PERPLEXITY-GUIDED TRAJECTORY REPAIR SHOWS NO ADVANTAGE OVER LENGTH-MATCHED CONTROLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovery from a failed agent trajectory can restart the task or keep a prefix and regenerate its suffix. We test whether stored-token perplexity helps choose where to regenerate, beyond positional controls, in ReAct-style QA agents. In a pre-specified 250-question HotpotQA study, two-step backtracking from the highest-perplexity step achieves 9.73% repair success versus 11.21% for restart (95% interval -4.13 to $+0.88$ points). A diagnosis/replay follow-up, a frozen two-model, three-dataset replication and a deadline study find no clear advantage in any pre-specified Holm family. A length-matched swap, which gives each failed question the uncertainty-chosen origin of an equal-length partner, holds origin distributions exactly equal by construction and is also null, including two confirmatory cells with enough discordant pairs to resolve a large effect: -0.17 points ($p = 1.000$) at 32B and $+1.48$ ($p = 0.397$) at 72B. All experiments pass replay, scoring and allowance audits. These results limit claims for this policy in offline QA; they establish neither equivalence nor an advantage under equal total compute.

1 INTRODUCTION

ReAct agents interleave reasoning, tool calls, and observations (Yao et al., 2023); a failed answer can reflect mistakes at several points. Recovery must decide both what to change and how much context to retain. Restart discards potentially useful work. Suffix regeneration preserves that work but constrains recovery to the retained reasoning and observations.

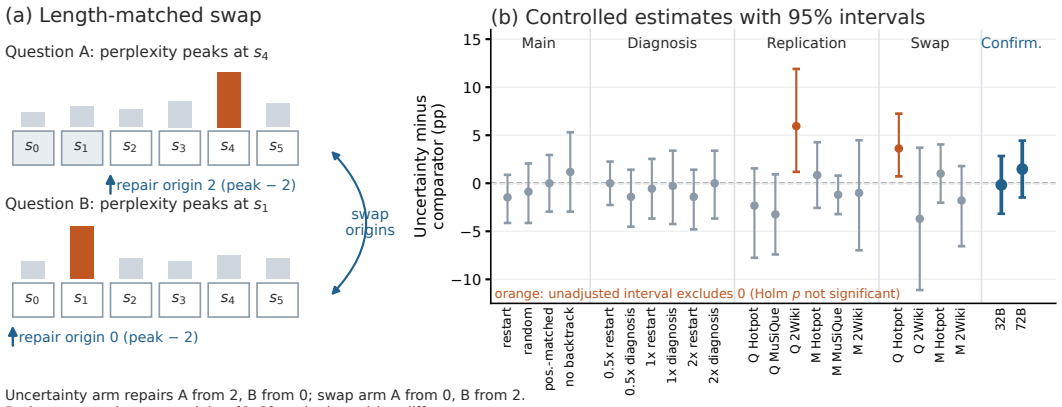

Figure 1: (a) Length-matched swap: equal-length failed trajectories exchange their uncertainty-chosen origins (highest-perplexity step minus two), so both arms repair from identical origins. (b) Uncertainty minus comparator in repair success, with 95% bootstrap intervals, for every main, diagnosis, swap and confirmatory primary contrast and for restart in each replication cell (Q: Qwen2.5-32B; M: Mistral-Nemo). No Holm-adjusted test is significant.

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.

Error attribution and recovery need not have the same objective. Attribution asks which step is erroneous, whereas recovery asks where a new rollout is most likely to succeed at an acceptable cost. A judge may identify an incorrect conclusion that is downstream of incomplete evidence gathering. Conversely, a locally uncertain action can be useful exploration. Neither uncertainty nor agreement with an error annotation directly establishes the value of intervening at that step.

We isolate origin selection with a common execution loop for targeted regeneration and restart. One stored-token policy is fixed before outcomes and compared with positional controls under the same recovery allowances. Separate experiments add a same-model diagnosis comparator, vary token allowances, replicate across two models and three datasets, and measure success by a completion deadline. They ask whether uncertainty adds useful selection information in a fixed tool environment, and whether an observed success rate justifies a resource advantage. Prior work already studies intervention-based debugging, localized QA repair with prefix reuse, and repair versus resampling (Ma et al., 2026; Jiao et al., 2026; Luan et al., 2026). We contribute a controlled test of a specific selector against these alternatives, not a new claim about prefix reuse or causal attribution.

Why this selector. Perplexity is the cheapest uncertainty signal a finished generation offers (Kadavath et al., 2022; Manakul et al., 2023), and argmax with a fixed backoff is the smallest rule turning it into an origin. It was our archived framework's default, where backing the judge's origin off two steps raised success on all three datasets and a perplexity variant beat restart on 2Wiki (Section 10) — an apparent benefit from an uncontrolled internal baseline, not a published claim, and exactly the kind of evidence a control can remove.

We make three contributions:

1. Controlled comparisons of one fixed selector, highest-perplexity step with two-step backtracking, spanning two model families and three QA datasets, with question-level paired inference, a development-fitted position control with measured replication imbalance, and a no-backtracking ablation. Neither the main study nor replication establishes an advantage for this policy in its prespecified comparison family.
2. An exactly balanced origin control: pairing failures by identical original length and swapping each pair's chosen origins removes positional imbalance by construction rather than by fitting, and the comparison stays null, including two confirmatory cells at 32B and 72B powered on discordant pairs (Figure 1). Alongside it, diagnosis, token-allowance and completion-deadline comparisons distinguish conditional repair success from resource savings and show which costs the allowances leave unmatched. The diagnosis comparator is an untrained, origin-only adaptation; its valid JSON and absent replication fallbacks establish parser coverage, not diagnosis quality.
3. Audited replay, reference scoring, complete trial coverage and frozen inputs supporting local reproduction of every completed comparison; older three-dataset summaries stay separate historical context.

2 Related Work

Agent recovery and self-correction. Reflexion uses verbal feedback from previous episodes to inform subsequent attempts (Shinn et al., 2023), while Self-Refine iteratively improves outputs using generated feedback (Madaan et al., 2023). Huang et al. (2024) document limitations of intrinsic reasoning self-correction without external feedback. Our benchmark gates on external correctness to select failed attempts. Conditional on that gate, we study suffix origins, not online failure detection.

Localized repair and replay. ReAgent uses local and global rollback in multi-agent QA (Xinjie et al., 2025). Doctor-RAG diagnoses failure type and location, then reuses prefixes and evidence for local repair on HotpotQA, MuSiQue, and 2WikiMultiHopQA (Jiao et al., 2026). This is direct task-level overlap, although its trained diagnoser, retrieval setting, and repair operators differ from our stored-token selector. DoVer evaluates intervention-driven debugging through task recovery (Ma et al., 2026). SymTrace reconstructs recorded prefixes to examine repair versus resampling (Luan et al., 2026); AgentRewind coordinates context and workspace rollback with memory of prior attempts (Zhuang et al., 2026); our offline tool replay is narrower than general environment rollback.

Step evaluation and failure attribution. Process supervision evaluates intermediate reasoning (Lightman et al., 2024), and Who&When studies multi-agent failure locations (Zhang et al., 2025). AgentRx uses executable constraints and violation logs to support diagnosis (Barke et al., 2026). REFLECT, Causal Agent Replay, and CausalFlow study interventions for attribution or repair (Lin et al., 2026; Shah, 2026; Bonagiri et al., 2026). AgenticRAG-FP injects faults and re-executes downstream reasoning to test attribution against known interventions (Pothuru, 2026). Recent systems replace passive step scoring with active investigation: FALAT searches typed dependencies to prune and verify candidates (Rafi et al., 2026), AgentDebugX runs multi-turn root-cause diagnosis in a detect–attribute–recover–rerun loop (Zhu et al., 2026), and tree-search agents revisit reasoning nodes during generation (Zhou et al., 2024a). All select an origin with far more information than the stored-token statistic tested here; none is evaluated below, and these nulls do not transfer to them. Our naturally failed traces lack such intervention-grounded causal labels.

Uncertainty signals. Self-consistency aggregates sampled reasoning outputs (Wang et al., 2023), and verbalized confidence can provide useful calibration information (Tian et al., 2023). ERGO already uses entropy changes for conversation resets (Mohammad Khalid et al., 2025). The archived framework includes action-level signals and stored-token statistics; the controlled study tests perplexity alone, without assuming calibration as error probabilities. Appendix I specifies the closest comparisons and information differences. No head-to-head advantage over them is established.

3 Problem and Evaluation Targets

An instance consists of a question x, an offline document collection D_x, and a reference answer y. An agent produces a trajectory $\tau = (s_0, \ldots, s_{T-1})$. Each step contains generated reasoning, a tool action, its argument, and the returned observation. The actions are `search`, `lookup` and `finish`.

Failure gate. The implemented QA success criterion accepts an answer when its normalized exact match is one or its token F1 is at least 0.5. This threshold is a study-specific recovery criterion, not the official aggregate metric for every benchmark. It is deliberately permissive: normalization already absorbs case, articles and punctuation, so an answer below 0.5 token F1 disagrees with the reference on most of its content rather than on formatting, and the cohort entering repair is one of substantive answer failures (Appendix H.5 quantifies how far the conclusions depend on this gate). Only trajectories that fail this gate enter repair. Evaluation therefore estimates recovery conditional on a known benchmark failure; the reference answer is not available to a deployed failure detector.

Repair origin. For an origin $k \in \{0, \ldots, T-1\}$, targeted repair retains $s_0, \ldots, s_{k-1}$ and generates a new suffix. Tool calls in the retained prefix are replayed to restore the environment state. Full restart uses $k = 0$. With backtracking offset b, the effective origin is

$$k' = \max(0, k - b). \tag{1}$$

Backtracking can make several nominal strategies share the same origin, and can turn targeted repair into a restart. Consequently, strategy counts are not counts of independent interventions.

Reference annotations. Historical diagnostics compare selected steps with a judge's error label k_J. The controlled studies use no such label for selection. Agreement with an error annotation does not identify the origin with the highest recovery success under a given agent, prompt and budget.

Success across seeds. For failed question q, strategy a, and generation seed r, write the binary repair outcome as Y_{qar}. Our primary descriptive statistic is

$$\widehat{R}_a = \frac{1}{N} \sum_{q=1}^{N} \frac{1}{S_q} \sum_{r=1}^{S_q} Y_{qar}. \tag{2}$$

Each question receives equal weight; S_q is written per question because the estimator needs no balanced design, though $S_q = 3$ throughout. Thresholding each question's mean success to a majority-over-seeds outcome measures repeatability, not the success rate of one sampled repair, and is reported separately.

4 UNCERTAINTY-GUIDED REGENERATION

4.1 PRESPECIFIED STORED-TOKEN POLICY

The controlled studies use step perplexity from stored sampled-token log probabilities. For log probability ℓ_{it} of generated token t in step i, define

$$U_i^{\mathrm{ppl}} = \exp\left(\min\left\{20, -\frac{1}{n_i}\sum_{t=1}^{n_i}\ell_{it}\right\}\right). \tag{3}$$

The policy selects the highest finite step score, breaking ties toward the earlier step, then subtracts two trajectory indices and clamps at zero. A step's measurement is unusable, and skipped, when it stores no generated tokens or when any stored token lacks a finite, non-positive log probability, either of which leaves U_i^{ppl} undefined; an empty usable profile falls back to origin zero. The $\min$ with 20 prevents overflow and does not winsorize the signal. The stored-token measure requires no additional sampled rollouts or diagnosis-model calls. The policy was prespecified before main-study outcomes, rather than chosen as the best-performing development or test variant; this is not an evaluation of the best possible uncertainty estimator.

The historical framework also evaluates entropy, sampled-token probability complement, action disagreement and verbalized uncertainty with several localization rules. Appendix A preserves their definitions. Those historical configurations are not conditions in the controlled experiment.

4.2 INTERVENTIONS AND INFORMATION ACCESS

Full restart discards the failed context; targeted policies retain a prefix. All six main-study conditions and all seven replication conditions receive the same generic retry hint. The uncertainty selector can inspect the complete failed trace's uncertainty profile, but receives no answer or supporting-fact labels. The recovery model sees only the question, retained prefix, and generic hint; discarded suffix observations are not inserted into its prompt. The same offline collection remains searchable. This is post-hoc selection after failure, not a policy that predicts the origin online. Judge-targeted and reference-informed repair belong only to the archived experiments. Their privileged information and unmatched historical restart hint are documented separately in Appendix B.

4.3 BUDGET SCOPE

The main study and replication cap recovery generation at the original failed trajectory's generated-token count. The diagnosis follow-up varies this cap across $0.5\times$, $1\times$ and $2\times$; diagnosis acquisition is charged separately. Each request is bounded by the remaining cap, with at most eight new steps and 512 generated tokens per step. Retained steps do not consume new-step opportunities. For an early failure, restart must regenerate a whole chain within the budget the original used, while repair spends it on a suffix; this asymmetry runs against restart, so it cannot explain a null in which no policy clearly outperforms restart. The runtime experiment retains these step limits, removes the total generated-token cap and evaluates measured completion deadlines. These designs do not equalize realized tokens, prompt processing or FLOPs. The historical runner counted retained steps toward its eight-step limit and could overshoot its token cap; its outcomes remain separate.

5 COMPLETED HOTPOTQA STUDY

Study organization. The main and diagnosis HotpotQA cohorts are distinct; replication's 300 questions are shared across two models (600 evaluations). Runtime reuses 25 Qwen HotpotQA replication failures for fresh timed executions, and the swap studies draw fresh cohorts (coverage in Table 8). Each component's primary tests form a separately frozen Holm family, with no pooled confirmatory test.

5.1 MODEL, COHORT AND HISTORICAL EXCLUSIONS

We evaluated Qwen2.5-32B-Instruct-AWQ on 250 HotpotQA questions (Yang et al., 2018) using the supplied distractor-context paragraphs: 248 questions have ten paragraphs, one has two, and one

has six. The offline `search` tool uses exact, substring and fuzzy title matching and returns the first five sentences; `lookup` advances through matching sentences in the current paragraph. Model and tokenizer were pinned to one snapshot (Appendix E). Initial generation was greedy; repairs used temperature 0.7 and seeds $0, 1, 2$. Five ordered batches of 50 completed on September 12, 2026. There was no adaptive sample-size change or policy selection after new outcomes.

The cohort was sampled uniformly without replacement from 6,856 sorted eligible IDs using seed 20260912, giving 196 bridge and 54 comparison questions. We excluded 549 distinct IDs: the union of a reconstructed 500-question historical pool and 60 AWS pilot/development questions, with 11 overlaps. The reconstruction reproduces the original sampler, seed and archived raw dataset. The original historical pool JSON was located but its full contents were not downloaded or compared. Thus disjointness is verified against the reconstructed documented history, not every possible undocumented run. This qualification limits any held-out claim.

5.2 Six Conditions and Shared Information

The treatment is perplexity/argmax with backtracking by two. Four primary controls are full restart, uniform random origin with the same backtracking, position-matched random, and the same uncertainty rule without backtracking. Fixed early origin $\min(1, T-1)$ is a descriptive sixth condition. All conditions receive the same generic hint, token cap and new-step allowance. The gate uses official HotpotQA answer EM or F1 at least 0.5; strict EM and continuous F1 are secondary outcomes on the same initial-failure cohort. Reference answers affect evaluation and failure selection, not origin choice or the recovery prompt. There is no diagnosis-model condition in this study.

The position control was fitted to 24 failed original development trajectories without using their repair outcomes. A trajectory has no effective origin independent of a selection policy, so the matched distribution is the one the prespecified policy of Section 4.1 produced on those development trajectories. The control samples one of those origins uniformly, normalizes by the development trajectory length and maps to the new length with nearest-position, half-up rounding. All 50 development source questions are excluded. The frozen profile reads no test uncertainty scores; matching in expectation does not guarantee identical finite-sample histograms. Appendix E specifies all origin formulas and profile provenance.

5.3 Paired Analysis and Execution Audit

Initial generation succeeded on 137/250 questions (54.8%), with 97 exact matches (38.8%) and mean answer F1 0.5164. The remaining 113 failures each received six conditions and three seeds: 339 trials per condition and 2,034 condition/seed rows. Identical effective origin/prompt/seed executions were shared, yielding 1,058 unique repair executions. The question, not the seed row, is the independent resampling unit. We average three seed outcomes within each question and use 10,000 paired question bootstrap resamples for individual 95% intervals. Paired sign-flip tests assume within-question exchangeability. The four prespecified treatment-versus-control comparisons form one Holm family; fixed early is not part of that family or a selected replacement policy.

All five archives passed independent checks against the frozen source, cohort, model and condition manifests. The audit replayed and reference-scored 1,308 trajectories, checked 8,552 observations and 1,742 retained prefix steps, and found zero replay/scoring mismatches or token/new-step violations. The pooled analysis reproduced locally within 10^{-12} of the AWS output. Appendix E reports batch counts, hashes, compute and audit details.

6 Controlled Repair Results

Prespecified uncertainty versus controls. The treatment achieves 9.73% repair success (33/339), compared with 11.21% for restart (38/339), 10.62% for matched-offset random (36/339), 9.73% for position-matched random (33/339), and 8.55% for uncertainty without backtracking (29/339). Fixed early achieves 11.50% (39/339), but its descriptive point estimate does not establish a superior policy. Table 1 also reports exact match and F1. The treatment does not exceed restart here: the apparent benefit that motivated this policy came from the uncontrolled archived runs of Section 10,

Table 1: Completed HotpotQA repair comparison on 113 initial failures, with three seeds per condition. Success is answer EM or F1 ≥ 0.5; EM and F1 use the same cohort. The final column reports effective origin zero. Fixed early is descriptive. Copied execution rows are not independent trials.

Condition	Successes	Success (%)	EM (%)	F1	$k' = 0$ (%)
Full restart	38/339	11.21	8.55	0.1442	100.00
Random + backtrack 2	36/339	10.62	6.78	0.1402	64.01
Fixed early (descriptive)	39/339	11.50	5.31	0.1369	0.00
Uncertainty + backtrack 2	33/339	9.73	7.08	0.1275	61.95
Position-matched random	33/339	9.73	5.60	0.1265	60.77
Uncertainty, no backtrack	29/339	8.55	4.13	0.1095	15.04

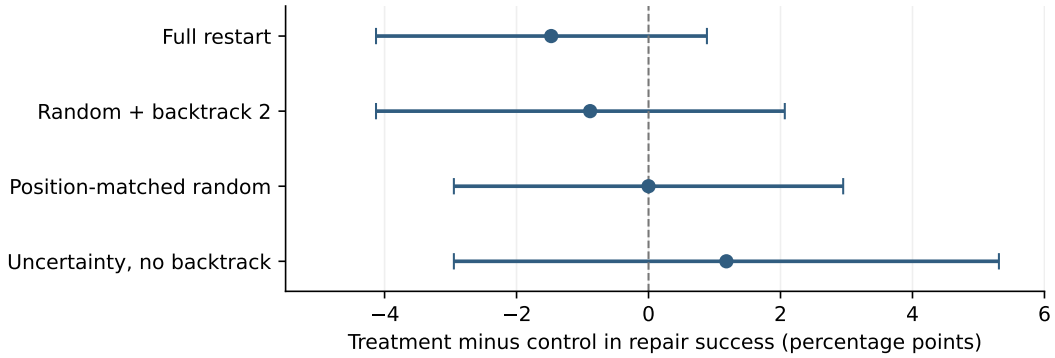

Figure 2: Treatment minus control in repair success. Points are paired question-level differences; bars are individual, unadjusted 95% bootstrap intervals over 113 questions. All four intervals include zero. All four Holm-adjusted sign-flip p-values are 1.000; the exact contrasts and raw p-values appear in Table 10.

and these matched controls do not reproduce it. Treatment equals the position-matched rate exactly, so this comparison shows no selection benefit beyond position; as below, that is not equivalence.

The treatment-minus-restart difference is -1.47 percentage points, with a 95% interval [-4.13, +0.88] (Figure 2). The differences versus offset-matched random, position-matched random and uncertainty without backtracking are -0.88, 0.00 and $+1.18$ points, with intervals $[-4.13, +2.06]$, $[-2.95, +2.95]$ and $[-2.95, +5.31]$, respectively. No primary comparison shows a clear treatment advantage. These intervals quantify uncertainty for this cohort and setting; they establish no equivalence, no universal rejection of uncertainty signals and no winning alternative.

How often backtracking becomes restart. The treatment selects effective origin zero in 210/339 rows (61.95%), compared with 64.01% for offset-matched random and 60.77% for the position-matched control. Without backtracking, uncertainty selects zero in 15.04% of rows. Thus many nominal targeted interventions reuse the same physical restart execution. The position control addresses an important alternative explanation, but matches a development-fitted distribution rather than every feature of each test trajectory, and the observed zero difference does not prove the policies equivalent.

Compute and measurement scope. Initial generation used 57,499 generated tokens; unique repairs used 303,700 generated and 5,292,100 prompt tokens, in about 81 minutes on one GPU (Appendix E). Equal caps and shared executions establish neither equal realized computation nor a wall-clock speedup.

7 Diagnosis and Allowance Follow-up

A separate frozen cohort of 250 HotpotQA questions excludes all 799 previously documented questions. Its 118 initial failures receive full restart, the same uncertainty policy, and a same-model

Table 2: Completed replication. Rates are conditional repair success (%). N_f counts failed questions; each policy receives three seeds per question. Δ is uncertainty minus restart in percentage points, with an individual 95% question-bootstrap interval. p_H belongs to the fixed 24-comparison Holm family. Failure cohorts differ between models.

Model	Dataset	N_f	Restart	Unc.	Diag.	Δ (pp)	95% interval (pp)	p_H
Qwen	HotpotQA	43	13.95	11.63	10.85	−2.33	[−7.75, +1.55]	1.000
Qwen	MuSiQue	72	7.87	4.63	7.41	−3.24	[−7.41, +0.93]	1.000
Qwen	2Wiki	28	15.48	21.43	16.67	+5.95	[+1.19, +11.90]	1.000
Mistral	HotpotQA	78	11.54	12.39	9.40	+0.85	[−2.56, +4.27]	1.000
Mistral	MuSiQue	83	7.23	6.02	7.63	−1.20	[−3.21, +0.80]	1.000
Mistral	2Wiki	67	16.92	15.92	14.43	−1.00	[−6.97, +4.48]	1.000

diagnosis/replay adaptation at $0.5\times$, $1\times$, and $2\times$ original generated tokens, with three seeds and eight new steps. The diagnosis sees the question and failed thoughts, actions, and observations, but no reference answer, support labels, or uncertainty scores. It selects a regeneration origin; recovery sees only the retained prefix and common retry hint. This is an untrained adaptation, not a reproduction of Doctor-RAG's trained system.

Tables 16 and 17 give every allowance. At $1\times$, uncertainty succeeds on 35/354 trials (9.89%), versus 37/354 (10.45%) for restart and 36/354 (10.17%) for diagnosis/replay. Uncertainty minus restart is -0.56 percentage points (individual 95% question-bootstrap interval $[-3.67, +2.54]$); versus diagnosis it is -0.28 points ($[-4.24, +3.39]$). All six primary Holm-adjusted p-values across allowances are 1.000. These results provide no clear advantage and do not establish equivalence. Appendix G gives all rates, intervals, costs and audits.

The allowances match recovery-generated tokens (Table 18); diagnosis adds a measured mean 40.31 generated tokens, 1,234.25 prompt tokens, and one model request per standalone policy attempt. Sharing that acquisition across study executions does not make diagnosis free for a deployed attempt.

8 Replication across Models and Datasets

A separately frozen extension evaluates Qwen2.5-32B-Instruct-AWQ and Mistral-Nemo-Instruct-2407 on HotpotQA, MuSiQue and 2WikiMultiHopQA. Each dataset supplies 100 new main and 20 disjoint development questions, shared across models. Seven policies receive three seeds and a recovery allowance equal to the original generated-token count, including positional controls and same-model diagnosis/replay. Its 371 initial failures yield 7,791 token-policy rows; runtime adds 225 attempts. All six cells pass independent replay, scoring and allowance audits, and reproduce locally.

Uncertainty-minus-restart differences range from -3.24 to $+5.95$ percentage points across cells (Table 2). All 24 primary Holm-adjusted p-values are 1.000, covering restart, development-fitted random, no-backtracking uncertainty and diagnosis/replay in every cell. That count overstates the evidence: only 3–23 questions per contrast have different three-seed success, and two cells cannot reach significance at any conventional level because their exact test has too few non-zero blocks. The Qwen 2Wiki interval excludes zero but rests on 4 wins, 0 losses and 24 ties, floor 0.125; its $p_H = 1.000$ gives no primary-test evidence. We retain both prespecified outputs; individual intervals are not simultaneous (Appendix H.6). These comparisons establish neither equivalence nor a model reliability ranking. The random position control fits only 6–19 development failures per cell. Its origin-zero frequency exceeds treatment by 19.38 points on Qwen HotpotQA and 26.19 on Qwen 2Wiki. This imbalance limits isolation of uncertainty information from positional preferences; Appendix H.3 reports full distributions and execution overlap. A separately frozen follow-up removes that imbalance by construction rather than by fitting: on 200 fresh questions it pairs failures by exact original length and gives each question its partner's uncertainty-chosen origin, so the two arms carry *identical* origin multisets. Both model families stay null across 102 pairs, with point estimates positive on HotpotQA and negative on 2Wiki for both models (Appendix H.7). A further frozen study re-tested the one suggestive cell on 1,000 fresh questions, pairing so that origins differ where possible: over 100 pairs, 40 of them discordant, the difference is -0.17 points ($[-3.17,$

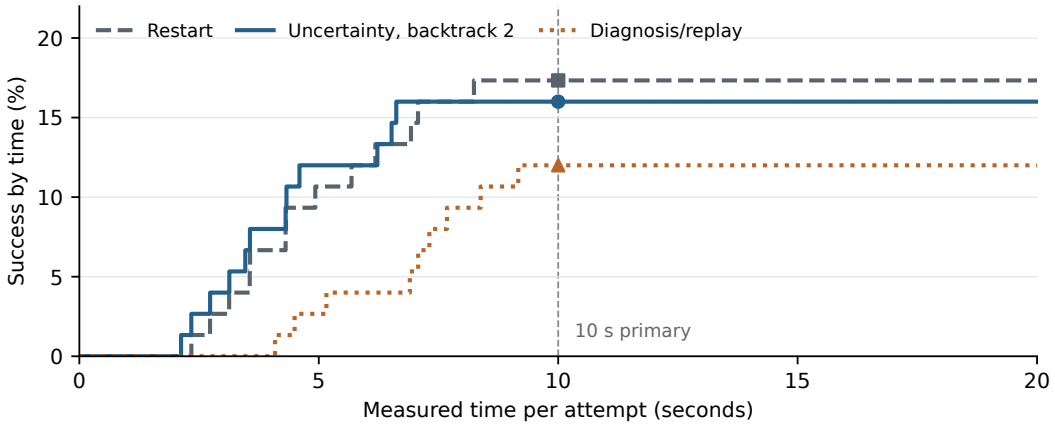

Figure 3: Completed runtime experiment: 25 questions and 75 attempts per policy. Each curve shows successful completions by the indicated time as a fraction of all attempts. The ten-second threshold is primary; every attempt ends within 11 s. Attempts execute fully before thresholds are applied, so the figure does not measure savings from cancellation.

+2.84], $p = 1.000$), a declared deviation since the run stopped at its time limit. A complete 72B rerun on further fresh questions gives +1.48 ([−1.48, +4.44], $p = 0.397$; Appendices H.8 and H.9). All policy outcomes, costs and contrasts appear in Appendix H.

9 Success under a Measured Time Limit

The runtime component uses the first 25 failed questions from the 100-question Qwen HotpotQA replication cell. Restart, uncertainty and diagnosis each receive three seeds, with separate executions in shuffled order after model warmup and with prefix caching disabled, which isolates per-attempt cost but charges prefix-preserving repair for prompt recomputation that a cache would serve (Appendix H.4). Attempts retain eight new steps and 512 tokens per request but have no total generated-token cap. Measured time includes selection, prefix replay and recovery. Diagnosis acquisition is measured once per question and charged to each attempt. The primary outcome is a correct completion by ten seconds; 5- and 20-second thresholds are exploratory.

At ten seconds (Figure 3), uncertainty succeeds on 12/75 attempts (16.0%), restart on 13/75 (17.3%), and diagnosis on 9/75 (12.0%). Uncertainty minus restart is -1.33 percentage points (individual 95% question-bootstrap interval $[-5.33, +2.67]$); versus diagnosis it is $+4.00$ points ($[-4.00, +13.33]$). Both Holm-adjusted p-values are 1.000. Every attempt ends within 11 s, so the threshold barely binds: success by 20 s equals success by 10 s for every policy, making this in effect success without a token cap. This small, ordered cohort provides no clear deadline-success advantage; it does not establish equivalence or population-wide latency rankings. Median measured time is 4.45, 4.93 and 4.50 s respectively. Appendices H.4 and H give the latency distribution, termination counts, thresholds and audit details.

10 Analysis and Discussion

Archived historical summaries. Older HotpotQA, MuSiQue and 2WikiMultiHopQA summaries cover 854 failures, but retain prompt/budget confounds, retrospective policy selection and unrecovered raw trials. They supply historical context and are not replications of the controlled result (Table 4). Appendix C contains their rates and diagnostics; these summaries support neither a causal error-cascade claim nor an established length-based recovery rule.

Review-stage diagnostics. Exploratory diagnostics added after the primary results were known qualify the null without replacing the frozen four-comparison family (Appendix F). Treatment and restart share all three executions on 70 of 113 failed questions, but the position-matched control

differs on 93, so the null is not merely collapse into restart. Treatment uses 29.5% more prompt tokens than restart at similar output, and 168/339 of its attempts end at a token or step cap without an answer. All 637 original failed-trace steps have usable perplexity, so missing signal does not explain the result.

11 LIMITATIONS

Scope and remaining comparisons. Two open-weight model families and three QA datasets, offline evidence and a permissive reference-answer gate do not establish transfer to live retrieval, models beyond 72B, closed frontier models or general agent tasks. At 72B, initial success and the null resemble 32B, and hosted frontier APIs do not reliably offer the token log probabilities, pinned weights and replay it needs (Appendices E and H.9). An EM-only gate would need more repairs. Prefix replay does not guarantee identical stochastic suffixes or restore arbitrary mutable state.

Comparator and resource scope. The diagnosis/replay adaptation omits trained diagnosis and diagnosis-specific repair operators. Runtime results use the first 25 failures, exclude startup and failure detection, and score completed attempts without active deadline cancellation. Diagnosis is timed once per question, leaving reacquisition variability unmeasured. Neither generation allowances nor deadline results establish an advantage at equal end-to-end latency, FLOPs or deployment cost.

Exclusions and statistical precision. Test IDs are frozen and disjoint from reconstructed exclusions, but the original historical pool was not downloaded or compared and undocumented runs cannot be ruled out. Fitted position controls use 24 main-study and 6–19 per-cell development failures, and their realized imbalance limits attribution beyond position; the length-matched swap holds origin distributions exactly equal instead. A nonsignificant difference is not evidence of equivalence, and bootstrap intervals lack simultaneous coverage. Sign-flip inference assumes within-question exchangeability, and shared seeds or executions add no independent questions. Precision, not effect size, binds these nulls: 83 of 102 swap pairs tie, so 3 of 4 swap cells and two of six replication cells could not have reached a conventional level whatever their outcomes. The floor falls only as non-tied pairs accumulate, so at this tie rate the cell with the most pairs shares the highest floor. Pairing for differing origins lifts that rate: with 40 and 51 discordant pairs the confirmatory floors sit far below any conventional level, so those nulls are not artifacts of the design (Appendices H.7–H.9).

What these limits imply for future work. Four steps follow, of which the first is taken here: power a resolving design on expected discordant pairs rather than on failures. Three remain. Test diagnosis-driven repair against a trained or search-based comparator (Rafi et al., 2026; Zhu et al., 2026) rather than the untrained adaptation used here. Measure latency with prefix caching enabled, so retained prefixes are served from cache as in deployment. Finally, offline QA reduces recovery to regenerating text; environments with irreversible external state (Zhou et al., 2024b) would test whether origin choice matters more when a prefix cannot simply be replayed.

12 CONCLUSION

One fixed selector, the highest-perplexity step with two-step backtracking, shows no clear advantage in any primary comparison family across two model families and three offline QA datasets. Main-study success is 9.73% versus 11.21% for restart, equal to the observed position-matched rate. Holding origin distributions exactly equal by construction leaves the result unchanged, and a confirmatory cell with 40 discordant pairs, the first with resolution to spare, returns -0.17 points ($p = 1.000$); at 72B it returns +1.48 ($p = 0.397$). Untrained-diagnosis, allowance and deadline comparisons remain inconclusive and say nothing about trained diagnosis, and earlier cells often could not have reached significance at all. Main-study prompt usage rises by 29.5% at similar output. These results support measuring acquisition and prompt costs alongside repair success, not equivalence, matched compute or a limit on uncertainty-guided repair as a method class, which we do not test.

REPRODUCIBILITY STATEMENT

All completed studies retain frozen cohorts, model/source identities, raw trajectories, checksums and independent audit outputs. Local analyses match the recovered results; paper assets are generated from checked records. Appendices E, G, H and H.7 give protocols and audit counts. The extension's 8,016 rows and 29 trial columns include its 225 runtime attempts. Exploratory diagnostics and historical summaries are identified separately.

A combined anonymous supplement covers every completed study. From an isolated extraction, with no workspace path or cloud credential, it reproduces seven primary families from their packaged trial exports: the main HotpotQA study's 4 tests over 2,034 rows, the diagnosis and allowance follow-up's 6 tests over 3,186 rows, the replication's 24 tests over 8,016 rows, the length-matched swap's two 2-test families over 738 (Qwen) and 1,098 (Mistral) rows, and the 32B confirmatory swap's single test over 1,200 rows, shipped with the deviation statement its truncated run carries, and the 72B confirmatory swap's single test over 1,488 rows. The runtime component's two tests are packaged in the replication analysis but not recomputed by the bundle. Each contrast's condition means and point estimate are recomputed from the rows, and each Holm adjustment is recomputed over its declared family; any altered file fails a checksum and any altered statistic fails the comparison. Three environments stay separate and are labeled as such in the bundle: statistical reproduction is included, while raw replay with reference scoring and GPU regeneration of trajectories are not, since those need the frozen raw trajectories and the pinned weights respectively. Human annotations and unrecovered historical raw trials are not supplied.

Independent verification is assigned but not complete. Two internal project auditors, commissioned by the authors and distinct from this venue's peer reviewers, were briefed before submission: one on the statistical argument, in particular whether the resolution floor is used fairly and whether these nulls are rigorous or underpowered, and the second on independently reproducing the packaged statistics and tracing reported claims back to records. Neither review had been received at submission, and no finding from either is reflected here. The historical appendix and pre-September material were not re-verified alongside the recent work and should be read as carrying their original, weaker guarantees. Human error-step annotation remains unavailable, which bounds this work to answer scoring rather than error localization. The 32B confirmatory swap stopped at its frozen time limit with 31 of its 131 informative pairs unrun; its completed pairs are reported as a declared deviation from that protocol's completion rule, not as the frozen analysis. The AI use statement records the authors' own verification of the assisted work and states the limits of that checking: it was carried out by the authors who also directed the assistance, so it is verification rather than independent replication, and it did not re-derive the historical appendix. Appendix J distinguishes verification from release availability.

ETHICS STATEMENT

This work involves no human participants, crowdworkers or personal data; the human error annotations discussed are unavailable and were not collected here. All questions come from public benchmarks used under their licenses: HotpotQA (CC BY-SA 4.0), MuSiQue (CC BY 4.0) and 2WikiMultiHopQA (Apache 2.0). The supplement redistributes question text only within these terms. Qwen2.5-32B-Instruct-AWQ and Mistral-Nemo-Instruct-2407 are Apache 2.0 models, and Qwen2.5-72B-Instruct-AWQ is used for evaluation only under its Qwen License Agreement; no weights are redistributed. The study evaluates recovery from agent failures on public question answering and adds no new capability. The two confirmatory cells used about 11 GPU-hours on one 96GB GPU; we report this so the compute cost of a resolving null is visible.

AI USE STATEMENT

Language models as the object of study. Language models are the subject of this work, not only a tool used to produce it. Qwen2.5-32B-Instruct-AWQ, Qwen2.5-72B-Instruct-AWQ and Mistral-Nemo-Instruct-2407 act as the ReAct agents whose failed trajectories are repaired, and a same-model prompt supplies the diagnosis/replay comparator. Their versions, revisions, quantization and decoding settings are pinned in each frozen protocol and reported in the appendices.

Generative AI assistance in producing this work. An AI coding assistant was used extensively. It contributed to: proposing and refining the hypotheses tested here; the design of the length-matched swap control and of the exact sign-flip analysis applied to it; the formulation of the resolution-floor claim that an exact two-sided sign-flip test over m non-zero blocks cannot report a p-value below $2/2^m$, together with the brute-force enumeration used to check it; implementation of the experiment runners, auditors, statistical code and paper-asset builders; orchestration of the cloud experiments, including budget gating, monitoring and result retrieval; reformatting and cleaning of per-trial records for analysis; the statistical diagnostics reported in the appendices; generation of the figures; literature lookup and related-work analysis; interpretation of results; and the drafting and revision of manuscript text, including sections of this manuscript and the appendix reporting the swap study. The accompanying test suite was likewise written with AI assistance. No synthetic dataset was generated: HotpotQA, 2WikiMultiHopQA and MuSiQue are used as published, and the trajectories analyzed here are experimental measurements of the agents under study rather than fabricated data.

What the safeguards do and do not establish. The safeguards in this project are structural rather than discretionary. Tables, figures and most reported statistics are generated from audited records by a build step rather than transcribed by hand, and the build refuses inputs whose checksums or provenance do not match. Some counts in the prose were entered by hand; each was checked against its source record before submission. Each study is frozen by checksum before execution and re-analyzed independently afterwards, and every reported cell reproduces with zero mismatches between the on-instance and local analyses. A supplement recomputes every primary test family except the two runtime tests from an isolated extraction. These safeguards constrain arithmetic and provenance tightly. They establish nothing about whether the design answers the question posed, whether the estimands are the right ones, or whether a null result is framed fairly. Much of the design and analysis was produced with AI assistance and then checked largely by the same process that produced it; self-consistency is not correctness.

Errors introduced during AI-assisted work. Eight errors reached execution and were caught before affecting any reported result. Five were faults in launching, monitoring or stopping cloud runs, none of which altered a recorded outcome; the others were an incorrect prediction that additional matched pairs would lower the resolution floor, when the tie rate binds instead; an edit that invalidated the checksum provenance of an already-completed study package; and three figures whose labels were obscured by legends, found only by inspecting the rendered images. One further error was caught before execution: the cohort sampler for the 72B cell did not exclude the 32B confirmatory questions and would have re-tested questions whose outcomes had been seen. Each was found by continued checking rather than by any process that guarantees detection. We report them because they bear on how much independent verification a reader should expect of work produced this way.

Independent verification. The authors checked the assisted work by four means. They ran the reproduction supplement themselves, from its own extraction rather than from the working tree, and confirmed that each packaged study returns the primary test counts and trial counts reported here. They read the core analysis implementation, which performs the length pairing, asserts exact balance of the two arms' origin multisets and evaluates the exact sign-flip null, closely enough to judge whether it does what this paper says it does. They traced reported quantities back to trial-level records rather than to the generated macros that place them in the text. They inspected the rendered figures against the numbers they report, a check that had already caught three figures whose labels were obscured by their legends.

This checking was performed by the authors who also directed the AI assistance, so it is verification rather than independent replication, and it shares whatever blind spots shaped the work. It was concentrated on the recent audited experiments. The historical material in the appendices and the pre-September results were not re-derived and should be read as unverified, unlike the recent audited experiments.

Internal audit. Two internal project auditors, commissioned by the authors and distinct from this venue's peer reviewers, were engaged before submission: one to examine the statistical argument — in particular whether the resolution-floor bound is correct and whether it is used fairly rather than to excuse an underpowered design — and one to examine whether each claim traces to evidence and

whether the reproduction bundle is genuinely self-contained. Both briefs, and the guidance that each auditor record what they did *not* check, will be released with the code after review.

Neither review had been received at submission. No finding from either auditor is reflected in this version of the manuscript, and no claim here should be read as having been independently confirmed by them. Anything they report will be addressed in revision.

Responsibility. The authors take full responsibility for all content of this submission, including any error introduced by AI assistance.

REFERENCES

Shraddha Barke, Arnav Goyal, Alind Khare, Avaljot Singh, Suman Nath, and Chetan Bansal. AgentRx: Diagnosing AI agent failures from execution trajectories. *arXiv preprint arXiv:2602.02475*, 2026. URL https://arxiv.org/abs/2602.02475v2. Version 2, August 31, 2026; reports acceptance to Findings of EMNLP 2026.

Akash Bonagiri, Devang Borkar, Gerard Janno Anderias, Setareh Rafatirad, and Houman Homayoun. CausalFlow: Causal attribution and counterfactual repair for LLM agent failures. *arXiv preprint arXiv:2605.25338*, 2026. URL https://arxiv.org/abs/2605.25338.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *International Conference on Computational Linguistics (COLING)*, pp. 6609–6625, 2020.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.01798.

Shuguang Jiao, Chengkai Huang, Shuhan Qi, Xuan Wang, Yifan Li, Quanchi Weng, Lingchuan Liu, Xunliang Cai, and Lina Yao. Doctor-RAG: A failure-aware repair framework for agentic retrieval-augmented generation. *arXiv preprint arXiv:2604.00865*, 2026. URL https://arxiv.org/abs/2604.00865v2. Version 2, June 12, 2026.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

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

Xiaofeng Lin, Yingxu Wang, Tung Sum Thomas Kwok, Daniel Guo, Sahil Arun Nale, Charles Fleming, and Guang Cheng. REFLECT: Intervention-supported error attribution for silent failures in LLM agent traces. *arXiv preprint arXiv:2606.09071*, 2026. URL https://arxiv.org/abs/2606.09071.

Zhongwen Luan, Xiaoyu Zhang, Ming Hu, Yue Yang, Jiongchi Yu, and Xiaohong Chen. Repair or resample? Rethinking failure debugging in LLM multi-agent systems. *arXiv preprint arXiv:2608.25920*, 2026. URL https://arxiv.org/abs/2608.25920v2. Version 2, August 29, 2026.

Ming Ma, Jue Zhang, Fangkai Yang, Yu Kang, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. DoVer: Intervention-driven auto debugging for LLM multi-agent systems. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=mrEK16Jy6h.

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 (NeurIPS)*, volume 36, pp. 46534–46594, 2023.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9004–9017, 2023.

Haziq Mohammad Khalid, Athikash Jeyaganthan, Timothy Do, Yicheng Fu, Vasu Sharma, Sean O'Brien, and Kevin Zhu. ERGO: Entropy-guided resetting for generation optimization in multi-turn language models. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pp. 273–286. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.uncertainlp-main.23. URL https://aclanthology.org/2025.uncertainlp-main.23/.

Lauren Pothuru. When failures propagate: Causal failure attribution in agentic retrieval-augmented generation. *arXiv preprint arXiv:2608.20627*, 2026. URL https://arxiv.org/abs/2608.20627v1.

Md Nakhla Rafi, Md Ahasanuzzaman, Dong Jae Kim, Zhijie Wang, and Tse-Hsun Chen. FALAT: Tracing failures in LLM agent trajectories via dependency-guided search. *arXiv preprint arXiv:2606.00765*, 2026. doi: 10.48550/arXiv.2606.00765.

Jaineet Shah. Causal agent replay: Counterfactual attribution for LLM-agent failures. *arXiv preprint arXiv:2606.08275*, 2026. URL https://arxiv.org/abs/2606.08275.

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 (NeurIPS)*, volume 36, pp. 8634–8652, 2023.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 809–819, 2018.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5433–5442, 2023. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single hop question composition. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 10, pp. 539–554, 2022.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhao Xinjie, Fan Gao, Xingyu Song, Yingjian Chen, Rui Yang, Yanran Fu, Yuyang Wang, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. ReAgent: Reversible multi-agent reasoning for knowledge-enhanced multi-hop QA. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4067–4089. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.202. URL https://aclanthology.org/2025.emnlp-main.202/.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2369–2380, 2018.

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 (ICLR)*, 2023.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and Qingyun Wu. Which agent causes task failures and when? On automated failure attribution of LLM multi-agent systems. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025. URL https://proceedings.mlr.press/v267/zhang25cq.html.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *International Conference on Machine Learning (ICML)*, 2024a. doi: 10.48550/arXiv.2310.04406.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024b. doi: 10.48550/arXiv.2307.13854.

Kunlun Zhu, Xuyan Ye, Zhiguang Han, Yuchen Zhao, Bingxuan Li, Weijia Zhang, Muxin Tian, Xiangru Tang, Pan Lu, James Zou, Jiaxuan You, and Heng Ji. AgentDebugX: An open-source toolkit for failure observability, attribution, and recovery in LLM agents. *arXiv preprint arXiv:2607.18754*, 2026. doi: 10.48550/arXiv.2607.18754.

Yu Zhuang, Kefei Chen, Yitong Duan, Shuxin Zheng, Jian Li, and Xu-Yao Zhang. AgentRewind: Recoverable execution for long-horizon LLM agents. *arXiv preprint arXiv:2608.14380*, 2026. URL https://arxiv.org/abs/2608.14380v1.

A Historical Signal and Rule Catalog

A.1 Signal Definitions

The historical framework computes five signal families, oriented so that larger values represent greater uncertainty. Token-based measurements use stored generation log probabilities. Let ℓ_{it} be the log probability of the sampled token t in step i.

Renormalized top-K entropy. For the K_{it} available returned alternatives, define

$$H_{it} = -\sum_{j=1}^{K_{it}} \widehat{p}_{itj} \log \widehat{p}_{itj}, \qquad \widehat{p}_{itj} = \frac{\exp(\ell_{itj})}{\sum_{u=1}^{K_{it}} \exp(\ell_{itu})}. \tag{4}$$

The grid uses maximum token entropy, requesting 20 alternatives per token; the returned set can also include the sampled token. This truncated, renormalized distribution does not recover full vocabulary entropy.

Sampled-token probability complement and perplexity. The probability-complement signal is $1 - \exp(\ell_{it})$, aggregated by its maximum within a step. The legacy code calls this metric `max_token_prob`, but the probability belongs to the sampled token. Step perplexity is

$$U_i^{\mathrm{ppl}} = \exp\left(\min\left\{ 20, -\frac{1}{n_i} \sum_{t=1}^{n_i} \ell_{it} \right\} \right). \tag{5}$$

Action disagreement and verbalized uncertainty. Action disagreement resamples the next step five times at temperature 0.7 and computes one minus the fraction matching the most common (action, argument) pair. Verbalized uncertainty is $1 - c_i/100$, where c_i is a confidence score elicited from the model. These signals require additional model calls. Their acquisition cost must be included when comparing total inference efficiency with a restart policy.

A.2 Localization Rules

For a profile $U_0, \ldots, U_{T-1}$, three standard rules select the largest score, the earliest step among the top three scores, or the earliest step at or above the trajectory's 75th percentile. Three additional heuristics test upstream and positional preferences:

- **Upstream:** move two scored positions earlier than the uncertainty maximum, clamping to the earliest scored step.
- **Gradient:** select the step with the largest increase in uncertainty between successive scored steps.
- **Position-weighted:** maximize $0.5\widetilde{U}_i + 0.5(1 - \widetilde{i})$, the min–max normalized uncertainty and position of Equation 6.

Ties favor earlier steps. Missing scores are skipped. Upstream localization and explicit backtracking can compound: selecting two steps before a peak and then backtracking two more may move four positions upstream. These heuristics encode preferences about where to intervene; their use alone does not establish that uncertainty peaks reveal causal error propagation.

The grid contains 30 metric–rule combinations. Crossing them with two backtrack offsets $(0, 2)$ and two hint types yields 120 configurations. Four judge-targeted variants plus random repair and full restart yield 126 configurations. Table 3 records the archived study's configuration. Three seeds are configured for each. Results with the same question, effective origin, hint, and seed can reuse the same generated suffix; their copied result rows are not independent samples.

Table 3: Recorded configuration for the archived QA study. A configured setting is not independent verification of the model actually loaded; runtime logs remain necessary for that check.

Component	Recorded setting
Agent	Qwen2.5-32B-Instruct-AWQ
Reference annotator	Qwen2.5-72B-Instruct-AWQ
Initial questions	500 per QA dataset
Initial generation	Temperature 0; at most 8 steps
Per-step generation limit	512 tokens
Saved token alternatives	Top 20
Repair generation	Temperature 0.7; seeds 0, 1, 2
Nominal recovery cap	Original generated-token count ($m = 1$)
Stored-token signals	Maximum entropy, probability complement, perplexity
Extra-call signals	Five-sample action disagreement; elicited confidence
Repair grid	5 signals, 6 rules, 2 offsets, 2 hint types
Additional configurations	4 reference-targeted variants, random, restart

B Execution and Annotation Details

B.1 Environment and Prefix Replay

The agent emits a thought, one action, and its argument. A `search` action queries an entity or title in the question's offline collection; `lookup` advances through matching sentences in the current page; `finish` supplies the final answer. The model is instructed not to generate the observation itself. The scratchpad contains the question followed by the retained thoughts, actions, arguments, and observations. Recovery replays the retained tool actions in a fresh environment before generating a suffix. It therefore retains both the language-model-visible prefix and the corresponding tool state, rather than merely truncating an answer string.

Replay audit and reproducibility boundary. The corrected runner re-executes retained tool actions and rejects a mismatch in the stored observation, tool-call status or retrieved title before further inference. The main-study independent audit additionally replayed all saved trajectories from frozen records and checked the retained prefix state. It found no mismatches across 8,552 observations and 1,742 retained steps. This audit applies to the new study; historical aggregate results do not acquire the same guarantee. Following the distinction between recorded-prefix reconstruction and live continuation in SymTrace (Luan et al., 2026), prefix fidelity is reported separately from suffix outcomes. Equal numeric seeds do not guarantee identical continuations across different prompts, batch compositions or hardware. Physical reuse of one execution is distinct from rerunning it twice.

Generated-token cost for the original prefix is sunk for a conditional recovery comparison. Rereading that prefix can still incur prompt cost, and replay can incur tool overhead. These costs should not be erased by excluding the prefix's previously generated tokens. The saved baseline summaries report recovery generation and tool calls, not this complete accounting. They also do not establish wall-clock speedups, which depend on batching, cache behavior, and hardware.

B.2 Prompt and Hint Information

The agent's output template is:

```
Thought: <reasoning for this step>
Action: <search, lookup, or finish>
Action Input: <argument>
```

The configured generic hint is: "Your previous attempt at this step may have been wrong. Reconsider carefully." It is appended to the scratchpad. The historical restart condition omits this hint. The controlled protocol uses identical generic wording for restart and targeted repair.

Informed hints depend on five annotation categories: wrong search query, wrong fact extraction, faulty reasoning, premature or wrong answer, and formatting/tool error. For example, the search

category suggests changing the entity or using more specific search terms, while the extraction category asks the agent to re-read the passage for the requested detail. The hint does not directly insert a replacement step, but its error type comes from a gold-aware reference process. It is therefore privileged even when the answer string is absent from the hint itself.

B.3 Reference Annotation and Fallbacks

The annotator receives the question, reference answer, supporting titles, the failed answer, and the numbered trajectory. Supporting-fact text is included when present. A programmatic check records whether required titles were retrieved and can flag missing pages. The judge is asked for a single earliest responsible step, one error class, and a short justification. Its displayed one-based step index is converted to the zero-based trajectory convention.

The implementation prefers a parsed judge response. If parsing fails and the retrieval check identifies a candidate, it uses the last search step as a retrieval-failure reference; otherwise it defaults to step zero and a reasoning-error label. Parsed out-of-range indices are clamped to the trajectory. These are labeling conventions, not evidence that a uniquely causal step exists. The missing raw annotations prevent reporting the fractions of judge, programmatic, default, and clamped labels. A complete validation must retain these sources rather than treating every reference as a successful judge judgment.

B.4 Localization and Offset Conventions

Score ranking breaks ties toward the earlier step. The top-three repair rule selects the earliest index among its three highest-scored positions, not the first index in a descending-score list. The percentile rule uses linear interpolation to form its threshold and selects the earliest qualifying position. The upstream rule moves two positions through the scored-step list, whereas the additional backtrack offset subtracts two trajectory indices. These operations can differ when scores are missing.

The position-weighted rule min–max normalizes uncertainty and position within the scored trajectory and gives each an equal weight. Writing $\widetilde{U}_i$ and $\widetilde{\imath}$ for those normalized values, the criterion maximized in Appendix A as $0.5\widetilde{U}_i + 0.5(1-\widetilde{\imath})$ is

$$C_i = \tfrac{1}{2}\underbrace{\frac{U_i - U_{\min}}{U_{\max} - U_{\min}}}_{\widetilde{U}_i} + \tfrac{1}{2}\left(1 - \underbrace{\frac{i - i_{\min}}{i_{\max} - i_{\min}}}_{\widetilde{\imath}}\right). \tag{6}$$

When either range is zero, the corresponding denominator is replaced by one. Empty score profiles fall back to origin zero in the repair selector. These choices create positional preferences even without a calibrated error signal; they motivate fixed-early and position-matched controls.

C Detailed Historical Recovery Analyses

C.1 Archived Study and Aggregate Rates

The historical study sampled 500 questions each from HotpotQA, MuSiQue (Trivedi et al., 2022), and 2WikiMultiHopQA (Ho et al., 2020), with Qwen2.5-32B as agent and Qwen2.5-72B as reference annotator. It contains 226, 425 and 203 failed questions, respectively, for a total of 854. These summaries describe different cohorts and a confounded historical execution regime; they are not pooled with the controlled study.

Full restart exceeds generic judge-targeted repair in the archived HotpotQA and MuSiQue means. On 2WikiMultiHopQA, moving the judge-selected origin two steps earlier changes success from 10.0% to 22.0%, versus 21.2% for restart (Figure 4). These observations motivate direct evaluation of recovery origins; they do not identify a causal error cascade. The retrospective uncertainty maxima include outcome-based selection and, on two datasets, judge-informed hints. They are not held-out policy estimates.

Table 4: Archived mean repair success (%). Judge-targeted rows use the generic retry hint. The last row is a retrospective maximum over the evaluated uncertainty grid, including informed variants; it is not a held-out policy. Historical budgets require auditing, and valid confidence intervals cannot be computed from the archived snapshot.

Strategy	HotpotQA	MuSiQue	2WikiMHQA
Random origin	8.1	3.8	19.0
Judge-targeted	10.9	4.9	10.0
Judge-targeted + backtrack 2	12.4	5.6	22.0
Full restart	14.2	5.9	21.2
Retrospective best uncertainty variant	12.8	5.6	25.5

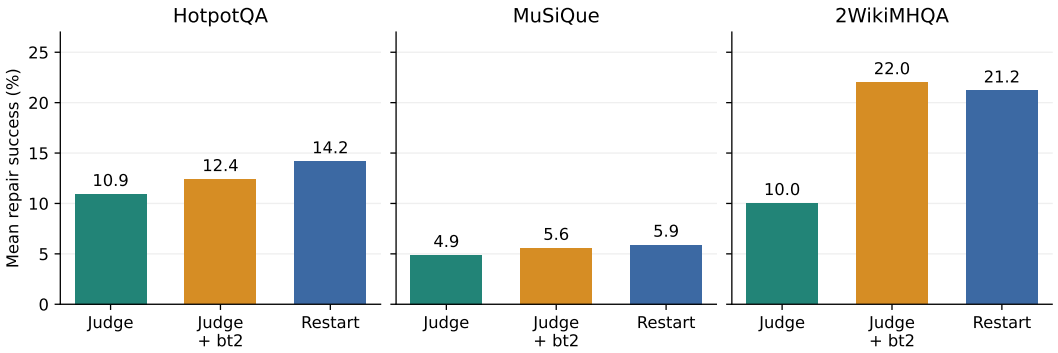

Figure 4: Descriptive archived rates for judge-targeted repair, two-step backtracking, and restart. All bars use mean seed success. The figure does not encode uncertainty intervals or establish a causal mechanism.

Source-hashed CSVs and indexed notebook displays recover baseline success, generated-token/tool costs and fixed-origin hint comparisons. The archived snapshot lacks the per-question logs needed for valid paired intervals; its obsolete row-wise bootstrap intervals are not reused. Human reference validation also remains unavailable. The subsections below preserve the detailed recovery, cost, hint and localization analyses, and Appendix D explains candidate coverage and snapshot limits.

C.2 RQ1: Targeted Repair and Restart

Full restart has higher archived mean success than judge-targeted repair on HotpotQA (14.2% versus 10.9%) and MuSiQue (5.9% versus 4.9%). It also exceeds the judge-targeted variants with two-step backtracking on those datasets. This indicates that selecting a judge-labeled error does not, by itself, guarantee a better recovery origin than the beginning of the trajectory. It does not establish why restart is better: annotation error, prefix quality, remaining steps, prompt differences, and budget overshoot remain alternative explanations.

The recovered random-origin means are 8.1%, 3.8%, and 19.0% in the same dataset order. These are mean-success summaries, unlike the lower majority-over-seeds values in an earlier cross-dataset table. On 2WikiMultiHopQA, random repair exceeds the unbacktracked judge-targeted rate. Thus the judge-selected point is not even an empirical upper bound on these baseline interventions. Random placement remains an incomplete control for uncertainty: it does not match the latter's origin distribution or its possible two-step backtracking.

C.3 RQ2: Moving the Regeneration Origin Upstream

Backtracking the judge-selected origin increases archived mean success by 1.5, 0.7, and 12.0 percentage points on HotpotQA, MuSiQue, and 2WikiMultiHopQA. The largest change occurs on 2WikiMultiHopQA: 10.0% to 22.0%, compared with 21.2% for restart. This motivates evaluating earlier recovery origins. It is not a direct test of error cascades, since backtracking simultaneously

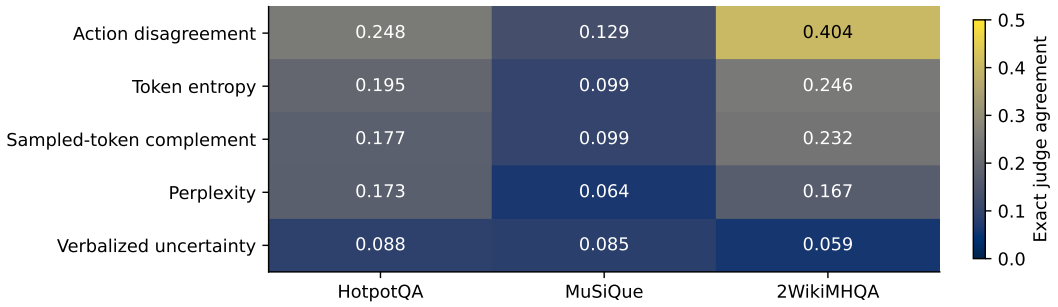

Figure 5: Exact agreement with the reference error step for argmax localization. Larger values indicate annotation agreement, not repair success or calibrated error probabilities. All entries are archived fractions; human validation is not available in the snapshot.

changes retained context, regeneration length, and whether the judge's selected step remains relevant.

There are two distinct questions within this comparison. An offset can improve a given localizer by preserving less context; it can also make many localizers operationally identical by clamping them to zero. The aggregate gains cannot distinguish these cases. A per-question analysis should report the selected origin before and after the offset, the fraction of interventions becoming restarts, and the gain among those retaining a nonempty prefix. These quantities require the original selection records.

C.4 Retrospective Uncertainty Selection

The aggregate maximum over uncertainty configurations is 12.8% on HotpotQA, 5.6% on MuSiQue, and 25.5% on 2WikiMultiHopQA. The corresponding saved notebook identifies an informed probability-complement/top-three variant on HotpotQA, an informed perplexity/position-weighted/backtrack-two variant on MuSiQue, and a generic perplexity/top-three/backtrack-two variant on 2WikiMultiHopQA. These are selected on the same questions used to report their success. Their rates describe the explored grid, not the expected performance of a policy chosen without looking at test outcomes. In particular, the first two maxima do not isolate uncertainty from judge-provided hints.

C.5 RQ3: Localization Is a Separate Outcome

Action disagreement has the largest exact agreement among the five reported argmax metrics in each dataset (Figure 5; the numeric table is in Appendix D). The corresponding values are 0.248, 0.129, and 0.404 for HotpotQA, MuSiQue, and 2WikiMultiHopQA (Table 7). This ordering concerns the reference label, not the success of every metric–rule repair combination. A fair random localization reference must use the eligible steps of each original trajectory, and requires the underlying records. Likewise, a comparison of uncertainty with random repair should match backtracking, hint access, budget, and question coverage, then evaluate a prespecified strategy.

The signal with the greatest annotation agreement need not be the cheapest to acquire: action disagreement requires five additional sampled actions per measured step, whereas stored-token signals reuse the initial rollout. Nor does the highest agreement establish the best downstream policy. The reported repair maximum varies both the signal and the localization rule, sometimes using privileged hints. Comparing that maximum with an argmax localization score would change multiple factors at once.

C.6 RQ4: Recovery Cost and Hint Information

Generated-token and tool costs. Table 5 restores the recorded recovery-only costs for the four baselines. These are rounded means from the same saved displays as their success rates. They exclude the reused prefix from generated-token counts but do not measure prompt processing or all

Table 5: Archived mean generated recovery tokens and recovery tool calls. HP: HotpotQA; MS: MuSiQue; 2W: 2WikiMultiHopQA. Prompt, uncertainty, diagnosis, and selection costs are not included.

Strategy	HP tokens	HP tools	MS tokens	MS tools	2W tokens	2W tools
Random origin	182	2.80	216	3.90	170	2.36
Judge-targeted	236	3.70	320	5.82	171	2.26
Judge + backtrack 2	272	4.51	347	6.39	234	3.68
Full restart	274	4.74	348	6.62	246	4.03

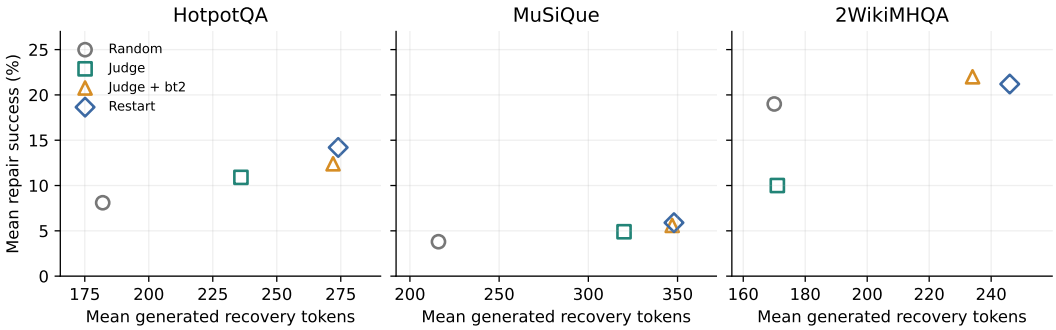

Figure 6: Recovery-token cost versus success for the four recoverable baseline summaries. Markers are historical point estimates, with no confidence intervals. This is not the full-grid Pareto frontier and does not include all inference costs.

policy overhead. They are therefore descriptive resource measurements, not an audited comparison at equal total computation.

Figure 6 plots these costs against success. On HotpotQA, two-step-backtracked judge repair uses 272 generated tokens and 4.51 tool calls on average, versus 274 and 4.74 for restart, while its success rate is lower. On MuSiQue the corresponding token means are 347 and 348. Retaining a prefix does not necessarily yield a substantial generated-token saving. On 2WikiMultiHopQA, backtracked judge repair has both a lower displayed token mean (234 versus 246) and a higher displayed success rate (22.0% versus 21.2%) than restart. This is a point-estimate comparison, not statistically established dominance.

Informed versus generic nudges. An origin-preserving hint comparison must keep backtracking fixed. Table 6 does so for the reference-targeted conditions. Figure 7 shows both descriptive changes. At offset zero, informed hints change the displayed rates by $+0.5$, 0.0, and $+2.2$ percentage points; at offset two the changes are -0.6, 0.0, and -1.0. Zero denotes equality at one-decimal display precision, not exact equality or statistical equivalence. Keeping origins fixed is necessary to distinguish a hint change from a backtracking change.

The direction varies with the origin, so these summaries support neither a blanket advantage nor a blanket penalty for more specific hints. An informed prompt might focus a useful correction, or direct the model toward an incorrect diagnosis. Distinguishing these explanations requires paired outcomes and diagnosis-quality annotations. In particular, the comparison does not demonstrate a gold-free self-correction capability.

D Expanded Archived Diagnostics

D.1 Candidate Coverage and Recoverable Costs

The saved "any uncertainty strategy succeeded" values are 36.0% on HotpotQA, 16.2% on MuSiQue, and 46.0% on 2WikiMultiHopQA. The evaluation code forms this pseudo-strategy by taking the maximum success within each question and seed over the zero-additional-backtrack uncertainty variants, including informed variants. It does not select an answer without the benchmark

Table 6: Reference-targeted hint comparison at fixed backtrack offsets. Delta is informed minus generic, in percentage points, calculated from rounded archived means. Informed hints use the gold-aware annotation.

Dataset	Backtrack	Generic (%)	Informed (%)	Delta (pp)
HotpotQA	0	10.9	11.4	+0.5
HotpotQA	2	12.4	11.8	−0.6
MuSiQue	0	4.9	4.9	+0.0
MuSiQue	2	5.6	5.6	+0.0
2WikiMHQA	0	10.0	12.2	+2.2
2WikiMHQA	2	22.0	21.0	−1.0

Table 7: Exact reference-step agreement for argmax, underlying Figure 5. Entries are fractions.

Metric	HotpotQA	MuSiQue	2WikiMHQA
Action disagreement	0.248	0.129	0.404
Token entropy (max)	0.195	0.099	0.246
Sampled-token complement (max)	0.177	0.099	0.232
Perplexity	0.173	0.064	0.167
Verbalized uncertainty	0.088	0.085	0.059

reference. Also, a rule with zero additional backtracking can still have an upstream preference built into the localizer.

The associated token totals sum nominal member rows. Because several members may select the same origin and hint and reuse one generated suffix, those totals can double-count physical executions. The earlier 47–70-fold compute claim is therefore not retained. These coverage values describe a candidate set under privileged retrospective selection; they should not be placed on an operational Pareto frontier alongside single-output policies. A valid ensemble comparison must freeze a gold-free selector and count unique candidates, acquisition calls, and selection.

D.2 Why Seed Aggregation Changes the Question

For three repair seeds, a question with outcomes $(1, 0, 0)$ contributes $1/3$ to the mean one-attempt repair rate and zero to majority repeatability. A question with $(1, 1, 0)$ contributes $2/3$ and one, respectively. Both summaries can be useful, but mixing them across rows changes the estimand. The restored main table consistently uses mean-success summaries. The earlier notebook's confidence intervals are not reused because its row-wise bootstrap does not preserve the question-level dependence.

The available displays show only ten strategy rows out of 127, including the coverage pseudo-strategy. The reconstruction parses those visible rows, rejects missing or duplicated required baselines, and checks three baseline rates against the cross-dataset CSVs. It does not infer the hidden rows, reconstruct their joint outcome distribution, or treat rounded costs as exact measurements. This distinction is why the cost plot shows baseline points rather than recreating a full-grid Pareto boundary.

D.3 Legacy Interpretation Limits

Context effects are hypotheses. The possibility that retained errors constrain recovery is consistent with the observed benefit of changing origins. However, the archived "trajectory length" statistic is the average number of tool calls in full-restart repairs. It is a post-intervention outcome, not the original failed trajectory's length. Correlating four dataset averages cannot justify a four-step routing threshold. Testing the role of prefix context requires original lengths, within-dataset comparisons, and interventions that manipulate the prefix while holding other factors constant.

Error propagation versus recovery opportunity. Backing up a rollout removes later reasoning and observations, but under the historical total-step limit it also grants more new agent turns. Both

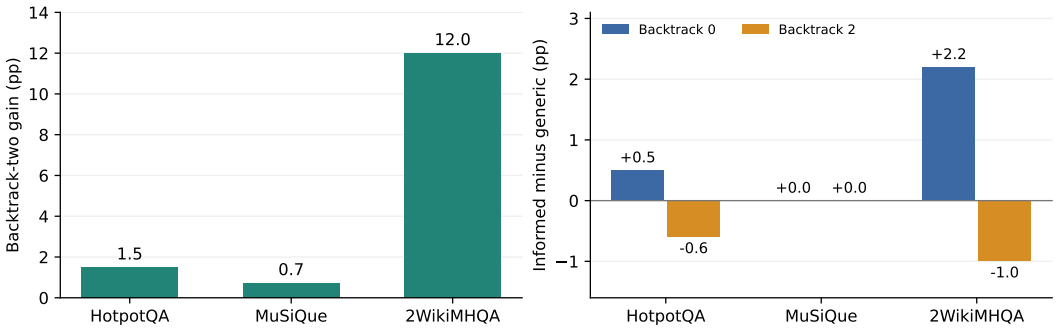

Figure 7: Two descriptive changes in the reference-targeted comparisons. Left: moving the origin two steps earlier, from archived mean-rate CSVs. Right: informed minus generic hints at fixed offsets, from rounded saved notebook displays. Neither panel provides paired confidence intervals.

changes can improve repair without proving that the first retained error caused a downstream uncertainty peak. An error-cascade explanation requires showing the ordering of reference errors and uncertainty peaks, then testing whether the proposed upstream intervention improves recovery beyond a position-matched control. A controlled new-step regime separates some of these factors; the historical total-step regime remains a useful, separately labeled operational setting.

Failure-mode analysis. Two hypotheses are an error preceding the uncertainty peak and legitimate uncertain exploration. The first requires the proportion of localization misses with $k_J < k_U$, its denominator, and annotation source. The existing exploration flag only identifies missed `search`/`lookup` actions, not their correctness. These overlapping hypotheses do not establish the originally asserted 70% and 20% prevalences.

Candidate coverage is not answer selection. An "any strategy succeeded" aggregate uses benchmark answers, so it measures retrospective coverage, not operational ensemble accuracy. Deployment requires a gold-free selector and costs for unique executions and selection; summing nominal strategy costs double-counts shared work.

From diagnosis to a repair policy. A policy must select an origin before observing its repair outcome. Predicting judge agreement does not establish repair utility, and selecting the best test rollout is retrospective. The completed study evaluates six prespecified conditions with matched hints and allowances. An all-origin matrix is deferred; neither an origin-wide optimum nor an all-origin cost frontier is estimated. Appendix E separates completed work from deferred extensions.

The practical implication concerns evaluation: measure recovery directly, include restart, and separate information access from origin selection. The results justify neither a length router, an advantage over diagnosis-and-replay, nor deployment of a gold-selected ensemble.

E Completed Controlled Protocol, Audit and Costs

This appendix records the completed HotpotQA experiment. The protocol and cohort were frozen before new outcomes, but were not publicly preregistered. Historical results elsewhere in the paper are not relabeled as its outcomes. Table 8 summarizes every completed controlled study.

E.1 Frozen Conditions and Model

The main model and tokenizer snapshot was `Qwen/Qwen2.5-32B-Instruct-AWQ`, revision `5c7cb76a268fc6cfbb9c4777eb24ba6e27f9ee6c`. The 72B confirmatory cell uses `Qwen/Qwen2.5-72B-Instruct-AWQ`, revision `698703eae6604af048a3d2f509995dc302088217`. Every agent in this paper is an open-weight model served locally, and the design requires it rather than merely preferring it. The tested policy is computed from the log probability of every sampled token, and the historical signals

Table 8: Completed study coverage, excluding development and historical runs. Repairs count unique executions; trial rows can reuse an execution across conditions. Runtime questions overlap replication, but its 225 executions are additional. Initial evaluations count model/question pairs.

Component	Initial evals.	Failures	Repairs	Trial rows	Primary tests
Main HotpotQA	250	113	1,058	2,034	4
Diagnosis/allowances	250	118	2,079	3,186	6
Replication: token	600	371	4,098	7,791	24
Runtime: subset	–	25	225	225	2
Length-matched swap	400	218	1,133	1,836	2 + 2
Confirmatory swap	1,000	452	1,200	1,200	1
Confirmatory swap, 72B	1,000	446	1,488	1,488	1

also use the top twenty alternatives at each position; many commercial chat APIs do not expose these for all models. Frozen protocols pin each model by revision hash so that replay audits can require zero mismatches, whereas a hosted model can change or be retired without a pinnable identifier, and the runtime comparison needs hardware and cache settings the study controls. The results therefore say nothing directly about closed frontier models, for which this policy cannot be computed as specified; that is also a practical limit on the policy itself.

Scale was tested as far as one family allows. Qwen2.5-72B-Instruct is the largest dense model in the Qwen2.5 family, so the 32B-to-72B contrast holds the family fixed, although each cell draws its own fresh cohort. A larger open model would change the model family as well, and the largest would need multi-GPU serving, which alters the runtime and cost comparison rather than only the model. The 72B cell needed 452 GPU-minutes for 1,000 originals and 1,488 repairs, and each repair took 12.8 s on average against 6.7 s at 32B for a similar number of generated tokens. Because a confirmatory cell must be sized on informative pairs rather than on questions, cost grows with that per-repair time, not with a smaller cohort.

The main study used one AWS London `g7e.2xlarge` with a 96GB RTX PRO 6000 GPU. Initial generation was greedy; repairs used temperature 0.7, seeds $0, 1, 2$ and batch size two. Every repair had at most eight new steps, 512 generated tokens per step, and total cap $B_q = G_q$, the original failed trajectory's generated-token count. Table 9 states each condition's effective origin. The complete failed trace's uncertainty profile selects the origin; the recovery prompt contains only the retained prefix and generic hint. The discarded suffix and reference labels are not inserted into that prompt.

Table 9: The six frozen conditions actually evaluated. k_U is the perplexity argmax, J a uniform draw over eligible original steps, and K_{dev} the mapped development-profile draw described below. Every condition has the same hint and recovery allowance.

Condition	Effective origin
Full restart	0
Offset-matched random	$\max(0, J-2)$, $J \sim \mathrm{Unif}\{0, \ldots, T-1\}$
Fixed early (descriptive)	$\min(1, T-1)$
Uncertainty + backtrack 2	$\max(0, k_U - 2)$
Position-matched random	K_{dev}
Uncertainty, no backtrack	k_U

The position profile contains 24 failed original development trajectories; its effective-origin counts are 14 at zero, two at one, five at three, one at four and two at five. It uses no development repair outcomes. Each k_j is the effective origin that the prespecified policy of Section 4.1 selected on development trajectory j: a trajectory carries no effective origin independent of the policy that selects one, so this is the distribution being matched. For a uniformly sampled development entry j with origin k_j and development trajectory length T_j, the new origin for an evaluated question q whose failed trajectory has T_q steps is

$$K_{\mathrm{dev}} = \left\lfloor \frac{k_j}{\max(1, T_j - 1)} (T_q - 1) + \frac{1}{2} \right\rfloor . \tag{7}$$

The profile and seeded sampling were frozen before test outcomes. This is a normalized-position mapping, not a fitted conditional model for every possible trajectory length. All 50 development source questions are excluded. The six conditions and three seeds allow at most $18N$ unique repair jobs for N failures; shared effective executions reduced 2,034 rows to 1,058 jobs.

E.2 Cohort Reconstruction and Immutable Evidence

The historical 500-question pool was reconstructed with the original sampler at git commit `7a73907`, seed 42, stratification by level and type, and the archived 7,405-record dataset. Its union with 60 AWS development IDs contains 549 distinct exclusions. The original pool file's displayed date, size and one visible question ID corroborate the reconstruction; they are not a comparison of its full contents. Undocumented historical runs cannot be excluded. The remaining 6,856 sorted IDs supplied the 250-question cohort, sampled without replacement using seed 20260912.

The freeze was recorded on September 12 at 03:27:45 UTC, before the 04:06:56 EC2 start request and 04:11:45 batch-runner start. The five batches were completed and pooled at 05:25:03 UTC without outcome-based changes. The immutable identities are:

- Study: `e4ea08c8291b0f797d05b15110d762634ec48d57ae53e05aaf4373c805986308`.
- Source: `b418bca20be1674d65dc76dc14c42f2b6304eee6856d34a6f083c3d4e6ef4eb9`.
- Raw dataset: `a9b4c5906587ac0d20eb83dcb29d0240e268e342c170f7f22988f59f13fa82c3`.
- Position profile: `3e4e834cca3857f409d0dae0baf71065825687883bc1ae5eeaf1ebb9c02bb8c`.

The local reproducibility record includes `study-freeze.json`, `test-ids-250.json`, `historical-reconstruction.json`, `archive-verification.json` and all five executed batch archives. The source identity above is distinct from the individual ZIP/archive checksums.

E.3 Question-Level Inference and Secondary Outcomes

All five batches are pooled with equal question weight after averaging seeds. Individual 95% intervals use 10,000 paired question bootstrap resamples with analysis seed zero. The sign-flip tests assume within-question exchangeability; four primary p-values receive one Holm correction. In the main, diagnosis, replication and runtime studies these p-values are Monte Carlo estimates from 10,000 random sign assignments with an add-one correction, so a value can sit slightly above the exact floor of Appendix H.6 (Qwen 2Wiki: 0.131 against 0.125); the swap studies enumerate the exact null. Fixed early remains descriptive. The intervals themselves are not simultaneous family-wise intervals.

Table 10: Prespecified treatment-minus-control comparisons in repair success, in percentage points. All use 113 paired question means. Intervals are individual 95% bootstrap intervals; p-values are paired sign-flip tests.

Control	Delta (pp)	95% interval (pp)	Raw p	Holm p
Full restart	−1.47	[−4.13, +0.88]	0.358	1.000
Random + backtrack 2	−0.88	[−4.13, +2.06]	0.695	1.000
Position-matched random	+0.00	[−2.95, +2.95]	1.000	1.000
Uncertainty, no backtrack	+1.18	[−2.95, +5.31]	0.682	1.000

Strict exact match and continuous answer F1 are exploratory sensitivities on the same 113 failures. Their full paired outputs are preserved in `pooled-analysis-local.json`; they are not substituted for the primary gate or its comparison family. Changing to an EM-only failure gate would include additional initially accepted questions for which this experiment did not run repairs. Three seeds per question do not triple the sample size. An ideal reference-gated overall success rate would

be $A_0 + (1 - A_0)R_a$ when initially successful answers are left untouched; this identity does not measure a deployed detector's false alarms or regressions.

E.4 Independent Artifact Audit

All 180 CPU tests and all 22 plan-only notebook cells passed before launch. Each completed batch was archived with a checksum and independently checked after retrieval. The auditor loaded the frozen source bundle, verified source/input hashes, reconstructed prespecified origins and prompts, checked full question/seed/condition coverage, replayed tool observations, and recomputed answer metrics with the official HotpotQA reference scorer.

Table 11: Completed batch coverage. Violations include replay/scoring mismatches and generated-token/new-step allowance violations; every count is zero.

Batch	Initial	Failed	Unique repairs	Condition/seed rows	Violations
1	50	20	195	360	0
2	50	21	212	378	0
3	50	25	223	450	0
4	50	21	196	378	0
5	50	26	232	468	0

Per-batch coverage is in Table 11. The audit covers 1,308 trajectories, 8,552 observations, 1,742 retained prefix steps, 241 input hashes and 2,034 origin/prompt checks. It found zero mismatches and zero token or new-step violations. Local pooling of the retrieved records agrees with the AWS analysis within 10^{-12}. These checks verify the saved evidence and local tool replay, not bitwise reproduction of stochastic continuations or independent human validation of reference judgments.

E.5 Compute

The study ran for about 81 minutes on the single GPU described above, from instance start to shutdown; records were retrieved in a separate session of about seven minutes without repeating any inference. Original generation used 57,499 generated tokens and unique repairs used 303,700 generated and 5,292,100 prompt tokens. Acquisition, prefix processing, tool latency, batching and selection must be accounted for to support a policy-efficiency claim; these totals alone do not supply that comparison.

E.6 Separate Follow-up and Deferred Scientific Work

The six-condition study does not include a diagnosis/replay comparator or additional token allowances. These comparisons were subsequently completed on a separate cohort (Section 7). Replication on the other two QA datasets, a second model and measured runtime are covered by the separately frozen extension in Appendix H. Its two-model replication and runtime component are complete (Sections 8 and 9). A full-origin matrix and new reference annotations from the 72B judge are outside these completed studies. FEVER (Thorne et al., 2018) remains excluded because its inspected loader substituted claim text for retrieved evidence. The extension independently audits MuSiQue conversion, answer aliases and dataset-specific reference scoring.

Historical localization claims still require independent human judgments. The existing proposal samples 50 failed traces per QA dataset independently of repair success, with two annotators blinded to judge labels and outcomes: 150 trajectories and 300 initial judgments. Individual judgments and adjudication would be preserved separately, with ambiguity and multiple-error cases allowed. No such annotations or annotator labor were supplied here. Context interventions and original-length analyses would require separately declared hypotheses; this paper does not infer error-cascade prevalence or a trajectory-length routing threshold from the archived averages.

Table 12: Exploratory execution overlap with the treatment. Shared counts are out of 339 question/seed pairs. Different-question counts are the questions, out of 113, with at least one distinct execution. Wins/losses/ties compare three-seed question means, not independent seed trials.

Control	Shared / 339	Different q. / 113	Wins / losses / ties
Full restart	210	43	5 / 8 / 100
Random + backtrack 2	173	68	9 / 9 / 95
Position-matched random	148	93	10 / 10 / 93
Uncertainty, no backtrack	51	96	15 / 10 / 88

F Review-Stage Diagnostics

These exploratory analyses were added after the primary study completed. They retain all 250 initial questions, 113 failures, six conditions and three seeds. The diagnostic loader verifies 398 input-to-archive comparisons against the five checksummed archives, rescoring original and repaired answers and rejecting missing trials, inconsistent shared executions or allowance violations. Exported trial and initial-question CSVs, the analysis-code hashes and the review-stage analysis plan accompany the diagnostic JSON. The original primary estimates and Holm family are unchanged.

F.1 Distinct Intervention, Resources and Metric Sensitivity

How much distinct intervention was tested. The treatment and restart share all three executions on 70 of the 113 failed questions. Only 43 questions have a nonzero treatment origin. Let D_q indicate such an origin and let Δ_D denote the mean treatment-minus-restart difference within those questions. Because identical origins, hints and seeds share an execution,

$$\widehat{\Delta} = \frac{N_D}{N}\widehat{\Delta}_D = \frac{43}{113}(-3.876 \text{ percentage points}) \simeq -1.47 \text{ percentage points}. \tag{8}$$

The subgroup interval is [−10.85, +2.33] points. This does not change the primary estimand or discard its zero differences: $N = 113$ remains the primary resampling sample. It shows how often the policies can disagree and why the subgroup estimate is imprecise. Across all questions, treatment has a higher seed mean on five, a lower mean on eight, and the same mean on 100 (Table 12). The position-matched comparison has 93 questions with at least one distinct execution, so the lack of a primary advantage cannot be attributed solely to collapse into restart.

Observed resources and unfinished repairs. Uncertainty repair uses 269.9 generated and 4083.3 prompt tokens per attempt, versus 271.6 and 3153.3 for restart: 29.5% more prompt tokens with similar output-token use. The paired differences are -1.7 output tokens (individual 95% interval $[-8.9, +5.7]$) and $+930.0$ prompt tokens ($[+579.2, +1321.9]$). Prefix reuse therefore does not imply lower recorded prompt usage. These counts do not measure cached-prefill savings, FLOPs or policy latency; Table 13 reports all six policies.

Original-failure token allowances range from 77 to 548 (median 344), and original failed trajectories have a median of six steps. The treatment ends 103/339 repairs at the token budget and 65/339 at the step limit; together, 168/339 attempts emit no final answer. All 637 original failed-trace steps have usable perplexity, with no empty-profile fallback. Thus missing signal records do not explain the observed result. Stopping at a cap does not identify what the same attempt would achieve with more tokens. Section 7 reports separate executions at three allowances on another frozen cohort.

Metric sensitivity and population interpretation. The eight EM/F1 comparisons in Table 14 and the descriptive fixed-early contrast are exploratory; they neither expand the primary family nor select a replacement policy. Keeping initially accepted answers and repairing only the 113 rejected ones yields 59.20% full-cohort gate success for treatment and 59.87% for restart (Table 15). These are reference-gated outcomes, not performance of a deployed failure detector. Forty initially accepted answers fail strict EM and were never repaired; an EM-only failure cohort would need additional executions.

F.2 Policy Costs and Termination

Costs in Table 13 average seeds within each question. Each standalone policy pays for the execution it would use, even if that execution was shared with another condition during this experiment. The study-level token totals instead deduplicate physical executions. All six policies use stored information and incur zero additional selection-model requests. CPU scoring, log-probability extraction, cache behavior and scheduling overhead are not measured by token totals.

Table 13: Exploratory mean recovery usage per attempted repair, on 113 questions and three seeds. F/B/S are counts terminating with an answer, at the token budget, or at the new-step limit; each row sums to 339. Prompt and output tokens are separate quantities, not an estimate of FLOPs.

Condition	Output tokens	Prompt tokens	Requests	Tool calls	F / B / S
Full restart	271.6	3153.3	5.26	4.54	162 / 99 / 78
Random + backtrack 2	268.4	4179.6	5.15	4.40	171 / 97 / 71
Fixed early (descriptive)	258.1	3720.6	4.82	4.07	193 / 75 / 71
Uncertainty + backtrack 2	269.9	4083.3	5.11	4.35	171 / 103 / 65
Position-matched random	264.5	3885.8	5.01	4.26	178 / 90 / 71
Uncertainty, no backtrack	255.6	5077.3	4.77	3.99	186 / 91 / 62

A repair can reach its token cap on a request that also emits a final answer or reaches the step limit. Consequently, the recorded termination reason and token-cap equality are different diagnostics: treatment has 103 budget-labeled stops but 152/339 executions use their complete token cap. There is no zero-token allowance in this study. Initial failures comprise 61 completed incorrect answers and 52 original step-limit stops. A different allowance requires separate executions; it cannot be reconstructed by reweighting these stopped traces.

F.3 Secondary Metrics and the Actual Failure Gate

The strict-EM treatment-minus-restart difference is -1.47 points (95% interval $[-3.83, +0.59]$). Against uncertainty without backtracking, the exploratory EM difference is $+2.95$ points ($[+0.29, +5.60]$); the corresponding primary-gate comparison remains inconclusive. The eight secondary EM/F1 intervals in Table 14 are unadjusted and exploratory. A separate descriptive comparison with fixed early gives -1.77 primary-gate points ($[-6.19, +2.65]$), without adding it to the prespecified family or selecting a replacement policy.

Table 14: Exploratory treatment-minus-control differences on the same 113 initial failures. Individual 95% question-bootstrap intervals are not adjusted for the eight secondary comparisons. EM is in percentage points; F1 differences are multiplied by 100. The original primary tests use the answer-success gate, not these metrics.

Control	EM delta [95% CI], pp	100ΔF1 [95% CI]
Full restart	−1.47 [−3.83, +0.59]	−1.67 [−4.28, +0.64]
Random + backtrack 2	+0.29 [−2.06, +2.65]	−1.27 [−4.14, +1.45]
Position-matched random	+1.47 [−0.59, +3.83]	+0.11 [−2.48, +2.62]
Uncertainty, no backtrack	+2.95 [+0.29, +5.60]	+1.80 [−1.35, +5.07]

The full-cohort calculation retains the same 40 F1-accepted but non-EM original answers for every policy. It never treats the reference gate as a learned detector and never uses majority-over-seeds voting or an oracle best-of-three selector. Question-bootstrap intervals for these descriptive rates are included in the diagnostic JSON.

F.4 Illustrative Traces and Selection Rule

For each of three outcome strata—higher treatment seed mean, higher restart seed mean, or zero success for both—we display the first sorted question ID and all three seeds. Questions are quoted verbatim from the dataset, including its spelling. These outcome-selected cases illustrate measured

Table 15: Outcomes over all 250 questions under the implemented reference gate: preserve initially accepted answers and apply one sampled repair only to the original 113 failures. Three seeds estimate that repair's mean outcome. Initial rates were 54.80% gate success, 38.80% EM and 0.5164 F1. These are not outcomes of an EM-only rerun or a deployed failure detector.

Recovery condition	Gate success (%)	EM (%)	F1
Full restart	59.87	42.67	0.5597
Random + backtrack 2	59.60	41.87	0.5579
Fixed early (descriptive)	60.00	41.20	0.5564
Uncertainty + backtrack 2	59.20	42.00	0.5522
Position-matched random	59.20	41.33	0.5517
Uncertainty, no backtrack	58.67	40.67	0.5440

behavior; they do not estimate prevalence, validate diagnoses or establish corrected causal mechanisms. F1 uses each full displayed answer; none of the 18 answers passes strict EM.

Case 1: Higher treatment seed mean. *In 1632, where was one of the most highly acclaimed English architects born?* Reference answer: East Knoyle. The original trace has 8 steps and a 439-token recovery allowance.

Condition	Origin	Gate; F1, seeds 0/1/2	Answers, seeds 0/1/2
Restart	0	0/0/0; 0.00/0.00/0.00	*no answer*; *no answer*; *no answer*
Uncertainty + bt2	3	0/0/1; 0.00/0.00/0.67	*no answer*; *no answer*; East Knoyle, Wiltshire, England

Question ID: `5a761976542992db9473728`.

Case 2: Higher restart seed mean. *Was the Metropolitan Life Insurance Company Tower [Met Life Tower] or the 15 Hudson Yards building designed by the firm of Napoleon LeBrun & Sons?* Reference answer: The Metropolitan Life Insurance Company Tower. The original trace has 8 steps and a 458-token recovery allowance.

Condition	Origin	Gate; F1, seeds 0/1/2	Answers, seeds 0/1/2
Restart	0	0/0/1; 0.31/0.32/0.62	The Metropolitan Life Insurance Company Tower was designed by Napoleon LeBrun & Sons, but there is no information to suggest that 15 Hudson Yards was designed by the same firm.; The Metropolitan Life Insurance Company Tower was designed by Napoleon LeBrun & Sons, but there is no information to suggest that 15 Hudson Yards was designed by them.; The Metropolitan Life Insurance Company Tower was designed by Napoleon LeBrun & Sons.
Uncertainty + bt2	4	0/0/0; 0.31/0.00/0.15	The Metropolitan Life Insurance Company Tower was designed by Napoleon LeBrun & Sons, but there is no information to suggest that Napoleon LeBrun & Sons designed the 15 Hudson Yards building.; *no answer*; The Met Life Tower was designed by Napoleon LeBrun & Sons, but the design firm for 15 Hudson Yards is not mentioned in the provided information.

Question ID: `5a7322a25542991f9a20c634`.

Case 3: Both policies unsuccessful. *Which movie is loosley based off the Brother Grimm's "Iron Henry"?* Reference answer: The Frog Prince. The original trace has 8 steps and a 382-token recovery allowance.

Condition	Origin	Gate; F1, seeds 0/1/2	Answers, seeds 0/1/2
Restart	0	0/0/0; 0.00/0.00/0.00	*no answer*; *no answer*; *no answer*
Uncertainty + bt2	5	0/0/0; 0.00/0.00/0.00	There is no; *no answer*; *no answer*

Question ID: `5a733b835542991f9a20c6b2`.

F.5 Precision and Further Evidence

The 43-question nonzero-origin comparison has a 95% interval $[-10.85, +2.33]$ percentage points. Its width limits any claim about targeted repair when a prefix is actually retained. For prospective planning, $n \simeq 1.96^2 \widehat{\sigma}_d^2 / h^2$ estimates the independent questions needed for an individual 95% interval half-width h, using the observed variance of paired question differences. For a two-point half-width, the four comparisons give approximately 179, 257, 248 and 476 failed questions, in the order of Table 10. These are design sensitivities, not observed power, guaranteed future precision, simultaneous intervals or an outcome-dependent stopping rule for the completed study.

Each additional experiment must declare its scope and affordable sample size before new outcomes. The separate follow-up now evaluates a same-model diagnosis/replay adaptation and three recovery-token allowances (Section 7); the replication and runtime components test transportability across a second model and two further datasets, and measured completion time (Sections 8 and 9). Human labels are required for new error-localization or mechanism claims, but are not substituted for the deterministic answer scorer in the present primary comparison.

G Diagnosis Follow-up: Complete Results and Costs

The follow-up freezes five batches of 50 new HotpotQA questions before outcomes. Its 250 IDs exclude the 799 documented earlier IDs. Initial generation succeeds on 132 questions (52.8%); 89 are exact matches (35.6%) and mean answer F1 is 0.4980. Each of the 118 failures receives three policies, three seeds, and three allowances, for 3,186 logical rows. The Qwen2.5-32B revision, common retry hint, and eight-new-step allowance match the earlier study. Initial generation uses temperature zero; repairs use 0.7.

Table 16: Separate diagnosis follow-up. Each policy/allowance has 354 seed trials on the same 118 initial failures. Success is answer EM or F1 at least 0.5. EM and F1 retain that failure cohort.

Allowance	Policy	Successes	Success (%)	EM (%)	F1
0.5×	Restart	10/354	2.82	2.54	0.0303
0.5×	Uncertainty, backtrack 2	10/354	2.82	2.54	0.0286
0.5×	Diagnosis/replay	15/354	4.24	3.67	0.0649
1×	Restart	37/354	10.45	7.63	0.1206
1×	Uncertainty, backtrack 2	35/354	9.89	7.06	0.1122
1×	Diagnosis/replay	36/354	10.17	7.91	0.1306
2×	Restart	44/354	12.43	9.32	0.1654
2×	Uncertainty, backtrack 2	39/354	11.02	8.19	0.1500
2×	Diagnosis/replay	39/354	11.02	9.32	0.1515

Diagnosis uses the same model with temperature zero, a 512-token maximum, and one deterministic seed. The response supplies an integer origin and a reason. Only the origin controls repair; neither its prose nor discarded suffix evidence is added to the retry prompt. A malformed response would fall back to restart and retain its cost; none of the 118 diagnoses required that fallback. Reference labels determine failure eligibility but do not enter diagnosis or recovery prompts. This is an untrained origin-selection adaptation, not a trained-method reproduction.

Table 17: Uncertainty minus control in the diagnosis follow-up. Intervals are individual, unadjusted 95% question-bootstrap intervals (10,000 resamples). All six two-sided question sign-flip tests share one Holm family.

Allowance	Control	Δ (pp)	95% interval (pp)	p	p_{Holm}
0.5×	Restart	+0.00	[−2.26, +2.26]	1.000	1.000
0.5×	Diagnosis/replay	−1.41	[−4.52, +1.41]	0.471	1.000
1×	Restart	−0.56	[−3.67, +2.54]	0.876	1.000
1×	Diagnosis/replay	−0.28	[−4.24, +3.39]	1.000	1.000
2×	Restart	−1.41	[−4.80, +1.41]	0.492	1.000
2×	Diagnosis/replay	+0.00	[−3.67, +3.39]	1.000	1.000

Table 18: Mean incremental resource use per standalone recovery attempt, including a complete diagnosis charge for diagnosis/replay. Generated-token allowances apply to recovery; prompt processing and diagnosis add cost.

Allowance	Policy	Generated	Prompt	Requests
0.5×	Restart	151.3	1463.2	3.49
0.5×	Uncertainty, backtrack 2	151.6	2146.9	3.41
0.5×	Diagnosis/replay	188.5	3289.1	4.19
1×	Restart	274.3	3311.8	5.38
1×	Uncertainty, backtrack 2	274.6	4504.8	5.36
1×	Diagnosis/replay	296.1	5371.0	5.84
2×	Restart	295.5	3522.6	5.56
2×	Uncertainty, backtrack 2	295.1	4710.7	5.53
2×	Diagnosis/replay	311.3	5555.6	5.97

Physical execution sharing yields 2,079 unique repairs and 118 diagnoses. Independent audits replay and rescore 2,329 original/repair trajectories, check 13,504 observations, 2,637 retained prefix steps, and 388 input hashes, and find zero mismatches. Unique repairs use 496,780 generated tokens, 7,763,726 prompt tokens, and 9,731 model requests; diagnoses use 4,756 generated tokens, 145,641 prompt tokens, and 118 requests. A finish action counts as a model request, not a retrieval call. The pooled local analysis matches all 3,186 remote trial rows and 455 numeric results within 10^{-12}.

An earlier development audit erroneously counted finish as retrieval; it stopped before any main question. The corrected auditor and a failing-first regression test changed neither the cohort nor its model, prompts, seeds, allowances, or planned comparisons. Failed attempts remain preserved. The completed records support token and request accounting, not equal runtime or FLOPs.

H Frozen Replication and Runtime Extension

This separate extension was frozen before its September 14, 2026 launch. The length-matched swap in Appendix H.7 and its two confirmatory cells are later, separately frozen studies (September 16–24), placed here because they extend the replication's position control. The extension's two-model replication and runtime component completed on September 15. All six cells and their local/remote analyses are independently reproduced. The protocol adds newly sampled HotpotQA, MuSiQue and 2WikiMultiHopQA cohorts to the completed studies above. Earlier partial archives remain preserved alongside the completed records.

H.1 Cohorts, Models and Scoring

Each dataset contributes 100 main and 20 disjoint development questions, with identical question IDs across the two models: 600 completed main model/question evaluations and 120 development evaluations. Declared historical exclusions contain 1,049 HotpotQA IDs, 500 MuSiQue IDs and 500 2WikiMultiHopQA IDs. Exclusion provenance retains the original historical pool limitation described above. The models are Qwen2.5-32B-Instruct-AWQ at revision `5c7cb76a268fc6cfbb9c4777eb24ba6e27f9ee6c` and Mistral-Nemo-Instruct-2407

in bfloat16 at revision `04d8a90549d23fc6bd7f642064003592df51e9b3`. Their failure cohorts are determined separately by the same reference gate; conditional recovery rates therefore do not directly rank model reliability.

MuSiQue retains full source paragraphs, answer aliases and repeated titles in their source order. 2Wiki scoring includes its source aliases and demonyms. Support annotations are available to evaluation alone. The input audit performed 174,191 checks against the dataset reference scorers with zero mismatches. The recovery gate remains answer EM or F1 at least 0.5, with strict EM and F1 reported on that same initial-failure cohort. Appendix H.5 quantifies how far the conclusion depends on that gate: 52 initial answers are accepted on the F1 threshold alone and so are never repaired, and 34.1% of accepted repairs fail strict EM.

H.2 Recovery Policies and Statistical Families

Seven policies receive three seeds, the same generic retry hint, a recovery allowance of one original generated-token count, eight new steps and a 512-token per-request maximum. They are restart, offset-matched random, fixed early, uncertainty with two-step backtracking, development-fitted random, uncertainty without backtracking, and same-model diagnosis/replay. The development-fitted random policy retains the frozen identifier `position_matched_random`; its name does not imply exact realized test matching. Each model/dataset fits the distribution using only failures among its 20 development questions. The explicit origin-zero fallback for no development failures was unused in all six cells. Diagnosis pays its full measured acquisition cost. Identical origin/prompt/seed executions are shared in the token experiment, while each policy is charged the cost of its own attempt.

Uncertainty is compared with restart, development-fitted random, no-backtracking uncertainty and diagnosis/replay in each of six cells. These form one prespecified 24-comparison Holm family. Questions are the paired resampling units after averaging their three seeds; individual 95% bootstrap intervals use 10,000 resamples. A zero-failure cell has no conditional repair estimate and contributes $p = 1$ to the fixed family. No sample-size increase or policy selection is permitted in response to extension outcomes.

H.3 Realized Position Balance

These post-results diagnostics describe the frozen policies; they were not prespecified balance tests and do not refit the control or change any primary estimate. Let k be the zero-indexed regeneration origin, which retains k original steps, and T the original trajectory length. We report raw retained steps and normalized origin $z = k/\max(1, T-1)$, with $z = 0$ for a one-step trace. Test distributions weight each failed question equally, then its three seeds equally. The repeated deterministic uncertainty origins do not supply three independent questions. Development distributions contain one origin per failed development question, before random sampling and rounding onto test lengths. Figure 8 shows all three distributions.

Table 19: Exploratory replication position diagnostics. N_d is the number of development failures used in the fit; N_f is the main-failure count. U denotes uncertainty and C development-fitted random. P_0 is the origin-zero percentage and $\bar{k}$ the mean retained-step count. $D = \sup_z |F_U(z) - F_C(z)|$ is the largest absolute gap between empirical normalized-origin cumulative distributions, on a 0–1 scale, with no significance test. Shared counts identical physical executions among $3N_f$ paired question/seed attempts.

Cell	N_d	N_f	P_0^U	P_0^C	$\bar{k}_U$	$\bar{k}_C$	D	Shared
Qwen / HotpotQA	6	43	62.79	82.17	0.91	0.25	0.194	65/129
Qwen / MuSiQue	16	72	33.33	29.63	1.65	1.75	0.125	48/216
Qwen / 2Wiki	6	28	60.71	86.90	1.07	0.49	0.262	45/84
Mistral / HotpotQA	13	78	67.95	75.21	0.59	0.43	0.085	123/234
Mistral / MuSiQue	19	83	54.22	48.59	1.10	1.00	0.129	82/249
Mistral / 2Wiki	11	67	59.70	53.23	0.88	0.86	0.129	87/201

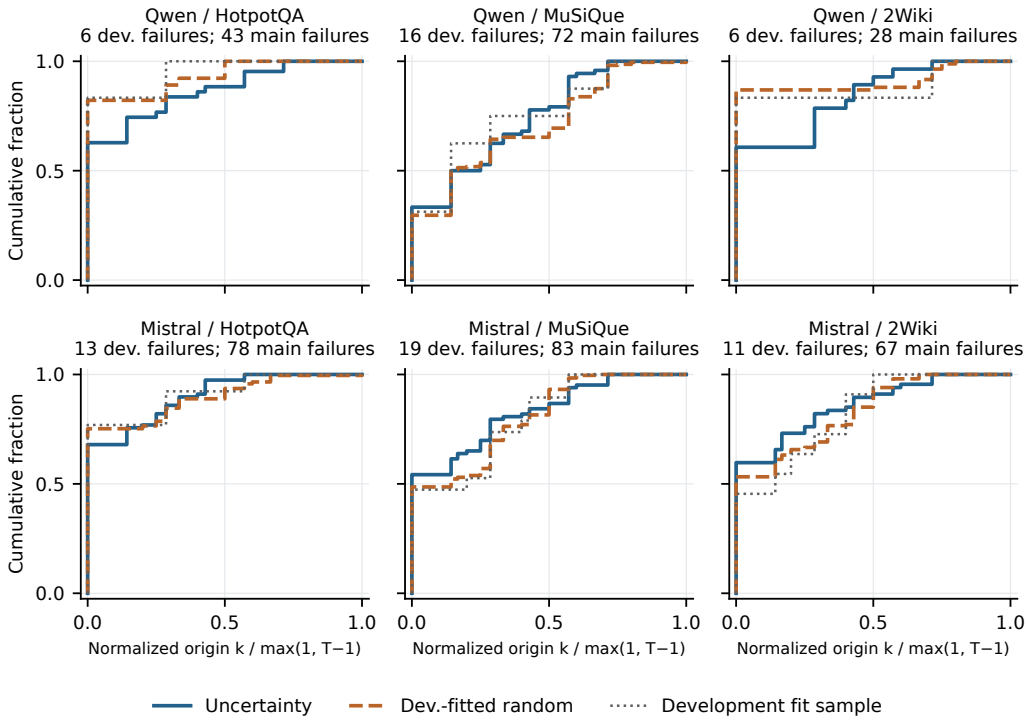

Figure 8: Full normalized-origin empirical cumulative distributions in all six replication cells. Solid and dashed curves describe main-failure policy attempts; dotted curves show the failed development traces used to fit each control. At a horizontal value z, the height is the fraction of origins at or below z. The vertical jump at zero counts full restarts. Differences between the dotted and dashed curves reflect test lengths, discrete rounding and the realized random draws; no test uncertainty is used to fit the control.

Qwen HotpotQA and Qwen 2Wiki each fit six development failures, five with origin zero. On main failures, their controls choose zero in 106/129 and 73/84 rows, versus 81/129 and 51/84 for uncertainty. Mean retained lengths also differ: 0.25 versus 0.91 steps on HotpotQA and 0.49 versus 1.07 on 2Wiki (control versus uncertainty). Origin-zero balance alone would miss other differences: Qwen MuSiQue differs by only 3.70 percentage points at zero but reaches a 0.125 cumulative-distribution gap elsewhere.

Shared token executions require the same question, seed, origin, prompt and allowance. The two policies share 22.2–53.6% of paired attempts (Table 19); the remaining attempts use different origins. Neither shared attempts nor similar marginal distributions prove equal outcomes or question-specific error localization. The saved row-level join and complete support counts retain each original length, raw origin and normalized origin. Their builder checks raw-record identities and hashes, reconstructs each development-only profile, and verifies every reported choice against the frozen policy and execution.

The comparison therefore evaluates these two fixed policies, but their realized imbalance limits attribution of any difference specifically to uncertainty information beyond position. The main study's closer zero-origin balance cannot be assumed for replication. Stronger evidence would require a larger development failure sample and a newly frozen study with declared position/length diagnostics; retuning on these completed test results would not repair the confirmatory comparison. Appendix H.7 reports such a study, run separately and frozen before execution.

H.4 Measured Runtime Scope

The runtime component answers a completion-time question for one execution configuration and is reported descriptively. Measured time covers selection, prefix replay and recovery generation for the attempt; diagnosis acquisition is measured once per question and charged to each of that question's attempts. A deadline plays no causal role during execution: every attempt runs to completion and the threshold is applied afterwards, so a deadline classifies completions and never cancels work or shortens an attempt.

Across 25 questions and 225 attempts, median measured time is 4.45 s for uncertainty, 4.93 s for restart and 4.50 s for diagnosis, with the distributions overlapping throughout (Figure 9, Table 20). 79 of 225 attempts stop at the new-step limit and 0 stop at a generated-token limit, because this component deliberately removes the token cap that the main and replication studies impose. The longest attempt takes 10.92 s, and no attempt that runs past ten seconds succeeds, so the ten-second threshold changes no policy's success count; it classifies completions rather than constraining them.

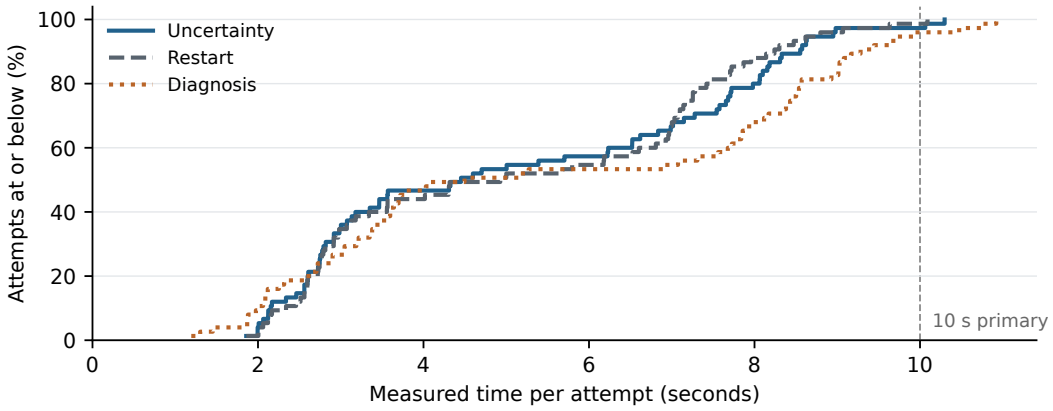

Figure 9: Empirical cumulative distribution of measured time per attempt. At a horizontal value t, the height is the percentage of that policy's attempts completing in at most t seconds. The dashed line marks the primary ten-second threshold, which is applied after execution rather than enforced during it.

These numbers describe this configuration only. Deployed prefix caching, concurrency, repeated acquisition and a matched total-resource comparator are all outside it, and none of them can be

Table 20: Measured time per attempt in seconds, with completion counts. Finished attempts emit an answer; step-limited attempts exhaust the new-step allowance. No attempt stops at a generated-token limit in this component.

Policy	n	Median	q_{75}	q_{90}	Max	Finished	Step-limited
Uncertainty	75	4.45	7.67	8.46	10.30	49	26
Restart	75	4.93	7.21	8.20	10.09	48	27
Diagnosis	75	4.50	8.44	9.24	10.92	49	26

inferred from these measurements. Disabling prefix caching is not neutral between the arms. Prefix-preserving repair resubmits a retained prefix that a served deployment would ordinarily hit in the key-value cache, and it carries 29.5% more prompt tokens than restart, so recomputing that prompt charges the targeted policy for work a cache would supply. Enabling reuse would therefore be expected to move latency in the targeted policy's favor, by an amount these measurements cannot bound; the deadline result reported here should be read as a measurement of this uncached configuration rather than as a latency ranking that would survive a cached one. A broader efficiency claim would need a separately frozen runtime design with an adequate sample and realistic cache and concurrency settings; active cancellation would be required only for a claim about resources saved by stopping late attempts.

H.5 Dependence on the Success Definition

The study gate accepts an answer on exact match *or* token F1 at least 0.5. Rescoring the completed records under strict EM is exploratory: it does not create a new failure cohort, because questions accepted under the gate were never offered repair and their repair outcomes do not exist. We write "accepted" throughout for accepted under the study gate, never for strict EM.

Of 600 initial evaluations, 229 are accepted and 177 match exactly, so 52 answers (8.67% of initial evaluations) are accepted on the F1 threshold alone and are therefore never repaired (Table 21 and Figure 10). This is a property of the gate, not an estimate of annotation error.

Table 21: Initial evaluations by success definition. Accepted uses the study gate; EM counts strict exact match; Gate only counts answers accepted on the F1 threshold alone, which are never offered repair.

Model	Dataset	N	Accepted	EM	Gate only
Qwen	HotpotQA	100	57	44	13
Qwen	MuSiQue	100	28	22	6
Qwen	2Wiki	100	72	64	8
Mistral	HotpotQA	100	22	13	9
Mistral	MuSiQue	100	17	13	4
Mistral	2Wiki	100	33	21	12
All		600	229	177	52

The dependence is larger among repairs. Across uncertainty and restart attempts, 84 of 246 accepted repairs (34.1%) fail strict EM. A reader who requires exact match would therefore see materially fewer successes for both policies. We do not report an EM-gated contrast as a result: the eligible failure set itself would change, and those additional repair outcomes were never generated.

Conditional repair rates describe failures only. Table 22 instead reports the full cohort, preserving each already-accepted answer and taking a repair outcome only on failures, which is what a deployment with a perfect failure detector would observe. Differences remain small in every cell, from -2.33 to +1.67 percentage points, and the cells disagree in direction. These are descriptive; they carry no separate test and no family-wise adjustment.

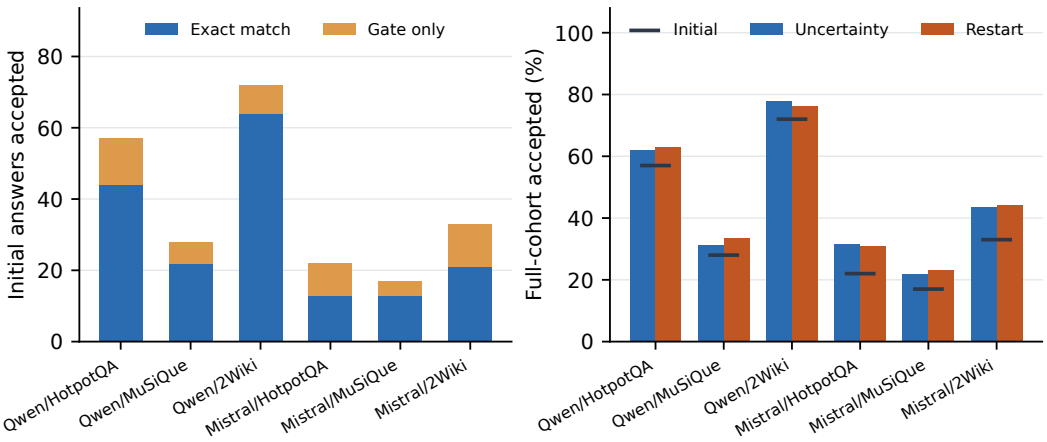

Figure 10: Left: initial accepted answers split into strict exact match and those accepted on the F1 threshold alone. Right: full-cohort accepted rate per policy, preserving answers already accepted; the dash marks the initial rate, so bar height above it is what repair adds.

Table 22: Full-cohort accepted percentage, preserving initially accepted answers and applying repair outcomes only to failures. Δ is uncertainty minus restart in percentage points.

Model	Dataset	Initial	Uncertainty	Restart	Δ
Qwen	HotpotQA	57.0	62.0	63.0	−1.00
Qwen	MuSiQue	28.0	31.3	33.7	−2.33
Qwen	2Wiki	72.0	78.0	76.3	+1.67
Mistral	HotpotQA	22.0	31.7	31.0	+0.67
Mistral	MuSiQue	17.0	22.0	23.0	−1.00
Mistral	2Wiki	33.0	43.7	44.3	−0.67

H.6 Paired Outcomes, Overlap, Termination and Practical Bounds

These are descriptive extensions of the frozen replication. Every primary Holm-adjusted output is reported unchanged; the exact values below are sensitivity checks, and no pooled confirmatory statistic is computed because question identifiers are shared across model cells.

A count of adjusted p-values is a poor summary of how much evidence each cell carries. Table 23 decomposes the uncertainty-minus-restart contrast into questions where the three-seed mean success is higher (W), lower (L) or equal (T). Ties are not discarded: every paired question, tied or not, belongs in the estimand and in the interval. They do drop out of the exact null, where a zero difference carries no sign to flip, and that is what governs resolution. With m non-zero blocks an exact two-sided sign-flip test cannot report below $2/2^m$. Two cells sit above 0.05 on that floor, so they could not have reached significance whatever their outcomes; this is a property of the design, not of the data (Figure 11).

Table 23: Paired uncertainty-minus-restart outcomes per cell. N_q is the paired question count; W, L and T are questions with higher, lower and equal three-seed mean success. Floor is the smallest two-sided sign-flip p attainable given the non-zero block count. Δ and its percentile interval are the frozen primary outputs, in percentage points.

Model	Dataset	N_q	W	L	T	Floor	Δ	95% CI
Qwen	HotpotQA	43	1	2	40	0.250	−2.33	[−7.75, +1.55]
Qwen	MuSiQue	72	3	8	61	<0.001	−3.24	[−7.41, +0.93]
Qwen	2Wiki	28	4	0	24	0.125	+5.95	[+1.19, +11.90]
Mistral	HotpotQA	78	5	2	71	0.016	+0.85	[−2.56, +4.27]
Mistral	MuSiQue	83	2	5	76	0.016	−1.20	[−3.21, +0.80]
Mistral	2Wiki	67	9	8	50	<0.001	−1.00	[−6.97, +4.48]

Qwen 2Wiki is the clearest example. Its $\Delta = +5.95$ points with interval [+1.19, +11.90] rests on 4 questions with a higher treatment mean, 0 with a lower mean and 24 ties. Its floor is 0.125. The interval and the exact test are not in conflict; they answer different questions, and the sparse paired evidence explains why a positive-looking interval carries no primary-test support.

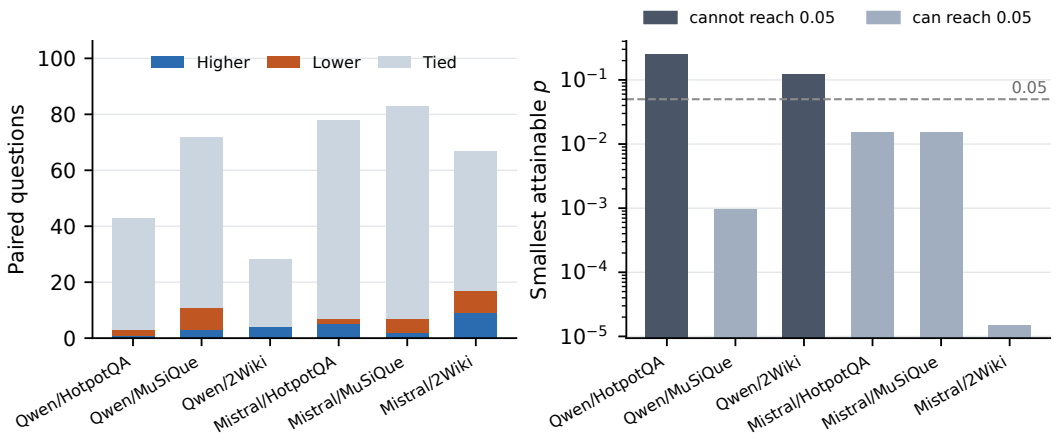

Figure 11: Left: paired uncertainty-minus-restart outcomes per cell, split into questions with higher, lower and equal three-seed mean success. Ties dominate every cell. Right: the smallest two-sided sign-flip p each cell could attain, $2/2^m$ for m non-zero blocks, on a logarithmic scale. Bars above the dashed 0.05 line mark cells that could not have reached significance whatever their outcomes.

Sharing explains part of the tie rate. Uncertainty and restart run the identical execution on 33.3–67.9% of paired attempts (Table 24); those attempts are tied by construction, not by outcome. Only $Q_{\neq}$ questions per cell have any differing execution, and that count bounds how many could express a difference.

Table 24: Execution overlap with restart. Attempts are question/seed pairs; Shared counts identical physical executions. $Q_{\neq}$ is the number of questions with at least one differing execution, which bounds how many questions could express any difference at all.

Model	Dataset	Attempts	Shared	Shared %	$Q_{\neq}$
Qwen	HotpotQA	129	81	62.8	16
Qwen	MuSiQue	216	72	33.3	48
Qwen	2Wiki	84	51	60.7	11
Mistral	HotpotQA	234	159	67.9	25
Mistral	MuSiQue	249	135	54.2	38
Mistral	2Wiki	201	120	59.7	27

Intervals are more informative than the test count as practical bounds. The completed main study rules out gains above $+0.88$ points, whereas Qwen 2Wiki permits a considerably larger effect. Treating all cells as equally informative loses that distinction. These bounds hold under the stated resampling assumptions and are individual, not simultaneous.

Finally, 621/1113 uncertainty rows (55.8%) stop at the generated-token or new-step limit rather than finishing. This is a descriptive bottleneck. It does not show that a larger allowance would raise success; the allowance study in Appendix G varied that budget directly and found no clear gain.

H.7 Length-Matched Swap: Design and Block Composition

Two separately frozen studies, one per model family, each evaluate a HotpotQA and a 2Wiki cell on the same 200 fresh questions, excluding every identifier used by the earlier studies and reusing those earlier questions only as development data for the secondary comparator. Across both studies, 400 model/question evaluations yield 218 failures, of which 204 pair by exact original length and 14 remain as odd stratum remainders. Each study fixes three repair seeds, three policies and its own 2-test primary family; execution produced 1,133 unique repairs and 1,836 trial rows. The studies share a cohort, which makes them a direct contrast and, for the same reason, forbids a pooled confirmatory statistic. All originals are generated before any repair, and pairing consumes only identifiers and lengths, so no recovery outcome can influence which questions are compared or which origins they receive.

Within a pair the swap control gives each question its partner's uncertainty-chosen origin. The odd remainder in each length stratum is excluded and reported.

Balance here is exact, not fitted. The audit re-derives every policy origin from independently recomputed uncertainty and requires the swapped pair to exchange origins exactly, the retained prefix lengths and per-question allowances to match, and execution sharing to track the selected origin. Both the on-instance and an independent local reproduction report zero mismatches. Because both arms keep the same seeds, hints and allowances, two policies that agree on an origin share one physical execution: 90/138 and 60/108 paired attempts for Qwen, and 108/198 and 84/168 for Mistral. Shared attempts do not imply equal outcomes elsewhere. The descriptive secondary comparison against development-fitted random, now fitted on a much larger development failure sample, is also inconclusive in every cell (unadjusted p from 0.23 to 0.75), although on Mistral HotpotQA its sign is opposite to the primary contrast (-2.02 points).

Under exact positional balance the comparison remains null (Table 25 and Figure 12): Holm-adjusted p-values of 0.250, 0.500, 1.000 and 1.000. Descriptively, the sign of each point estimate tracks the dataset, not the model: +3.62 and +1.01 percentage points on HotpotQA against -3.70 and -1.79 on 2Wiki. Mistral's point estimates are smaller: about a quarter of Qwen's on HotpotQA and half on 2Wiki.

The Qwen HotpotQA percentile interval excludes zero but should not be read as a positive effect: only 4 of 23 pairs are non-zero blocks, and an exact two-sided sign-flip test cannot report below $2/2^m$ for m non-zero blocks, a floor of 0.125 here. That cell could not have reached significance whatever the data showed, so the exact test does not corroborate the interval.

Table 25: Length-matched swap experiment, two model families on the same 200 fresh questions. N_f is the main-failure count and Pairs the number of length-matched blocks. U and S are mean per-seed success percentages for uncertainty and the swapped-origin control. $\Delta = U - S$ with a percentile interval over resampled pairs; p is an exact two-sided pair-block sign-flip test and p_H its Holm-adjusted value within the prespecified 2-test family.

Model	Dataset	N_f	Pairs	U	S	Δ	95% CI	p	p_H
Qwen	HotpotQA	47	23	13.77	10.14	+3.62	[+0.72, +7.25]	0.125	0.250
Qwen	2Wiki	40	18	15.74	19.44	−3.70	[−11.11, +3.70]	0.500	0.500
Mistral	HotpotQA	71	33	9.60	8.59	+1.01	[−2.02, +4.04]	0.750	1.000
Mistral	2Wiki	60	28	9.52	11.31	−1.79	[−6.55, +1.79]	0.688	1.000

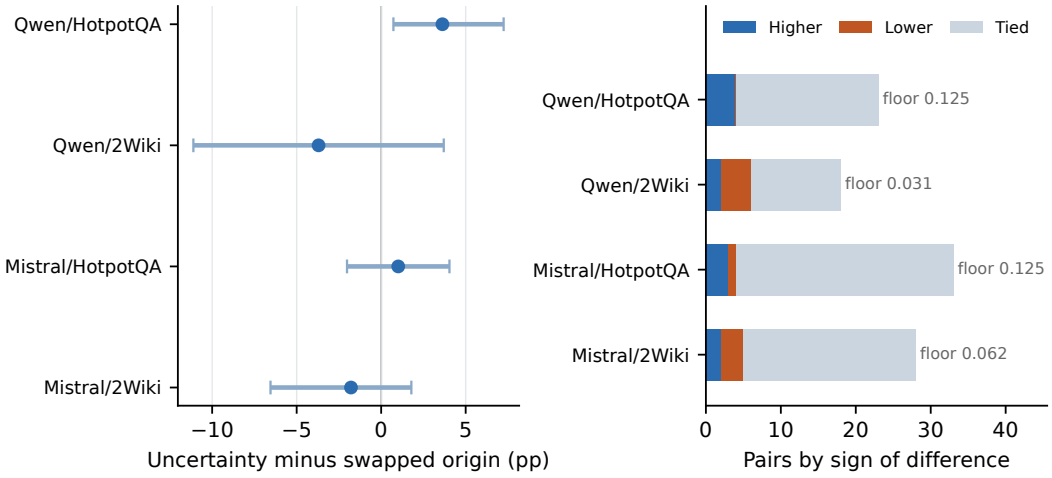

Figure 12: Left: uncertainty minus the swapped-origin control in percentage points, with the percentile interval over resampled pairs; the two datasets' point estimates fall on opposite sides of zero. Right: pairs by the sign of their seed-summed difference, annotated with the resolution floor. Most pairs tie, which is what bounds the exact test rather than the size of the effect.

Table 26: Sign-flip block composition. $+$, $-$ and 0 count pairs whose seed-summed success difference is positive, negative or tied. Floor is the smallest two-sided p attainable given the non-zero block count, $2/2^m$; a cell cannot report below its floor whatever the outcomes.

Model	Dataset	Pairs	$+$	$-$	0	Floor
Qwen	HotpotQA	23	4	0	19	0.125
Qwen	2Wiki	18	2	4	12	0.031
Mistral	HotpotQA	33	3	1	29	0.125
Mistral	2Wiki	28	2	3	23	0.062

Most pairs tie: 83 of 102 across the four cells (Table 26). That, not the effect size, is the binding precision limit, and 3 of 4 cells sit above 0.05 on their own floor. Against the prespecified Holm adjustment the bound is stricter: the one cell with a floor below 0.05, Qwen 2Wiki at 0.031, could report an adjusted value no smaller than 0.0625 within its two-test family, so no swap cell could have reached an adjusted 0.05. The floor is $2/2^m$ in the number of non-zero blocks m, so it is not fixed by the cohort size directly: m grows with the cohort only at the rate at which pairs fail to tie, and clearing 0.05 at all requires $m \geq 6$. At the tie rate observed here that rate is low enough that the cell with the most pairs shares the highest floor, because its extra pairs were ties; adding questions at this rate buys resolution far more slowly than the failure count suggests, and a design targeting a conventional level must be powered on the expected discordant rate rather than on failed trajectories. The exact test's resolution is governed by the non-zero block count, so all four cells are underpowered by construction: a reader should treat these as bounded diagnostics and not as evidence of equivalence. Retaining every tied pair in the point estimate and interval, while excluding ties from the exact null, is the prespecified behavior of the test, not a post-hoc choice. The swap borrows its partner's uncertainty, so it is a diagnostic comparator and not a policy anyone could run online; holding origins exactly equal leaves the tested rule no measurable advantage, but this establishes neither equivalence nor a power-backed absence of effect.

H.8 Confirmatory Length-Matched Swap

The swap above left one suggestive cell, Qwen HotpotQA, whose interval excluded zero while its exact test could not report below 0.125. A separately frozen study re-tested that cell on 1,000 fresh HotpotQA questions, drawn from the 6,136 identifiers no earlier study had used, with design, cohort size and analysis fixed before any outcome existed.

Two changes target the resolution limit rather than the effect. Within each exact original length, pairing now maximizes the number of pairs whose uncertainty-chosen origins differ, and only those informative pairs are repaired: a pair that shares an origin is swapped into itself, so both arms run identical executions and tie before any outcome exists. Such pairs are counted and reported but never run, which makes the primary estimand conditional on informative pairs; the frozen protocol declares it as such. The descriptive development-fitted control is dropped, leaving one primary test at two-sided $\alpha = 0.05$ with no multiplicity adjustment. The cohort was fixed in advance for power 0.80 at a discordant-pair win probability of 0.75, using the yield and discordance measured in the completed cells.

Of 452 initial failures (54.8% initial success), 224 pairs formed: 131 informative, 93 structural ties and 4 unmatched remainders.

The run stopped at its frozen time limit. Repairs ran slower than the completed cells predicted, because an informative pair always retains a non-zero prefix and prefix caching is disabled, so 100 of the 131 informative pairs finished and 31 did not. Pairs execute in a seeded order fixed before any outcome existed, so the completed pairs are a leading run of that order rather than a selection, and the analysis verifies that shape: any gap or later start is refused. Reporting them nonetheless departs from the prespecified completion rule, which marks a partial run incomplete, and we report it as a deviation rather than as the frozen confirmatory analysis. The comparison covers 1,200 unique repairs and 1,200 trial rows with zero replay, scoring or allowance mismatches and exactly balanced origins.

Uncertainty succeeds on 11.50% of seed attempts against 11.67% for the swapped origin: $\Delta = -0.17$ points, 95% interval [−3.17, +2.84], exact two-sided $p = 1.000$ from 22 positive and 18 negative blocks with 60 ties (Table 27 and Figure 13).

Unlike every earlier cell, this one could have resolved an effect. 40 pairs disagree, so the floor is $2/2^{40}$ rather than the 0.031–0.125 of the earlier cells, and the completed design retains power 0.90 at a win probability of 0.75 and 0.70 at 0.70. The earlier +3.62-point estimate lies outside this interval, and the discordant pairs split 22/18. Under exact positional balance, with resolution sufficient to detect a large effect, the tested policy shows no effect of that size. This does not establish equivalence: a positive effect below the interval's upper bound of +2.84 points remains possible.

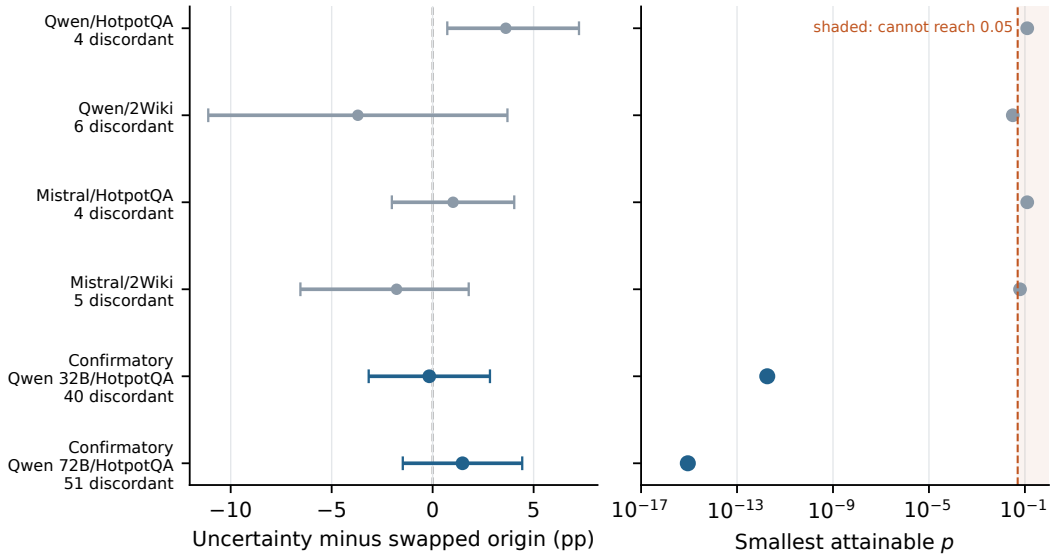

Figure 13: The two confirmatory cells, 32B and 72B, beside the four earlier swap cells. Left: uncertainty minus the swapped origin in percentage points, with the percentile interval over resampled pairs. Right: the smallest two-sided sign-flip p each cell could attain, $2/2^m$ for m discordant pairs, on a logarithmic scale; points in the shaded region cannot reach 0.05 whatever the data show. The earlier cells are bounded by their discordant counts; pairing for differing origins moves both confirmatory cells far out of that region, so their nulls are not limits of the design.

Table 27: Confirmatory length-matched swap, Qwen HotpotQA. Informative counts pairs whose origins differ, the only pairs repaired; Completed counts those that finished before the frozen time limit. U and S are mean per-seed success percentages for uncertainty and the swapped origin; $+/-/0$ splits the completed pairs by the sign of their seed-summed difference; p is the single exact two-sided sign-flip test.

N_f	Pairs	Informative	Completed	U	S	Δ	95% CI	$+/-/0$	p
452	224	131	100	11.50	11.67	−0.17	[−3.17, +2.84]	22/18/60	1.000

H.9 Confirmatory Length-Matched Swap at 72B

The same frozen design was repeated with Qwen2.5-72B-Instruct-AWQ as the agent, pinned by revision and served locally like every other model here, on 1,000 HotpotQA questions drawn from the 5,136 identifiers no earlier study had used, including the 32B confirmatory cohort. Pairing, repair scope, estimand, seeds and the single primary test are unchanged; only the model and a larger frozen time limit differ, the latter sized so that the slower model would not repeat the 32B cell's truncation.

Of 446 initial failures (55.4% initial success, close to the 32B cell's 54.8%), 223 pairs formed: 124 informative, 99 structural ties and 0 unmatched remainders. All 124 informative pairs finished within the frozen limit, so this cell satisfies its prespecified completion rule and needs no deviation. The comparison covers 1,488 unique repairs and 1,488 trial rows with zero replay, scoring or allowance mismatches and exactly balanced origins, and the on-instance and local analyses agree on every reported statistic.

Uncertainty succeeds on 12.90% of seed attempts against 11.42% for the swapped origin: $\Delta = +1.48$ points, 95% interval [−1.48, +4.44], exact two-sided $p = 0.397$ from 30 positive and 21 negative blocks with 73 ties (Table 28 and Figure 13). With 51 discordant pairs the floor is $2/2^{51}$, and the completed design has power 0.96 at a win probability of 0.75 and 0.84 at 0.70, so this null too is not an artifact of the design's resolution. Scaling the agent from 32B to 72B within one model family changes neither the initial success rate materially nor the conclusion: under exact positional balance the tested policy shows no resolvable advantage over its own origins given to a length-matched partner. As before, this does not establish equivalence: a positive effect below the interval's upper bound of +4.44 points remains possible.

Table 28: Confirmatory length-matched swap at 72B, Qwen HotpotQA. Columns as in Table 27; every informative pair completed.

N_f	Pairs	Informative	Completed	U	S	Δ	95% CI	$+/-/0$	p
446	223	124	124	12.90	11.42	+1.48	[−1.48, +4.44]	30/21/73	0.397

H.10 Measured Completion Deadlines

The runtime cohort is the first 25 frozen Qwen HotpotQA failures, or all failures if fewer. Restart, uncertainty with backtracking and diagnosis/replay each run separately with three seeds in shuffled order, after model warmup and with prefix caching disabled. Measured time includes origin acquisition, prefix replay and recovery. Diagnosis acquisition is measured once per question and charged in full to every diagnosis-policy attempt; its latency is not remeasured for every seed.

Recovery retains eight new steps and the per-request token cap but has no total generated-token cap. Success by 10 seconds is primary, with two contrasts in a separate Holm family; 5 and 20 seconds are exploratory. Full attempts execute before applying these deadlines to measured outcomes; Tables 29 and 30 give every threshold. This design measures completion-time success, not savings from active cancellation, equal FLOPs or end-to-end latency including initial failure detection and model startup.

Table 29: Completed runtime component. Counts and rates use all 75 attempts per policy. Mean time includes unsuccessful attempts and full execution beyond a threshold. Ten seconds is primary; 5 and 20 seconds are exploratory.

Policy	By 5 s (%)	By 10 s (%)	By 20 s (%)	Mean time (s)
Restart	8/75 (10.7)	13/75 (17.3)	13/75 (17.3)	5.13
Uncertainty, backtrack 2	9/75 (12.0)	12/75 (16.0)	12/75 (16.0)	5.12
Diagnosis/replay	2/75 (2.7)	9/75 (12.0)	9/75 (12.0)	5.60

H.11 Completed Replication Results

All six model/dataset cells have complete execution and independent audits. The 600 main evaluations contain 371 initial failures, with all seven policies and three seeds represented for every failure.

Table 30: Runtime treatment-minus-control differences in percentage points. Intervals are individual 95% question-bootstrap intervals over 25 questions. Only the two ten-second comparisons receive the prespecified Holm correction; unadjusted exploratory intervals are not confirmatory evidence.

Deadline (s)	Control	Δ (pp)	95% interval (pp)	p	p_H
5	Restart	+1.33	[−5.33, +9.33]	1.000	–
5	Diagnosis/replay	+9.33	[+1.33, +18.67]	0.095	–
10	Restart	−1.33	[−5.33, +2.67]	1.000	1.000
10	Diagnosis/replay	+4.00	[−4.00, +13.33]	0.567	1.000
20	Restart	−1.33	[−5.33, +2.67]	1.000	–
20	Diagnosis/replay	+4.00	[−4.00, +13.33]	0.567	–

Mistral contributes 2,473 unique repairs and 4,788 trial rows; together with Qwen and the runtime component, the extension has 4,323 unique repairs and 8,016 trial rows. Every cell has zero audit mismatches, and Figure 14 shows every cell's uncertainty-minus-restart interval. Initial reference-gate accuracy, strict EM and F1 are shown in Table 31. Every policy's outcomes appear in Tables 32 and 33, its contrasts in Tables 34 and 35, and its costs in Tables 36 and 37.

Table 31: Initial main-cohort outcomes before repair. Each cell also has 20 separate development questions. Success uses the reference gate; EM and F1 use the corresponding dataset's scorer.

Model	Dataset	N	Failures	Success (%)	EM (%)	F1
Qwen	HotpotQA	100	43	57.0	44.0	0.5439
Qwen	MuSiQue	100	72	28.0	22.0	0.2852
Qwen	2Wiki	100	28	72.0	64.0	0.7120
Mistral	HotpotQA	100	78	22.0	13.0	0.2343
Mistral	MuSiQue	100	83	17.0	13.0	0.1867
Mistral	2Wiki	100	67	33.0	21.0	0.3478

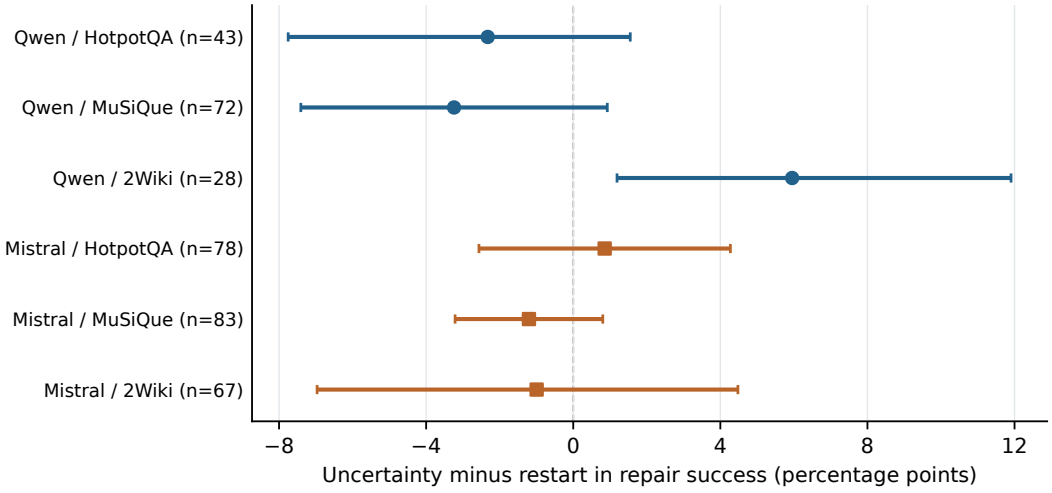

Figure 14: Uncertainty-minus-restart repair-success differences across all six completed cells. Points average three seeds per failed question; bars are individual 95% question-bootstrap intervals. They are not simultaneous intervals or inversions of the sign-flip test. Every contrast has $p_H = 1.000$ in the prespecified 24-comparison family.

Table 32: All Qwen replication policies. Success, strict EM and F1 retain the same initially failed questions; each policy has three seeds per question.

Dataset	Policy	Successes	Success (%)	EM (%)	F1
HotpotQA	Restart	18/129	13.95	8.53	0.1515
HotpotQA	Random, backtrack 2	22/129	17.05	11.63	0.1789
HotpotQA	Fixed early	11/129	8.53	5.43	0.1148
HotpotQA	Uncertainty, backtrack 2	15/129	11.63	6.20	0.1273
HotpotQA	Dev.-fitted random	18/129	13.95	7.75	0.1485
HotpotQA	Uncertainty, no backtracking	11/129	8.53	5.43	0.1160
HotpotQA	Diagnosis/replay	14/129	10.85	7.75	0.1358
MuSiQue	Restart	17/216	7.87	6.02	0.0992
MuSiQue	Random, backtrack 2	11/216	5.09	3.70	0.0775
MuSiQue	Fixed early	14/216	6.48	6.48	0.0873
MuSiQue	Uncertainty, backtrack 2	10/216	4.63	2.31	0.0597
MuSiQue	Dev.-fitted random	14/216	6.48	5.56	0.0846
MuSiQue	Uncertainty, no backtracking	14/216	6.48	3.70	0.0888
MuSiQue	Diagnosis/replay	16/216	7.41	5.56	0.0940
2Wiki	Restart	13/84	15.48	10.71	0.1732
2Wiki	Random, backtrack 2	15/84	17.86	11.90	0.1912
2Wiki	Fixed early	17/84	20.24	19.05	0.2334
2Wiki	Uncertainty, backtrack 2	18/84	21.43	15.48	0.2267
2Wiki	Dev.-fitted random	13/84	15.48	9.52	0.1675
2Wiki	Uncertainty, no backtracking	19/84	22.62	17.86	0.2337
2Wiki	Diagnosis/replay	14/84	16.67	8.33	0.1814

Table 33: All Mistral replication policies, evaluated on its own frozen initial-failure cohorts. Conditional rates do not compare overall model reliability.

Dataset	Policy	Successes	Success (%)	EM (%)	F1
HotpotQA	Restart	27/234	11.54	7.69	0.1288
HotpotQA	Random, backtrack 2	24/234	10.26	6.84	0.1230
HotpotQA	Fixed early	31/234	13.25	9.40	0.1775
HotpotQA	Uncertainty, backtrack 2	29/234	12.39	9.83	0.1497
HotpotQA	Dev.-fitted random	31/234	13.25	8.12	0.1470
HotpotQA	Uncertainty, no backtracking	25/234	10.68	6.84	0.1449
HotpotQA	Diagnosis/replay	22/234	9.40	7.69	0.1238
MuSiQue	Restart	18/249	7.23	5.62	0.0951
MuSiQue	Random, backtrack 2	20/249	8.03	5.62	0.0930
MuSiQue	Fixed early	24/249	9.64	6.43	0.1080
MuSiQue	Uncertainty, backtrack 2	15/249	6.02	4.42	0.0769
MuSiQue	Dev.-fitted random	16/249	6.43	4.02	0.0851
MuSiQue	Uncertainty, no backtracking	15/249	6.02	4.02	0.0803
MuSiQue	Diagnosis/replay	19/249	7.63	4.02	0.0911
2Wiki	Restart	34/201	16.92	9.45	0.1854
2Wiki	Random, backtrack 2	33/201	16.42	10.45	0.1976
2Wiki	Fixed early	36/201	17.91	10.45	0.2430
2Wiki	Uncertainty, backtrack 2	32/201	15.92	8.96	0.1832
2Wiki	Dev.-fitted random	36/201	17.91	12.94	0.2160
2Wiki	Uncertainty, no backtracking	28/201	13.93	9.45	0.1884
2Wiki	Diagnosis/replay	29/201	14.43	10.95	0.2201

Table 34: Qwen primary uncertainty-minus-control contrasts, in percentage points. Individual 95% intervals and unadjusted p accompany the Holm-adjusted p_H from the full 24-comparison family.

Dataset	Control	Δ (pp)	95% interval (pp)	p	p_H
HotpotQA	Restart	−2.33	[−7.75, +1.55]	0.744	1.000
HotpotQA	Dev.-fitted random	−2.33	[−7.75, +3.10]	0.587	1.000
HotpotQA	Uncertainty, no backtracking	+3.10	[−3.10, +8.53]	0.439	1.000
HotpotQA	Diagnosis/replay	+0.78	[−6.20, +7.75]	1.000	1.000
MuSiQue	Restart	−3.24	[−7.41, +0.93]	0.218	1.000
MuSiQue	Dev.-fitted random	−1.85	[−5.56, +2.31]	0.521	1.000
MuSiQue	Uncertainty, no backtracking	−1.85	[−6.02, +2.31]	0.546	1.000
MuSiQue	Diagnosis/replay	−2.78	[−7.41, +1.85]	0.366	1.000
2Wiki	Restart	+5.95	[+1.19, +11.90]	0.131	1.000
2Wiki	Dev.-fitted random	+5.95	[+0.00, +13.10]	0.196	1.000
2Wiki	Uncertainty, no backtracking	−1.19	[−8.33, +5.95]	1.000	1.000
2Wiki	Diagnosis/replay	+4.76	[−3.57, +14.29]	0.471	1.000

Table 35: Mistral primary contrasts, using the same full-family correction and interval construction as Table 34.

Dataset	Control	Δ (pp)	95% interval (pp)	p	p_H
HotpotQA	Restart	+0.85	[−2.56, +4.27]	0.810	1.000
HotpotQA	Dev.-fitted random	−0.85	[−5.13, +2.99]	0.834	1.000
HotpotQA	Uncertainty, no backtracking	+1.71	[−3.42, +6.84]	0.631	1.000
HotpotQA	Diagnosis/replay	+2.99	[−1.28, +7.26]	0.255	1.000
MuSiQue	Restart	−1.20	[−3.21, +0.80]	0.461	1.000
MuSiQue	Dev.-fitted random	−0.40	[−2.81, +2.01]	1.000	1.000
MuSiQue	Uncertainty, no backtracking	+0.00	[−1.61, +1.61]	1.000	1.000
MuSiQue	Diagnosis/replay	−1.61	[−4.42, +0.80]	0.356	1.000
2Wiki	Restart	−1.00	[−6.97, +4.48]	0.869	1.000
2Wiki	Dev.-fitted random	−1.99	[−5.97, +1.99]	0.488	1.000
2Wiki	Uncertainty, no backtracking	+1.99	[−3.48, +6.97]	0.597	1.000
2Wiki	Diagnosis/replay	+1.49	[−3.48, +6.47]	0.693	1.000

Table 36: Qwen mean incremental tokens and model requests per policy attempt, including measured origin acquisition and diagnosis. Identical executions may be shared by the study, but each standalone policy is charged its full cost.

Dataset	Policy	Generated	Prompt	Requests
HotpotQA	Restart	243.1	2574.8	4.89
HotpotQA	Random, backtrack 2	238.6	3047.0	4.73
HotpotQA	Fixed early	228.8	2956.5	4.36
HotpotQA	Uncertainty, backtrack 2	236.9	3333.6	4.70
HotpotQA	Dev.-fitted random	238.9	2751.2	4.79
HotpotQA	Uncertainty, no backtracking	223.3	4304.7	4.28
HotpotQA	Diagnosis/replay	253.3	4250.0	5.13
MuSiQue	Restart	314.4	4618.8	6.26
MuSiQue	Random, backtrack 2	319.2	6987.5	6.33
MuSiQue	Fixed early	316.7	5759.0	6.28
MuSiQue	Uncertainty, backtrack 2	320.9	6564.9	6.35
MuSiQue	Dev.-fitted random	314.2	7187.8	6.23
MuSiQue	Uncertainty, no backtracking	296.0	8111.3	5.80
MuSiQue	Diagnosis/replay	344.7	8434.7	6.99
2Wiki	Restart	268.5	2952.1	5.20
2Wiki	Random, backtrack 2	260.2	3593.6	4.93
2Wiki	Fixed early	242.5	3266.4	4.54
2Wiki	Uncertainty, backtrack 2	248.7	3553.2	4.75
2Wiki	Dev.-fitted random	262.8	3278.7	5.05
2Wiki	Uncertainty, no backtracking	251.7	4790.5	4.57
2Wiki	Diagnosis/replay	271.2	5149.7	5.06

Table 37: Mistral mean incremental policy costs under the same accounting as Table 36. Matching output allowances does not match prompt processing, runtime or total computation.

Dataset	Policy	Generated	Prompt	Requests
HotpotQA	Restart	170.7	2043.3	5.02
HotpotQA	Random, backtrack 2	164.0	2218.1	4.66
HotpotQA	Fixed early	150.8	2207.8	4.35
HotpotQA	Uncertainty, backtrack 2	164.9	2311.0	4.86
HotpotQA	Dev.-fitted random	164.0	2143.9	4.78
HotpotQA	Uncertainty, no backtracking	147.6	2534.0	4.06
HotpotQA	Diagnosis/replay	204.9	3466.0	5.13
MuSiQue	Restart	185.9	2834.6	5.36
MuSiQue	Random, backtrack 2	183.0	3802.8	5.16
MuSiQue	Fixed early	173.8	3246.0	5.02
MuSiQue	Uncertainty, backtrack 2	182.3	3505.3	5.18
MuSiQue	Dev.-fitted random	182.4	3379.5	5.20
MuSiQue	Uncertainty, no backtracking	180.5	4463.9	4.94
MuSiQue	Diagnosis/replay	226.8	5230.2	5.69
2Wiki	Restart	181.8	2172.4	5.23
2Wiki	Random, backtrack 2	170.5	2302.1	4.84
2Wiki	Fixed early	159.4	2283.5	4.52
2Wiki	Uncertainty, backtrack 2	169.2	2349.6	4.80
2Wiki	Dev.-fitted random	172.9	2400.1	4.91
2Wiki	Uncertainty, no backtracking	158.0	2507.8	4.28
2Wiki	Diagnosis/replay	202.9	3609.8	4.89

H.12 Recovery and Audit Provenance

All Qwen executions completed before an audit failure prevented Mistral from starting. A bounded retrieval recovered the original archive. JSON conversion of integer token-ID keys to strings changed sorting during checksum verification; restoring their original types reproduced the saved hashes without changing probability values or raw records. Cross-platform recomputation also produced tiny uncertainty differences: 3,542 values differed, with maximum absolute error 1.78×10^{-15}. The amended auditor permits at most 10^{-12} absolute roundoff and requires exactly identical independently recomputed position profiles, policy origins, prompts, seeds and allowances. The statistical protocol and frozen input bytes are unchanged.

The three Qwen audits cover 300 main and 60 development evaluations, 1,850 unique repairs and 3,228 trial rows with zero mismatches. Their local and remote trial exports match across all 29 columns; all 4,739 original Qwen result files remain byte-identical. Platform-specific roundoff counts are validated and preserved separately from compared outcomes. Failing-first regressions cover token-key restoration, altered probabilities, excessive roundoff and even tolerated roundoff that changes a repair origin. Mistral continued on September 15 under the unchanged frozen design and amended audit source. Its partial archive was retrieved and checksum-verified without new inference, and the continuation completed the same day with every previously unfinished Mistral repair included; it preserved the frozen scientific protocol and all 8,536 previously saved result files. The original and continued archives and the audit amendments remain separate provenance artifacts, and no failed execution or partial archive is discarded.

I Related-Work Comparison and Required Controls

This focused literature update uses primary sources available on September 9, 2026; it is not an exhaustive survey. Table 38 compares designs, not repair rates across incompatible cohorts.

Table 38: Closest design overlap and the corresponding evaluation requirement. The last column distinguishes evaluation designs; completion of individual controls is specified in the text. It does not attribute shortcomings to the cited work.

Work	Relevant design	Required distinction here
Doctor-RAG	Trained diagnosis and error-specific QA repair	Compare a diagnosis-based policy; separate origin choice, hints, evidence access, and diagnosis cost.
ReAgent	Local/global rollback in multi-agent QA	Do not claim rollback itself as new; our experiment changes a single agent's recovery origin.
DoVer	Intervention-driven multi-agent debugging	Recovery, not label agreement alone, is already an established evaluation target.
SymTrace	Recorded-prefix reconstruction and live suffix generation	Audit replay fidelity independently of repair success and suffix randomness.
AgentRewind	Context/workspace restoration with rewind memory	Do not generalize read-only tool replay to external side-effect recovery.
AgentRx; AgenticRAG-FP	Constraint-supported diagnosis; injected-fault attribution	Keep diagnosis evidence and causal identification separate from recovery utility.

Nearest-method comparator. Doctor-RAG's reasoning-repair operator can reuse documents retrieved throughout the failed trace, and its study uses an EM-only failure gate (Jiao et al., 2026). A direct comparison must reconcile these differences, not copy published rates into our table. Our same-model diagnosis-and-replay implementation is an origin-only adaptation with the same generic hint and prefix-only recovery information as uncertainty. The six-condition main study omits this arm; the separate follow-up and all six replication cells evaluate it with its full acquisition charge. They do not evaluate diagnosis-specific retry guidance or reproduce the trained Doctor-RAG system. A future full-method comparison would need to declare any additional evidence and guidance, include diagnosis, tool and selection costs, and report training or distillation costs separately if incurred.

Position-distribution control. The main study and six replication cells include offset-matched random and a development-fitted position control that does not read test uncertainty. In the main study, origin-zero frequencies are 60.77% for the latter and 61.95% for treatment; their paired success difference is zero with a 95% interval of $[-2.95, +2.95]$ points. These observations establish neither exact distributional matching nor equivalence. Replication matching is weaker: Appendix H.3 gives full distributions, including the 26.19-point origin-zero gap on Qwen 2Wiki.

Scope of further evidence. The completed manifests distinguish the main six-policy study, three-policy diagnosis follow-up, seven-policy replication, three-policy runtime experiment, three-policy length-matched swap and two-policy confirmatory swap cells at 32B and 72B. New selectors, a trained-method reproduction, human error-step validation and end-to-end resource comparisons require additional evidence. Any confirmatory extension must freeze its hypotheses and analysis before new outcomes. Selecting a policy on these completed cohorts would require fresh evaluation data for an untouched test.

J Artifact Availability and Verification Scope

Table 39 distinguishes records checked locally from a package ready for independent release. Statistical reproduction recomputes results from saved trials; the raw-trajectory audit additionally checks scoring, retained tool state, execution coverage and allowances. Neither check is a fresh model-generation rerun.

Table 39: Verification and packaging status of the evidence used in this submission. Retained records are not a claim that all artifacts have been anonymously released. Replication counts here include the runtime component.

Evidence	Completed verification	Packaging and scope
Main HotpotQA	Five batch audits; 1,058 repairs and 2,034 trial rows; local statistical reproduction.	In the combined anonymous supplement; statistics reproduced in isolation. Full generation environment is separate.
Diagnosis follow-up	Five main-batch audits; 2,079 repairs and 3,186 rows; matching local/remote analyses.	Six archives, including development, retained. In the combined supplement; statistics reproduced in isolation.
Replication and runtime	Six cells; 4,323 repairs and 8,016 rows; matching local/remote analyses and preserved prior records.	Two model archives and checked exports retained. In the combined supplement; the 24 replication tests are recomputed, the two runtime tests are not.
Length-matched swap	Four cells in two per-model studies, audited on-instance and locally; 1,133 repairs and 1,836 rows; zero mismatches.	Both archives retained. In the combined supplement; both 2-test families reproduced in isolation.
Confirmatory swap	One cell, 100 of 131 informative pairs completed before the frozen time limit; 1,200 repairs and rows; zero mismatches.	Records and the derived analysis package retained. In the combined supplement; its single test reproduced in isolation.
Confirmatory swap, 72B	One cell, all 124 informative pairs completed; 1,488 repairs and rows; zero mismatches; matching local/remote analyses.	Records and package retained. In the combined supplement; its single test reproduced in isolation.
Historical summaries	Source-hashed aggregates and identified notebook confounds.	Raw trials incomplete; original pool uncompared; no human validation.

The combined anonymous supplement covering every completed study passes isolated-extraction checks: every packaged file is checksummed, every primary family except the two runtime tests is recomputed from its trial export, and a scan rejects workspace paths, account identifiers and private keys before packaging. Raw-record replay with the frozen reference scorers, and regeneration under the pinned model weights, remain separate environments that the bundle does not claim. A release

package must additionally retain third-party notices and remove identifying notebook and operational metadata. Statistical reproduction and fresh GPU runs need separate environment checks. Packaging cannot supply missing human labels, historical-pool comparisons or further experiments.