SEARCH-AGENT SCALING GAINS SURVIVE ANSWER-WITNESS DELETION

Anonymous authors
Paper under double-blind review

ABSTRACT

Search agents are evaluated by answer accuracy and by whether retrieved evidence supports their answers, but neither reveals how an answer responds when that evidence is removed. We introduce counterfactual evidence accounting: at every retrieval call, passages matched as containing an accepted-answer alias are deleted and replaced by lower-ranked candidates, and the complete search policy is rerun. Paired rollouts partition ordinary accuracy exactly into correct answers that become incorrect after this deletion (D-failing) and those that remain correct (D-surviving). On 623 matched questions, the released 14B Search-R1 checkpoint is 10.3 percentage points more accurate than the 3B checkpoint, and 9.5 of those points are D-surviving; the analysis detects no increase in D-failing correctness. Five released Qwen2.5 search-agent recipes, prompted Qwen checkpoints without search-agent reinforcement learning, and an SSRL Llama comparison show the same concentration, which persists under residual-witness, parsing, and exposure sensitivity analyses. With matched weights, 7.9 of the 9.5 points overlap with closed-book correctness. Deleting passages that match the agent's own answer yields a label-free diagnostic, self-ablation: with one additional rollout, it lowers D-surviving correctness among released answers from 8.1% to 2.4% at 81% coverage, at a cost of 8.3 points of total correct yield. Search-agent evaluations should report how accuracy responds to targeted changes in retrieved evidence, alongside accuracy and support.

1 INTRODUCTION

A search agent can answer correctly after retrieving a passage that states the answer, and still give the same answer when that passage is removed. Accuracy records that the answer is correct, and evidence-support evaluations that the retrieved context suffices for it; neither reveals how the answer responds when the retrieved evidence changes. The distinction matters when retrieval is meant to update, correct, or justify an answer. It also affects how we interpret accuracy gains from scaling. A larger agent may become more accurate because it uses retrieved evidence more effectively, or because its language model answers more questions in ways that do not depend on the retrieved answer witnesses; both routes may operate at once.

We audit the composition of the accuracy gain with answer-witness deletion (Figure 2). Relative to the ordinary world O, world D removes every retrieved candidate matched as containing an accepted-answer alias, backfills the display from lower-ranked candidates, and reruns the complete search policy, reapplying the edit at every retrieval call. An ordinary-correct answer is *D-failing* if it is incorrect after deletion and *D-surviving* if it remains correct; the two exactly partition ordinary accuracy. Neither label is a verdict on quality: a D-surviving answer may be correctly retained despite missing evidence, and a D-failing answer may reflect fragility rather than good use of evidence. The partition measures how accuracy responds to a specified retrieval edit.

Applied to model scale, the partition reveals a pattern that accuracy conceals (Figure 1). On 623 matched questions, released Search-R1 checkpoints gain $10.27\,\mathrm{pp}$ of accuracy from 3B to 14B, and $9.47\,\mathrm{pp}$ $[+6.74, +12.20]$ of that gain is D-surviving. The D-failing change is $+0.80\,\mathrm{pp}$ $[-3.05, +4.49]$: the analysis does not detect an increase in this component, although it remains substantial at every rung. Across five released Qwen2.5 search-agent recipes, every 3B-to-7B D-surviving interval excludes zero, while every D-failing point estimate is negative. Prompted Qwen

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

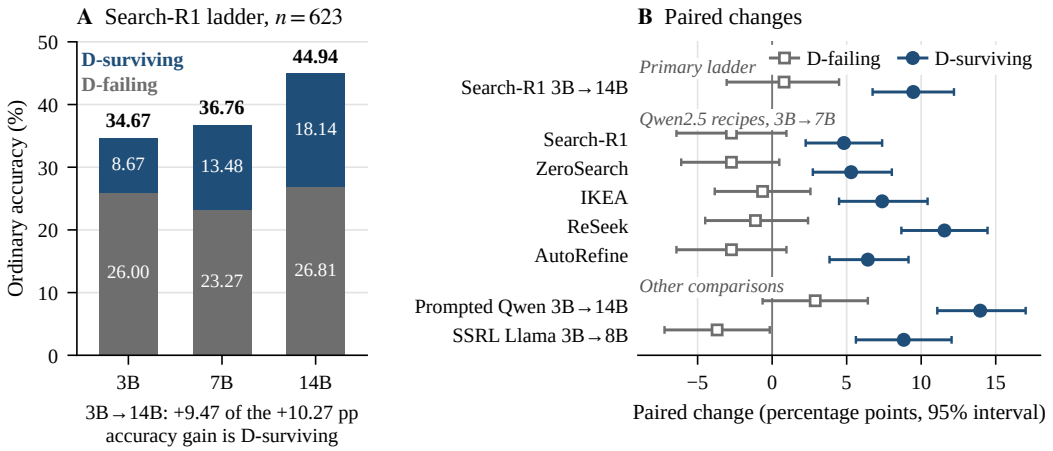


Figure 1: **Most of the accuracy gain lies in D-surviving correctness.** (A) Ordinary accuracy on the matched Search-R1 ladder ($n = 623$), partitioned into D-failing (correct in O, incorrect after deletion) and D-surviving (correct in both worlds) correctness. (B) Paired larger-minus-smaller changes with 95% bootstrap intervals (ITT, same questions). Rows share questions, base-model families, and in one case a checkpoint, so they are related rather than independent comparisons; all are observational, and the Llama pair also changes model version.

checkpoints without search-agent reinforcement learning and a self-search reinforcement learning (SSRL) Llama comparison (Fan et al., 2025) show the same concentration. These released-checkpoint comparisons are observational, so we treat them as associations with model scale.

With matched weights, 7.87 pp of the 9.47 pp D-surviving increase overlaps with closed-book correctness, and when retrieved evidence is rewritten to a same-type substitute, larger prompted Qwen checkpoints repeat their original answer more often without a comparable rise in substitute adoption. Replacing the reference answer with the agent's own answer turns the intervention into a label-free diagnostic, self-ablation, whose selective-release costs we report alongside its benefits. The repairs we tested reduce D-surviving output mainly through abstention or suppression rather than conversion into D-failing correctness.

Our contributions are:

1. We apply answer-witness deletion at every retrieval call of a complete policy replay and decompose accuracy using paired outcomes. We characterize what this procedure identifies and when label-free self-ablation recovers the same target (Section 3).
2. We find that accuracy gains concentrate in D-surviving correctness across the primary ladder and related comparisons. We test sensitivity to residual witnesses, exposure, parsing, execution, and deletion volume, and examine closed-book overlap and responses to substituted evidence (Sections 5 and 6).
3. We show that label-free self-ablation tracks the gold-indexed D label and differs from support judgments. We measure risk, coverage, compute cost, and total correct yield under selective release. We also test interventions intended to convert D-surviving into D-failing correctness (Sections 7 and 8).

2 Related Work

Search agents. Retrieval-augmented generation (RAG) combines parametric generation with retrieved memory (Lewis et al., 2020). WebGPT and ReAct interleave search with reasoning (Nakano et al., 2021; Yao et al., 2023), and Search-R1 and related systems train models to search iteratively (Jin et al., 2025; Sun et al., 2025a; Huang et al., 2025; Li et al., 2025; Shi et al., 2025); they are compared primarily by answer accuracy.

Support versus response to intervention. Grounding and attribution evaluations ask whether retrieved evidence supports an answer. Wallat et al. (2025) distinguish correct citation from causal influence, Worledge et al. (2023) distinguish corroborative from contributive attribution, and evidence-sufficiency judges assess whether a context is adequate for an answer (Joren et al., 2025). Rationale-faithfulness work tests selected text with sufficiency and comprehensiveness interventions (DeYoung et al., 2020). We measure how the answer changes when the retrieval environment is edited.

Evidence interventions in interactive search. ContextCite attributes a fixed generation to spans of a fixed context (Cohen-Wang et al., 2024); document deletion and partial replay measure causal utility at other granularities (Mukhopadhyay et al., 2026); and RAGONITE and CUE-R intervene on evidence for single-shot RAG responses (Saha Roy et al., 2025; Jain & Vedam, 2026). Entity substitution tests responses to conflicting context (Longpre et al., 2021), while context-faithfulness work tracks or trains reliance on supplied context (Goyal et al., 2025; Bi et al., 2025). These studies edit or attribute a given context, or train a reader. We apply the edit at every retrieval call of an acting search policy and rerun it, so the agent can issue new queries and retrieve new passages.

Confidence and selective prediction. Selective classification trades coverage against conditional risk (Geifman & El-Yaniv, 2017). Sequence likelihood, semantic uncertainty, repeated sampling, and self-evaluation predict answer correctness (Kadavath et al., 2022; Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024). We use the risk–coverage vocabulary with a different loss, D-surviving correctness among released answers, and test these signals against it.

Our focus is the paired answer outcome after editing retrieval throughout a new search trajectory, and how that outcome partitions search-agent accuracy across checkpoints.

3 Counterfactual Evidence Accounting

The procedure is simple: run the ordinary agent; edit the retrieved candidates according to the selected world; rerun the agent from the beginning, allowing new queries, and apply the edit again at every retrieval call. The ordinary and edited answers to a question form a paired outcome.

3.1 Setting and counterfactual worlds

A unit $x = (q, g, E)$ consists of a question q, its evaluator-accepted answer class $[g]$, and a retrieval environment E; strings satisfy $a \sim b$ when the evaluator accepts them as the same answer. A possibly stochastic policy π maps a question and environment to a search trajectory and final answer. A matcher returns the candidates $\mathcal{W}_E([a])$ matched as containing an accepted alias of $[a]$, and $T_{[a]}(E)$ removes these candidates whenever they are retrieved and backfills the display from lower-ranked candidates. We evaluate four worlds (Figure 2):

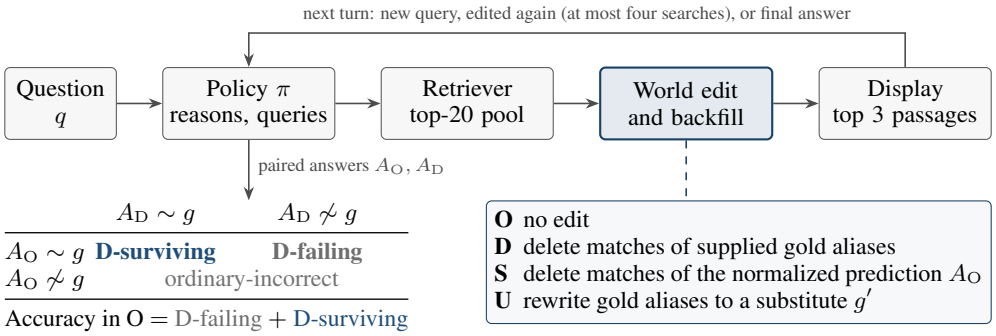


Figure 2: **Counterfactual evidence accounting.** Each world's edit is applied to the ranked candidates at every retrieval call of a complete rerun, so passages retrieved by new queries are edited in turn. D deletes candidates matching supplied gold aliases, and S candidates matching the agent's normalized ordinary prediction; both backfill the display. U rewrites gold aliases. Paired O and D answers partition ordinary accuracy.

O, ordinary: the unedited environment E.

D, gold deletion: $T_{[g]}(E)$ removes candidates matched as containing a gold alias.

S, self-ablation: after observing A_{O}, $T_{[A_{\mathrm{O}}]}(E)$ removes candidates matched as containing the agent's own answer, implemented by matching the normalized ordinary prediction; S needs no reference answer at deployment time.

U, substitution: recognized gold aliases in displayed passages are rewritten to a same-type substitute, testing response to conflicting evidence.

Exact alias matching can miss paraphrases and indirect clues (Section 5 reports how much answer evidence remains); Appendix B gives the implementation.

3.2 Exact accuracy decomposition

With paired answers A_{O} and A_{D}, $Y_{\mathrm{O}} = \mathbf{1}\{A_{\mathrm{O}} \sim g\}$, and $Y_{\mathrm{D}} = \mathbf{1}\{A_{\mathrm{D}} \sim g\}$, ordinary correctness splits pointwise, and so does any accuracy contrast between two checkpoints:

$$Y_{\mathrm{O}} = \underbrace{Y_{\mathrm{O}}(1 - Y_{\mathrm{D}})}_{=:\, L^{\mathrm{fail}}} + \underbrace{Y_{\mathrm{O}} Y_{\mathrm{D}}}_{=:\, L^{\mathrm{surv}}}, \tag{1}$$

$$\Delta\mathrm{Acc} = \Delta\,\mathbb{E}\big[L^{\mathrm{fail}}\big] + \Delta\,\mathbb{E}\big[L^{\mathrm{surv}}\big]. \tag{2}$$

We call L^{fail} and L^{surv} *D-failing* and *D-surviving* correctness. The identity needs no causal assumption, but each label is a paired rollout outcome that depends on the execution protocol coupling randomness across O and D (Appendix A). Reading the terms as evidence-dependent and evidence-independent correctness would also require complete removal of answer evidence, so we use the operational terms throughout.

3.3 Identification scope and self-ablation

Could a detector that inspects the ordinary answer and its evidence recover the D outcome without rerunning the agent? Consider detectors that observe q, A_{O}, the fixed evidence payload C, and any number of model queries whose evidence remains C. If a policy class contains two policies that behave identically on C but extend differently to the D environment, fixed-evidence observations do not uniformly identify the intervention outcome over that class (Proposition 2), a specialization of observational equivalence (Pearl, 2009; Bareinboim et al., 2022). Such signals may still predict D outcomes on a particular benchmark; we test this in Section 7.

Self-ablation changes the environment and so lies outside this class. With $Z_{\mathrm{S}} = \mathbf{1}\{A_{\mathrm{S}} \sim A_{\mathrm{O}}\}$, on every ordinary-correct unit

$$Y_{\mathrm{O}} = 1 \;\Rightarrow\; [A_{\mathrm{O}}] = [g] \;\Rightarrow\; T_{[A_{\mathrm{O}}]}(E) = T_{[g]}(E) \;\Rightarrow\; Z_{\mathrm{S}} = Y_{\mathrm{D}}, \tag{3}$$

exactly when S and D execute the same code path with shared exogenous randomness, and in distribution under independent randomness (Proposition 3). Our implementation supplies gold aliases to D but indexes S by the prediction string, and greedy replays are not perfectly repeatable, so we evaluate S/D correspondence empirically. "Label-free" refers to running S at deployment; evaluating its risk still requires reference answers.

4 Experimental Setup

Questions, retrieval, and matching. The scaling panel contains 623 questions from seven open-domain question-answering (QA) datasets, after input-level validity exclusions and restriction to questions observed in every compared arm. Agents search a pinned Wikipedia-2018 corpus with an E5-base-v2 dense retriever (Wang et al., 2022). Each agent can issue up to four searches, and each call displays three passages from a top-20 candidate pool. Exact accepted aliases define answer matching and the D transformation; S matches the normalized ordinary prediction. Appendix C specifies datasets, prompts, retrieval, and parsing.

Checkpoints and comparisons. The primary ladder compares released Search-R1 Qwen2.5 checkpoints at 3B, 7B, and 14B with matched questions, worlds, prompts, and generation budgets (Jin et al., 2025; Qwen Team, 2024); the 7B and 14B runs used tensor parallelism across two GPUs, whereas the 3B run did not. Five released Qwen2.5 recipes (Search-R1, ZeroSearch, IKEA, ReSeek, and AutoRefine) are compared from 3B to 7B (Sun et al., 2025a; Huang et al., 2025; Li et al., 2025; Shi et al., 2025). Two further comparisons use public prompted Qwen2.5-Instruct 3B and 14B checkpoints without search-agent reinforcement learning, and an SSRL pair from Llama-3.2-3B to Llama-3.1-8B that changes model version as well as parameter count (Llama Team, AI at Meta, 2024; Meta AI, 2024).

Cohorts. Scaling contrasts use the 623-question panel; detector analyses use ordinary-correct Search-R1 questions, for which D-survival is defined; and selective release uses a 2,591-question Search-R1 panel that includes incorrect answers. Appendix C.6 lists every cohort and its rule.

Uncertainty and parsing. O and D were generated in separate temperature-zero jobs initialized with the same nominal seed and then matched by question; random-number streams were not token-aligned. Labels therefore refer to this matched-execution protocol, not a shared-randomness counterfactual. Checkpoint contrasts use 10,000 percentile-bootstrap replicates for a fixed execution and exclude training- and repeated-decoding variation. ITT counts unparsed outputs as incorrect; common-parse and D-nonparse imputations assess sensitivity (Appendix D). An independent 14B execution checks that endpoint only, not the ladder's configuration difference.

5 Scaling Results

5.1 Primary Search-R1 ladder

Applying Eq. (2), from 3B to 14B Search-R1 accuracy rises by $10.27\,\mathrm{pp}$ $[+6.90, +13.64]$. D-surviving correctness rises by $9.47\,\mathrm{pp}$ $[+6.74, +12.20]$, 92% of this net increase, while the D-failing change is $+0.80\,\mathrm{pp}$ $[-3.05, +4.49]$ (Figure 1A). Most of the measured accuracy increase therefore lies in correctness that survives the specified deletion, and the analysis does not detect an increase in the D-failing component. D-failing correctness nonetheless accounts for between 23.27 and 26.81 points of accuracy at every rung. Search rate and exposure to answer-bearing passages do not decrease with model size, and the D-surviving change is positive in each of the seven datasets. The per-dataset results are descriptive (Appendix D).

5.2 Related checkpoint comparisons

The same concentration appears in related comparisons on the identical 623 questions (Figure 1B). In all five Qwen2.5 recipe ladders, the 3B-to-7B D-surviving interval excludes zero under ITT, in four of five common-parse analyses, and in three of five conservative imputations of unparsed D outputs; every D-failing point estimate is negative. The equal-weight mean changes are $+7.09\,\mathrm{pp}$ in D-surviving and $-1.99\,\mathrm{pp}$ in D-failing correctness. From 3B to 14B, prompted Qwen2.5-Instruct checkpoints gain $16.85\,\mathrm{pp}$ of accuracy, $13.96\,\mathrm{pp}$ of it D-surviving; the SSRL Llama pair gains $8.83\,\mathrm{pp}$ of D-surviving correctness while D-failing correctness falls by $3.69\,\mathrm{pp}$. These named systems are a set of released checkpoints rather than a random sample of training methods. The Qwen comparisons stay within Qwen2.5; the Llama comparison also changes model version (Appendix D).

5.3 Construct checks

Table 1 collects the checks that bear most directly on what the deletion removes and how labels are produced; each qualifies the result in a different way.

Some D-surviving cases retain direct witnesses, preventing a witness-free interpretation of those outcomes; even under the adverse reassignment, more than half of the accuracy gain remains D-surviving. The pattern also holds where ordinary trajectories displayed a witness at every rung, and, for a fixed retrieved set, targeted deletion changes answers more often than count-matched random deletion (Appendix E).

Table 1: Construct checks for the primary Search-R1 3B-to-14B D-surviving change ($+9.47$ pp, $n = 623$). Each row addresses a different threat and has a different scope; none is an unconditional bound. Full analyses appear in Appendices D and E.

Threat	Evidence	Scope
Incomplete deletion	Both primary judges flag a surviving direct witness in 28.5% of D-surviving and 6.6% of D-failing cases. Moving every case flagged by either primary judge, and the three unannotated 14B cases, against the contrast leaves $+5.94$ pp (57.8% of the accuracy gain).	Sensitivity scenario, not a lower bound
Witness exposure	$+16.18$ pp $[+11.65, +20.71]$ on the $n = 309$ questions whose ordinary trajectory displayed a matched witness in all three arms.	Restricted population
Parser failures	$+11.93$ pp $[+7.95, +15.90]$ on the $n = 327$ questions parsed in both worlds at all three scales.	Complete-case analysis
Execution	An independent 14B execution reproduces each aggregate component within 0.16 pp.	One endpoint only
Deletion volume	On $n = 500$ questions with a fixed retrieved set, targeted S deletion lowers survival of the original answer by 21.2 pp; count-matched random deletion lowers it by 4.4 pp.	Fixed-context S check, not end-to-end D

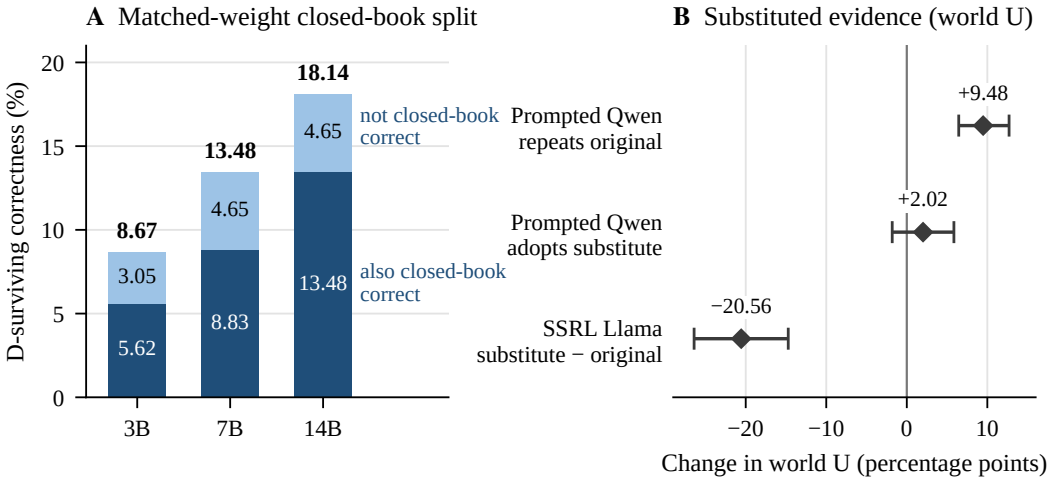


Figure 3: **Closed-book overlap and substituted evidence.** (A) D-surviving correctness on the matched Search-R1 ladder ($n = 623$), split by whether the same checkpoint also answers correctly closed-book. (B) Changes in world U with 95% paired intervals on 496 substitution items: prompted Qwen from 3B to 14B, and the substitute-minus-original margin for SSRL Llama from 3B to 8B.

6 Closed-Book Overlap and Response to Conflicting Evidence

A D-surviving increase can arise because larger checkpoints more often produce the answer without retrieval, because they respond less to conflicting evidence, or both. We examine each with the behavioral measurements in Figure 3.

Closed-book overlap. Closed-book evaluation asks whether the same weights answer correctly under a no-search prompt. On the matched-weight Search-R1 ladder, 7.87 pp $[+5.46, +10.43]$ of the 9.47 pp D-surviving increase overlaps with closed-book correctness; the remaining $+1.61$ pp has interval $[-0.32, +3.53]$. Across seven agents, 58.2–75.0% of D-surviving answers are also closed-book correct on a common complete-case panel. Most of the increase therefore overlaps with answers the same checkpoint also produces correctly without retrieval. This overlap does not identify

which information produced the original answer or provide a mediation estimate. A no-search prompt may also fail to elicit a correct answer (Appendix F).

Substituted evidence. World U rewrites displayed occurrences of the accepted answer to a same-type substitute. From 3B to 14B, prompted Qwen repeats the original answer 9.48 pp $[+6.45, +12.70]$ more often, while substitute adoption changes by $+2.02$ pp $[-1.81, +5.85]$. For SSRL Llama, the substitute-minus-original margin changes by -20.56 pp $[-26.41, -14.72]$ from 3B to 8B. World U measures responses to artificial conflicts created with deliberately false substitutes. Exposure depends on each policy's retrieval trajectory: 265 of 496 Llama-3B and 307 of 496 Llama-8B episodes display a substitute. These contrasts therefore reflect both differences in exposure and differences in response.

Attention heads. Held-out ablations of selected heads shift the first-token substitute-versus-original margin more than layer-matched controls in Search-R1 and IKEA. Readouts from these heads, however, do not reliably predict whole-policy D-survival. The causal finding is limited to the first-token margin (Appendix G).

7 Label-Free Diagnosis by Self-Ablation

Gold deletion defines the target but requires a reference answer; self-ablation applies the same deletion to the agent's own answer. We ask whether S tracks the D label, whether it differs from judging support, and what a release rule based on S achieves and at what cost.

7.1 Does self-ablation track the D label?

On 1,120 ordinary-correct Search-R1 questions, the binary S label predicts the gold-indexed D label with balanced accuracy 0.843 (F1 0.767; Table 21). Aggregate S- and D-survival rates differ by $+0.89$ points $[-1.25, +3.04]$, and item-level labels agree on 87.1% of questions. Because D and S share the alias matcher, a witness missed in both worlds can produce agreement, so this correspondence does not independently validate the matcher. An independent Search-R1-14B execution reproduces both the D correspondence and the separation in U (Appendix H).

Replay variability is a natural alternative explanation. A post-hoc placebo on 951 ordinary-correct questions with five matched O and S replays compares one ordinary replay's survival of the original answer with one S replay: across the five replicates, balanced accuracy for the D label is 0.541–0.555 for the ordinary replay and 0.759–0.779 for S, and each paired difference has a question-bootstrap confidence interval excluding zero (Appendix H). On 598 U-eligible cases, S-surviving answers repeat the original value in 36.4% of episodes, against 5.2% for S-changing answers. U provides an additional check, although it shares the alias inventory used by S.

7.2 Comparison with support and confidence signals

Table 2 compares signals on two targets: whether the displayed context suffices to answer (all answers) and D-survival (ordinary-correct answers). Sufficiency judges discriminate support better than D-survival, while self-ablation shows the opposite profile. The two targets are evaluated on different cohorts. On the D-survival cohort, fixed-evidence confidence signals reach direction-free AUROC between 0.532 and 0.572, and the strongest noninterventional score we tested, a commitment score, reaches 0.757, compared with 0.842 for self-ablation.

7.3 Selective release: benefits and costs

We evaluate selective release for settings in which correct answers are required to change when their retrieved answer witnesses are removed. The gate releases an answer when it changes under self-ablation, $R_{\mathrm{S}} = \{i : Z_{\mathrm{S},i} = 0\}$. For a release set R over n questions, the survival risk and the D-failing-correct yield are

$$\mathrm{risk}(R) = \frac{1}{|R|} \sum_{i \in R} L_i^{\mathrm{surv}}, \qquad \mathrm{yield}(R) = \frac{1}{n} \sum_{i \in R} L_i^{\mathrm{fail}}; \tag{4}$$

coverage $|R|/n$ and total correct yield are reported separately.

Table 2: AUROC on two targets. Support uses all answers ($n = 2{,}548$); D-survival uses ordinary-correct answers ($n = 946$). The difference column is support minus D-survival AUROC with a 95% bootstrap interval. The cohorts differ, so the comparison is descriptive.

Signal	Support	D-survival	Difference
GPT-4.1-mini sufficiency	.745	.510	$+.234$ $[+.200, +.255]$
GPT-4.1 sufficiency	.774	.560	$+.214$ $[+.168, +.257]$
Llama-70B sufficiency	.758	.513	$+.245$ $[+.204, +.267]$
Self-ablation	.516	.842	$-.327$ $[-.362, -.289]$

Table 3: Selective release on $n = 2{,}591$ Search-R1 questions. Risk is D-surviving-correct share of released answers. Comparator rows minimize risk subject to D-failing-correct yield of at least 26.48%, the gate's own yield; thresholds chosen retrospectively. Self-ablation is fixed binary rule.

Rule	Survival risk	Coverage	Extra evaluation
Release all	8.14%	100.00%	none
Self-ablation	**2.44%**	80.78%	1 policy rollout
GPT-4.1 sufficiency	6.82%	95.14%	1 judge call
Support self-report	7.35%	96.57%	1 model call
SelfCheckGPT	7.41%	95.87%	5 samples
Mean sequence log-probability	8.03%	98.03%	none

On 2,591 Search-R1 questions, self-ablation lowers survival risk from 8.14% under release-all to 2.44% at 80.78% coverage (Table 3). It requires one additional whole-policy rollout, withholds 214 correct answers, and lowers total correct yield by 8.26 points. Comparator thresholds are chosen retrospectively by exact rank sweeps that minimize risk subject to at least the gate's 26.48% D-failing-correct yield, and comparators differ in compute. The strongest, a GPT-4.1 sufficiency judge, reaches 6.82%; its paired risk gap to self-ablation is $+4.36$ points $[+3.46, +5.34]$.

If a release is invalid when it is either wrong or D-surviving correct, answer-all has 71.44% invalidity, a free string-containment gate 64.39%, and self-ablation 67.22%: the string rule filters wrong answers, whereas self-ablation targets D-surviving correctness. The same fixed gate lowers D-panel risk for all five evaluated agents, but its advantage over the closest noninterventional comparator is smaller under world U, where its interval excludes zero for only two of five agents (Appendix H).

8 Repair Experiments

We next test whether interventions can convert D-surviving correctness into D-failing correctness while preserving ordinary correctness and coverage. A successful repair should also follow substituted evidence in U and pass sham and output-validity controls. Answer-only rewards do not distinguish policies on this target: policies with the same ordinary answer distribution receive the same expected reward however they respond to edited evidence (Proposition 1). The scaling pattern also appears in prompted checkpoints without search-agent reinforcement learning.

Table 4 summarizes the tested interventions. Answer-conditioned continuation often retrieves corroboration for the candidate being checked. Quote-and-regenerate pipelines reduce D-surviving output mainly through abstention. The evidence-only reader benefits substantially from a retrieved witness, but we detect no full-panel improvement in D-failing correctness at the tested operating point. In the single-seed training pilot, the high-exposure penalty reduces D-surviving correctness with an accompanying accuracy loss and no detected increase in D-failing correctness.

9 Discussion and Limitations

Implications. *Evaluation:* search-agent accuracy can improve mostly in a component that persists under a specified retrieval edit, so claims of improved evidence-responsive behavior should report the paired O/D decomposition alongside accuracy, with the deletion rule and replay protocol stated. *Diagnosis:* support sufficiency and deletion response answer different questions, whether the dis-

Table 4: Tested repair implementations, with the worlds and cohort behind each reported quantity. Changes are in percentage points. The training row is a single-seed pilot; its intervals resample questions, not training runs. S-evaluated gate and bonus pilots and O-only context-aware decoding are reported in Appendix I.

Implementation	Worlds and cohort	Reported outcome
Answer-conditioned continuation	O/D; six agent–depth settings	One unique D-surviving to D-failing conversion across settings; the query retrieves apparent support in 32.54% of the 2,056 triggered continuations whose answer stays unchanged
QUOTE / REGEN	O/D; fixed 400-question panel	D-surviving 5.75% $\rightarrow$ 2.25% / 1.75%; D-failing 24.00% $\rightarrow$ 24.25% / 23.75%; 32 / 34 abstentions
Evidence-only GPT-4.1 reader	O/D/U; replaces outputs on 67 routed cases	Gain of $+54.29$ $[40.00, 68.57]$ points when shown the retrieved witness, relative to witness-empty input; full-panel D-failing change 0.00 $[-1.50, +1.25]$; ordinary answer coverage 62.69%
High-exposure penalty vs. control	O/D; $n = 505$	D-surviving -3.17 $[-5.35, -0.99]$; D-failing -0.40 $[-4.16, +3.56]$; accuracy -3.56 $[-7.13, 0.00]$

played context suffices and whether the answer moves when witnesses are removed; they may be complementary. *Training:* a lower D-surviving rate can reflect abstention or degraded output rather than conversion, so objectives should be assessed jointly on ordinary correctness, the O/D partition, coverage, and conflict behavior.

Limitations. *Construct validity and replay coupling.* The matcher-defined deletion can miss paraphrases, indirect clues, and knowledge outside the retrieved set, so D-survival refers to the specified intervention; each label comes from one coupled replay, and greedy executions are not repeatable at the item level. *Checkpoint and configuration inference.* Scale comparisons are observational across separately trained checkpoints, and the intervals cover questions rather than training runs. The primary ladder changes tensor-parallel execution above 3B, the independent execution checks only the 14B endpoint, and the Llama pair changes model version. *Diagnostic and deployment scope.* The closed-book result is associative, and the head intervention is causal only for its first-token substitution margin. S shares matcher error with D, has not been prospectively validated under deployment shift, and has a smaller advantage under world U. Its release rule targets D-surviving correctness rather than factuality and costs correct yield and compute.

10 CONCLUSION

From 3B to 14B, 9.47 of the 10.27-point net accuracy increase of released Search-R1 checkpoints lies in correctness that survives answer-witness deletion, and related comparisons show the same concentration. Much of it overlaps with closed-book correctness, and self-ablation makes the intervention a label-free diagnostic with measurable costs. Search-agent evaluations should report how accuracy responds to targeted changes in retrieved evidence, alongside accuracy and support.

REPRODUCIBILITY STATEMENT

The supplementary material provides row-level derived results, provenance manifests, analysis and figure scripts, and an executable verifier for every reported table and figure. Appendix J documents model revisions, data construction, inference settings, uncertainty, and known historical-provenance limitations.

ETHICS STATEMENT

The study uses public QA data and no private user data. Counterfactual retrieval editing can support selective evidence-response screening, but can also be repurposed for evidence suppression or corpus probing. Appendix J discusses these risks, mitigations, and coverage disparities.

AI USE STATEMENT

Generative AI assisted with code drafting, literature discovery, manuscript editing, and critical review. The authors verified retained claims against local artifacts and primary bibliographic records and remain responsible for the study and submitted material.

REFERENCES

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556, 2022.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. Context-DPO: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10280–10300, 2025. doi: 10.18653/v1/2025.findings-acl.536.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. Contextcite: Attributing model generation to context. In *Advances in Neural Information Processing Systems*, 2024.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020. doi: 10.18653/v1/2020.acl-main.408.

Yuchen Fan, Kaiyan Zhang, Heng Zhou, Yuxin Zuo, Yanxu Chen, Yu Fu, Xinwei Long, Xuekai Zhu, Che Jiang, Yuchen Zhang, Li Kang, Gang Chen, Cheng Huang, Zhizhou He, Bingning Wang, Lei Bai, Ning Ding, and Bowen Zhou. SSRL: Self-search reinforcement learning. *arXiv preprint arXiv:2508.10874*, 2025.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Sachin Goyal, Christina Baek, J. Zico Kolter, and Aditi Raghunathan. Context-parametric inversion: Why instruction finetuning can worsen context reliance. In *International Conference on Learning Representations*, 2025.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of COLING*, pp. 6609–6625, 2020. doi: 10.18653/v1/2020.coling-main.580.

Ziyang Huang, Xiaowei Yuan, Yiming Ju, Jun Zhao, and Kang Liu. Reinforced internal-external knowledge synergistic reasoning for efficient adaptive search agent. *arXiv preprint arXiv:2505.07596*, 2025.

Siddharth Jain and Venkat Narayan Vedam. CUE-R: Beyond the final answer in retrieval-augmented generation. *arXiv preprint arXiv:2604.05467*, 2026.

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

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of ACL*, pp. 1601–1611, 2017. doi: 10.18653/v1/P17-1147.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Shiyu Li, Yang Tang, Yifan Wang, Peiming Li, and Xi Chen. Reseek: A self-correcting framework for search agents with instructive rewards. *arXiv preprint arXiv:2510.00568*, 2025.

Llama Team, AI at Meta. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.565.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of ACL*, pp. 9802–9822, 2023. doi: 10.18653/v1/2023.acl-long.546.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of EMNLP*, 2023.

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, 2024.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Debayan Mukhopadhyay, Utshab Kumar Ghosh, and Shubham Chatterjee. Bridge evidence: Static retrieval utility does not predict causal utility in multi-step agentic search. *arXiv preprint arXiv:2607.15253*, 2026.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of EMNLP*, pp. 5687–5711, 2023. doi: 10.18653/v1/2023.findings-emnlp.378.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Rishiraj Saha Roy, Joel Schlotthauer, Chris Hinze, Andreas Foltyn, Luzian Hahn, and Fabian Kuech. Evidence contextualization and counterfactual attribution for conversational QA over heterogeneous data with RAG systems. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2025. arXiv:2412.10571.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of NAACL*, 2024.

Yaorui Shi, Sihang Li, Chang Wu, Zhiyuan Liu, Junfeng Fang, Hengxing Cai, An Zhang, and Xiang Wang. Search and refine during think: Facilitating knowledge refinement for improved retrieval-augmented reasoning. *arXiv preprint arXiv:2505.11277*, 2025.

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. Zerosearch: Incentivize the search capability of llms without searching. *arXiv preprint arXiv:2505.04588*, 2025a.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *International Conference on Learning Representations*, 2025b.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the ACM SIGIR International Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pp. 22–32, 2025. arXiv:2412.18004; ICTIR 2025 Best Paper Honorable Mention.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Theodora Worledge, Judy Hanwen Shen, Nicole Meister, Caleb Winston, and Carlos Guestrin. Unifying corroborative and contributive attributions in large language models. *arXiv preprint arXiv:2311.12233*, 2023.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of EMNLP*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Appendix Roadmap

Appendices follow the order of the main text and retain the complete numerical evidence behind it.

Appendix	Contents	Supports
A	Replay coupling, reward equivalence, non-identification, S/D pivot, repair bounds	Section 3
B	World transformations, substitutes, matching, no-op cases, replay variation	Section 3
C	Datasets, checkpoints, retrieval, parsing, configuration, cohort map	Section 4
D	Full scaling tables, parser readings, per-dataset and control comparisons	Section 5
E	Residual witnesses, reassignment, sham, exposure, shared matcher	Table 1
F	Closed-book overlap across agents, matched-weight partition, world U	Section 6
G	Attention-head selection, held-out ablations, scope	Section 6
H	Detector correspondence, support versus survival, release fronts, replays	Section 7
I	Repair protocols, bounds, and training pilots	Section 8
J	Artifacts, ethics, and AI assistance	Statements

A Formal Results and Proofs

A.1 Potential Outcomes and Replay Coupling

Let R collect every exogenous source of rollout randomness and let κ be the declared joint law of the randomness used in O and D. The potential answers are

$$A_O(x, R), \qquad A_D(x, R) = A_\pi(q, T_{[g]}(E); R).$$

Marginal correctness under D is an interventional distribution. By contrast, the joint event $Y_O Y_D$ is a cross-world quantity and depends on κ. A one-replay label is therefore a realized paired outcome whose interpretation depends on the coupling. Shared randomness, matched independent randomness, and unmatched execution settings support different interpretations. Under coupling κ, we write the D-surviving and D-failing labels of Section 3 as

$$L_\kappa^{\text{surv}} = Y_O Y_D, \qquad L_\kappa^{\text{fail}} = Y_O(1 - Y_D).$$

Lemma 1 (Paired accuracy decomposition)**.** *For every realized O/D pair,*

$$Y_O = Y_O(1 - Y_D) + Y_O Y_D.$$

Proof. Because $Y_D \in \{0, 1\}$, $1 = (1 - Y_D) + Y_D$. Multiplication by Y_O gives the result. □

The lemma is an accounting identity and requires no causal assumption. Interpreting D-survival as an absence of evidence use additionally requires complete witness removal; the matcher may leave a witness or the policy may recover semantically equivalent evidence.

A.2 Answer-Only Reward Blindness

Let $r(q, [a], [g])$ be a reward depending on the question, emitted answer class, and accepted answer class, but not on the evidence or trajectory.

Proposition 1 (Outcome-reward blindness)**.** *If policies π_1 and π_2 induce the same conditional distribution of accepted answer classes given q in O, then they receive the same expected reward in O.*

Proof. Conditioning on q and summing over answer classes gives

$$\mathbb{E}_{\pi_j}[r \mid q] = \sum_{[a]} P_{\pi_j}([a] \mid q) r(q, [a], [g]).$$

The summands agree for $j = 1, 2$ by hypothesis, and hence so do their expectations. □

The proposition establishes reward equivalence. Characterizing every optimum additionally requires a policy class that can realize an observation-blind policy with the same ordinary answer law. The released Search-R1 correctness scorer illustrates the premise operationally: a correct answer receives reward 1 with complete, empty, irrelevant, cross-paired, or absent evidence, while a wrong answer receives 0 with complete evidence.

A.3 Fixed-Evidence Non-Identification

Fix a question q and the exact evidence payload C at commitment. Let $\mathcal{U}_C$ contain detector prompts whose instructions and candidate fields may vary while C remains unchanged, and define

$$\Sigma_C(\pi) = \{P_\pi(\cdot \mid u) : u \in \mathcal{U}_C\}.$$

A fixed-evidence detector is a possibly randomized functional of q, C, A_O, and $\Sigma_C(\pi)$.

Proposition 2 (Fixed-evidence non-identification)**.** *Suppose a policy class contains π_R and π_M such that, on a positive-probability set of questions: (i) their ordinary trajectory laws agree; (ii) their fixed-evidence signatures agree; and (iii) after the full D intervention, π_R emits a non-gold answer while π_M emits an accepted gold answer. No fixed-evidence detector uniformly identifies D-survival or D-failure over this class.*

Proof. Construct the two policies to agree on every history reachable in O and on every prompt in $\mathcal{U}_C$. Every fixed-evidence detector then has the same distribution under both policies. Extend the policies differently only on histories reached after $T_{[g]}(E)$: let π_R emit a fixed non-gold answer and π_M emit an accepted gold answer. Their D labels differ on a set of positive probability, so no common detector output can identify both. The construction may be deterministic and does not rely on a particular stochastic coupling. □

This is the familiar observational-equivalence argument specialized to a search policy (Pearl, 2009; Bareinboim et al., 2022). Nuisance correlations may still support useful prediction on a particular benchmark despite the absence of identification. Table 5 classifies the evaluated signals by whether they remain inside this fixed-evidence information class.

Table 5: Membership of evaluated signals in the fixed-evidence class, defined by the conditionals available to each score.

Signal	Fixed evidence?	Reason
Mean sequence log probability	yes	token conditionals under original C
Normalized-answer entropy	yes	repeated samples under unchanged C
Exact-answer consistency	yes	repeated samples under unchanged C
Support self-report	yes	new instruction, same evidence payload
Retrieval count	no	reads actions across changing contexts
Self-ablation	no	reruns in $T_{[A_O]}(E)$

A.4 Equivalence of Gold and Self-Ablation

Define $Z_S = \mathbf{1}\{A_S \sim A_O\}$.

Proposition 3 (Self-ablation pivot)**.** *Assume (i) $a \sim g$ implies $T_{[a]}(E) = T_{[g]}(E)$; (ii) S and D execute the same policy and code path; and (iii) they share exogenous randomness. On every ordinary-correct unit,*

$$A_S = A_D, \qquad Z_S = Y_D,$$

and therefore $Y_O Z_S = L_\kappa^{\text{surv}}$ and $Y_O(1 - Z_S) = L_\kappa^{\text{fail}}$.

Proof. If $Y_O = 1$, then $[A_O] = [g]$, so assumption (i) makes the S and D environments identical. Assumptions (ii) and (iii) give the same policy input and exogenous state, hence $A_S = A_D$. Because $A_O \sim g$, survival of A_O is equivalent to correctness of A_D. □

Under matched independent randomness, the conditional laws and population expectations agree, although realized item labels may differ. In the implementation, D receives a supplied alias set whereas S is indexed by the prediction string. We therefore report empirical S/D correspondence.

A.5 Limits of Post-Hoc Repair

Observation 1 (Worldwise frozen emission)**.** *If a postprocessor produces (B_O, B_D) with $B_O \sim A_O$ and $B_D \sim A_D$ for every unit, then its output-defined D-failing and D-surviving labels equal those of the base policy.*

The statement follows immediately because both labels depend only on the accepted answer classes in O and D. Its scope excludes regeneration, abstention, changed retrieval, and world-dependent revision.

Observation 2 (Logged-pool inherited-label ceiling)**.** *Let a fixed pool for unit i contain candidates with inherited labels $\ell_{ij} \in \{0, 1\}$. Any selector σ evaluated by the chosen candidate's inherited label satisfies*

$$\frac{1}{n}\sum_i \ell_{i,\sigma(i)} \le \frac{1}{n}\sum_i \max_j \ell_{ij}.$$

The inequality holds pointwise. In our 28-unit witness-contrastive pool, only two units contain a different candidate that is both correct and potentially D-failing, giving a 7.1% oracle ceiling. The bound applies to this logged pool and inherited labels; it does not extend to larger generated pools or to a selector rerun separately in O and D.

B Counterfactual Worlds and Implementation

B.1 Full-Policy Transformations

At search step t, the base retriever returns a ranked pool $P_t = (p_{t1}, \ldots, p_{t20})$. The agent sees the first three passages after the world-specific operation:

- O returns $p_{t1:3}$ unchanged.
- D removes every candidate whose normalized title or text matches a supplied gold alias, then returns the first three remaining candidates.
- S applies the same filter using the normalized ordinary prediction.
- U retains rank order but rewrites recognized occurrences of the gold aliases to an input-defined same-type substitute before displaying the top three.

After each response, the policy may issue a different next query. D and S therefore operate on every new retrieval call. U maintains local consistency in displayed evidence without rebuilding the complete corpus.

B.2 Substitute Construction and Rewriting

Substitutes are assigned before model evaluation with random-number-generator seed 20260919. PopQA draws another object from the same relation. Other items start from a heuristic candidate: years are shifted by a uniformly sampled 3–17 years, and remaining questions draw a same-dataset answer matched on question type and answer length. A fixed GPT-4.1-mini rewrite at temperature zero then replaces the non-PopQA candidates with plausible false answers of the same type, format, and granularity. Candidates are rejected if they normalize to a gold alias, contain or are contained by one, have normalized length below two, or occur in the question. The final 3,125-row evaluation artifact contains 2,138 usable substitutes and retains both the initial candidate and the filtered value. The analysis uses this frozen substitution artifact; because the final filtering implementation is unavailable, that stage cannot currently be regenerated end to end.

At runtime, accepted aliases are accent-folded, ordered longest first, escaped into case-insensitive whole-alphanumeric-boundary expressions with flexible whitespace, and replaced in both the title and body of every displayed passage. PopQA additionally uses its object aliases. The same substitute is applied after every query. Each episode records whether an original alias or substitute was displayed. Head-intervention eligibility further requires that the commitment prefix contain the substitute, contain no original alias, and use distinct first tokens for the substitute and original answer.

B.3 Answer Matching and Exclusions

Exact match uses SQuAD-style normalization, accent folding, and the accepted alias list. We exclude yes/no items and questions in which an accepted answer already appears in the question, because removing retrieved passages cannot identify dependence there. A passage is treated by the matcher as stating an answer only when a normalized alias boundary matches. This prioritizes precision and does not detect arbitrary paraphrases, a limitation quantified by the residual-witness analysis.

B.4 Treatment and No-Op Cases

Deletion can be a no-op when no matched witness occurs in the candidate pool. Such units remain in the ITT analysis because they are part of the deployable procedure; treatment-received and ordinary-witness-exposed subsets appear separately in Appendix E. Backfill holds the displayed passage count fixed whenever the finite pool contains enough untreated passages.

B.5 Replay Variation

Generation uses nominally greedy decoding, but batching, kernels, and multi-turn sampling make executions empirically non-identical. Five ordinary replays disagree on at least one answer for 53.1% of the full panel, while five executions of the self-ablation verdict disagree for 15.4%. The primary analysis therefore treats D-survival as the outcome of a declared paired replay and uses population-level contrasts. Repeated-replay results are reported in Appendix H.

C Datasets, Models, and Evaluation

C.1 Datasets and Common Panels

We draw up to 100 questions per dataset from 2WikiMultiHopQA (Ho et al., 2020), Bamboogle (Press et al., 2023), HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), Natural Questions (NQ) (Kwiatkowski et al., 2019), PopQA (Mallen et al., 2023), and TriviaQA (Joshi et al., 2017). Input-defined validity exclusions remove yes/no, empty-answer, and answer-in-question cases. We then intersect question IDs present in every required arm and exclude any D trace that re-exposes the gold answer. In the realized comparison, every arm contains all 700 scheduled questions and no unit is removed by the D-trace rule, leaving a 623-question common valid-witness panel. Selective-release analyses use the larger 2,591-question Search-R1 panel; detector validation uses 1,120 ordinary-correct questions; cohort sizes are always reported with results. Appendix C.6 lists every cohort.

C.2 Models and Provenance

The primary endpoints are released Search-R1 Qwen2.5 instruction-tuned checkpoints trained with proximal policy optimization (PPO) at 3B, 7B, and 14B (Jin et al., 2025; Qwen Team, 2024). The recipe panel adds ZeroSearch, IKEA, ReSeek, and AutoRefine at 3B and 7B. Prompted controls use public Qwen2.5-Instruct 3B and 14B checkpoints under one search scaffold. The SSRL Llama comparison uses pinned SSRL Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct revisions (Fan et al., 2025; Llama Team, AI at Meta, 2024; Meta AI, 2024).

C.3 Retriever and Inference

Agents search a pinned Wikipedia-2018 corpus of 21,015,324 passages through an E5-base-v2 dense retriever (Wang et al., 2022) and the released index. Each call retrieves a pool of 20 and displays three passages. Agents may make four searches and generate at most 500 new tokens per turn. Recipe-specific native chat templates and parsers are retained. Generation uses temperature zero and seed zero. Questions, world sets, per-dataset quotas, prompts, and generation budgets are matched in the primary ladder, but the historical 7B and 14B runs use tensor parallelism (TP) across two A4000s whereas the 3B run does not.

C.4 Parsing and Estimands

The parser reads the final `<answer>` span and tolerates a missing closing tag. Unparsed outputs are incorrect under ITT. Common-parse estimates retain only units parsed in every compared arm. Conservative D imputation credits a D nonparse only to the smaller checkpoint, favoring the null direction; the primary Search-R1 imputation row instead credits every D-world nonparse as survival (Appendix D). Checkpoint contrasts are paired by question ID; 95% intervals use 10,000 paired percentile-bootstrap resamples unless another count is stated.

C.5 Configuration Matching

Primary behavioral contrasts fix questions, per-dataset batching, prompt family, retriever, worlds, and turn and token budgets. Engine configuration is matched where the artifacts permit; the primary 3B-to-7B/14B ladder retains the tensor-parallel difference stated above. This matters empirically: a repair-disabled comparison that jointly changes the context window from 8192 to 10240 tokens and tensor parallelism from two to one shifts accuracy by 8.57 points. The independent 14B closed-book and replay sensitivities deliberately compare TP$=$ 1/A100 execution with historical TP$=$ 2/A4000 artifacts and are labeled as configuration sensitivities rather than hardware-matched scale experiments. They reproduce aggregate conclusions but show substantial item-level variation, reinforcing the distinction between item-bootstrap and run-level uncertainty.

C.6 Cohort Map

Table 6 lists the cohorts used in the paper, their eligibility rules as stated in the relevant appendix, and where they are used. Subset sizes are those reported with each result.

Table 6: Cohorts, eligibility, and use.

Cohort	n	Eligibility	Used in
Scaling panel	623	Up to 100 questions per dataset after yes/no, empty-answer, and answer-in-question exclusions; present in every compared arm (App. C)	Figure 1, Sections 5–6
Common parse	327; 523	Scaling-panel questions parsed in every compared arm (Search-R1; prompted Qwen)	Tables 8, 12
Exposure subsets	415; 309; 304; 589	Scaling-panel questions satisfying each restriction in all three Search-R1 arms	Table 16
Residual-witness census	945	Eligible finite population of 207 D-surviving and 738 D-failing control cases (App. E)	Tables 14, 15
Endpoint witness check	187	Questions eligible at both exact 3B and 14B endpoints	App. E
Sham check	500	Questions with a fixed retrieved set; S and count-matched random deletion	Table 1
U substitution items	496	Input-defined substitution items (App. F)	Figure 3
Head intervention	461; 260	Held-out halves of eligible U episodes (Search-R1; IKEA)	App. G
Detector panel	1,120	Ordinary-correct Search-R1 questions	Section 7
Replay placebo	951	Ordinary-correct questions with five matched O and S replays	Section 7
S-stratified U	598	U-eligible cases	Section 7
Support task	2,548	All answers; target is context sufficiency	Table 2
D-survival task	946	Ordinary-correct answers	Table 2
Release panel	2,591	Search-R1 questions, correct or incorrect	Table 3
D/U release panels	1,025–1,816	Each agent's D/U-eligible questions	Table 22
Repair panels	400; 505; 560; 518	Method-specific (App. I)	Table 4

D Additional Scaling Results

D.1 Search-R1 Ladder and Parser Sensitivity

Table 7 gives the levels underlying Figure 1A.

Table 7: Question-, world-, prompt-, and budget-matched Search-R1 ladder on the common panel ($n = 623$). The 7B and 14B runs use tensor parallelism while the 3B run does not. The final column is the paired 14B-minus-3B change in percentage points with a 95% interval.

	3B	7B	14B	Δ_{14B-3B}
Accuracy	34.67	36.76	44.94	$+10.27$ $[+6.90, +13.64]$
D-failing correctness	26.00	23.27	26.81	$+0.80$ $[-3.05, +4.49]$
D-surviving correctness	8.67	13.48	18.14	$+9.47$ $[+6.74, +12.20]$

Table 8 brackets the primary 3B-to-14B contrast under three parser treatments. ITT counts every nonparse as incorrect. Common parse retains only questions parsed in both worlds at all three scales. D-nonparse imputation credits every D-world nonparse as survival, a check on whether parser failures create the primary contrast. The panel falls from 623 to 327 under common parse, but the primary D-surviving contrast remains positive under all three readings.

Table 8: Search-R1 14B-minus-3B changes under parser sensitivity. Values are percentage points with 95% paired bootstrap intervals.

Analysis	n	Δaccuracy	ΔD-failing	ΔD-surviving
ITT	623	$+10.27$ $[+6.90, +13.64]$	$+0.80$ $[-3.05, +4.49]$	$+9.47$ $[+6.74, +12.20]$
Common parse	327	$+10.09$ $[+5.50, +14.68]$	-1.83 $[-7.03, +3.36]$	$+11.93$ $[+7.95, +15.90]$
D-nonparse imputation	623	$+10.27$ $[+6.90, +13.64]$	-2.57 $[-6.26, +1.12]$	$+12.84$ $[+9.63, +16.21]$

The D-surviving increase is also positive in every dataset on the ITT panel (Table 9). These cells are descriptive disaggregations of the registered panel rather than seven separately powered hypothesis tests.

Table 9: Primary Search-R1 3B-to-14B D-surviving-correctness change by dataset, in percentage points.

	2Wiki	Bamboogle	HotpotQA	MuSiQue	NQ	PopQA	TriviaQA
n	55	99	79	99	97	97	97
ΔD-surviving	+3.64	+7.07	+5.06	+3.03	+15.46	+11.34	+17.53

D.2 Scaling Across Five Qwen2.5 Recipes

Table 10 gives the ITT contrasts from the five released 3B-to-7B ladders plotted in Figure 1B. The questions and panel fingerprint are identical across rows. Recipe names identify the released checkpoints evaluated here.

All five ITT D-surviving intervals exclude zero. Four remain positive on common-parse panels; ZeroSearch, IKEA, and ReSeek remain positive under conservative D-nonparse imputation. Search-R1 and AutoRefine include zero under that analysis. The equal-weight changes are $+7.09$ pp $[+5.62, +8.64]$ for D-surviving correctness and -1.99 pp $[-3.76, -0.16]$ for D-failing correctness. These averages summarize the five named recipes without treating them as a random sample of training methods.

D.3 Per-Dataset Recipe Results

Table 11 disaggregates the 3B-to-7B D-surviving contrast. Thirty-two of 35 recipe–dataset cells are positive, one is zero, and two are negative. Dataset sample sizes are fixed across recipes: 55 for

Table 10: Paired 3B-to-7B changes for five released Qwen2.5 search-agent recipes on the identical $n = 623$ panel. Entries are ITT percentage-point changes with 95% paired intervals. Every D-surviving interval excludes zero, while every D-failing point estimate is negative.

Recipe	Δaccuracy	ΔD-failing	ΔD-surviving
Search-R1	$+2.09\ [-1.28, +5.30]$	$-2.73\ [-6.42, +0.96]$	$+4.82\ [+2.25, +7.38]$
ZeroSearch	$+2.57\ [-0.64, +5.78]$	$-2.73\ [-6.10, +0.48]$	$+5.30\ [+2.73, +8.03]$
IKEA	$+6.74\ [+3.53, +10.11]$	$-0.64\ [-3.85, +2.57]$	$+7.38\ [+4.49, +10.43]$
ReSeek	$+10.43\ [+7.22, +13.64]$	$-1.12\ [-4.49, +2.41]$	$+11.56\ [+8.67, +14.45]$
AutoRefine	$+3.69\ [+0.32, +7.06]$	$-2.73\ [-6.42, +0.96]$	$+6.42\ [+3.85, +9.15]$

2Wiki, 99 for Bamboogle, 79 for HotpotQA, 99 for MuSiQue, and 97 each for NQ, PopQA, and TriviaQA.

Table 11: D-surviving-correctness change from 3B to 7B by recipe and dataset, in percentage points. Cells are descriptive and unadjusted for multiplicity.

Recipe	2Wiki	Bamboogle	HotpotQA	MuSiQue	NQ	PopQA	TriviaQA
Search-R1	+1.82	+4.04	+7.59	+2.02	+5.15	+3.09	+9.28
ZeroSearch	+1.82	+16.16	+5.06	+5.05	−1.03	−1.03	+9.28
IKEA	+5.45	+14.14	+1.27	+1.01	+6.19	0.00	+21.65
ReSeek	+10.91	+14.14	+6.33	+7.07	+13.40	+4.12	+23.71
AutoRefine	+10.91	+13.13	+1.27	+4.04	+3.09	+5.15	+8.25

D.4 Prompted Qwen Without Search-Agent Reinforcement Learning

Table 12 reports ITT and parser-sensitivity readings for the prompted-Qwen comparison.

Table 12: Prompted Qwen2.5-Instruct 3B-to-14B changes. Conservative imputation credits a D nonparse only to the smaller rung.

Reading	n	Δaccuracy	ΔD-surviving	ΔD-failing
ITT	623	$+16.85\ [+13.16, +20.55]$	$+13.96\ [+11.08, +17.01]$	$+2.89\ [-0.64, +6.42]$
Common parse	523	$+17.59\ [+13.77, +21.41]$	$+15.49\ [+12.24, +18.93]$	$+2.10\ [-1.53, +5.74]$
Conservative D parse	623	$+16.85\ [+13.16, +20.55]$	$+12.68\ [+9.63, +15.73]$	$+4.17\ [+0.64, +7.54]$

Common parse retains 83.9% of the panel and at least 78.5% within every dataset. The D-surviving change is positive in all seven datasets. The D intervention for this control has not received a separate blinded residual-witness analysis, so we describe the outcome as D-surviving rather than evidence-independent.

D.5 SSRL Llama Comparison

Table 13 reports the cross-family Llama comparison under the same three parser readings.

The D-surviving difference is positive for all seven datasets and under every reported parser analysis. The survival-minus-failure contrast is positive in six of seven datasets, with HotpotQA as the exception. These results do not establish a dataset-uniform interaction or a causal parameter-count effect.

D.6 Independent 14B Replay

The independent TP=1/A100 Search-R1-14B replay and historical TP=2/A4000 endpoint have identical aggregate accuracy, 44.94%, on the common panel. Their answer strings agree on 81.5% in O and 74.2% in D. Independent-minus-historical changes are $+0.16$ pp $[-1.93, +2.25]$ for D-

Table 13: Preregistered Llama larger-minus-smaller contrasts. Survival−failure is the change in D-surviving correctness minus the change in D-failing correctness. The comparison is Llama-3.2 3B to Llama-3.1 8B and therefore changes model version and parameter count.

Analysis	Accuracy	D-failing	D-surviving	Survival−failure
ITT	$+5.14$ $[+1.93, +8.51]$	-3.69 $[-7.22, -0.16]$	$+8.83$ $[+5.62, +12.04]$	$+12.52$ $[+6.74, +18.46]$
Common parse	$+6.31$ $[+2.78, +9.83]$	-5.38 $[-9.09, -1.67]$	$+11.69$ $[+8.35, +15.03]$	$+17.07$ $[+10.95, +23.19]$
Conservative D parse	$+5.14$ $[+1.93, +8.51]$	-3.21 $[-6.74, +0.32]$	$+8.35$ $[+5.30, +11.72]$	$+11.56$ $[+5.78, +17.50]$

failing and -0.16 pp $[-1.77, +1.44]$ for D-surviving correctness. This supports aggregate endpoint robustness while showing substantial unit-level execution variation.

E Sensitivity Analyses

E.1 Residual-Witness Annotation

Each adjudicator receives the question, accepted answer, and every passage displayed in D. They do not see the agent's answer, correctness, operational label, or the other adjudicator's verdict. The 945 cases comprise the eligible finite population; Table 14 gives the primary agreement counts.

Table 14: Residual-witness annotations under agreement of both primary judges. Counts are in parentheses.

What survives D	D-surviving ($n = 207$)	D-failing control ($n = 738$)
Directly states the answer	**28.5%** (59)	6.6% (49)
Answer remains derivable	34.8% (72)	10.8% (80)
Passages appear over-deleted	6.8% (14)	9.1% (67)

The direct-witness estimate is sensitive to judge pooling. On the 204 D-surviving cases with all three verdicts, individual rates range from 20.1% to 39.7%; all-three and any-of-three pooling give 14.7% and 46.1%. Because the annotation covers the finite population, this spread is judge sensitivity rather than binomial sampling uncertainty.

E.2 Adverse Endpoint Reassignment

The annotation covers 51/54 D-surviving cases at 3B and 110/113 at 14B. Table 15 moves every flagged case to D-failing and then assigns all three unannotated 14B cases, but none of the unannotated 3B cases, against the estimated contrast. These are sensitivity scenarios rather than observed relabelings or unconditional causal bounds.

Table 15: Remaining Search-R1 3B-to-14B D-surviving gain under adverse endpoint reassignment. These are sensitivity scenarios, not lower bounds.

Judge pooling	Annotated cases	Plus adverse unannotated assignment
Both primary judges	$+8.67$ (84.4%)	$+8.19$ (79.7%)
GPT-4.1-mini	$+8.51$ (82.8%)	$+8.03$ (78.1%)
Llama-3.3-70B	$+6.58$ (64.1%)	$+6.10$ (59.4%)
Either primary judge	$+6.42$ (62.5%)	$+5.94$ (57.8%)

On the 187 questions eligible at both exact endpoints, strict residual-witness presence is 19.3% at 3B and 15.5% at 14B, a paired -3.74 pp interval $[-8.02, +0.53]$ (exact McNemar $p = 0.143$). This diagnostic is inconclusive and does not establish that matcher fidelity changes with checkpoint size.

E.3 Size-Matched Sham

On 500 questions, targeted S deletion and a random sham each remove 1.95 passages per question. The target edit reduces survival of the original answer from 30.8% to 9.6%, a 21.2 pp change; the

count-matched sham gives 26.4%, a $4.4\,\mathrm{pp}$ change. Thus most of the response depends on which passages leave rather than deletion volume alone. This validation uses a fixed retrieved set and therefore checks the deletion mechanism, not the end-to-end S policy.

E.4 Exposure and Finite-Pool Sensitivity

The positive D-surviving change remains when restricting to cases where D removes at least one passage, where O displays a matched witness, and where sufficient untreated passages exist to backfill the displayed set. These restrictions address no-op treatment and finite-pool failure but change the target population. Table 16 reports each restriction as a separate sensitivity estimand.

Table 16: Search-R1 14B-minus-3B sensitivity analyses. Each restriction must hold in all three scale arms. Changes and intervals are in percentage points.

Panel	n	Δaccuracy	ΔD-surviving
Canonical	623	$+10.27$ $[+6.90, +13.64]$	$+9.47$ $[+6.74, +12.20]$
D edit received	415	$+11.57$ $[+7.47, +15.66]$	$+13.25$ $[+9.40, +17.11]$
O displays a matched witness	309	$+11.00$ $[+6.15, +15.86]$	$+16.18$ $[+11.65, +20.71]$
O witness and D edit	304	$+10.86$ $[+6.25, +15.46]$	$+16.12$ $[+11.51, +20.72]$
No D pool exhaustion	589	$+10.70$ $[+7.30, +14.26]$	$+9.51$ $[+6.62, +12.39]$

E.5 Dependence Induced by a Shared Matcher

D and S use the same alias matcher, so a witness missed in both worlds can induce agreement between their labels. D/S agreement therefore does not independently validate semantically complete deletion. U shares the alias inventory but tests a different behavioral response. The residual-witness annotation directly measures this shared limitation.

F Parametric Availability and Evidence Substitution

F.1 Closed-Book Overlap Across Seven Agents

Table 17 compares the overlap on a shared complete-case panel and on each agent's own parsed panel.

Table 17: Share of D-surviving correct answers that are also closed-book correct. Complete-case columns compare the same fully parsed questions across all seven agents.

Agent	Complete-case overlap	Own-panel overlap	Complete-case remainder
Search-R1-3B	71.2%	72.7%	28.8%
ZeroSearch-3B	58.2%	50.9%	41.8%
IKEA-3B	60.3%	56.2%	39.7%
ZeroSearch-Llama-3B	75.0%	75.2%	25.0%
Qwen-3B prompted	68.6%	69.2%	31.4%
ReSeek-3B	72.9%	62.0%	27.1%
AutoRefine-3B	66.1%	50.7%	33.9%

The closed-book prompt fails to elicit 25.0–41.8% of the complete-case D-surviving answers. Because closed-book prompting can have false negatives, this remainder does not exclude additional parametric availability. Overlap exceeds 58% for every agent, indicating a substantial measured component.

F.2 Search-R1 Scaling with Matched Weights

Table 18 partitions D-surviving correctness across the Search-R1 ladder. Each closed-book arm uses the same checkpoint weights as its corresponding open-book arm.

Table 18: Descriptive matched-weight partition on the identical $n = 623$ Search-R1 panel. Overlap and remainder partition D-surviving correctness itemwise.

Rung	D-surviving	Closed-book	Joint overlap	Remainder
3B	8.67	13.32	5.62	3.05
7B	13.48	17.82	8.83	4.65
14B	18.14	26.16	13.48	4.65
Δ3B$\rightarrow$14B	$+9.47$	$+12.84$	$+7.87$	$+1.61$

The joint increase is $+7.87$ pp $[+5.46, +10.43]$ and the remainder $+1.61$ pp $[-0.32, +3.53]$. Pairing the same 14B closed-book run with the independent TP=1 open replay gives $+7.22$ pp $[+4.98, +9.63]$ overlap and $+2.09$ pp $[+0.16, +4.01]$ remainder. Under both execution settings, most of the shift co-occurs with closed-book correctness, while the remainder remains sensitive to execution configuration.

F.3 Substitution Analysis

The U panel contains 496 input-defined substitution items. Because retrieval is policy-dependent, not every episode displays a substituted passage. We therefore report aggregate adoption and original repetition as diagnostics and retain exposure counts. In the Llama comparison, 265/496 3B and 307/496 8B episodes expose the substitute. The -20.56 pp margin change consequently mixes a change in exposure with a change in response. The prompted-Qwen and five-recipe results show the same qualitative asymmetry between original repetition and substitute adoption. These observational comparisons do not identify either component causally.

For prompted Qwen, original-answer repetition increases by 9.48 pp $[+6.45, +12.70]$ from 3B to 14B, while substitute adoption changes by 2.02 pp $[-1.81, +5.85]$. For SSRL Llama, the substitute-minus-original margin changes by -20.56 pp $[-26.41, -14.72]$ from 3B to 8B.

G Attention-Head Interventions

This appendix reports the head analysis summarized in Section 6. Component ablation and causal mediation test whether attention heads or neurons affect model behavior (Michel et al., 2019; Vig et al., 2020). Retrieval-head analyses identify sparse components that copy or recover information from long contexts (Wu et al., 2025), and ReDeEP contrasts internal signals for parametric and contextual knowledge in fixed-context RAG (Sun et al., 2025b). Our held-out interventions establish an effect on a constructed first-token substitution margin; they do not identify mediation of whole-policy D-survival.

G.1 Intervention Definition and Head Selection

For each eligible U episode, we retain the trajectory through the last retrieval block, append `<answer>`, keep the final 4,600 tokens, and define the next-token margin $m = \ell(t_{\text{sub}}) - \ell(t_{\text{gold}})$. Direct logit attribution maps this vocabulary direction through the fixed final root-mean-square normalization layer and assigns the resulting margin to each head's input block before the attention output projection. Items are split deterministically by the Secure Hash Algorithm 1 (SHA-1) hash of the question ID modulo 2: Search-R1 has 547 discovery and 461 evaluation items; IKEA has 303 and 260. Heads are ranked by mean signed contribution on discovery items, and the top 12 substitute-favoring heads are selected before evaluation on the disjoint half.

The intervention is a forward pre-hook on each selected layer's attention output projection: it zeros that head's pre-projection block at every token position and then runs a full uncached forward pass. Thirty controls independently sample the same number of non-selected heads within each selected layer, using seeds $20260923 + k$ for $k = 0, \ldots, 29$, and are evaluated on the same episodes. Paired bootstrap intervals use 10,000 draws and seed 1. Search-R1 and IKEA each derive their own head set, making IKEA an independent application of the selection procedure. Table 19 records the selected coordinates.

Table 19: Selected zero-indexed (layer, head) coordinates, ordered by discovery-set mean direct logit attribution.

Agent	Selected heads
Search-R1	(33,11), (31,7), (32,7), (33,3), (32,3), (34,14), (33,15), (33,9), (33,1), (33,2), (33,4), (34,12)
IKEA	(33,11), (32,3), (31,7), (32,7), (33,3), (33,15), (33,9), (34,14), (33,4), (33,2), (33,6), (33,1)

G.2 Held-Out Intervention Results

Table 20 reports the selected-head effects and their layer-matched controls.

Table 20: Held-out constructed-prefix intervention. Δm is the change relative to the unablated forward pass; the last column subtracts the mean layer-matched control effect.

Agent	n	Selected-head Δm	Difference from layer controls
Search-R1	461	-6.45 $[-7.26, -5.65]$	-4.81 $[-5.61, -4.01]$
IKEA	260	-5.55 $[-6.48, -4.66]$	-4.21 $[-5.19, -3.27]$

No layer-matched Search-R1 control reaches the selected-head effect. Depth nevertheless explains a nontrivial part of the raw effect, which is why the controlled differences carry the claim. Figure 4 shows the selected-head effects against the control sets.

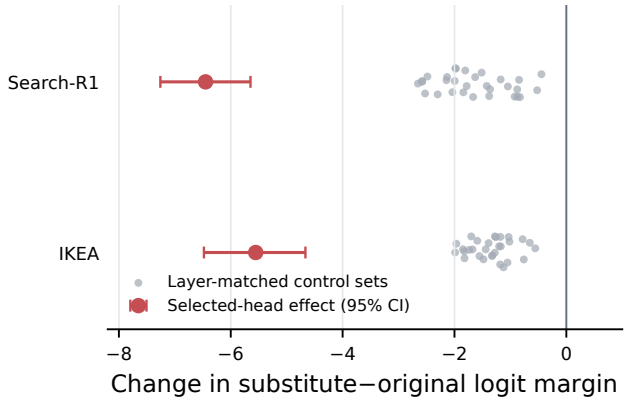


Figure 4: Held-out head ablation. Selected-head ablations shift the local U substitution margin more than each of 30 layer-matched control sets; the selected-minus-control effects are -4.81 for Search-R1 and -4.21 for IKEA (Table 20). The intervention does not localize whole-policy D-survival.

G.3 Scope and Control Analyses

Scaling selected heads during full generation produces no monotone change in substitute adoption and no useful repair. Readouts from the selected groups score area under the receiver operating characteristic (AUROC) values of 0.531 in Search-R1 and 0.599 in IKEA for behavioral D-survival. The causal interpretation is therefore limited to the defined first-token margin; the analysis does not localize full-policy D-survival, establish mediation of scaling, or provide a deployable predictor.

H Label-Free Baselines and Selective Prediction

H.1 Correspondence Between Self-Ablation and Gold Deletion

Table 21 compares self-ablation with two inexpensive trajectory signals on the same $n = 1120$ Search-R1 questions. The positive class is correctness that survives world-D deletion. Self-ablation is evaluated as a binary rule fixed before this comparison; no test-label threshold is selected.

Table 21: Prediction of D-surviving correctness on a shared panel. The interval is the paired question-bootstrap interval for F1; MCC denotes Matthews correlation coefficient.

Signal	Precision	Recall	F1	MCC
Self-ablation	.755	.780	.767 [.729, .803]	.678
String witness rule	.710	.217	.332 [.272, .393]	.297
Number of searches	.289	.319	.303 [.257, .349]	.025

Aggregate S- and D-survival rates differ by $+0.89$ points $[-1.25, +3.04]$, while the item-level labels agree on 87.1% of questions. These estimates describe correspondence; no equivalence margin was specified. Pooled over four released agents and 2842 questions, F1 is 0.929, with a per-agent range of 0.909–0.942. The independent Search-R1-14B analysis gives balanced accuracy 0.881 $[0.841, 0.919]$ and F1 0.858 $[0.805, 0.904]$ on its leak-free panel.

On 598 U-eligible cases, S-surviving answers repeat the original value in 36.4% of episodes, compared with 5.2% for S-changing answers, a $+31.15$ pp contrast $[+21.73, +41.25]$. Substitute adoption is 41.4% versus 67.1%, respectively.

H.2 Support Sufficiency and Deletion Survival

Table 2 in Section 7 reports a crossed-task comparison. The support task contains all answers and asks whether the displayed context suffices to answer; the deletion-survival task is restricted to ordinary-correct answers. The methods therefore specialize to different targets, but the differing cohorts preclude a same-cohort statistical double dissociation.

On the same $n = 946$ D-survival panel, support self-report, exact-answer sample consistency, semantic entropy, and mean sequence log-probability achieve direction-free discrimination $\max(\mathrm{AUROC}, 1 - \mathrm{AUROC})$ of 0.572, 0.554, 0.532, and 0.539, respectively. The strongest tested noninterventional commitment score reaches 0.757, compared with 0.842 for self-ablation. These values compare discrimination of the operational D label and are separate from ordinary-correctness prediction.

H.3 Selective-Prediction Operating Points

For a release set R, we report

$$\mathrm{risk}(R) = \frac{\sum_i \mathbf{1}\{i \in R\} Y_{O,i} Y_{D,i}}{\sum_i \mathbf{1}\{i \in R\}}, \qquad \mathrm{yield}_{\mathrm{fail}}(R) = \frac{1}{n} \sum_i \mathbf{1}\{i \in R\} Y_{O,i}(1 - Y_{D,i}).$$

Coverage and total correct yield are reported separately. Table 3 in Section 7 gives the principal Search-R1 operating point. For each continuous comparator, exact rank enumeration in both score directions selects the minimum-risk point subject to D-failing-correct yield at least that of the gate, with fixed question-ID tie breaking. The yield constraint controls for trivial risk reductions through near-total abstention.

The paired risk gap between GPT-4.1 sufficiency and self-ablation is $+4.36$ points $[+3.46, +5.34]$. Relative to releasing everything, however, the gate withholds 214 correct answers and loses 8.26 points of total correct yield. Rollout count provides a descriptive compute measure because an agent rollout, a sampled completion, and an API judge call differ in latency and floating-point operations (FLOPs).

This objective differs from general factuality. If an invalid release is defined as either wrong or D-surviving correct, answer-all has 71.44% joint invalidity, a free string-containment gate has 64.39%, and self-ablation has 67.22% on the same panel. The string rule has lower joint invalidity because it filters wrong answers, while self-ablation has lower D-surviving-correct risk. The two objectives should therefore be reported separately.

The same fixed gate lowers D-panel risk in every evaluated agent. The resulting risks are 2.44% for Search-R1, 1.64% for AutoRefine, 1.53% for IKEA, 1.71% for ReSeek, and 1.35% for ZeroSearch. Table 22 compares the gate's advantage over the closest tested noninterventional comparator on the

shared D/U-eligible panels. The advantage is smaller under independently constructed world U and its interval excludes zero for two of five agents. Twelve of 30 agent-by-dataset U-fronts contain a baseline point that dominates a gate point. We therefore treat the D-front as the primary operational result and U as a stricter external check.

Table 22: Risk advantage of self-ablation over the closest tested noninterventional comparator at matched yield. D is recomputed on each agent's U-eligible panel. Positive values favor self-ablation; intervals are paired-bootstrap 95% intervals.

Agent	n	D advantage	U advantage
Search-R1	1816	$+3.68$ $[+2.79, +4.73]$	$+1.32$ $[+0.68, +1.91]$
IKEA	1306	$+8.52$ $[+6.54, +10.06]$	$+0.93$ $[-0.36, +2.54]$
ReSeek	1294	$+5.83$ $[+4.48, +7.10]$	$+0.10$ $[-0.51, +0.77]$
AutoRefine	1025	$+7.31$ $[+5.77, +9.32]$	$+1.87$ $[+0.84, +3.03]$
ZeroSearch	1306	$+5.20$ $[+3.81, +6.47]$	$+0.35$ $[-0.34, +1.17]$

H.4 Replay Variability

Across the full 2,591-question panel, five nominally greedy ordinary replays disagree on at least one answer for 53.1% of questions; the corresponding S verdict disagrees for 15.4%. For predicting D-surviving correctness on this full panel, averaging S survival over five replays raises AUROC from 0.756 for one replay to 0.840 $[0.809, 0.869]$. Relative to the published one-rollout gate on the same items, the paired difference is $+0.032$ $[-0.004, +0.068]$, so the five-rollout score does not clearly improve the instrument.

Conditioning instead on the 951 ordinary-correct questions, we compare S with a same-compute O/O placebo. Across the five replicates, ordinary-replay survival attains balanced accuracy 0.541–0.555 for the gold D label, whereas the matched S replay attains 0.759–0.779. Every paired difference lies between 0.206 and 0.236, with all five question-bootstrap intervals excluding zero. Averaging survival over five replays gives AUROC 0.560 for O/O and 0.848 for S, a paired difference of 0.289 $[0.250, 0.328]$. This post-hoc correct-only analysis covers 84.9% of the primary detector cohort and shows that the S signal is not reducible to ordinary replay instability.

I Interventions to Improve Evidence Dependence

I.1 Evaluation Criteria and Scope

We define successful repair as conversion of correct D-surviving cases into correct D-failing cases while preserving ordinary accuracy and answer coverage. A repaired policy should also follow a substituted witness in U. Deletion fragility, abstentions, malformed outputs, and ordinary-world errors do not satisfy this criterion. Selective release is evaluated separately because it deliberately withholds risky answers.

Two narrow classes have direct limits. A transformation that preserves the accepted answer class separately in O and D cannot change an output-defined O/D label. A selector over a fixed logged candidate pool also cannot exceed the number of questions for which that pool already contains a positively labeled candidate; the observed oracle ceiling is 2 of 28 cases (7.1%). New generation, changed retrieval, and training remain outside these bounds.

Table 23 summarizes the evaluated protocols. Reported endpoints use the counterfactual environments run for each method, keeping S-based training outcomes distinct from gold-D outcomes.

I.2 Answer-Conditioned Verification Search

The continuation method appends a verification search when the current answer lacks a matched witness. Its query contains the question and the answer being verified. Table 24 reports within-configuration comparisons. The conversions at $K = 2$ and $K = 3$ occur on the same question, leaving one unique conversion across the table.

Table 23: Configurations of completed repair experiments. "Eval." lists the worlds used for the reported held-out outcome.

Method	Protocol	Eval.
Answer-seeded continuation	Native agent continues after querying with the question plus its own candidate answer; temperature zero; up to $K = 1, 2, 3$ extra searches; first and final answers are compared within the same rollout.	O/D, six settings
QUOTE/REGEN	Fixed 400-question panel; 67 S-surviving cases selected by the self-ablation router; recorded pre-answer queries replayed in both worlds. REGEN adds a question-only relation query that cannot read the prior answer. The Search-R1-3B reader must emit a supporting sentence and short answer or abstain.	O/D, $n = 400$
Context-aware decoding	Search-R1-3B greedy answer plus deduplicated sampled candidates, ranked by $(1 + \alpha) \log p(a \mid c) - \alpha \log p(a \mid \varnothing)$ for $\alpha \in \{0, .25, .5, 1, 1.5, 2, 3, 5\}$.	O only, $n = 518$
Fixed evidence-only reader	GPT-4.1 at temperature zero; the same 67 S-surviving cases; answer-filtered pre-answer queries retrieved at depth 20 and eight passages displayed after O/D/U editing. The old answer and world labels are hidden; an identical empty-context prompt is the control.	O/D/U, 350 primary calls
Hard gate / shaped reward	Search-R1-3B low-rank adaptation (LoRA; rank 32, $\alpha = 64$), AdamW, learning rate 3×10^{-5}, 40 steps, six questions and six rollouts per step on 12,000 NQ/HotpotQA training questions.	O/S, $n = 560$
High-exposure penalty	Same model family; 100 steps on 677 no-search-answerable NQ/HotpotQA questions, with six questions and six rollouts per step.	O/D $n = 505$; U $n = 393$

Table 24: Witness-conditioned continuation. The last column counts correct D-surviving to correct D-failing transitions.

Agent and depth	Trigger rate	Change in D-failing correctness	Conversions
Search-R1-3B, $K = 1$	12.7%	$-.06$ $[-.18, .00]$	$0/184$
Search-R1-3B, $K = 2$	12.8%	$.00$ $[-.18, +.18]$	$1/183$
Search-R1-3B, $K = 3$	12.8%	$+.12$ $[-.12, +.35]$	$1/184$
ZeroSearch-3B, $K = 1$	19.0%	$+.06$ $[-.18, +.35]$	$0/164$
IKEA-3B, $K = 1$	19.4%	$.00$ $[-.24, +.24]$	$0/239$
Search-R1-7B, $K = 1$	14.9%	$+.32$ $[-.32, +.96]$	$0/78$

The implementation helps explain the result. Among 2056 triggered continuations whose answer stays unchanged, the answer-conditioned query retrieves apparent support in 32.54% of cases. Those cases have no detected correctness advantage, -3.61 points $[-8.27, +1.10]$. The query therefore often retrieves evidence selected by the candidate it is intended to test.

I.3 Evidence-Conditioned Regeneration

Evidence-conditioned methods reduce D-surviving output primarily by withholding it. On the fixed 400-question panel in Table 25, D-failing correctness remains essentially unchanged while ordinary accuracy falls and 32–34 questions are abstained.

Table 25: Evidence-conditioned regeneration. All percentages use the same 400-question denominator.

Pipeline	Accuracy O	D-failing	D-surviving	Abstentions
No repair	29.75%	24.00%	5.75%	0
QUOTE	26.50%	24.25%	2.25%	32
REGEN	25.50%	23.75%	1.75%	34

Answer-agnostic context-aware decoding (Shi et al., 2024) loses 9.3 accuracy points at $\alpha = 2$. Because its selected outputs were not rerun under D, this experiment measures ordinary accuracy and retention rather than conversion. Replacing routed outputs with a fixed GPT-4.1 evidence-only reader is a stronger positive control: over witness-empty input, the same reader gains 54.29 points $[40.00, 68.57]$ when shown the retrieved witness. Yet its full-panel D-failing correctness changes by 0.00 points $[-1.50, +1.25]$, its ordinary answer coverage is 62.69%, and only 1 of the 11 routed baseline D-surviving cases whose ordinary evidence lacks the designated substitute has the joint O-correct, D-failing, U-substitute signature. This result applies to the tested router, reader, retrieval procedure, and operating point; evidence-conditioned regeneration as a broader class remains open.

I.4 Counterfactual Training Objectives

Table 26 summarizes the evaluated dependence-shaped training objectives. The hard gate and shaped bonus were evaluated against self-ablation S rather than gold deletion D and leave S-failing correctness, ordinary-correct answers that become incorrect after self-ablation, unresolved around zero. Raising treatment exposure makes the penalty effective at suppressing D-surviving correctness, but does not produce a positive D-failing contrast and lowers ordinary accuracy.

Let c denote ordinary correctness and ℓ denote acceptance under the online self-ablation check. The hard gate uses $r = \mathbf{1}[c \wedge \ell]$; the shaped arm uses $r = \mathbf{1}[c](1 + \mathbf{1}[\ell])$. In the high-exposure study, the control uses $r = \mathbf{1}[c]$ and the penalty arm uses $r = \mathbf{1}[c \wedge \ell] - \mathbf{1}[c \wedge \neg\ell]$. Restricting training questions to those answerable without search raises penalty exposure from 7.5% to 35–37%, providing a high-exposure evaluation of the objective.

Table 26: Held-out training contrasts in percentage points. "World" names the counterfactual used to define failing and surviving correctness. Each is a single-seed pilot; question-bootstrap intervals do not measure seed variation.

Intervention	World	Accuracy O	Failing	Surviving
Hard counterfactual gate	S	—	$+1.25$ $[-2.14, +4.82]$	—
Shaped counterfactual bonus	S	—	$+.18$ $[-3.39, +3.93]$	—
High-exposure penalty vs control	D	-3.56 $[-7.13, .00]$	$-.40$ $[-4.16, +3.56]$	-3.17 $[-5.35, -.99]$
High-exposure penalty vs base	D	-1.78 $[-4.95, +1.39]$	$-.79$ $[-4.16, +2.57]$	$-.99$ $[-3.37, +1.39]$

Across two identically configured runs of counterfactual-paired group relative policy optimization (CP-GRPO), the estimated effect changes sign, so the data do not support a single-seed effect claim. Across completed tests, the observed outcomes are suppression, abstention, or no detected conversion. Training objectives that reward recovery of newly retrieved evidence remain an open direction.

J Reproducibility, Ethics, and AI Assistance

J.1 Artifacts and Reproduction

Reported results are backed by row-level or derived artifacts and analysis scripts. Runs with complete provenance include configuration records and SHA-256 manifests; other released-checkpoint evaluations retain the available launch metadata. Large raw episodes are excluded from version control, while the submitted artifact contains the derived files needed to reproduce the reported tables and figures. Running `powershell -File gen/build_rewrite.ps1` compiles the manuscript and checks reported values, cohort sizes, references, and page count. Figures 1 and 3 are drawn by `figs/make_figures.py` from values reported in Tables 7, 10, 12, 13, and 18 and Appendix F, and Figure 2 is defined in the LaTeX source.

For runs with complete provenance we record the model repository and revision, prompt contract, runner hash, world-transform hash, input manifest, dataset quota, decoding parameters, and software and hardware metadata. Before estimating each contrast, we verify ordered question IDs and the configuration fields available for its constituent artifacts; any known mismatch is disclosed with the result. Nominally greedy decoding is not bitwise stable across batch shapes, motivating these checks. Confidence intervals use paired question resampling and do not represent training-seed or execution-level uncertainty unless explicitly stated.

The primary evaluation uses seven open-domain QA datasets and the fixed wiki-18 corpus described in Appendix C. We will release the manuscript generator, numerical verifier, artifact manifests, analysis scripts, prompts, and all distributable row-level outputs. Model weights and datasets remain under their original licenses. Where recorded, scripts identify the external repositories and pinned revisions required to recover model weights; other arms are distributed with their preserved identifiers and metadata.

J.2 Ethical Considerations

The measurement can help operators detect answers that appear supported but survive removal of their apparent witnesses. Its direct use is selective release, which may reduce misleadingly grounded answers at the cost of lower coverage. That cost can be distributed unevenly when retrieval quality differs by language, topic, or population, so deployments should report per-domain coverage and error rates rather than a pooled operating point alone.

The intervention is also dual-use. It edits retrieval results and could be repurposed to remove unwanted evidence, probe corpus membership, or develop retrieval-poisoning strategies. We mitigate this risk by evaluating on public QA data, reporting aggregate statistics rather than sensitive queries, preserving the unedited answer and corpus for provenance, and treating the deletion output as a diagnostic rather than a factual source. No human participants or private user data are used.

J.3 Use of Generative AI

Generative AI assisted with code drafting, literature discovery, manuscript editing, and critical review. The authors verified retained claims against local artifacts and primary bibliographic records and remain responsible for the study and submitted material.