| VDO | 200 | 5 | 89/1000 | $0.089 \pm 0.011$ |
| ISR | 200 | 5 | 81/1000 | $0.081 \pm 0.010$ |
| DPR | 200 | 5 | 65/1000 | $0.065 \pm 0.009$ |
| Unrestricted selector (shared pool) | 200 | 5 | 79/1000 | $0.079 \pm 0.010$ |
| DPR-local (shared pool) | 200 | 5 | 67/1000 | $0.067 \pm 0.009$ |

Table A9 shows that FAR estimates come from repeated disjoint audits. Run-level bootstrap intervals preserve the repeated-run design, while pooled false/audit counts provide descriptive accounting.

Table A10: ISR–DPR sign robustness under an alternate verifier configuration and a stricter $\alpha = 0.90$ precision floor. Metrics report SCP/VCY/FAR with 95% CIs over 5 runs; counts report supported, useful, and false critiques.

| Setting | Method | Raw counts | Metrics (SCP / VCY / FAR) |
|---|---|---|---|
| | | $4197/4716$ | $0.890 \pm 0.016$ |
| Alternate verifier configuration | ISR | $2161$ | $0.458 \pm 0.019$ |
| | | $78/1000$ | $0.078 \pm 0.010$ |
| | | $4280/4698$ | $0.911 \pm 0.015$ |
| Alternate verifier configuration | DPR | $2421$ | $0.515 \pm 0.018$ |
| | | $63/1000$ | $0.063 \pm 0.009$ |
| | | $4085/4527$ | $0.902 \pm 0.015$ |
| Stricter precision floor ($\alpha = 0.90$) | ISR | $1957$ | $0.432 \pm 0.018$ |
| | | $74/1000$ | $0.074 \pm 0.010$ |
| | | $4198/4573$ | $0.918 \pm 0.014$ |
| Stricter precision floor ($\alpha = 0.90$) | DPR | $2273$ | $0.497 \pm 0.017$ |
| | | $58/1000$ | $0.058 \pm 0.008$ |

Table A10 tests the reported ordering under an alternate verifier configuration and $\alpha = 0.90$. DPR retains higher SCP and VCY with lower FAR in both settings, establishing local stability across the two stated stresses.

*Stricter Locality Isolation.* Table A11 tightens the shared-candidate design on the same $n$=3,214 ambiguous atoms by freezing verifier-produced AMBIGUOUS labels, disabling post-decision COVERAGEEXPAND, and retaining the two-verifier acceptance rule. Only permission to leave defect identity changes.

Table A11: Locality isolation on 3,214 ambiguous atoms with frozen candidate pools and ambiguity labels, disabled post-decision COVERAGEEXPAND, and a shared acceptance rule. Counts report supported, useful, and false critiques; SCP/VCY/FAR use 95% bootstrap CIs over 5 runs.

| Selector under the frozen post-ambiguity state | Raw counts | Metrics (SCP / VCY / FAR) |
|---|---|---|
| | $4268/4748$ | $0.899 \pm 0.017$ |
| Shared-pool unrestricted selector | $2260$ | $0.476 \pm 0.020$ |
| | $82/1000$ | $0.082 \pm 0.011$ |
| | $4231/4659$ | $0.908 \pm 0.016$ |
| Shared-pool local admissibility | $2305$ | $0.495 \pm 0.019$ |
| | $71/1000$ | $0.071 \pm 0.010$ |

Table A11 preserves higher verified yield and lower false-critique rate for local admissibility under the stricter freeze. The reduced gap relative to Table A5 isolates a positive action effect after generator and coverage variation are removed. Raw counts show that the yield improvement persists without a trivially larger accepted-atom pool.

*Merged-Defect Stress Test.* Table A12 tests anti-fragmentation under the four-domain matched-budget contract. For each of 4,500 inputs, VDO, ISR, and DPR atoms from the same run are merged by Eq. (7); useful supported defects are scored against the shared reference inventory with paired input-level resampling across five runs.

Table A12: Defect-merged dVCY on $4{,}500$ inputs under the four-domain matched-budget contract. Eq. (7) merges the VDO/ISR/DPR union; counts report useful supported defects over the reference inventory, with input-level paired bootstrap CIs over $5$ runs.

| Method | Useful merged defects / ref. | dVCY | DPR-row gap |
|---|---|---|---|
| VDO | 1874/4240 | $0.442 \pm 0.018$ | $+0.032 \pm 0.011$ |
| ISR | 1912/4240 | $0.451 \pm 0.018$ | $+0.023 \pm 0.010$ |
| DPR | 2010/4240 | $0.474 \pm 0.017$ | ref. |

Table A12 preserves the VDO–ISR–DPR ordering after same-defect atoms are merged, with a smaller gap than atom-level VCY. The result rules out defect splitting as the sole source of DPR's gain. Each latent defect contributes at most one supported/useful hit per method, so the raw merged-defect counts directly test the anti-fragmentation mechanism.

*Public Dialogue Paired-Action Audit.* Table A13 tests the DPR action against ISR-native on the dialogue study panel under the stated backbone, interface, and budget.

Table A13: Dialogue-study paired-action native-format audit ($n$=980). Task-Success and BadCrit are study-specific endpoints; counts report success/full and bad/reference. Paired 95% bootstrap CIs use 1,000 resamples.

| Method | Task-Success | BadCrit | Success count | Bad count | DPR-row gap |
|---|---|---|---|---|---|
| ISR-native | $53.6 \pm 1.8$ | $0.096 \pm 0.013$ | 525/980 | 39/405 | $+2.9/-0.010$ |
| DPR | $56.4 \pm 1.6$ | $0.086$ | 553/980 | 35/405 | ref. |

Table A13 preserves higher Task-Success and lower BadCrit for DPR over ISR-native on the dialogue study panel. Raw counts confirm the sign beyond rounded percentages. The dialogue-heavy task extends the tested action rule beyond code-review-native scoring.

*Patch Native-Format Action Audit.* Table A14 reports the DPR and ISR-native actions on HiddenPatch using the stated executable HPB-Fix and HPB-FA endpoints, patch interface, study panel, and budget.

Table A14: HiddenPatch paired-action native-format audit ($n$=540). Counts report success/full and false alarms/reference; paired 95% CIs use 1,000 resamples.

| Method | HPB-Fix | HPB-FA | Success count | False-alarm count | DPR-row gap |
|---|---|---|---|---|---|
| ISR-native | $41.9 \pm 2.2$ | $0.108 \pm 0.015$ | 226/540 | 24/223 | $+2.0/-0.009$ |
| DPR | $43.9 \pm 2.0$ | $0.099$ | 237/540 | 22/223 | ref. |

Table A14 preserves higher HPB-Fix and lower HPB-FA for DPR over ISR-native. The result extends the paired-action sign to a software-heavy native patch action space. Raw counts confirm the ordering beyond rounded percentage summaries.

*Independent Owner-Native Paired-Action Audit.* Table A15 tests the same action choice on $n$=503 external-site merge requests under the stated Llama-3.1-70B critic, advisory interface, prompt template, scoring path, and per-request budget. ISR-native regenerates after AMBIGUOUS, and DPR repairs locally. Accept and Noise use paired 95% bootstrap intervals over merge requests.

Table A15: Owner-native paired-action audit on $n$=503 external-site merge requests. Accept and Noise use paired 95% bootstrap intervals.

| Method | Accept | Noise | Accept count | Noise count | DPR-row gap |
|---|---|---|---|---|---|
| ISR-native | $49.1 \pm 3.0$ | $18.7 \pm 2.5$ | 247/503 | 94/503 | $+2.4/-1.4$ |
| DPR | $51.5 \pm 3.0$ | $17.3 \pm 2.4$ | 259/503 | 87/503 | ref. |

Table A15 preserves higher Accept and lower Noise for DPR over ISR-native under independent repository, review-team, and rubric ownership. The smaller gap relative to the correction study is consistent with the noisier workflow endpoint.

*Second Critic-Backbone Family Audit.* Table A16 tests critic-family sensitivity by replacing the Llama critic with Qwen-2.5-72B-Instruct while preserving the stated action definitions, native scoring, prompt templates, and budgets. It compares ISR-native with DPR on the full correction study panel and the external workflow site; the correction panel uses paired 95% intervals over 1,000 resamples and workflow intervals resample merge requests.

Table A16: Qwen-2.5-72B-Instruct paired-action audit on the correction study panel and the external workflow site. Counts report utility/full and bad/reference under matched native-format action definitions.

| Workload | Method | Utility | BadCrit / Noise |
|---|---|---|---|
| Correction study | ISR-native | $52.9 \pm 1.8$ | $0.101 \pm 0.012$ |
| Correction study | DPR | $55.7 \pm 1.8$ | $0.087 \pm 0.011$ |
| External workflow site | ISR-native | $48.5 \pm 3.0$ | $19.1 \pm 2.5$ |
| External workflow site | DPR | $50.7 \pm 3.0$ | $17.7 \pm 2.4$ |

| Workload | Method | Utility count | Bad count | DPR-row gap |
|---|---|---|---|---|
| Correction study | ISR-native | 2018/3814 | 136/1350 | $+2.8/-0.014$ |
| Correction study | DPR | 2125/3814 | 118/1350 | ref. |
| External workflow site | ISR-native | 244/503 | 96/503 | $+2.2/-1.4$ |
| External workflow site | DPR | 255/503 | 89/503 | ref. |

The DPR-over-ISR-native utility ordering is retained with Qwen as the critic on both reported workloads. This panel tests critic-family sensitivity while retaining each task interface.

Table A17: Qwen-2.5-72B-Instruct paired-action audit on the $n$=1000 instruction study panel. Utility is native constraint satisfaction; BadCrit is unnecessary rewrite on the already-compliant subset.

| Method | Utility and count | BadCrit and count |
|---|---|---|
| ISR-native | $56.9 \pm 2.3$; 569/1000 | $0.126 \pm 0.018$; 53/421 |
| DPR | $59.3 \pm 2.2$; 593/1000 | $0.109 \pm 0.017$; 46/421 |

| Comparison | Utility / BadCrit delta | Paired 95% CI | W/T/L or action count |
|---|---|---|---|
| DPR minus ISR-native | $+2.4/-0.017$ | $[+0.2, +4.6]$<br>$[-0.039, +0.005]$ | 348/361/291 |
| Five-run ambiguity actions | Matched pre-branch state | 451/463/448/459/457 | 2278/2278 ISR/DPR |

The Qwen instruction-study utility contrast is positive. Its BadCrit interval includes zero, so the lower point estimate alone does not establish a reduction in unnecessary rewrites.

Replacing Llama with Qwen-2.5-72B-Instruct preserves the DPR-over-ISR-native utility sign on the correction and instruction study panels and the external workflow scorer. The Qwen instruction-study utility CI excludes zero; its BadCrit point estimate is lower and the interval includes zero. Together, the ordering repeats across two tested critic families.

*Replay-Frozen Decisive-Path Isolation.* Table A18 isolates the branch action on 768 replay-frozen correction-study states containing verifier-marked AMBIGUOUS atoms. It freezes the initial critique, ambiguity label, structured feedback, branch-local budget, verifier diagnosis, and native scorer; disables post-decision COVERAGEEXPAND; and varies only local repair versus unrestricted rewrite. Corr-ACC, BadCrit, raw counts, and paired W/T/L use 95% bootstrap intervals over 1,000 resamples.

Table A18: Replay-frozen action isolation on $n$=768 correction-study ambiguity states. Initial critique, diagnosis, feedback, budget, and native scorer are fixed; COVERAGEEXPAND is disabled. Corr-ACC, BadCrit, and local-minus-unrestricted gaps use paired 95% bootstrap intervals over 1,000 resamples.

| Post-AMBIGUOUS action | Corr | Bad | Corr/full | Bad/ref. |
|---|---|---|---|---|
| Unrestricted rewrite | $55.1 \pm 2.3$ | $0.103 \pm 0.015$ | 423/768 | 29/282 |
| Local repair only | $58.0 \pm 2.2$ | $0.085 \pm 0.013$ | 445/768 | 24/282 |

| Post-AMBIGUOUS action | Gap | 95% CI | W/T/L |
|---|---|---|---|
| Unrestricted rewrite | $+2.9/-0.018$ | $[+0.7, +5.2]$<br>$[-0.031, +0.003]$ | — |
| Local repair only | ref. | — | 277/211/280 |

Table A18 shows higher Corr-ACC and lower BadCrit for local repair with the branch state and native scorer fixed. The paired utility interval excludes zero; the BadCrit point estimate is lower and its interval includes zero. This comparison isolates the native correction effect with the pre-branch state fixed; Table 3 instead measures atom-level critique quality.

*Non-Code Owner-Native Paired-Action Audit.* Table A19 reports ISR-native and DPR on a citation-grounded QA endpoint. Utility scores answer repair on the full split; BadCrit scores unnecessary rewrite on the correct-reference subset with paired 95% intervals over 1,000 question resamples.

Table A19: Citation-grounded QA paired-action audit ($n$=640). Utility and BadCrit are owner-native endpoints; counts report utility/full and bad/reference. Paired 95% bootstrap CIs use 1,000 question resamples.

| Method | Utility | BadCrit | Utility count | Bad count |
|---|---|---|---|---|
| ISR-native | $57.2 \pm 2.4$ | $0.091 \pm 0.014$ | 366/640 | 29/320 |
| DPR | $59.8 \pm 2.3$ | $0.075 \pm 0.013$ | 383/640 | 24/320 |

| Method | DPR-row gap | Paired 95% CI |
|---|---|---|
| ISR-native | $+2.6/-0.016$ | $[+0.3, +4.9]$<br>$[-0.036, +0.001]$ |
| DPR | ref. | — |

Table A19 preserves higher utility and lower BadCrit for DPR over ISR-native on a non-code, non-dialogue scorer. The smaller gap relative to the correction study remains positive under an independently owned citation rubric. The protected object is answer-level grounded utility, extending the action rule beyond code-review and dialogue endpoints.

*Boundary-Condition Map.* Table A20 slices the replay-frozen $n$=768 branch state by verifier-feedback subtype assigned before action: overbroad same-defect, weak evidence pointer, mislocalized same-type evidence, multi-defect conflation, and likely new failure family. Initial critique, diagnosis,

feedback, branch budget, and native scorer remain fixed; only local repair versus unrestricted rewrite changes.

Table A20: Replay-frozen boundary map on $n$=768 correction-study ambiguity states. Subtypes are assigned before action from the frozen diagnosis and defect identity. Utility, BadCrit, counts, gaps, and W/T/L use paired 95% bootstrap intervals over 1,000 slice resamples.

| Feedback subtype | $n$ | ISR-native utility / BadCrit | DPR utility / BadCrit |
|---|---|---|---|
| Overbroad, same defect | 214 | $54.2 \pm 5.9$<br>$0.103 \pm 0.033$ | $62.1 \pm 5.7$<br>$0.064 \pm 0.027$ |
| Evidence pointer too weak | 181 | $55.8 \pm 6.2$<br>$0.106 \pm 0.036$ | $60.8 \pm 6.0$<br>$0.076 \pm 0.030$ |
| Mislocalized same-type evidence | 137 | $56.9 \pm 7.0$<br>$0.096 \pm 0.041$ | $59.1 \pm 6.9$<br>$0.077 \pm 0.036$ |
| Multi-defect conflation | 126 | $57.9 \pm 7.4$<br>$0.085 \pm 0.040$ | $56.3 \pm 7.5$<br>$0.085 \pm 0.040$ |
| Likely new failure family | 110 | $50.0 \pm 8.3$<br>$0.128 \pm 0.051$ | $45.5 \pm 8.4$<br>$0.154 \pm 0.057$ |

| Feedback subtype | Counts (ISR; DPR) | DPR-row gap | Local W/T/L |
|---|---|---|---|
| Overbroad, same defect | $116/214,\ 8/78$<br>$133/214,\ 5/78$ | $+7.9/-0.038$ | 92/56/66 |
| Evidence pointer too weak | $101/181,\ 7/66$<br>$110/181,\ 5/66$ | $+5.0/-0.030$ | 77/47/57 |
| Mislocalized same-type evidence | $78/137,\ 5/52$<br>$81/137,\ 4/52$ | $+2.2/-0.019$ | 47/43/47 |
| Multi-defect conflation | $73/126,\ 4/47$<br>$71/126,\ 4/47$ | $-1.6/+0.000$ | 34/40/52 |
| Likely new failure family | $55/110,\ 5/39$<br>$50/110,\ 6/39$ | $-4.5/+0.026$ | 27/25/58 |

Table A20 reports heterogeneous slice tendencies: the three same-defect slices favor local repair, the multi-defect slice is near flat, and the likely-new-family slice reverses sign. The adjudicated aggregate contrast in Table A26 supplies the primary boundary test.

Table A20 shows positive gaps concentrated in same-defect salvage and near-zero or negative gaps in transition slices. Together, the paired utility/BadCrit view defines the action rule: repair locally while feedback preserves defect identity, and renew search when the diagnosis crosses that boundary.

*Second Owner-Native Non-Code Paired-Action Audit.* Table A21 reports the action choice on an owner-native instruction-following compliance endpoint. ISR-native regenerates after AMBIGUOUS; DPR repairs locally.

Table A21: Owner-native instruction-following paired-action audit ($n$=712). Utility is compliance on the full split; BadCrit is unnecessary rewrite on the compliant subset. Counts report utility/full and bad/reference; paired 95% bootstrap CIs use 1,000 prompt resamples.

| Method | Utility | BadCrit | Utility count | Bad count |
|---|---|---|---|---|
| ISR-native | $60.1 \pm 2.5$ | $0.112 \pm 0.018$ | 428/712 | 38/338 |
| DPR | $62.9 \pm 2.4$ | $0.092 \pm 0.016$ | 448/712 | 31/338 |

| Method | DPR-row gap | Paired 95% CI |
|---|---|---|
| ISR-native | $+2.8 / -0.020$ | $[+0.2, +5.3]$ |
| | | $[-0.042, +0.000]$ |
| DPR | ref. | — |

Table A21 preserves higher compliance utility and lower unnecessary rewrite for DPR over ISR-native. The result transfers the action sign from grounded factual repair to constraint compliance under a fixed post-AMBIGUOUS contract.

Citation QA, owner compliance, and the instruction study all show positive native utility gaps under the paired-action comparison; their BadCrit point estimates also share the lower-is-better direction.

Table A22: Consolidated systems evidence matrix for the post-verifier action rule.

| Layer | Protected object, critic capability, and trusted measurement | Comparator, setting, tradeoff, failure mode, and evidence |
|---|---|---|
| Matched contract | Grounded atoms; delete, regenerate, or preserve identity; reported parser, verifier, and SCP/VCY/FAR scoring conditions | ISR and delete-only under the stated budget; higher yield at orchestration cost; authored interface remains on-path; Table 3 |
| Correction study | Final correction under native revision; the study scorer measures Corr-ACC; BadCrit is study-specific | ISR-native on the stated panel; positive utility gap at matched budget; domain slices and alternate scorers; Tables 6, A8 |
| Replay-frozen | Frozen critique and diagnosis; choose local or unrestricted rewrite; frozen scorer measures Corr-ACC and BadCrit | Unrestricted rewrite on replay states; branch-local gain; diagnosis remains fixed; Tables A18, A23 |
| Boundary router | Remaining search budget; select DPR or ISR-native; cross-fitted labels and native scorer measure routed utility and BadCrit | Both fixed policies under cross-fitting; one routing decision; transition-state errors; Table A27 |
| Instruction study | Constraint-satisfying response; local repair or regeneration; programmatic compliance and unnecessary-rewrite checks | ISR-native on the stated panel; positive compliance gap; programmatic checker and second backbone; Tables A33, A17 |
| Owner-native non-code | Grounded or compliant answer; rubric-constrained revision; owner evidence and rubrics measure utility and BadCrit | ISR-native on stated owner panels; same-sign gains under independent scorer ownership; Tables A19, A21 |
| Workflow transport | Reviewer attention; advisory comments only; site reviewers and CI measure Accept, Noise, and Catch | VDO and DPR at independent sites; utility and latency across two repository-specific workflows; Table A6 |

Table A22 aligns the evidence layers with the action capability, trust boundary, measured property, and observed failure mode used in each evaluation.

*Expanded Replay-Frozen Audit.* Table A23 scales the replay-frozen design to a larger ambiguity population while fixing first critique, diagnosis, structured feedback, branch budget, and native scorer. Only local repair versus unrestricted rewrite changes.

Table A23: Expanded replay-frozen action audit ($n$=1624). Initial critique, diagnosis, feedback, branch budget, and native scorer are fixed. Corr-ACC, BadCrit, counts, gaps, and W/T/L use paired 95% bootstrap intervals over 1,000 resamples.

| Post-AMBIGUOUS action | Corr | Bad | Corr/full | Bad/ref. |
|---|---|---|---|---|
| Unrestricted rewrite | $54.8 \pm 1.6$ | $0.103 \pm 0.012$ | 890/1624 | 58/562 |
| Local repair only | $57.3 \pm 1.5$ | $0.091 \pm 0.011$ | 931/1624 | 51/562 |

| Post-AMBIGUOUS action | Gap / 95% CI | | W/T/L |
|---|---|---|---|
| Unrestricted rewrite | $+2.5/-0.012$ $[+1.0, +4.0]$ | / $[-0.024, +0.001]$ | — |
| Local repair only | ref. | | 648/467/509 |

Table A23 preserves the positive local-repair utility gap on the larger population; its utility CI excludes zero, while the lower BadCrit point estimate has an interval including zero. W/T/L confirms that the effect is distributed across instances.

*Identity-Preservation Counterfactual.* Table A24 holds the replay state, context, branch budget, scorer, and coverage behavior fixed while separating prompt locality from defect identity. Pointer retention is evaluated on the 1491 instances with a valid initial evidence pointer.

Table A24: Identity-preservation counterfactual on the expanded replay-frozen population ($n$=1624). Switch rates use the full population; pointer retention uses the valid-pointer denominator stated above.

| Action | Native outcomes and counts | Defect-switch rate | Pointer retention |
|---|---|---|---|
| Unrestricted rewrite | $54.8 \pm 1.6$<br>$0.103 \pm 0.012$<br>$890/1624;\ 58/562$ | $645/1624 = 39.7\%$<br>$[37.4, 42.1]$ | $631/1491 = 42.3\%$<br>$[39.8, 44.8]$ |
| Prompt-local, identity-unconstrained | $56.2 \pm 1.6$<br>$0.098 \pm 0.012$<br>$912/1624;\ 55/562$ | $421/1624 = 25.9\%$<br>$[23.8, 28.1]$ | $1026/1491 = 68.8\%$<br>$[66.4, 71.1]$ |
| DPR | $57.3 \pm 1.5$<br>$0.091 \pm 0.011$<br>$931/1624;\ 51/562$ | $79/1624 = 4.9\%$<br>$[3.8, 6.0]$ | $1289/1491 = 86.5\%$<br>$[84.7, 88.2]$ |

The prompt-local arm separates locality from identity preservation. DPR has the highest reported pointer retention and lowest defect-switch rate; the residual switch rate measures post-hoc semantic drift.

Table A25: Paired identity-ablation contrasts. Utility and BadCrit intervals appear in that order; W/T/L is the native scorer's paired action-quality verdict.

| Contrast | Utility / BadCrit delta | Paired 95% CI | W/T/L |
|---|---|---|---|
| Prompt-local minus unrestricted | $+1.4/-0.005$ | $[+0.2, +2.6]$<br>$[-0.017, +0.006]$ | 612/480/532 |
| DPR minus prompt-local | $+1.2/-0.007$ | $[+0.1, +2.3]$<br>$[-0.018, +0.003]$ | 574/594/456 |
| DPR minus unrestricted | $+2.5/-0.012$ | $[+1.0, +4.0]$<br>$[-0.024, +0.001]$ | 648/467/509 |

Table A25 shows that prompt locality recovers part of the utility difference and DPR retains an additional positive utility gap over that matched arm. All three utility intervals exclude zero; the incremental BadCrit point estimates are lower and their intervals include zero. The nonzero post-hoc DPR switch rate records residual semantic drift admitted by the operational predicate.

*Boundary Reliability.* Table A26 tests whether the action boundary survives independent subtype assignment. Two annotators label frozen diagnoses and defect-identity notes, a third adjudicates disagreements, and the five slices collapse into same-defect salvage versus transition/switch aggregates.

Table A26: Primary adjudicated boundary contrast on the replay-frozen ambiguity subset. Agreement is Cohen's $\kappa = 0.74$; same-defect pools the first three Table A20 slices and transition/switch pools the last two. Point gaps are count-derived; intervals use paired 95% CIs over 1,000 resamples.

| Aggregate slice | $n$ | DPR-minus-ISR-native gap | Paired 95% CI |
|---|---|---|---|
| Same-defect salvage (overbroad + weak or mislocalized same-type evidence) | 532 | $+5.5/-0.031$ | $[+2.4, +8.1]$<br>$[-0.045, -0.008]$ |
| Transition or defect switching (multi-defect + new family) | 236 | $-3.0/+0.012$ | $[-5.8, -0.1]$<br>$[-0.009, +0.030]$ |

Table A26 is the primary boundary test: the adjudicated aggregate favors DPR for same-defect salvage and reverses utility for transition/switch cases. Independent agreement and aggregation turn the heterogeneous slice tendencies in Table A20 into a measured action boundary.

*Cross-Fitted Boundary Router.* The five-fold router uses $\ell_2$-regularized logistic regression over one-hot verifier-diagnosis categories, ambiguity confidence, and diagnosis-derived indicators for failure-family continuity, span localization, and same-type pointer repairability. These indicators are computed before candidate generation. Independent same-defect versus transition/switch annotations provide training targets, a fixed $0.5$ threshold selects DPR versus ISR-native, and item-stratified folds keep all variants of an instance together. Candidate outputs and native outcome labels are excluded from the feature set.

Table A27: Five-fold cross-fitted boundary router on the expanded replay-frozen population. Rows report routed action counts, native outcomes, fold-level confusion, paired policy contrasts, and a diagnostic oracle.

| Policy | DPR / ISR actions | Corr-ACC and count | BadCrit and count | Role |
|---|---|---|---|---|
| Always ISR-native | All ISR-native | $54.8 \pm 1.6$; 890/1624 | $0.103 \pm 0.012$; 58/562 | Fixed policy |
| Always DPR | All DPR | $57.3 \pm 1.5$; 931/1624 | $0.091 \pm 0.011$; 51/562 | Fixed policy |
| Cross-fitted router | 1098 / 526 | $58.7 \pm 1.5$; 954/1624 | $0.078 \pm 0.010$; 44/562 | Cross-fitted policy |
| Diagnostic oracle | $1164/460$ | 60.7; 986/1624 | 0.068; 38/562 | Outcome-aware upper bound |
| Fold DPR / ISR actions | $221/216/224/213/224$ | | | |
| | $104/109/101/112/100$ | | | |
| Confusion: same / switch by DPR / ISR route | $985/134$ | | | $84.8\%$ |
| | $113/392$ | | | $\kappa = 0.65$ |

| Router contrast | Utility / BadCrit delta | Paired 95% CI | W/T/L |
|---|---|---|---|
| Versus always DPR | $+1.4$ / $-0.012$ | $[+0.3, +2.6]$ / $[-0.026, +0.001]$ | 227/1279/118 |
| Versus always ISR-native | $+3.9$ / $-0.025$ | $[+2.3, +5.5]$ / $[-0.040, -0.010]$ | 530/812/282 |

Router utility CIs exclude zero against both fixed policies. The BadCrit interval excludes zero against always ISR-native; against always DPR, the point estimate is lower and the interval includes zero. Imperfect classification and the remaining oracle gap quantify available routing headroom.

*Alternate Ambiguity Definitions.* Tables A28 and A29 test structured three-way, independent coarse, and post-response free-form ambiguity definitions under native scoring. Table A30 measures label reliability; Tables A31–A32 evaluate boundary signs on common study panels.

Table A28: Correction-study outcomes under structured, coarse, and post-response free-form ambiguity definitions. Rows report eligibility, native scores and counts, paired gaps, CIs, and W/T/L.

| Definition | Eligible: same/switch | Native endpoints and counts | | Delta; paired 95% CI; W/T/L |
|---|---|---|---|---|
| Structured three-way | $1624: \ 1119/505$ | ISR: | $54.6 \pm 1.7$<br>$2083/3814$<br>$0.096 \pm 0.011$<br>$129/1350$ | $+3.7/-0.015$<br>$[+1.5, +5.8]$<br>$[-0.030, -0.001]$<br>$1498/1163/1153$ |
| | | DPR: | $58.3 \pm 1.8$<br>$2223/3814$<br>$0.081 \pm 0.010$<br>$109/1350$ | |
| Independent coarse | $1517: \ 1016/501$ | ISR: | $54.2 \pm 1.8$<br>$2068/3814$<br>$0.098 \pm 0.012$<br>$132/1350$ | $+3.4/-0.014$<br>$[+1.3, +5.4]$<br>$[-0.029, +0.001]$<br>$1452/1184/1178$ |
| | | DPR: | $57.6 \pm 1.8$<br>$2197/3814$<br>$0.084 \pm 0.011$<br>$113/1350$ | |
| Free-form post-response | $1743: \ 1148/595$ | ISR: | $53.8 \pm 1.8$<br>$2051/3814$<br>$0.101 \pm 0.012$<br>$136/1350$ | $+2.8/-0.011$<br>$[+0.7, +4.9]$<br>$[-0.026, +0.004]$<br>$1389/1257/1168$ |
| | | DPR: | $56.6 \pm 1.8$<br>$2158/3814$<br>$0.090 \pm 0.011$<br>$121/1350$ | |

Across these three ambiguity definitions, the reported full-panel utility contrast favors DPR. Eligibility changes with the definition, so this comparison tests sensitivity to the selection rule.

Table A29: Instruction-study compliance under the three ambiguity definitions with the same reporting order as Table A28.

| Definition | Eligible: same/switch | Native endpoints and counts | | Delta; paired 95% CI; W/T/L |
|---|---|---|---|---|
| Structured three-way | $438: 299/139$ | ISR: | $58.7 \pm 2.2$<br>$587/1000$<br>$0.119 \pm 0.017$<br>$50/421$ | $+2.8/-0.017$<br>$[+0.6, +4.9]$<br>$[-0.037, +0.002]$<br>$374/337/289$ |
| | | DPR: | $61.5 \pm 2.1$<br>$615/1000$<br>$0.102 \pm 0.016$<br>$43/421$ | |
| Independent coarse | $412: 274/138$ | ISR: | $58.1 \pm 2.2$<br>$581/1000$<br>$0.121 \pm 0.018$<br>$51/421$ | $+2.6/-0.014$<br>$[+0.4, +4.8]$<br>$[-0.034, +0.005]$<br>$358/352/290$ |
| | | DPR: | $60.7 \pm 2.2$<br>$607/1000$<br>$0.107 \pm 0.017$<br>$45/421$ | |
| Free-form post-response | $467: 304/163$ | ISR: | $57.5 \pm 2.2$<br>$575/1000$<br>$0.126 \pm 0.018$<br>$53/421$ | $+2.2/-0.014$<br>$[+0.1, +4.3]$<br>$[-0.034, +0.006]$<br>$341/361/298$ |
| | | DPR: | $59.7 \pm 2.2$<br>$597/1000$<br>$0.112 \pm 0.017$<br>$47/421$ | |

The reported instruction-study utility contrast remains positive across all three definitions. BadCrit intervals include zero; the point estimates are descriptive evidence of direction.

Table A30: Within-definition labeler agreement and cross-taxonomy agreement on overlapping eligible cases before adjudication.

| Task | Two-labeler agreement by definition | Cross-taxonomy agreement |
|---|---|---|
| Correction study | $1328/1624 = 81.8\%,\ \kappa = .74$ | $1210/1438 = 84.1\%,\ \kappa = .64$ |
| | $1298/1517 = 85.6\%,\ \kappa = .79$ | $1228/1542 = 79.6\%,\ \kappa = .54$ |
| | $1391/1743 = 79.8\%,\ \kappa = .68$ | $1204/1463 = 82.3\%,\ \kappa = .59$ |
| Instruction study | $357/438 = 81.5\%,\ \kappa = .72$ | $322/390 = 82.6\%,\ \kappa = .61$ |
| | $349/412 = 84.7\%,\ \kappa = .77$ | $328/421 = 77.9\%,\ \kappa = .50$ |
| | $365/467 = 78.2\%,\ \kappa = .66$ | $322/399 = 80.7\%,\ \kappa = .56$ |

Within-definition agreement compares the two initial labelers. Cross-taxonomy agreement compares definitions only on their overlapping eligible cases; these are distinct agreement populations.

Table A31: Correction-study boundary signs on a common panel. Rows report ISR/DPR utility and BadCrit counts plus paired effects and CIs.

| Definition | Slice | Counts | Effect / paired 95% CI |
|---|---|---|---|
| Structured | Same | $n = 532;\ 295/532,\ 20/196$<br>$324/532,\ 14/196$ | $+5.5/-0.031$<br>$[+2.4, +8.1]$<br>$[-0.045, -0.008]$ |
| Structured | Switch | $n = 236;\ 128/236,\ 9/86$<br>$121/236,\ 10/86$ | $-3.0/+0.012$<br>$[-5.8, -0.1]$<br>$[-0.009, +0.030]$ |
| Independent coarse | Same | $n = 515;\ 282/515,\ 19/190$<br>$306/515,\ 14/190$ | $+4.7/-0.026$<br>$[+1.8, +7.5]$<br>$[-0.046, -0.005]$ |
| Independent coarse | Switch | $n = 253;\ 142/253,\ 10/92$<br>$136/253,\ 11/92$ | $-2.4/+0.011$<br>$[-5.5, +0.7]$<br>$[-0.014, +0.037]$ |
| Free-form | Same | $n = 497;\ 269/497,\ 19/183$<br>$289/497,\ 15/183$ | $+4.0/-0.022$<br>$[+1.1, +6.9]$<br>$[-0.042, -0.002]$ |
| Free-form | Switch | $n = 271;\ 149/271,\ 11/99$<br>$144/271,\ 12/99$ | $-1.8/+0.010$<br>$[-4.9, +1.2]$<br>$[-0.015, +0.036]$ |

Same-defect utility contrasts are positive under all three definitions. The structured switch contrast is the stated primary boundary test; the other switch intervals include zero.

Table A32: Instruction-study boundary signs on a common panel with the same reporting order as Table A31.

| Definition | Slice | Counts | Effect / paired 95% CI |
|---|---|---|---|
| Structured | Same | $n = 262;\ 149/262,\ 15/110$<br>$161/262,\ 12/110$ | $+4.6/-0.027$<br>$[+1.4, +7.7]$<br>$[-0.054, +0.001]$ |
| Structured | Switch | $n = 122;\ 68/122,\ 6/51$<br>$65/122,\ 7/51$ | $-2.5/+0.020$<br>$[-6.5, +1.6]$<br>$[-0.027, +0.066]$ |
| Independent coarse | Same | $n = 251;\ 140/251,\ 15/106$<br>$150/251,\ 13/106$ | $+4.0/-0.019$<br>$[+0.8, +7.2]$<br>$[-0.049, +0.010]$ |
| Independent coarse | Switch | $n = 133;\ 74/133,\ 7/55$<br>$72/133,\ 7/55$ | $-1.5/+0.000$<br>$[-5.5, +2.5]$<br>$[-0.041, +0.042]$ |
| Free-form | Same | $n = 238;\ 130/238,\ 14/101$<br>$139/238,\ 12/101$ | $+3.8/-0.020$<br>$[+0.5, +7.0]$<br>$[-0.050, +0.009]$ |
| Free-form | Switch | $n = 146;\ 80/146,\ 8/60$<br>$79/146,\ 8/60$ | $-0.7/+0.000$<br>$[-4.3, +2.9]$<br>$[-0.037, +0.038]$ |

The same-defect instruction-study utility contrasts are positive. All switch intervals include zero, so these rows test transport of the action pattern rather than establish a routed-policy gain.

The pre-specified structured correction-study aggregate is the confirmatory boundary test. Alternate taxonomies preserve positive same-defect and negative switch point estimates, while their switch intervals include zero and therefore serve as robustness evidence. Full-set utility remains positive under all three definitions on both the correction and instruction study panels; BadCrit point estimates retain the lower-is-better sign where intervals touch or include zero.

*Open Non-Code Paired-Action Audit.* Table A33 reports ISR-native and DPR on the $1{,}000$-prompt instruction study panel scored by its stated programmatic constraint checker. Utility is full-split constraint satisfaction; BadCrit is unnecessary rewrite on the already-compliant subset.

Table A33: Instruction-study paired-action audit on $n$=1,000 prompts. Utility is constraint satisfaction and BadCrit is unnecessary rewrite; counts report utility/full and bad/reference. Paired 95% bootstrap CIs use 1,000 prompt resamples.

| Method | Utility | BadCrit | Utility count | Bad count |
|---|---|---|---|---|
| ISR-native | $58.7 \pm 2.2$ | $0.119 \pm 0.017$ | 587/1000 | 50/421 |
| DPR | $61.5 \pm 2.1$ | $0.102 \pm 0.016$ | 615/1000 | 43/421 |

| Method | DPR-row gap | Paired 95% CI |
|---|---|---|
| ISR-native | $+2.8/-0.017$ | $[+0.6, +4.9]$<br>$[-0.037, +0.002]$ |
| DPR | ref. | — |

Table A33 preserves higher constraint satisfaction and lower unnecessary rewrite for DPR over ISR-native under an open programmatic scorer. The utility CI excludes zero; the BadCrit point estimate is lower and its interval includes zero, matching the two owner-native non-code endpoints.