RAG CONTROL IS NOT ANSWER FIDELITY: AN AUDITABLE PROTOCOL FOR THE SUFFICIENCY–FIDELITY GAP IN MULTI-HOP RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Evidence accumulation and stopping judgments need not track answer correctness. We audit this mismatch with PR-RAG, a training-free bounded protocol exposing dependency state, verification, typed failures, and per-call cost. Five-seed paired experiments on HotpotQA distractor motivate a coverage–fidelity control plane. At seed 42, 208/500 answers are EM-incorrect despite a positive sufficiency judgment; 153 of these have full gold-chain coverage. Across five seeds, 34.32% of PR-RAG outcomes have complete coverage and incorrect EM, versus 22.32% for NaiveRAG. None of six reported EM contrasts survives Holm correction. A blind two-annotator audit finds a 43.2% false-negative rate for the gold-chain proxy. These observations motivate separate evaluation of coverage and answer support: coverage and EM are measured here, while human-rated fidelity remains a distinct target. An offline calibrator and second-model replication probe the audit interface. A fidelity-aware gate, open-corpus validation, matched-budget comparisons, and an expanded human fidelity audit remain outstanding.

1 INTRODUCTION

Retrieval-augmented generation (RAG) grounds a language model in external passages and is widely used for knowledge-intensive tasks (Lewis et al., 2020; Gao et al., 2023). Most pipelines issue one query and then generate. This design is fragile for multi-hop questions: a raw query can name the first entity while the second supporting paragraph is indexed under an intermediate answer.

We measure this premise before proposing a controller. On the official HotpotQA distractor validation split (Yang et al., 2018), each question supplies ten candidate paragraphs, including two gold supports. With BM25, a single top-3 query retrieves both supports for only 52.9% of bridge questions and 52.7% of comparison questions. A dense retriever improves comparison coverage to 95.8% but reaches only 62.4% on bridge questions. Thus semantic matching alone does not remove the dependency-conditioned retrieval gap.

PR-RAG's deployed loop is training-free and has four explicit interfaces. Planning decomposes the question into dependent sub-questions; retrieval instantiates placeholders with intermediate answers; verification decides whether the accumulated evidence is sufficient to answer; and typed reflection attributes unresolved items to retrieval, dependency, reading, or verifier failures and selects a recovery action. The deployed controller is a training-free heuristic loop; the learned verifier is an offline *calibrator* (a re-thresholded sufficiency score), not a learned controller, and no online RL policy is trained. Main-table results, the coverage premise, mechanism ablations, and calibrator evaluation are *offline* measurements on logged trajectories; the learned-gate replay is the only *online* (live-decoding) intervention and is reported as single-seed exploratory evidence, not a learned-control claim. The paper's object of study is therefore the control interface itself: what each primitive changes relative to a shared baseline, what the sufficiency signal does and does not carry, and where the audit instrument's own blind spots lie. The audit yields a sharper conclusion than a system comparison would: the evaluated primitives move evidence coverage reliably, yet 208/500 stopping decisions had positive sufficiency judgments and EM-incorrect answers, and no reported EM contrast survives correction.

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.

Our contributions are findings first; the protocol exists to make them measurable: We study an audit interface separating evidence coverage, stopping judgments, answer correctness, and computation. We do not claim SOTA EM, learned control, matched-budget efficiency, open-corpus generalization, or cross-dataset generalization.

- a five-seed diagnosis of coverage and answer correctness: complete-chain, EM-incorrect outcomes persist in 34.32% of PR-RAG outcomes, and 0/6 reported EM contrasts survive Holm correction; these observations motivate, but do not validate, separate fidelity control;
- an auditable state-action protocol that makes intermediate evidence, dependency information, stopping decisions, typed failures, recovery actions, and call costs first-class audit outputs, released with per-question traces as a reusable measurement instrument and a scaffold for future learned controllers;
- a reproducible evaluation protocol with fixed questions, shared retrievers, five decoding seeds, question-level paired bootstrap, Holm correction, a complete second-model replication, and a two-annotator proxy audit that quantifies sample-conditional disagreement with human sufficiency.

2 Related Work

RAG combines parametric generation with sparse or dense retrieval (Lewis et al., 2020; Guu et al., 2020; Karpukhin et al., 2020; Robertson & Zaragoza, 2009); iterative reasoning/search, decomposed prompting, critique, complexity routing, passage grading, query rewriting, and confidence-triggered retrieval extend this pattern (Trivedi et al., 2023; Shao et al., 2023; Yao et al., 2023; Press et al., 2023; Khot et al., 2023; Asai et al., 2024; Jeong et al., 2024; Yan et al., 2024; Ma et al., 2023; Jiang et al., 2023).

Multi-hop reranking and graph retrieval also evaluate downstream QA: PromptRank ranks document paths by question likelihood (Khalifa et al., 2022), while SPRIG uses seeded graph propagation (Wang, 2025). Our focus is joint coverage–stopping–correctness–cost diagnosis, rather than the claim that prior work measures coverage alone.

Iterative retrieval, learned search policies, sufficiency judges, and repair systems address complementary control decisions (Trivedi et al., 2023; Shao et al., 2023; Jin et al., 2025; Li et al., 2025; Guan et al., 2025; Wu et al., 2026; Zhu et al., 2025; Jiang et al., 2025; Shi et al., 2026; Qiu et al., 2026; Li et al., 2026a; Jiao et al., 2026; Asl et al., 2025). Confidence and telemetry also inform retrieval (Jiang et al., 2023; Yao et al., 2025; Vinod et al., 2026). Our deployed loop is training-free, and the learned component is an offline calibrator. The empirical contribution concerns jointly observable state, outcomes, and costs, rather than novelty of individual primitives.

PlanRAG, Plan*RAG, and PyRAG expose plans, dependencies, or executable programs (Lee et al., 2024; Verma et al., 2024; Sun et al., 2026); PRISM and PAR^2-RAG structure evidence selection and accumulation (Nahid & Rafiei, 2025; Li et al., 2026b); Stop-RAG learns stopping decisions (Park et al., 2025); and CIRAG and BridgeRAG address evidence integration and later-hop retrieval (Wei et al., 2026; Bacellar, 2026). These are relevant alternatives, not systems proven inferior here. Our comparisons are limited to the shared-model implementations documented in Appendix I; the scientific question is how coverage, stopping judgments, correctness, and cost relate on paired trajectories.

3 Auditable Controller

3.1 State and plan

Let x be a question, $\rho(q, k)$ a retriever, and $\mathcal{G}$ an instruction-tuned LLM. The controller maintains cumulative evidence E_t, intermediate answers A_t, an unresolved set M_t, and iteration index $t \leq T$. A one-time planner emits sub-questions $P = \{p_i\}_{i=1}^{m}$ and dependencies $\mathrm{dep}(i) \subseteq \{1, \ldots, i-1\}$. Placeholders are instantiated as

$$q_t^{(i)} = \phi\Big(p_i, \{a_t^{(j)} : j \in \mathrm{dep}(i)\}\Big), \tag{1}$$

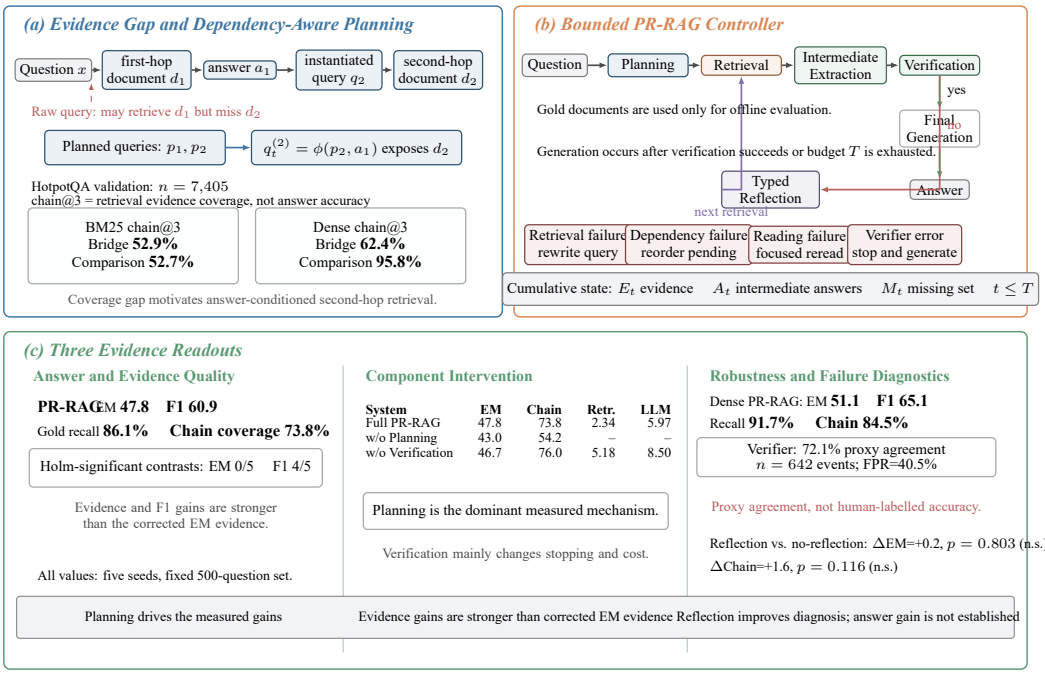


Figure 1: PR-RAG overview: dependency-conditioned planning, bounded retrieval and verification, typed reflection, and evidence/cost diagnostics. Numbers come from real-model logs; verifier agreement uses an offline gold-chain proxy.

with the original question as fallback when a dependency is unresolved. Retrieval updates $E_t \leftarrow E_{t-1} \cup \rho(q_t^{(i)}, k)$ and an extractor updates A_t from the shared pool.

3.2 Verification and Reflection

The verifier emits $s_t = V(x, E_t, A_t) \in \{0, 1\}$ and the unresolved set $M_t = \{i : a_t^{(i)} \text{ unknown}\}$. In our implementation, V produces three structured labels: evidence-chain complete, intermediate answers available, and answerable. Generation starts at $\tau = \min(\{t : s_t = 1\} \cup \{T\})$:

$$\hat{y} = \mathcal{G}(x, E_\tau). \tag{2}$$

Gold documents are never provided to the controller; they are used only offline to score retrieval and verifier decisions. If $s_t = 0$, reflection classifies each unresolved item without gold labels. Retrieval failures trigger query rewriting, dependency failures reorder pending items, reading failures trigger focused rereading, and verifier errors stop and generate. The loop is bounded by T; there is no abstention.

3.3 Audit Trace

Every action is logged with question id, seed, iteration, query, retrieved titles, intermediate answer, verifier labels, failure type, next action, retrieval calls, LLM calls, latency, and token count. This trace defines reproducible interfaces for module ablations and offline diagnostics. It also prevents a common ambiguity: evidence-chain coverage is a retrieval measure, not answer accuracy.

3.4 Learned Sufficiency Calibration

Let z_t contain only inference-time observable trace features: question type, iteration, plan size, dependency indicator, missing-item count, calls and unique titles so far, retrieval-score summaries, extraction-success counts, and the prompt verifier's three labels. A logistic calibrator learns

$$\Pr(s_t^\star = 1 \mid z_t) = \sigma(w^\top z_t), \tag{3}$$

where $s_t^\star$ is offline gold-chain completeness. Gold labels are targets and never features. The threshold is selected within each training fold for balanced accuracy. This tests whether audit state contains a learnable stopping signal; logged trajectories cannot reveal counterfactual evidence after a changed stop, so we do not claim end-to-end gains without replay.

For completeness, the released harness also implements a question-disjoint online replay mode (‘prrag_lver‘): at each verification event it loads only the calibrator fold that excludes the current question, replaces the prompt gate with the learned decision, and continues the loop until stopping or T. We treat this as an evaluation protocol rather than a new training objective; its single-seed end-to-end results are reported as exploratory evidence in the mechanism analysis.

4 Experimental Design

4.1 Data, systems, and serving

We use the 7,405-question HotpotQA distractor validation split, with ten supplied paragraphs per question and no open-corpus indexing. The evaluation set is fixed and stratified (400 bridge, 100 comparison). Six main systems share the same metric implementation and BM25 interface: NaiveRAG, AdvancedRAG, Iter-RetGen-style, Self-Ask-style, CRAG-style, and PR-RAG with $T = 3$; PR-RAG, NaiveRAG, and AdvancedRAG are also evaluated with BAAI/bge-small-en-v1.5 (Xiao et al., 2023). Baselines follow published descriptions and are identified as reference-inspired implementations, not exact reproductions. The main-table rows are therefore a *method-performance* comparison under a common protocol (shared candidate pool, retriever, generator, and decoding), not an *implementation-protocol* (official-reproduction) comparison; implementation status and full provenance are in Table 1 and Appendix I.

All systems use Qwen2.5-7B-Instruct (Yang et al., 2024), Q4_K_M quantization, local llama.cpp serving, temperature 0.7, top-p 0.8, and a 6144-token context. Seeds 42–46 vary only decoding. Main systems and the dense variants use five seeds; most mechanism variants use seed 42.

4.2 Metrics and inference

We report normalized EM and token-level F1, supporting-document recall, full-chain coverage, retrieval calls, LLM calls, latency, and token counts. Point estimates average seeds within each question, then questions. Paired bootstrap confidence intervals use 10,000 question-level resamples; Holm correction is applied within metric families. The reported Holm family contains six comparisons (five main baselines and the single-seed k=6 control). The dated expansion protocol follows the original data collection and therefore does not establish prospective registration of these 500-question analyses. We retain the reported family and treat these comparisons as retrospective; planned expansion has a separate protocol. A power analysis only sets the target size of the planned expansion; it is design basis, not an experimental result. Verifier agreement is against an offline gold-chain proxy, not human-labelled accuracy; a two-annotator audit of the proxy is reported in Section 6.6. For learned verification, all 3173 events from the same question across five decoding seeds remain in one of five folds, and every reported prediction is strictly out of fold; thresholds are selected on training folds only for balanced accuracy. No budget-matched composite controller was run, so cost comparisons identify the tested protocol and its own ablations, not a causal advantage of verification over an equally expensive alternative.

5 Results

Coverage–fidelity control plane. We separate evidence coverage from answer support as evaluation targets. Our measured projection uses full gold-title chain coverage and normalized EM; neither certifies human-rated fidelity. Across five seeds, complete-chain/EM-incorrect outcomes constitute 34.32% for PR-RAG versus 22.32% for NaiveRAG, while incomplete-chain/EM-incorrect outcomes fall from 34.36% to 17.84%. Appendix G reports the four quadrants. These descriptive associations motivate a two-axis controller whose efficacy requires an intervention.

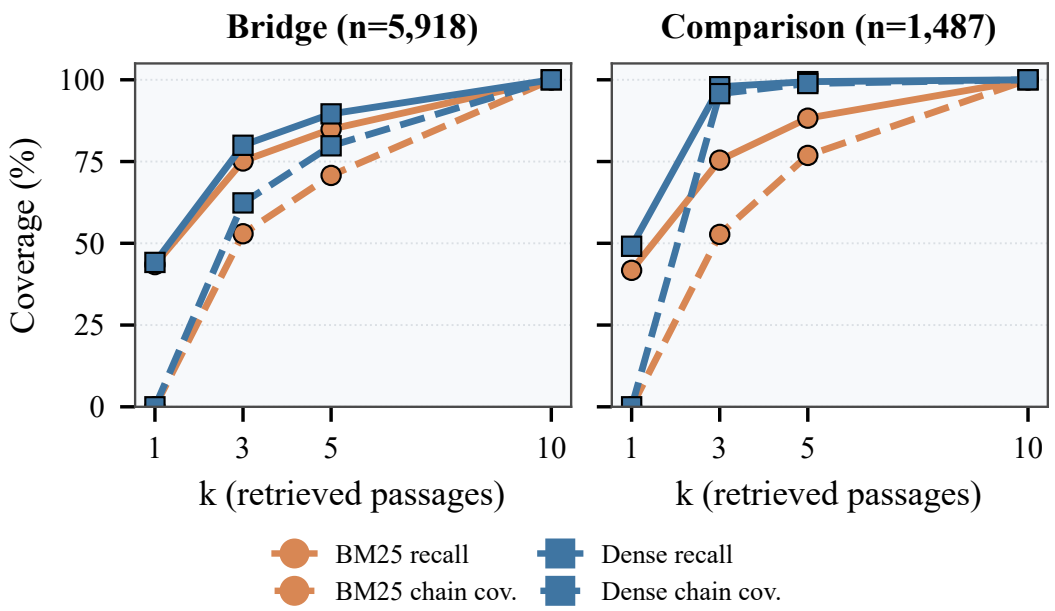


Figure 2: Real HotpotQA retrieval coverage. Solid lines show gold recall and dashed lines full-chain coverage; this is evidence coverage, not answer accuracy.

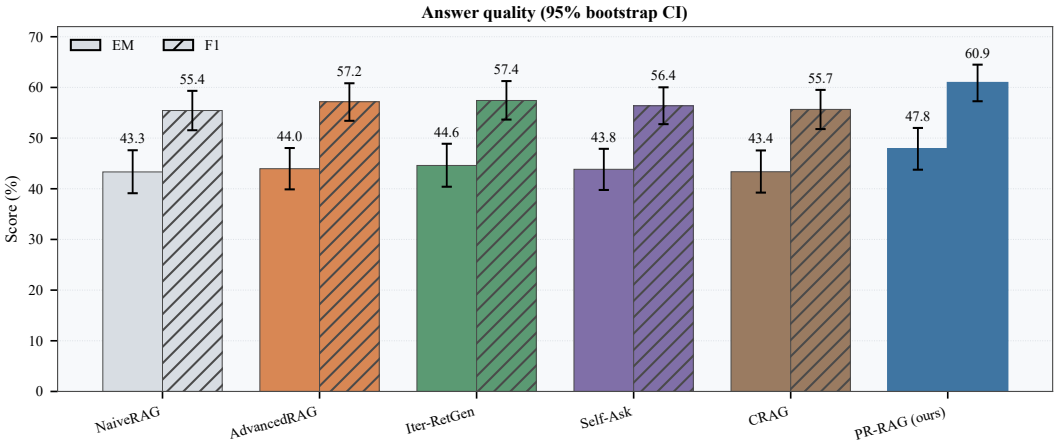


Figure 3: EM and F1 for the six main systems; five decoding seeds are aggregated per question before bootstrap intervals.

5.1 Coverage premise

Across all 7,405 questions, the BM25 bridge/comparison chain rates are 52.9%/52.7%. Dense retrieval yields 62.4%/95.8%. The bridge gap persists because the second support is often indexed under an intermediate answer.

5.2 Benchmark diagnostics

The central diagnostic is negative: across the six reported contrasts (five main baselines plus the single-seed k=6 retrieval-budget control), 4/6 raw EM comparisons are significant but 0/6 survive Holm correction; the raw-significant contrasts narrow-miss the threshold rather than failing decisively, but we treat EM as unestablished. In contrast, 4/6 F1 contrasts survive correction (against NaiveRAG, AdvancedRAG, Self-Ask-style, and CRAG-style; not against Iter-RetGen-style or the

Table 1: Main comparison on the fixed 500-question set. Rec. is gold-document recall and Chain is full-chain coverage; brackets give 95% question-level bootstrap CIs (seeds aggregated per question); calls are means per question and Seeds is the number of decoding seeds. The Status column marks implementation provenance (F = faithful reimplementation, R = reference-inspired, † = non-official alternative); see Appendix I.

System	Status	Seeds	EM	F1	Rec.	Chain	Retr.	LLM	Lat.
NaiveRAG	F	5	43.3 [39,48]	55.4 [52,59]	75.9 [73,78]	54.2 [50,59]	1.00	1.00	1
AdvancedRAG	R	5	44.0 [40,48]	57.2 [53,61]	84.4 [82,86]	69.8 [66,73]	2.00	2.00	4
Iter-RetGen-style	R	5	44.6 [40,49]	57.4 [54,61]	82.4 [80,85]	66.8 [63,71]	2.00	2.00	1
Self-Ask-style	R	5	43.8 [40,48]	56.4 [53,60]	83.6 [82,86]	67.9 [64,72]	2.21	3.21	6
CRAG-style	R†	5	43.4 [39,48]	55.7 [52,59]	77.7 [75,80]	57.4 [53,62]	1.23	1.23	3
PR-RAG (ours)	ours	5	47.8 [44,52]	**60.9 [57,64]**	**86.1 [84,88]**	**73.8 [70,77]**	2.34	5.97	12

Impl.: F = faithful reimplementation, R = reference-inspired (not an official reproduction), † = non-official alternative to an official release that could not be run in this environment (see Appendix). Seeds = number of decoding seeds aggregated per question; brackets give 95% question-level bootstrap CIs. Bold marks only the columns whose Holm-corrected paired contrasts support the difference (F1, gold-document recall, chain coverage); no EM contrast survives Holm correction (significance audit, Appendix D.2), so the EM and cost columns are not bolded.

k=6 control; Table 7). Read against the 208/500 gate-positive, EM-incorrect stops (Section 6.4), this pattern is consistent with a sufficiency signal only weakly coupled to answer correctness: more control over evidence does not convert into established answer-level gains. The table's remaining rows calibrate the instrument rather than rank systems: PR-RAG reaches 47.8 EM and 60.9 F1 with 86.1% recall and 73.8% chain coverage, the highest point estimates at $k = 3$, with EM differences versus the five baselines ranging from +3.2 to +4.5points, so the primary claims rest on evidence coverage, F1, mechanism effects, and the cost-quality trade-off. Table 7 and Figure 4 report the full paired-bootstrap contrasts: mean differences, 95% CIs, raw and Holm-adjusted p, Cohen's d, and per-question win/tie/loss counts.

Stopping-controller context. A limited, single-seed, reference-inspired S2G-RAG-style contrast is reported in Appendix E. Its prompted untrained judge is not the official LoRA model, and its paired differences are not statistically established; it is therefore excluded from the reported main comparison and contributes no superiority claim.

5.3 Dense retrieval and budget

The headline table uses BM25 because all six main systems share that retrieval interface; dense variants are reported separately as robustness checks. Replacing BM25 with the dense retriever gives PR-RAG 51.1 EM, 65.1 F1, 91.7% recall, and 84.5% chain coverage — the strongest evaluated point estimate — but the dense-vs-BM25 EM contrast is not Holm-significant, while recall and chain coverage improve.

The $T = 1$ PR-RAG control uses 1.95 retrieval calls and has 47.4 EM, so it does not isolate control flow from a two-query rewrite at this scale. Removing planning reduces chain coverage to 54.2% and EM to 43.0(single-seed exploratory; Table 9), suggesting planning as the dominant measured mechanism. The single-seed NaiveRAG $k = 6$ control reaches 79.0% chain coverage but only 44.8 EM, exploratory evidence that buying coverage does not guarantee answer quality. Explicit-placeholder traces are associated with higher coverage but lower EM: across five seeds, explicit-dependency traces reach 79.4% chain coverage versus 75.1% for fallback, while the EM association runs the other way (46.8 vs. 49.9%; Appendix E). We therefore use "dependency-aware" for the control interface, not as an established answer-quality gain from explicit dependencies.

5.4 Cross-model five-seed replication

All six systems were rerun with Llama-3.1-8B-Instruct under the identical fixed-question protocol for seeds 42–46; all 30 system–seed cells contain 500 valid questions, with retry records deduplicated by question id and no imputation. PR-RAG reaches 47.6 EM and 61.7 F1, exceeding the strongest baseline by +2.8 EM and +3.5 F1. Contrast directions agree with the Qwen run across EM, F1, recall, and chain coverage (5/5, 5/5, 5/5, 5/5). This is a complete five-seed second-model replication on the same HotpotQA subset, not a claim of generalization beyond the two evaluated models;

seed-wise SDs, question-level CIs and Holm tests, costs, and audit provenance are in Appendix A and the released CSV.

The cross-model results use different multiplicity families: six Qwen contrasts versus five Llama contrasts. A post-hoc sensitivity excluding Qwen's single-seed k=6 control yields 4/5 Holm-significant EM contrasts (all except Iter-RetGen-style) and 5/5 F1 contrasts (Appendix D.1). This recalculation uses unchanged raw p-values and does not replace the retained six-contrast results. Neither family was established as prospectively registered for these data; the sensitivity is non-confirmatory and does not resolve model or compute confounding.

5.5 Second dataset: 2WikiMultiHoQA

We treat this as a single-seed directional prefix, not completed replication or generalization. We also port the protocol to 2WikiMultiHoQA (Ho et al., 2020). The completed prefix and full table are in Appendix B; because it is incomplete, we treat the result as an exploratory directional observation.

6 Mechanism and Failure Analysis

6.1 Verification as a cost controller

Without verification, PR-RAG uses 5.18 retrieval and 8.50 LLM calls per question, versus 2.34 and 5.97 with verification. The EM difference is +1.1 points (Holm $p = 0.5370$, not significant), while chain coverage changes by -2.2 points (Holm $p < 0.0001$). This identifies stopping and cost changes relative to the no-verification ablation. PR-RAG remains the most expensive main-table system in LLM calls (5.97 versus 1.00–3.21). Whether reinvesting verification costs in retrieval or generation yields better quality requires an unrun matched-budget comparison; trace-level diagnosis does not establish absolute efficiency.

6.2 Offline learned-verifier probe

Table 6 (Appendix D.2) shows a real but asymmetric trade-off: the calibrator learns a more conservative stop boundary rather than uniformly dominating the prompt verifier. It lowers proxy-FPR from 44.0% to 30.8% and raises agreement by +1.3 points, indicating that observable audit state contains a learnable sufficiency signal. Its F1 is 78.3% versus 78.5%, while its FNR increases from 21.7% to 26.2%.

6.3 End-to-end learned-gate replay

The single-seed replay provides integration evidence, not an established quality or efficiency improvement. The completed replay replaces the prompt gate with the question-disjoint out-of-fold calibrator at every verification event, retaining the same retriever, extractor, reflection handlers, and generation budget. On the fixed 500-question set at seed 42, EM is 48.200% (+0.400 points versus the paired prompt-gate run, $p = 0.370363$) and F1 is 60.610% (-0.080 points, $p = 0.462354$). Replay p-values are paired single-sample bootstrap over questions, the same resampling unit as the main protocol; the replay is a separate exploratory protocol, not a re-run of the reported main contrasts. Retrieval calls increase by +0.086 to 2.480 and LLM calls by +0.272 to 6.332, while latency changes by -0.521 s ($p < 0.001$); the verifier LLM call still executes in both arms because its label is a gate feature, so latency reflects stopping time and machine load rather than eliminated calls. The replay demonstrates integration and a mixed cost trade-off, but not a statistically reliable quality improvement or absolute efficiency gain. Gate decisions over 681 replay events have 71.659% proxy agreement, 27.160% FPR, and 28.995% FNR; these remain proxy measurements rather than human-labelled accuracy.

6.4 What the trace can and cannot identify

The logged trajectory records observed retrieval and decisions, but cannot reveal the answer under an unexecuted intervention. If required gold support is absent from the candidate pool, a controller restricted to that pool cannot retrieve it; if support remains available, its presence does not establish

that a changed policy would retrieve or use it correctly. We therefore make no oracle-EM or regret claim. The 208/500gate-positive, EM-incorrect stops motivate separate evaluation of stopping judgments and answer support; they do not establish a second gate as necessary or effective.

We distinguish four objects: the prompt-based heuristic controller; the question-grouped offline calibrator, which only re-thresholds sufficiency; an offline policy estimated from logs (none is trained); and an online RL policy (none is present). The logs contain no training reward or counterfactual outcome, so we make no learned-control, RL, regret, or convergence claim.

Selective prediction and error propagation. Logged verifier decisions support a selective boundary: a positive gate covers 84.2% of questions at 50.59% EM versus 47.8% when answering all, while the 79 abstained questions reach 32.91% EM. Conditional traces show 53.6% EM when the first verification has full gold-chain coverage versus 33.1% otherwise, with only 4.9% recovery from first-hop-missing cases. Retrieval failures dominate the 277 reflection events (66.4%), followed by dependency (18.8%) and reading (3.6%). These seed-42 logged associations are not live abstention or causal recovery estimates; complete tables remain in Appendix E.

6.5 Why reflection is currently weak

Reflection is retained because its trace isolates a downstream bottleneck: retrieval-failure rewriting can recover evidence, yet answer-level recovery remains weak and the net EM change is not statistically established. The failure distribution also points to concrete, log-supported next actions: retrieval-failure events dominate (184/277) and the dominant residual is a missing second hop, for which bridge-conditioned rescoring of the *existing* candidate pool (Bacellar, 2026) would cost no additional retrieval calls; dependency-failure events (52/277) motivate entity-swap rewriting; and reading failures plus the 208 gate-positive, EM-incorrect stops motivate a separate answer-fidelity gate. All three are unevaluated; the released per-question traces make each one targetable without re-running the controller. Detailed event counts, recovery rates, typed variants, and distractor diagnostics are moved to Appendix E.

6.6 Human audit of the gold-chain proxy

Because every verifier and calibrator metric in this paper is measured against the gold-chain proxy, the proxy itself is audited with two blind annotators on a fixed stratified sample of 60 seed-42 final stop decisions (sampling rules and seed frozen; annotators saw only the question, retrieved evidence, intermediate and final answers, never the proxy label). Inter-annotator agreement is 98.3% raw (Cohen's $\kappa = 0.97$, Krippendorff's $\alpha = 0.97$), although agreement does not establish label validity or rule out shared annotator bias. Table 5 (Appendix D.2) shows the proxy's dominant error is false negative: under the lenient reference its unweighted audit-sample FNR is 43.2% (not a population estimate), and 55.2% of proxy-negative events contain alternate retrieved evidence that annotators judged sufficient. Two consequences follow. First, proxy false negatives can label justified stops as errors; the direction and size of bias in verifier FPR require the joint verifier–proxy–human distribution and cannot be inferred from the proxy FNR alone. Second, the learned calibrator's FPR reduction (Section 6.2) is measured against the same proxy and may partly reflect fitting the proxy's insufficiency criterion, including its blind spots, rather than true sufficiency; this bias cannot be resolved without human relabeling beyond the audit sample. All verifier and calibrator classification metrics are proxy-conditional; human fidelity relabeling beyond the completed 60-event sufficiency audit remains future work.

7 Limitations

The study covers one primary dataset, two 7B/8B configurations, mostly two-hop plans, a fixed 500-question generation subset, and reference-inspired baselines. Coverage is also checked on all 7,405 questions. The dense variant (Section 5.3) is our only retrieval-stack variation; rerankers and sentence- or triple-level retrieval granularity are unevaluated. The proxy audit is small and proxy-conditional; most mechanism variants and k=6 are single-seed exploratory; 2Wiki and learned-gate replay are incomplete; and no imputation is used. Official strong baselines remain unavailable

because of missing Hub/PyPI access, checkpoints, API keys, and CUDA resources. Thus we make no universal superiority, absolute-efficiency, verifier-generalization, regret, or convergence claim.

The logged selective-prediction boundary and failure taxonomy (Section 6.4) are re-derived from logged seed-42 verifier decisions—not live abstention runs, matched-budget comparisons, or causal recovery estimates—and open-corpus retrieval is not run, so no open-corpus claim is made. After Holm correction, 0/6 EM contrasts but 4/6 F1 contrasts survive, so claims center on coverage, F1, mechanisms, and cost-quality trade-offs. Each within-table comparison shares one generator; official CRAG, S2G-RAG, SURE-RAG, and Self-RAG releases could not be run (Appendix I), so conclusions are restricted to the shared-model comparison.

8 Discussion and Conclusion

Established. The dependency-conditioned coverage gap is measurable before any controller is built; planning is the clearest mechanism and acts on coverage; verification changes stopping and cost without an established EM effect; typed reflection is diagnostic with no established net EM gain; audit state contains a learnable sufficiency signal; evidence sufficiency and answer correctness are separate control problems (208/500 gate-positive, EM-incorrect stops); computation is a first-class axis (k=6 buys coverage without EM); and directions replicate on a second model. **Not established.** Universal or SOTA superiority (EM contrasts fail Holm correction); open-corpus behavior; learned control; causal recovery effects; absolute efficiency; cross-dataset generalization (2WikiMultiHoQA is a single-seed prefix); and deployed abstention. The strongest claim is a reproducible interface separating coverage, answer quality, cost, and stopping—not SOTA EM or learned control.

Reproducibility Statement

The anonymous supplement contains per-question JSONL traces, fixed question ids, retrieval and serving configurations, model/data hashes, seed lists, table-generation scripts, and the exact commands used to regenerate all reported macros and figures. No simulator output contributes to the reported results.

Ethics Statement

The work uses a public question-answering benchmark and does not collect personal data. Potential misuse is limited to automation of factual question answering; retrieval errors and verifier bias are documented explicitly.

AI Use Statement

Generative AI tools assisted with manuscript organization, language editing, and LaTeX formatting. AI assistance also contributed to analysis scripts, code corrections, and audit preparation. Authors remain responsible for validating the final code, evidence, citations, and submission.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Mohammad Aghajani Asl, Majid Asgari-Bidhendi, and Behrooz Minaei-Bidgoli. Fair-rag: Faithful adaptive iterative refinement for retrieval-augmented generation. *arXiv preprint arXiv:2510.22344*, 2025.

Andre Bacellar. Bridgerag: Training-free bridge-conditioned retrieval for multi-hop question answering. *arXiv preprint arXiv:2604.03384*, 2026.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Xinyan Guan, Jiali Zeng, Fandong Meng, Chunlei Xin, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, and Jie Zhou. Deeprag: Thinking to retrieve step by step for large language models. *arXiv preprint arXiv:2502.01142*, 2025.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning*, 2020.

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 the 28th International Conference on Computational Linguistics*, pp. 6609–6625. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *North American Chapter of the Association for Computational Linguistics*, 2024.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Empirical Methods in Natural Language Processing*, 2023.

Zhouyu Jiang, Mengshu Sun, Lei Liang, and Zhiqiang Zhang. Retrieve, summarize, plan: Advancing multi-hop question answering with an iterative approach. In *Companion Proceedings of the ACM Web Conference (WWW 2025)*, 2025.

Shuguang Jiao, Chengkai Huang, Shuhan Qi, Xuan Wang, Yifan Li, Quanchi Weng, Lingchuan Liu, Xunliang Cai, and Lina Yao. Doctor-rag: Failure-aware repair for agentic retrieval-augmented generation. *arXiv preprint arXiv:2604.00865*, 2026.

Bowen Jin et al. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Empirical Methods in Natural Language Processing*, 2020.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. Few-shot reranking for multi-hop qa via language model prompting. arXiv:2205.12650, 2022. URL https://arxiv.org/abs/2205.12650.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *International Conference on Learning Representations*, 2023.

Myeonghwa Lee, Seonho An, and Min-Soo Kim. Planrag: A plan-then-retrieval augmented generation for generative large language models as decision makers. *arXiv preprint arXiv:2406.12430*, 2024. NAACL 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, 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*, 2020.

Minghan Li, Junjie Zou, Xinxuan Lv, Chao Zhang, and Guodong Zhou. S2g-rag: Structured sufficiency and gap judging for iterative retrieval-augmented qa. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25846–25862, 2026a. doi: 10.18653/v1/2026.acl-long.1185.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025.

Xingyu Li, Rongguang Wang, Yuying Wang, Mengqing Guo, Chenyang Li, Tao Sheng, Sujith Ravi, and Dan Roth. Par^2-rag: Planned active retrieval and reasoning for multi-hop question answering. *arXiv preprint arXiv:2603.29085*, 2026b.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting for retrieval-augmented large language models. In *Empirical Methods in Natural Language Processing*, 2023.

Md Mahadi Hasan Nahid and Davood Rafiei. Prism: Agentic retrieval with llms for multi-hop question answering. *arXiv preprint arXiv:2510.14278*, 2025.

Jaewan Park, Solbee Cho, and Jay-Yoon Lee. Stop-rag: Value-based retrieval control for iterative rag. *arXiv preprint arXiv:2510.14337*, 2025. NeurIPS 2025 MTI-LLM Workshop.

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 the Association for Computational Linguistics: EMNLP*, 2023.

Jingxi Qiu, Zeyu Han, and Cheng Huang. Sure-rag: Sufficiency and uncertainty-aware evidence verification for selective retrieval-augmented generation. *arXiv preprint arXiv:2605.03534*, 2026.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Zhihong Shao, Yeyun Cui, Wentao Wang, et al. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Yuling Shi, Maolin Sun, Zijun Liu, Mo Yang, Yixiong Fang, Tianran Sun, and Xiaodong Gu. Reasoning in trees: Improving retrieval-augmented generation for multi-hop question answering. In *Companion Proceedings of the ACM Web Conference (WWW 2026)*, pp. 626–635, 2026. doi: 10.1145/3774905.3794658. arXiv:2601.11255.

Jiashuo Sun, Jimeng Shi, Yixuan Xie, Saizhuo Wang, Jash Rajesh Parekh, Pengcheng Jiang, Zhiyi Shi, Jiajun Fan, Qinglong Zheng, Peiran Li, Shaowen Wang, Ge Liu, and Jiawei Han. Retrieval is cheap, show me the code: Executable multi-hop reasoning for retrieval-augmented generation. *arXiv preprint arXiv:2605.12975*, 2026.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Association for Computational Linguistics*, 2023.

Prakhar Verma, Sukruta Prakash Midigeshi, Gaurav Sinha, Arno Solin, Nagarajan Natarajan, and Amit Sharma. Plan*rag: Efficient test-time planning for retrieval augmented generation. *arXiv preprint arXiv:2410.20753*, 2024.

Adithya Vinod, Yuyang Ding, and Elias Stengel-Eskin. Calvert: Augmenting agents with calibrated verifier telemetry improves action and learning in knowledge-intensive tasks. *arXiv preprint arXiv:2606.21777*, 2026.

Qizhi Wang. Democratizing graphrag: Linear, cpu-only graph retrieval for multi-hop qa. arXiv:2602.23372, 2025. URL `https://arxiv.org/abs/2602.23372`. Date follows the source citation metadata.

Zili Wei, Xiaocui Yang, Yilin Wang, Zihan Wang, Weidong Bao, Shi Feng, Daling Wang, and Yifei Zhang. Cirag: Construction-integration retrieval and adaptive generation for multi-hop question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. arXiv:2601.06799.

Yaqi Wu, Xiaolei Guo, Chenyu Zhou, Jiaqi Huang, Xianfa Zhang, Junxu Zhang, Zhuo Yu, Zhubo Shi, Jianghao Lin, and Dongdong Ge. Dynakrag: A unified framework for learnable evidence control in multi-hop retrieval-augmented generation. *arXiv preprint arXiv:2607.06507*, 2026.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Empirical Methods in Natural Language Processing*, 2018.

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

Zheng Yao, Weizhi Qi, Liangming Pan, Shuyang Cao, Linmei Hu, Weichao Liu, Lei Hou, and Juanzi Li. Seakr: Self-aware knowledge retrieval for adaptive retrieval augmented generation. In *Association for Computational Linguistics*, 2025.

Rongzhi Zhu, Xiangyu Liu, Zequn Sun, Yiwei Wang, and Wei Hu. Mitigating lost-in-retrieval problems in retrieval augmented multi-hop question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

A Cross-Model and Cross-Dataset Tables

The Llama-3.1-8B results below aggregate all five planned decoding seeds after averaging seeds within question. The bootstrap intervals and Holm values resample questions; they do not separately model uncertainty over a population of possible decoding seeds, so seed-wise standard deviations are reported alongside them.

Table 2: Cross-model replication on the fixed 500-question set (Llama-3.1-8B-Instruct; all 5 planned seeds complete). * denotes question-level Holm-adjusted $p < 0.05$; seeds are first averaged within question.

	Llama-3.1-8B-Instruct (prefix seeds 42,43,44,45,46)				Qwen2.5-7B (5 seeds)	
System	EM	F1	Rec.	Chain	EM	F1
NaiveRAG	41.2*	52.8*	75.9*	54.2*	43.3	55.4
Adv.RAG	44.8*	58.2*	82.4*	67.3*	44.0	57.2
Iter-RetGen	44.6*	56.7*	82.9*	68.0*	44.6	57.4
Self-Ask	42.4*	54.2*	81.3*	64.3*	43.8	56.4
CRAG	41.4*	53.1*	76.6*	55.6*	43.4	55.7
PR-RAG	**47.6**	**61.7**	**91.9**	**84.7**	47.8	60.9

Llama cost per question: retrieval / LLM calls, latency (s), tokens

System	Retr.	LLM	Lat.	Tok.
NaiveRAG	1.00	1.00	1.5	449
Adv.RAG	1.98	2.00	2.4	744
Iter-RetGen	2.00	2.00	2.7	929
Self-Ask	2.25	3.25	9.2	1356
CRAG	1.04	1.04	4.9	905
PR-RAG	3.17	6.45	10.0	4022

B Second-Dataset Exploratory Prefix

The 2WikiMultiHoQA table reports only complete cells from the current seed prefix; incomplete cells are omitted without imputation. This is a single-seed exploratory prefix, not completed replication or generalization.

C Exact Controller Contracts

The planner receives the original question and is instructed to emit two or three single-hop subquestions, one unnumbered line each. If line i requires an earlier answer, it must use `<1>` or `<2>`

Table 3: 2WikiMultiHoQA dev subset under the common protocol (current completed prefix 1 of five planned seeds). Bold marks a PR-RAG column only when its Holm-corrected advantage holds against a majority of baselines; EM is not bolded because its Holm advantage holds only vs. Adv.RAG (p=0.004) and not vs. the strongest baseline Iter-RetGen (p=0.84).

System	EM	F1	Rec.	Chain	Retr.	LLM
NaiveRAG	32	36.5	74.6	51	1	1
Adv.RAG	28.6	33.6	80	61.2	2	2
Iter-RetGen	34.8	39.5	80	61.4	2	2
Self-Ask	32.4	36.7	78.4	57.6	2.19	3.19
CRAG	32.8	37.1	76.6	54.2	1.46	1.46
PR-RAG	35.2	**43.3**	**90.4**	**82.8**	2.58	6.5

rather than guessing the entity. Parsing retains at most three non-empty lines. A placeholder induces a directed dependency on its referenced earlier line; when a multi-line plan contains no placeholder, the implementation uses the auditable sequential fallback $\mathrm{dep}(i) = \{i-1\}$ for $i > 1$. An empty parse falls back to the original question.

For each planned item, the extractor must return exactly two lines: a `FOUND`/`NOT_FOUND` verdict followed by a short entity or date, or `UNKNOWN`. Only non-unknown answers instantiate downstream placeholders. The verifier receives the original question, every planned item with its current extracted answer, and the accumulated passages. It must return exactly three labels: `evidence_chain: COMPLETE/INCOMPLETE`, `intermediate_answers: AVAILABLE/MISSING`, and `answerable: YES/NO`. The prompt gate stops only when all three labels are positive; otherwise the typed handler is selected from the logged missing items, extraction verdicts, duplicate retrievals, and remaining budget. The released JSONL records the raw plan, parsed dependencies, instantiated queries, retrieved ranks and scores, extraction outputs, all three verifier labels, reflection type, next action, and call counts. These contracts match the evaluated implementation; they are specifications, not claims of planner accuracy.

D Additional Audit Tables

D.1 Post-hoc Holm-family sensitivity

Excluding the single-seed $k = 6$ control changes the multiplicity family after observing the outcomes. We report the resulting sensitivity transparently, without replacing the six-contrast family or interpreting it as confirmatory. The unchanged raw bootstrap probabilities yield 4/5 significant EM contrasts and 5/5 significant F1 contrasts. Family composition affects threshold decisions; it does not change effect sizes, resource imbalance, or the unmeasured fidelity outcome.

Table 4: Post-hoc sensitivity using archived raw probabilities and cumulative-maximum Holm correction. Five-contrast results exclude $k = 6$ and are non-confirmatory.

Metric	Baseline	Raw p	Holm (6)	Holm (5)
EM	selfask	0.0085	0.0510	0.0425
EM	naive	0.0088	0.0510	0.0425
EM	crag	0.0100	0.0510	0.0425
EM	advanced	0.0106	0.0510	0.0425
EM	itergen	0.0536	0.1072	0.0536
F1	crag	0.0008	0.0048	0.0040
F1	naive	0.0018	0.0090	0.0072
F1	selfask	0.0020	0.0090	0.0072
F1	advanced	0.0130	0.0390	0.0260
F1	itergen	0.0344	0.0688	0.0344

Table 5: Human sufficiency audit of the gold-chain proxy (n =60 stop decisions, two blind annotators). H_{strict} counts only evidence humans judged sufficient; H_{lenient} additionally counts cases where alternate retrieved evidence suffices. Small-sample audit; estimates are indicative, not full gold labels.

	H_{strict}	H_{lenient}
Proxy–human agreement	76.3	59.3
Proxy FPR (%)	22.2	33.3
Proxy FNR (%)	25.0	43.2

D.2 Retained results and measurement audit

The complete paired-comparison table is generated from the same logs and is included here to make multiplicity control inspectable. The significance audit reports, for each of the six reported contrasts, the mean difference, 95% bootstrap CI, raw and Holm-adjusted p, Cohen's d, and per-question win/tie/loss counts. The human sufficiency audit table supporting Section 6.6 and the offline calibrator table supporting Section 6.2 are also included here.

Table 6: Question-grouped five-fold out-of-fold stop classification on 3173 real events from 500 questions. Values are percentages against the offline gold-chain proxy.

Verifier	Agreement	F1	FPR	FNR
Prompt gate	71.0	78.5	44.0	21.7
Learned calibrator	**72.3**	78.3	**30.8**	26.2

Table 7: Significance audit: paired bootstrap contrasts for PR-RAG on the fixed 500-question set. *: Holm-adjusted $p < 0.05$.

Metric	vs	Δ	95% CI	raw p	Holm p	Cohen's d	W/T/L
EM	naive	+4.52	[+1.08, +7.96]	0.0088	0.0510	0.115	90 / 344 / 66
EM	advanced	+3.88	[+0.88, +6.80]	0.0106	0.0510	0.114	90 / 349 / 61
EM	itergen	+3.24	[-0.08, +6.60]	0.0536	0.1072	0.086	85 / 350 / 65
EM	selfask	+4.00	[+1.04, +7.04]	0.0085	0.0510	0.115	97 / 338 / 65
EM	crag	+4.48	[+1.08, +7.84]	0.0100	0.0510	0.117	94 / 341 / 65
EM	naive$_{k=6}$	+3.04	[-0.12, +6.28]	0.0570	0.1072	0.083	77 / 367 / 56
F1	naive	+5.45	[+2.13, +8.81]	0.0018	0.0090	0.143	117 / 297 / 86 *
F1	advanced	+3.72	[+0.75, +6.61]	0.0130	0.0390	0.111	117 / 295 / 88 *
F1	itergen	+3.47	[+0.26, +6.69]	0.0344	0.0688	0.094	109 / 302 / 89
F1	selfask	+4.52	[+1.65, +7.46]	0.0020	0.0090	0.136	134 / 276 / 90 *
F1	crag	+5.23	[+1.95, +8.53]	0.0008	0.0048	0.142	122 / 293 / 85 *
F1	naive$_{k=6}$	+1.90	[-1.23, +5.03]	0.2314	0.2314	0.052	95 / 321 / 84
gold-rec.	naive	+10.24	[+7.82, +12.62]	¡0.0001	$< 10^{-4}$	0.376	163 / 268 / 69 *
gold-rec.	advanced	+1.72	[-0.52, +3.94]	0.1310	0.1310	0.068	117 / 275 / 108
gold-rec.	itergen	+3.74	[+1.44, +6.06]	0.0022	0.0088	0.143	115 / 290 / 95 *
gold-rec.	selfask	+2.56	[+0.46, +4.64]	0.0140	0.0280	0.107	120 / 285 / 95 *
gold-rec.	crag	+8.40	[+6.04, +10.76]	¡0.0001	$< 10^{-4}$	0.311	155 / 264 / 81 *
gold-rec.	naive$_{k=6}$	-3.06	[-5.18, -0.92]	0.0066	0.0198	-0.126	69 / 303 / 128 *
chain	naive	+19.60	[+15.24, +24.08]	¡0.0001	$< 10^{-4}$	0.390	160 / 278 / 62 *
chain	advanced	+4.04	[-0.04, +8.20]	0.0524	0.0524	0.085	118 / 281 / 101
chain	itergen	+6.96	[+2.76, +11.20]	0.0008	0.0032	0.145	111 / 299 / 90 *
chain	selfask	+5.88	[+2.00, +9.84]	0.0022	0.0066	0.132	119 / 291 / 90 *
chain	crag	+16.44	[+12.08, +20.88]	¡0.0001	$< 10^{-4}$	0.328	154 / 273 / 73 *
chain	naive$_{k=6}$	-5.20	[-9.32, -1.08]	0.0130	0.0260	-0.111	69 / 309 / 122 *

E Mechanism and Failure Diagnostics

This appendix contains the exploratory diagnostics removed from the main narrative to meet the page limit. They use the same real-model logs and macros; single-seed rows are not seed-averaged.

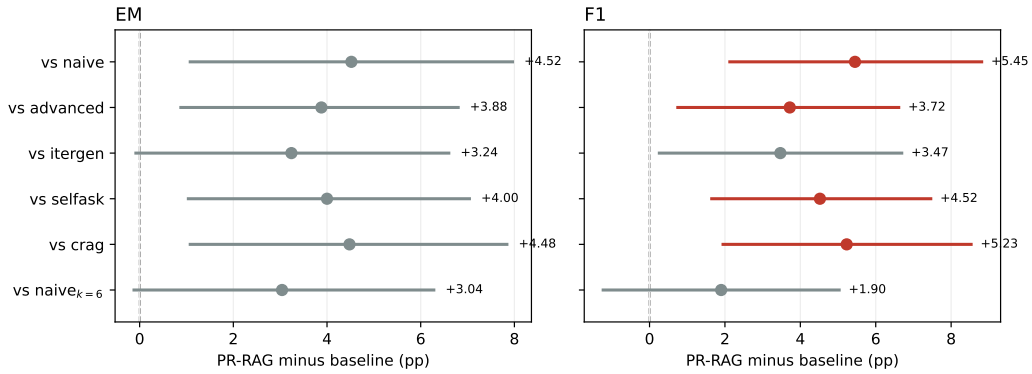


Figure 4: Effect-size forest plot: PR-RAG minus each baseline for EM and F1 with 95% bootstrap CIs. Grey intervals do not survive Holm correction; red intervals do.

Table 8: Exploratory ablation and dense-retrieval contrasts.

Contrast	Metric	Δ	d	95% CI	p	p_{Holm}
PR-RAG T=1 vs AdvancedRAG	EM	+3.4	0.09	[+0.1, +6.9]	0.0461	0.6454
PR-RAG T=1 vs AdvancedRAG	Rec.	+0.3	0.01	[-2.1, +2.7]	0.8146	1.0000
PR-RAG T=1 vs AdvancedRAG	Chain	+1.8	0.04	[-2.6, +6.2]	0.4140	1.0000
PR-RAG vs w/o verifier	EM	+1.1	0.09	[+0.0, +2.2]	0.0358	0.5370
PR-RAG vs w/o verifier	Chain	-2.2	-0.17	[-3.4, -1.1]	< 0.0001	< 0.0001
PR-RAG vs w/o reflection	EM	+0.2	0.01	[-1.8, +2.4]	0.8030	1.0000
PR-RAG vs w/o reflection	Chain	+1.6	0.07	[-0.4, +3.7]	0.1158	1.0000
PR-RAG vs blind retry	EM	+0.4	0.02	[-1.6, +2.5]	0.6894	1.0000
PR-RAG vs blind retry	Chain	+2.0	0.09	[+0.0, +4.0]	0.0490	0.6370
PR-RAG vs reading-only	EM	+0.4	0.02	[-1.6, +2.5]	0.6714	1.0000
PR-RAG vs reading-only	Chain	+2.0	0.09	[+0.0, +4.0]	0.0494	0.5928
PR-RAG vs retrieval-only	EM	+0.0	0.00	[-2.0, +2.2]	0.9360	0.9360
PR-RAG vs retrieval-only	Chain	+0.8	0.04	[-1.0, +2.8]	0.4096	1.0000
PR-RAG vs dep-only	EM	+0.4	0.02	[-1.6, +2.5]	0.6626	1.0000
PR-RAG vs dep-only	Chain	+2.0	0.09	[+0.0, +4.1]	0.0498	0.5478
NaiveRAG dense vs BM25	EM	+1.6	0.04	[-2.3, +5.3]	0.4330	1.0000
NaiveRAG dense vs BM25	Rec.	+7.7	0.22	[+4.6, +10.7]	< 0.0001	< 0.0001
NaiveRAG dense vs BM25	Chain	+15.0	0.24	[+9.4, +20.6]	< 0.0001	< 0.0001
AdvancedRAG dense vs BM25	EM	+3.6	0.11	[+0.8, +6.3]	0.0137	0.2329
AdvancedRAG dense vs BM25	Rec.	+4.1	0.17	[+1.9, +6.2]	0.0008	0.0144
AdvancedRAG dense vs BM25	Chain	+8.6	0.19	[+4.7, +12.6]	< 0.0001	< 0.0001
PR-RAG dense vs BM25	EM	+3.2	0.10	[+0.5, +6.0]	0.0210	0.3360
PR-RAG dense vs BM25	Rec.	+5.6	0.24	[+3.6, +7.6]	< 0.0001	< 0.0001
PR-RAG dense vs BM25	Chain	+10.7	0.26	[+7.0, +14.4]	< 0.0001	< 0.0001

E.1 Budget and component interventions

Table 9: Budget and component interventions. † denotes single-seed (42) exploratory rows; unmarked rows use five seeds.

System	EM	F1	Chain	Retr.	LLM
PR-RAG (full, T=3)	**47.8**	**60.9**	**73.8**	**2.34**	**5.97**
PR-RAG (T=2)†	47.8	60.7	72.8	2.22	5.68
PR-RAG (T=1)†	47.4	60.2	71.6	1.95	5.12
w/o Planning†	43.0	54.4	54.2	1.43	3.19
w/o Verification	46.7	59.8	76.0	5.18	8.50
NaiveRAG (k=6)†	44.8	59.0	79.0	1.00	1.00

E.2 Selective prediction and logged error structure

The following tables are re-derived from the logged seed-42 PR-RAG trajectories; they are selective-prediction boundary analyses and logged conditional frequencies, not new live runs and not causal estimates.

Table 10: Selective answering on logged seed-42 verifier decisions: answering only questions the verifier judged sufficient (the deployed loop answers everything). This is a logged-trace re-derivation, not a live abstain run.

	N	EM (%)	Coverage (%)
Answer everything (deployed)	500	47.8	100.0
Selective: answer iff sufficiency	421	50.6	84.2
Abstain those questions	79	32.9	–

Table 11: Logged error propagation: final-answer EM and gold-chain state conditioned on the first logged verification (seed 42). Frequencies are associations over the fixed 500-question set, not causal recovery estimates.

Conditional (seed 42, N=500)	n	%
P(final correct | first-verification gold)	358	53.6
P(final correct | first-verification missing)	142	33.1
P(final gold | first gold)	358	100.0
P(gold recovered | first missing)	142	4.9
P(chain coverage | first missing)	142	4.9

E.3 Observed budget envelopes (descriptive)

The following table conditions each system's five-seed, question-aggregated records on its own realized number of LLM calls. It is included to expose the cost-quality support of the logs, not to create a matched-budget comparison: the eligible question sets differ by system (for example, only 318 PR-RAG questions satisfy $B \leq 5$, versus 500 for every main baseline). Therefore these conditional summaries are descriptive and non-causal; a controller re-run under an identical per-question budget remains missing.

E.4 Reflection variants and typed recovery

Reflection fires on 112 questions. Retrieval-failure rewriting treats 184 events and recovers a gold document in 60.3% of treated events, but only 19.6% end with an extracted sub-answer; the net EM change against no reflection is only +0.2 points ($p = 0.803$), with 1 net EM flip. These numbers identify an evidence-to-answer bottleneck rather than a retrieval-only failure. Entity-swap rewriting and a separate answer gate are natural next interventions; neither is evaluated here.

E.5 Verifier reliability and errors

The seed-42 prompt verifier has 72.1% proxy agreement over 642 events and FPR 40.5%. On seed 42, 208/500questions stop with judged-sufficient evidence but an incorrect final answer; retrieval failures (53) dominate reflected errors, followed by dependency (19), verifier (18), and reading (4) failures. These observations motivate separately measuring evidence sufficiency and answer support; a second gate remains an unvalidated intervention.

E.6 Planner, proxy, and distractor diagnostics

Across five seeds, explicit placeholder dependencies occur in 52.4–54.8% of questions; the remainder use sequential fallback or collapse to one sub-question. Explicit-placeholder traces have 79.4% chain coverage versus 75.1% for fallback, but the EM association is opposite (46.8 vs. 49.9): these are associations, not planner-accuracy or causal estimates, so explicit dependencies are unproven for EM. As a proxy stress test, 209 of 2,500 seed-question pairs are answered correctly without complete gold-title coverage, and there are 455 prompt-positive/proxy-negative verification events. For seed-42 retrieval-failure rewrites, 49.4% of treated questions introduce a new non-gold title (0.7 on average); paired with no reflection, 3 questions improve and 2 worsen in EM, with the remainder unchanged. These diagnostics motivate distractor-aware recovery; the human sufficiency audit is reported in Section 6.6.

Table 12: Failure taxonomy over logged reflection events (seed 42). Final-EM-ok is the share of questions with that failure type whose final EM is correct (association, not recovery rate).

Failure type	N	%	Final-EM-ok %
retrieval_failure	184	66.4	32.1
dependency_failure	52	18.8	26.9
reading_failure	10	3.6	10.0
verifier_failure	0	0.0	–
unknown	0	0.0	–

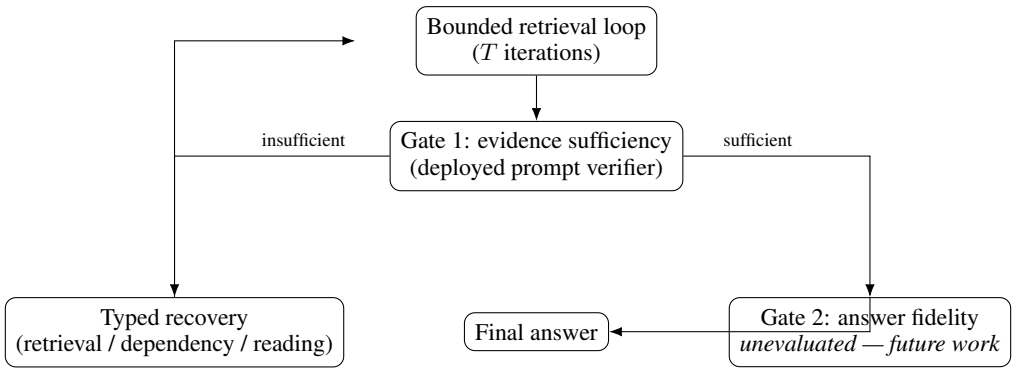


Figure 5: Conceptual separation of the two control problems. PR-RAG deploys and audits Gate 1 (evidence sufficiency); logged traces show that passing Gate 1 does not imply a correct answer (208/500seed-42 stops), motivating an answer-support Gate 2 whose benefit remains untested. The harness contains a preliminary answer-support variant, but no validated Gate 2 experiment is reported.

E.7 Additional figures

F Plan Stability and Three-Label Verification

These appendix analyses respond to reviewer questions about (i) the stability of planning outputs across decoding seeds and (ii) the internal interdependence of the three verification labels. Both are zero-cost re-derivations from the released per-question JSONL traces (`results_main/logs/seed{42..46}/prrag.jsonl`, 5×500 trajectories); no model was decoded. All numbers below are logged associations, not causal estimates.

F.1 Plan stability across seeds

We define a plan signature as the tuple (#sub-questions, #non-empty place-holder dependency sets, explicit-dependency flag). Across the 500 questions and 5 decoding seeds, $100 \cdot 0.628\%$ of questions produce the *identical* signature in all five seeds, and $100 \cdot 0.82\%$ are stable in at least four of five seeds; the average match rate to the modal signature is 0.883. Plan stability is not orthogonal to outcome: on seed 42, plan-stable questions reach 46.59 EM and 74.15% chain coverage, versus 53.33 and 67.78% for plan-unstable questions. We emphasize this is an association, not causal: unstable plans tend to be the multi-hop, explicit-dependency traces, which gain evidence coverage while their EM association runs the opposite way (Section 6.4 and the explicit-dependency diagnostics). Plan attribution is a plan-shape statistic, not a validated planner-accuracy estimate, because we do not hold gold labels for correct dependency structure.

F.2 Three-label verification interdependence

The verifier emits three labels — evidence chain (complete/incomplete), intermediate answers (available/missing), and answerability (yes/no). Over the last logged verification per question (seed 42, 500 questions), only four of the eight possible combinations occur (Table 17): the la-

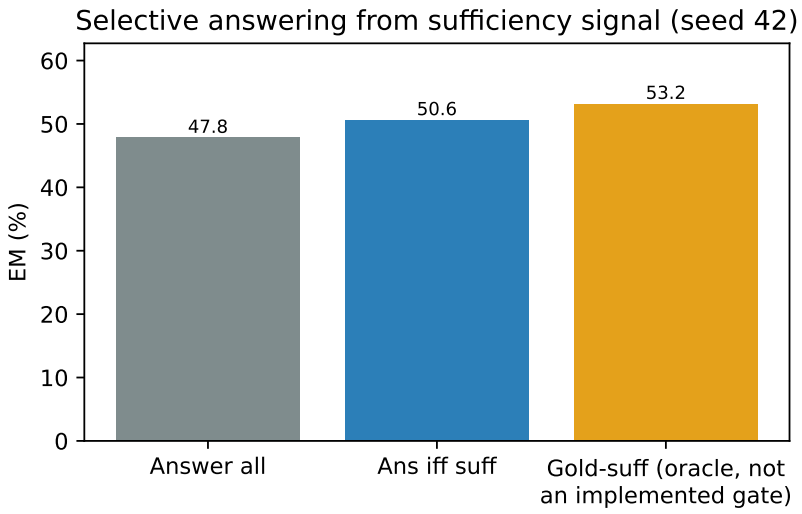


Figure 6: Selective answering from the sufficiency signal (logged seed-42 decisions; the gold-suff bar is an oracle upper reference, not an implemented gate).

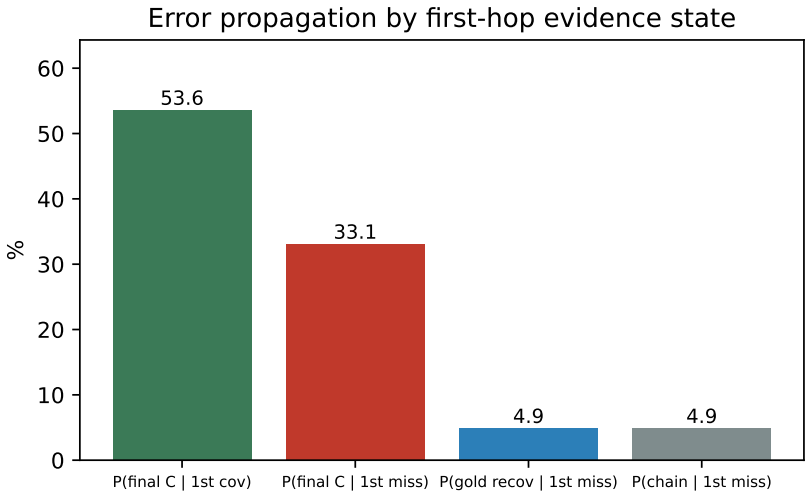


Figure 7: Logged error propagation by first-verification evidence state (seed 42; conditional frequencies, not causal estimates).

bels are strongly correlated in practice — $I+$ never co-occurs with an incomplete chain, and answerable=YES collapses to the all-positive combination (421/500). The informative contradictions are rare but real: 6 questions receive a COMPLETE chain label yet a NO answerability label (disagreement between verifier outputs, not independently established evidence sufficiency), and 11 have incomplete chains and are not answerable. For the calibrator boundary, Table 18 shows that as single features, `evidence_chain=COMPLETE` and `answerable=YES` reach precision 0.789 with recall 0.923/0.91 against the gold proxy; `intermediate_answers=AVAILABLE` has the highest recall of gold sufficiency (0.942) at the lowest specificity (0.304). Adding `answerable` sharpens precision marginally without shifting recall materially, a descriptive pattern that does not by itself establish which features causally determine the calibrator's boundary.

Table 13: Observed conditional budget envelopes on the complete HotpotQA main run. Values are percentages except n; seeds are averaged within question before conditioning. This table is not matched-budget evidence.

Budget B	System	n	EM	F1	Rec.	Chain
4	naive	500	43.3	55.4	75.9	54.2
4	advanced	500	44.0	57.2	84.4	69.8
4	itergen	500	44.6	57.4	82.4	66.8
4	selfask	474	43.9	56.3	83.0	66.7
4	crag	500	43.4	55.7	77.7	57.4
4	prrag	24	53.3	65.7	71.7	43.3
5	naive	500	43.3	55.4	75.9	54.2
5	advanced	500	44.0	57.2	84.4	69.8
5	itergen	500	44.6	57.4	82.4	66.8
5	selfask	500	43.8	56.4	83.6	67.9
5	crag	500	43.4	55.7	77.7	57.4
5	prrag	318	51.9	66.2	88.4	77.7
6	naive	500	43.3	55.4	75.9	54.2
6	advanced	500	44.0	57.2	84.4	69.8
6	itergen	500	44.6	57.4	82.4	66.8
6	selfask	500	43.8	56.4	83.6	67.9
6	crag	500	43.4	55.7	77.7	57.4
6	prrag	388	51.3	64.9	88.2	77.7
10	naive	500	43.3	55.4	75.9	54.2
10	advanced	500	44.0	57.2	84.4	69.8
10	itergen	500	44.6	57.4	82.4	66.8
10	selfask	500	43.8	56.4	83.6	67.9
10	crag	500	43.4	55.7	77.7	57.4
10	prrag	459	49.2	62.6	87.5	76.4

Table 14: Reflection variants on seed 42 (single decoding seed; exploratory, not seed-averaged).

System	EM	F1	Chain	Retr.	LLM
Typed reflection (full)	**47.8**	**60.7**	**73.0**	**2.39**	**6.06**
No reflection (whole-plan retry)	47.6	60.3	72.2	2.73	6.14
Blind retry (no rewrite)	47.4	60.1	71.8	2.35	5.67
Reading-failure only	47.4	60.1	71.8	2.35	5.69
Retrieval-failure only	47.8	60.6	73.0	2.32	5.98
Dependency-failure only	47.4	60.1	71.8	2.35	5.67

G Coverage–Fidelity Control Landscape

An exploratory reanalysis of all 30 complete main system–seed cells classifies each outcome before averaging seeds within question and then questions. High coverage means full gold-title chain coverage; correct means normalized EM equal to one. A: incomplete/wrong; B: complete/wrong; C: complete/correct; D: incomplete/correct. D may reflect alternate evidence or parametric knowledge; B includes normalization discrepancies and cannot be labelled unsupported without a fidelity audit. Of 2,500 paired seed–question outcomes, 301 move from NaiveRAG A to PR-RAG B; 39 move in reverse. These are paired observations, not independent samples or causal mediation estimates. At seed 42, 208 of 421 positive final sufficiency judgments have incorrect EM; 153 of those have full chain coverage. Across all 500 seed-42 outcomes, 171 have complete coverage and incorrect EM. These denominators must not be interchanged. Complete gold-chain coverage is neither necessary nor sufficient for EM correctness in these logs. Reproduction: `analyze_control_plane.py`; outputs include input hashes and question-bootstrap intervals.

H Offline Decision-Problem Formalization

This appendix describes the scope of an offline diagnostic utility and the limits of logged-state analysis.

Table 15: Seed-42 mechanism diagnostics (single decoding seed; exploratory, not seed-averaged). Agreement is with an offline gold-chain proxy, not human-labelled accuracy.

Diagnostic	Value
Plans with explicit dependencies	54.0%
Avg. sub-questions per plan	1.95
Verifier precision (vs. proxy)	78.9%
Verifier recall (vs. proxy)	78.7%
Verifier F1 (vs. proxy)	78.8%
Verifier FPR	40.5%
Verifier FNR	21.3%
Verifier stop agreement (vs. proxy)	72.1%
Oracle: premature stops	0
Oracle: wasted iterations	0
EM loss upper bound	$\sim$11.0 pts
Dep.-failure events	52 (recovery 5.8%)
Retr.-failure events	184 (recovery 19.6%)
Reading-failure events	10 (recovery 10.0%)

Table 16: Limited-scale S2G-style stopping-controller contrast on the fixed 500-question set (seed 42, same BM25 k=3 protocol). The S2G-style row uses a prompted "Untrained Judge (w/o SFT)" proxy, not the official LoRA judge; p-values are paired bootstrap over questions and are exploratory, outside the six reported Holm-corrected contrasts.

System	EM	F1	Chain (%)	Retr.	LLM
S2G-RAG-style (ref.-inspired, 1 seed)	44.4	54.97	66.8	2.91	4.23
PR-RAG (paired seed-42 set)	47.8	60.69	73.0	2.39	6.06

H.1 Offline utility and identification limits

For descriptive analysis, the released audit computes a bounded terminal utility

$$u = Q - \lambda C, \qquad Q = \max(\mathrm{F1}, \mathrm{EM}), \qquad C = \min\left(1, \frac{\texttt{llm_calls}}{12}\right), \quad \lambda = 0.5.$$

The normalization and weight are analysis choices. This quantity is not a training reward, a validated human preference, or a learned objective. Since $Q, C \in [0, 1]$, the arithmetic range is $[-\lambda, 1]$.

Logged state summaries include dependency depth, unresolved-item count, retrieval scores, verifier labels, and remaining iterations. They are audit features, not proven sufficient Markov states. Query text, evidence contents, pending items, and stochastic model outputs also influence subsequent actions. A deterministic conjunction over three observed gate labels does not establish a deterministic policy over this compressed state, and budget exhaustion or fallback can terminate the loop without a positive gate.

The iteration limit bounds retrieval rounds, not every internal action or model call. Retrieval, extraction, verification, rewriting, and rereading may all occur within one round; a 4^T action-sequence bound therefore does not describe the implemented protocol. Independent controller state across questions also does not prove statistical independence of questions or disjoint corpus content.

The logs expose factual trajectories under the executed protocols. They do not identify the answers that unexecuted actions would have produced, optimal stopping, causal recovery effects, or regret. A changed policy must be evaluated by a controlled rerun. The offline calibrator re-thresholds a proxy sufficiency signal; no offline reinforcement-learning policy or online RL policy is trained.

I Baseline Provenance

The main-table rows are method-performance comparisons under a common protocol (shared data pool, retriever, generator, and decoding). The following tables record implementation provenance for each row — status, origin, retrieval API, budget, decoding, and extra-training-data use — and, for the four strong baselines whose official releases could not be run here, the verified repository/commit, required checkpoints, and the primary technical blocker.

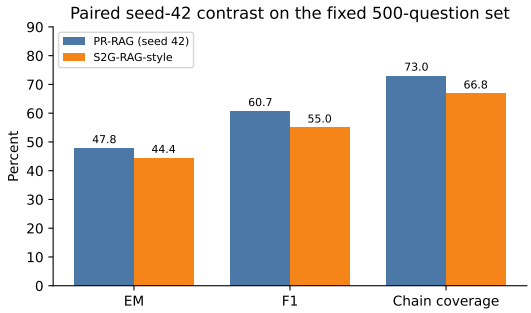


Figure 8: Paired seed-42 comparison on the fixed 500-question set: PR-RAG versus a reference-inspired S2G-RAG-style stopping controller. Differences are not statistically established under paired bootstrap.

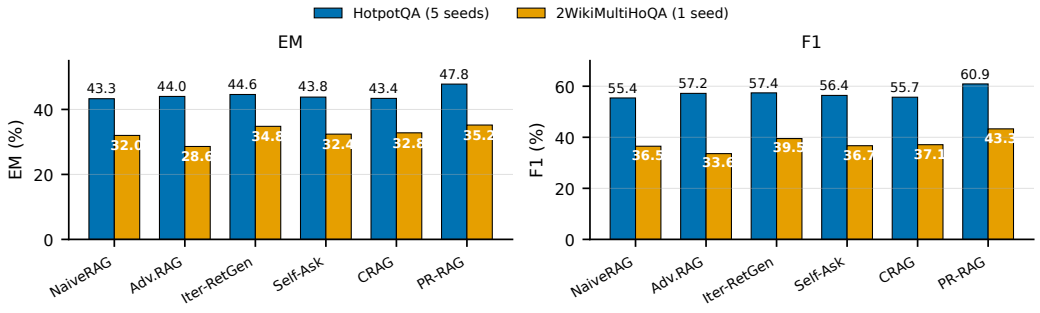


Figure 9: EM and F1 on HotpotQA (five decoding seeds, 95% bootstrap CIs in Table 1) and the 2WikiMultiHoQA subset (one completed seed, seed 42). All values are generated from the released logs; the 2Wiki series is a single-seed exploratory prefix and is not used for generalization claims.

Machine-readable fields with full blocking reasons and model-difference notes are released as `baseline_provenance.json`; `baseline_audit.md` records the audit narrative and the distinction between method-performance and implementation-protocol comparison. No reference-inspired or non-official row is described as an official baseline.

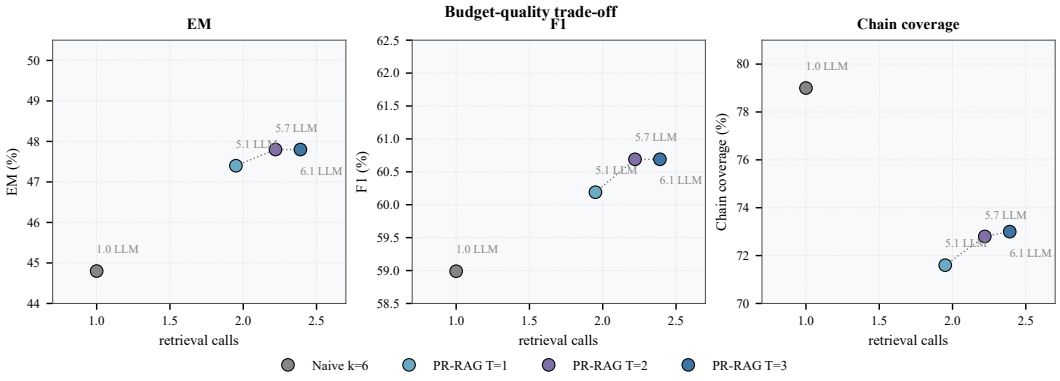


Figure 10: Budget-quality trade-off on seed 42 (single decoding seed, exploratory). Labels expose mean LLM calls; no calls are hidden. Table 1 supplies the shared baseline coordinate system for quality, calls, and latency.

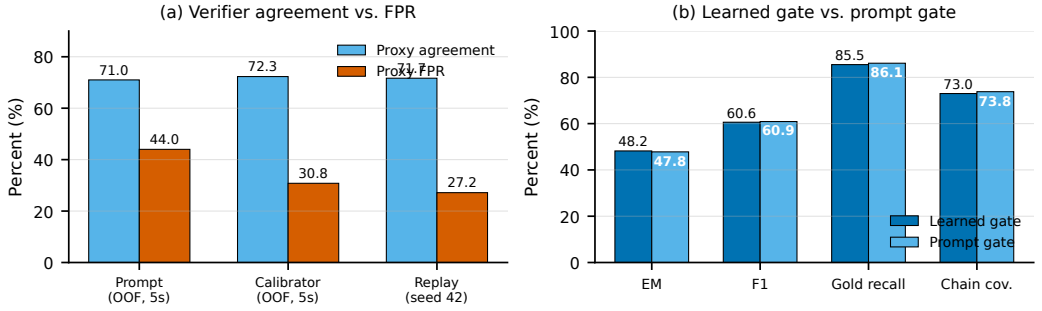


Figure 11: Composite gate-control evidence. (a) Verifier agreement and FPR against the offline gold-chain proxy for the prompt verifier (question-grouped out-of-fold, five seeds), the learned calibrator (out-of-fold, five seeds), and the learned-gate replay (seed 42). Agreement is with the gold-chain proxy, not human-labelled accuracy; the completed blind human sufficiency audit is reported in Section 6.6. (b) Learned-gate versus prompt-gate paired metrics on the fixed 500-question set at seed 42 (the only completed seed of the multi-seed replay; seeds 43–46 are not imputed). Differences are within noise for EM/F1/recall/coverage while call counts increase, so no learned-control quality claim is made. Values are generated from the released logs.

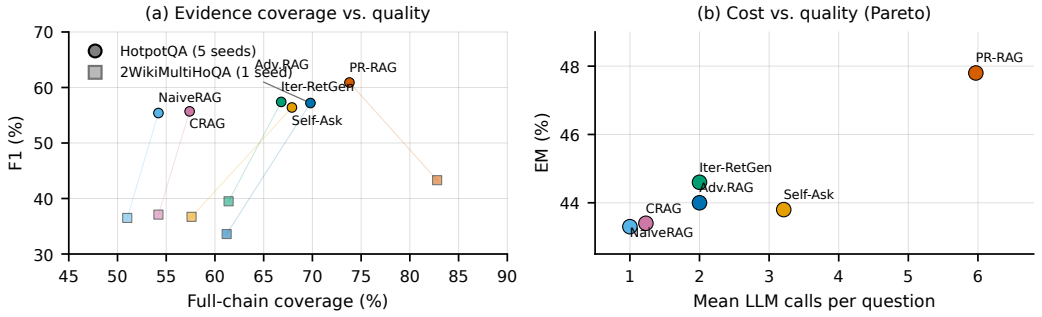


Figure 12: Composite quality trade-offs. (a) Full-chain coverage versus F1 per system on HotpotQA (five seeds) and the 2WikiMultiHoQA subset (one seed). Coverage is a retrieval measure, not answer accuracy; the systems with higher evidence coverage also tend to have higher F1, and the 2Wiki series is a single-seed exploratory prefix. (b) Cost–quality trade-off on HotpotQA (five seeds): mean LLM calls per question versus EM. PR-RAG is the most expensive row in Table 1 and has the highest EM point estimate; the figure makes the exchange explicit and is not a claim of absolute efficiency. Values are generated from the released logs.

Table 17: Observed three-label combinations at the last logged verification (seed 42). Chain $\in \{\mathrm{C}, \mathrm{NC}\}$; Intermed. $\in \{I+, I-\}$; Answerable $\in \{A+, A-\}$. Only four of eight combinations occur.

Chain	Intermed.	Answerable	#
C	I+	A+	421
NC	I-	A-	62
NC	I+	A-	11
C	I+	A-	6

Table 18: Single-feature separation of gold-chain sufficiency (seed 42), versus the combined prompt gate (precision 0.789, recall 0.91). These are the features available to the offline calibrator.

Feature	Prec(v1|gold1)	Recall	Spec
evidence_chain=COMPLETE	0.789	0.923	0.333
intermediate_answers=AVAILABLE	0.785	0.942	0.304
answerable=YES	0.789	0.910	0.341

Table 19: Coverage/correctness decomposition (%). Categories sum to 100; five seeds repeat the same 500 questions.

System	A	B	C	D
naive	34.4	22.3	31.9	11.4
advanced	23.3	32.7	37.0	6.9
itergen	25.4	30.0	36.9	7.7
selfask	25.1	31.1	36.8	7.0
crag	32.1	24.5	32.8	10.5
PR-RAG	17.8	34.3	39.5	8.4

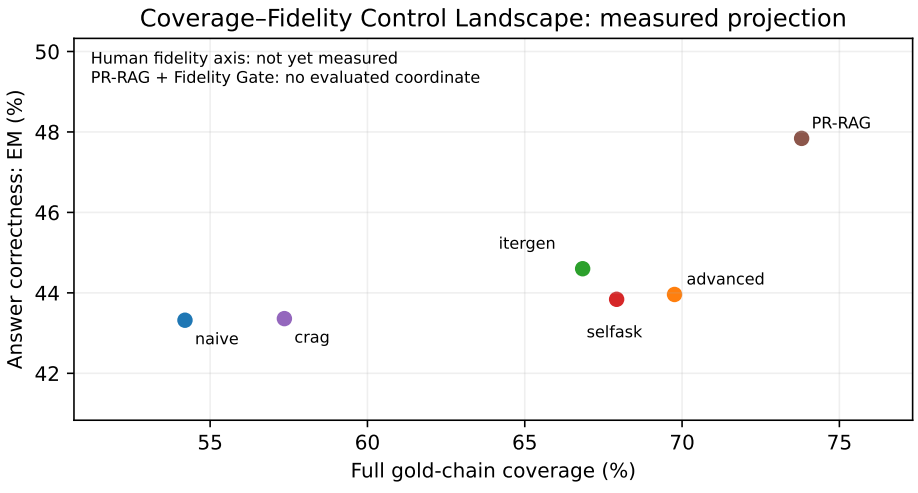


Figure 13: Coverage–fidelity landscape: measured projection. The vertical coordinate is EM, not human-rated fidelity. PR-RAG plus a fidelity gate has no measured coordinate. Costs differ between systems.

Table 20: Closest-work positioning by research emphasis. These categories are not claims that other capabilities are absent, and do not imply official reproductions.

Work	Research emphasis
Stop-RAG	Learned value-based stopping
PAR^2-RAG	Evidence coverage and commitment
CIRAG / BridgeRAG	Evidence integration and later-hop retrieval
SURE-RAG / S2G-RAG	Evidence sufficiency judgments
PR-RAG	Joint audit of observable state, stopping, answer correctness, and cost

Table 21: Implementation provenance of the six main-table systems. Status: F = faithful reimplementation, R = reference-inspired (not an official reproduction), † = non-official alternative to an official release that could not be run in this environment.

Method	Status	Origin	Retrieval API	Budget	Decoding	Extra training data
NaiveRAG	faithful reimplementation	Standard single-shot RAG recipe (Lewis et al., 2020); no per-method official release exists	BM25 over the per-question 10-paragraph pool, k=3	1 retrieval call, 1 LLM call per question	temperature 0.7, top-p 0.8, seeds 42-46	none
AdvancedRAG	reference-inspired	Generic query-rewriting RAG (no canonical official release)	BM25 k=3 per rewritten query over the same pool	2 retrieval calls, 2 LLM calls per question	temperature 0.7, top-p 0.8, seeds 42-46	none
Iter-RetGen-style	reference-inspired	Shao et al., 2023 (Iter-RetGen, arXiv:2305.15294); official code targets open-domain dense retrieval + text-davinci-003, not run here	BM25 k=3; round 2 query = draft——question over the same pool	2 retrieval calls, 2 LLM calls per question (T=2 protocol)	temperature 0.7, top-p 0.8, seeds 42-46	none
Self-Ask-style	reference-inspired	Press et al., 2023 (Self-Ask, arXiv: 2210.03350); official code uses text-davinci-003 + open-domain retrieval, not run here	BM25 k=3 per follow-up sub-question over the same pool	1 + n_sub retrieval calls, 2 + n_sub LLM calls ($n_sub_{¡}$=3)	temperature 0.7, top-p 0.8, seeds 42-46	none
CRAG-style	reference-inspired	Yan et al., 2024 (CRAG, arXiv: 2401.15884); non-official alternative (official release not run - see strong-baselines provenance table)	BM25 k=3; corrective query re-retrieved over the same pool when no passage is graded correct	1-2 retrieval calls, 1-2 LLM calls per question	temperature 0.7, top-p 0.8, seeds 42-46	none
PR-RAG (ours)	this work (PR-RAG)	This work (PR-RAG)	BM25 k=3 per instantiated sub-query over the same pool	T=3 iterations; bounded by the verifier stop or T	temperature 0.7, top-p 0.8, seeds 42-46	learned calibrator uses only the controller's own logged audit state (no extra corpus); trained offline with question-grouped folds

Table 22: Official strong baselines and the technical reason each could not be run in this environment. All four are recorded as `unavailable`; reproducible alternatives are noted in `baseline_audit.md`.

Method	Git commit / release	Repository	Required checkpoint(s)	Primary blocker
CRAG (official release)	`de7c2961ae624a1483a138c5798e1f6d0c4fb0e0` (2024-10-08)	`https://github.com/HuskyInSalt/CRAG`	fine-tuned T5-large retrieval evaluator (weights on Google Drive); generator = selfrag/selfrag_llama2_7b (HuggingFace)	No HuggingFace Hub access
S2G-RAG (official release)	not verified (no author-released code found)	no public official repository verified (arXiv/ACL pages list none)	trained S2G-Judge (backbone not identified in the release record); no public checkpoint verified	No verified author-released code or checkpoint
SURE-RAG (official release)	not verified	no public official repository verified (`github.com/microsoft/sure-rag` does not resolve; arXiv record lists no code)	trained claim-evidence verifier and (in ablation) a DeBERTa concat cross-encoder; no public checkpoint verified	No verified author-released code or checkpoint
Self-RAG (official release)	`1fcdc420e48f50a7d7ab1ece5494221b93252e99` (2024-03-19)	`https://github.com/AkariAsai/self-rag`	selfrag/selfrag_llama2_7b (or 13b) - critic+generator fine-tuned from Llama-2-7B with reflection tokens (HuggingFace); contriever-msmarco encoder + Wikipedia embeddings (facebook/contriever-msmarco + fbaipublicfiles)	No HuggingFace Hub access