

When Failures Propagate: Causal Failure Attribution in Agentic Retrieval-Augmented Generation

Lauren Pothuru

Anote

laurenpothuru100@gmail.com

Abstract

Agentic retrieval-augmented generation (RAG) interleaves retrieval, reasoning, and answer generation across multiple hops. A retrieval error at hop 1 can surface only as a wrong answer at hop 3, while later retrieval can also repair the trajectory. This paper introduces AgenticRAG-FP, an interventional benchmark for causal failure attribution in agentic RAG. The benchmark injects a certified fault at a specified hop, re-executes the downstream trajectory, and evaluates diagnosers against the known intervention. Its central question is whether a post-hoc trace still identifies the injected hop after the suffix changes. In the completed strict dense Claude Haiku 4.5 sweep on 80 three-hop MuSiQue questions, coverage-based diagnosis is 0.91 at hop 1 and 0.00 at hops 2 and 3 ($n=43, 36, 21$ failed trajectories). A smaller content-corruption study changes an answer-bearing or bridge fact in topically intact evidence. At depth 2, where 18 failed cases remain after filtering, coverage-based diagnosis is 0.00 and a frozen-hop counterfactual probe is 0.67 in an exploratory pooled comparison. Depth-3 content estimates are descriptive only because they contain three failed cases. These results make propagation depth an explicit evaluation axis for diagnosing agentic RAG failures while distinguishing broad evidence of post-hoc signal loss from small-sample method comparisons.

1 Introduction

Retrieval-augmented generation (RAG) conditions generation on retrieved evidence rather than only model parameters (Lewis et al., 2020). Agentic RAG extends this process across several decisions: a language model selects a sub-query, observes evidence, and decides whether to search again or answer, as in ReAct-style reasoning and acting (Yao et al., 2023). This enables multi-hop

question answering, but it also separates a visible end-to-end error from its cause.

Consider a three-hop trajectory. If the evidence at hop 1 is irrelevant, the agent can form a drifted sub-query at hop 2, retrieve evidence on the wrong premise, and produce an ungrounded final answer. An evaluator that inspects only the final trace may identify a late low-coverage hop or label the failure as answer generation, while missing the earliest causal fault. A different agent may retrieve compensating evidence at a later hop and answer correctly. Standard answer accuracy, retrieval recall, and final-answer correctness conflate an injected fault, its propagation, and its recovery.

Existing diagnostic methods can inspect traces or attempt local repair, but post-hoc traces do not provide a certified root-cause label. This leaves two evaluation questions unresolved: can a diagnoser recover a known intervention hop after the failure has propagated, and when do counterfactual repair probes supply information missing from the final trace? Answering either question requires an intervention that allows the agent to react and retains the intervention label.

AgenticRAG-FP provides this setting. It corrupts a trace prefix at hop h , runs the remaining trajectory with the same agent, and scores diagnoses against the certified intervention. The paper centers on how propagation changes the information available for exact-hop attribution. Its three contributions are:

1. An interventional benchmark that supports structural retrieval faults and certified content corruption. The latter changes a known fact span in otherwise relevant evidence and records the span pair, enabling deterministic absorbed, resisted, and derailed outcome labels without an LLM judge.
2. A depth-conditioned analysis of post-hoc signal. In the completed strict dense MuSiQue

sweep, coverage-based attribution falls from 0.91 at hop 1 to 0.00 at hops 2 and 3. This pattern concerns information available after suffix re-execution rather than end-to-end task accuracy.

3. A bounded comparison of counterfactual probes on content faults. Frozen local repair and suffix regeneration ask different questions about downstream dependence. The reported depth-2 content results illustrate this distinction, while the low-count depth-3 results are retained only as descriptive evidence.

2 Related Work

RAG and active retrieval. RAG combines a parametric generator with a non-parametric retrieval module (Lewis et al., 2020). Active and reflective variants decide when to retrieve during generation (Jiang et al., 2023; Asai et al., 2023). These approaches aim to improve answer quality by making retrieval responsive to uncertainty or the evolving generation. That same responsiveness creates the attribution problem studied here: a change to an early retrieval can alter the questions the agent asks later, rather than simply changing one fixed evidence set.

Agentic RAG and diagnosis. ReAct-style agents interleave reasoning traces with external actions (Yao et al., 2023). Their trajectories can be inspected, localized, and repaired, but inspection alone cannot establish whether a suspicious hop caused the final answer. Doctor-RAG uses coverage-gated localization and prefix reuse for diagnosis and repair (Jiao et al., 2026); Doctor-RAG instantiates its coverage-localization idea as a post-hoc baseline. The present evaluation differs in its target: it asks whether a diagnoser recovers a certified injected cause after an agent has generated a new suffix. This requires observing both the intervention and the resulting trajectory.

Multi-hop benchmarks. HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2021) provide multi-hop questions with supporting facts. FRAMES (Krishna et al., 2024) evaluates factuality, retrieval, and reasoning, and CRAG (Yang et al., 2024) includes dynamic, long-tail, and false-premise questions. These datasets provide question and corpus substrates, but their standard answer labels do not identify the hop that caused

a failed trajectory. The benchmark in this paper supplies that missing causal label through a controlled intervention.

Causal evaluation. The intervention perspective follows causal evaluation (Pearl, 2009): the fault and its hop are set before the resulting trajectory is observed. The object of intervention is not a model parameter or an internal activation. It is the trajectory-level context available to a retrieving agent. This framing makes recovery informative rather than inconvenient, because a correct answer after intervention is evidence that the agent did not propagate the fault to its output.

3 Task and Trace Model

Each example is a question-answer pair (q, y^*) together with a retrieval corpus $\mathcal{C}$. An agent produces a trace

$$\tau = (q, \{(q_h, D_h)\}_{h=1}^H, A, y, c),$$

where q_h is the sub-query issued at hop h , $D_h \subseteq \mathcal{C}$ is the retrieved evidence, A is the final answer, y is the reference answer, and c is total token cost. The trace records the information available at each decision and is sufficient to resume execution from an observed prefix.

Failure stages. Each hop has a stage

$$s \in \{\text{retrieval, tool, answer, none}\}.$$

Retrieval failures include empty or irrelevant evidence, query drift, false-premise evidence, stale evidence, and a wrong fact in otherwise relevant documents. Tool failures include missing actions and premature termination. Answer failures include empty answers, incorrect answers, and hallucinations that are unsupported by the retrieved evidence. The *none* stage denotes a trace without an identified failure at that hop. A diagnosis record contains a predicted stage $\hat{s}$, predicted hop $\hat{h}$, propagation flag, severity estimate, and root-cause description.

Identifiability. A diagnoser d is identifiable at depth h when it reliably recovers $\hat{h} = h$ for faults injected at hop h . This is a stricter property than detecting that the final answer is wrong or naming the broad failure stage. When an injected context changes the later trajectory, the decisive question is whether the final trace still contains information

Intervention	Effect at the selected hop
Empty retrieval	Replace retrieved evidence with $\emptyset$.
Irrelevant documents	Replace evidence with off-topic documents.
Query drift	Substitute a drifted sub-query, then retrieve.
False premise	Add a confidently wrong factual claim to evidence.
Stale evidence	Add outdated temporal evidence.
Content corruption	Change a certified fact span in otherwise relevant evidence.
Early termination	Force an answer from the preceding evidence.

Table 1: Interventions used by AgenticRAG-FP. In every live intervention, the downstream suffix is generated after the selected-hop corruption.

that distinguishes the injected hop from its downstream consequences.

The benchmark treats the selected failure type, stage, and hop as certified labels. Certification matters because a late low-coverage hop can be a cause, a downstream effect, or a benign consequence of an earlier intervention. Without a known intervention, these possibilities are observationally similar. The live comparison fixes the cause before execution and then evaluates the diagnosis only after the agent has had an opportunity to propagate or repair that cause.

4 Interventional Benchmark

Resumable live intervention. Given an executed hop prefix, the agent resumes at the next hop and continues to generate sub-queries, retrieve evidence, and decide when to answer. A live intervention replaces evidence or a sub-query at hop h , then resumes the agent at hop $h + 1$. The downstream suffix is therefore an agent response to the corrupted context, not a post-hoc trace edit. This distinction separates the causal effect of the intervention from the behavior of a static edited trace, which cannot reveal whether an agent would propagate or recover from a fault.

For every live intervention, the evaluation retains the intended failure family and selected hop as the target label. The intervention can be applied to retrieval evidence, a sub-query, or the decision to terminate. Structural faults test whether a visible retrieval anomaly remains localizable after the suffix changes. Content faults test the harder case in which the document still looks relevant and only its factual content is wrong. The same trace model therefore supports both readily observable and semantically hidden failures.

Intervention semantics. Empty and irrelevant retrieval remove answer support or replace it with off-topic evidence. Query drift changes the search intent before retrieval. False-premise and stale-evidence interventions preserve retrieved material but introduce a specific misleading claim. Early termination changes the available trajectory by requiring an answer before a later retrieval could occur. These families span missing evidence, misleading evidence, and altered agent decisions. They also differ in how visible they are to a post-hoc inspection, which makes them useful tests of whether a diagnoser is recognizing a local symptom or identifying an intervention after propagation.

All interventions preserve the selected-hop target while allowing the suffix to change. An answer error following an intervention is thus not assumed to be caused by an edited final answer. It is an observed outcome of the agent’s reaction to the modified prefix. This construction also permits a recovered trajectory to remain part of the evaluation: recovery is evidence about the agent’s resilience, while failed trajectories are the subset on which a root cause can be localized.

Certified content corruption. Structural interventions replace whole documents or queries, so they can be visible in the trace. Content corruption instead changes one fact in retrieved documents while preserving topical relevance. This models an error such as a stale fact, a corrupted document, or an upstream extraction error without making the selected retrieval obviously off-topic.

The changed span is selected by priority. An *answer fact* is selected when the gold answer occurs in the retrieved evidence and can be replaced by a certified wrong value. Otherwise, a *bridge entity* is selected when it occurs both in the hop evidence and in a later sub-query, but not in the original question. This identifies information that the agent carried forward. A *salient entity* or number supplies a fallback when neither condition holds. Numeric spans receive deterministic perturbations; entity spans receive in-domain distractors with no token overlap with the original span.

Corruption is seeded per trace and hop, so the selected replacement is reproducible. Samples with no certifiable span are skipped and counted rather than receiving an unverified change. The recorded original and corrupted spans certify the intervention and make deterministic answer-level

evaluation possible. They also make it possible to distinguish copying a wrong fact from producing another unrelated error.

Diagnosers. The main comparisons use a coverage-based localizer (Doctor-RAG), a trace-reading LLM judge (LLM-Judge), frozen-hop counterfactual repair (Propagation-Aware), and suffix-regeneration repair (Suf-Regen). The frozen-hop probe repairs one candidate hop, holds all other hops fixed, and tests whether the answer becomes correct. The suffix-regeneration probe repairs the candidate hop and then regenerates the remaining trajectory. Thus the two active probes differ only in their treatment of downstream evidence: one preserves the observed suffix and the other permits the agent to construct a new suffix in the repaired context. The judge can inspect semantic trace content but incurs additional token cost; coverage gating requires no re-execution but uses answer support as its signal. A rule-based diagnostic control and fuller implementation-neutral descriptions appear in Appendix A.

5 Metrics

Attribution identifiability. For injected traces that remain incorrect after suffix re-execution, exact-hop accuracy for diagnoser d is

$$\text{Acc}_d(h) = \frac{1}{|\mathcal{F}_h|} \sum_{\tau_i \in \mathcal{F}_h} 1[\hat{h}_{d,i} = h_i],$$

where $\mathcal{F}_h$ is the set of failed injected traces at depth h . Recovered trajectories are excluded because there is no final failure to attribute. Accuracy is accompanied by bootstrap 95% confidence intervals using $B=1,000$ resamples. Cells with fewer than 10 failed traces, including many CRAG cells, are reported as descriptive estimates rather than used for comparative conclusions. In particular, the main content-fault comparison excludes its depth-3 method ranking because only three failed cases remain after filtering.

The denominator is deliberately the set of traces that remain failed after live re-execution. Including recovered trajectories as incorrect diagnoses would conflate root-cause localization with the separate question of whether a fault reached the answer. Conversely, reporting only answer accuracy would discard the information carried by a certified intervention. This conditional metric makes the two outcomes visible: recovery mea-

sures resilience, and attribution measures localization among failures that actually propagated.

Localization beyond exact hop. Exact-hop accuracy is intentionally demanding when faults are causally entangled across hops. Stage accuracy asks whether $\hat{s} = s$. Hop-tolerance accuracy accepts $|\hat{h} - h| \leq 1$. Ancestor-hit rate counts a prediction when $\hat{h} \leq h$ and $\hat{s} = s$, giving partial credit when a predicted earlier hop is a causal ancestor of the injected fault. Mean absolute hop error summarizes the distance between $\hat{h}$ and h . These metrics separate broad failure recognition from localization of the intervention. They are secondary diagnostics and are not used for the central empirical claim.

Counterfactual recovery. Recovery is the fraction of live interventions whose corrupted trajectory still produces a correct answer:

$$\text{Recovery}(h) = \Pr[\text{correct}(A, y) \mid \text{do}(f, h)].$$

It is a separate robustness outcome. High recovery means that an intervention did not propagate into an end-to-end failure, not that a diagnoser localized it successfully.

Deterministic generation outcomes. For content faults, the certified span pair supports a judge-free answer-level classification:

absorbed	A contains changed-span tokens,
resisted	if $\text{correct}(A, y)$,
derailed	otherwise.

Correctness takes precedence when an answer contains both the gold and corrupted values. Query contamination records whether the corrupted span appears in a later sub-query, providing an observable channel for propagation. The outcome labels answer a different question from exact-hop attribution. They characterize the effect of a known content change on the generated answer even when the resulting trajectory is correct or has too little evidence for a localization comparison. Because the labels are computed from the certified span pair, a value can be checked consistently across backbones without introducing a second model’s interpretation of whether the error was copied.

Cost per correct diagnosis. For token-spending diagnosticians, total diagnosis tokens are divided by

the number of correct localizations. This distinguishes a method’s localization quality from deployability: an accurate method can still be unsuitable for online diagnosis if each correct localization requires a disproportionate amount of re-execution or judging.

6 Experimental Setup

Datasets and corpora. HotpotQA and MuSiQue are anchor multi-hop question-answering datasets with annotated supporting facts. FRAMES supplies variable-depth RAG questions with Wikipedia passage corpora, and CRAG contributes dynamic, long-tail, and false-premise questions. Headline FRAMES conditions use fetched and cached Wikipedia passage text. Conditions that use only link titles as a stand-in corpus are excluded from headline comparisons and flagged separately because that substitute changes retrieval fidelity.

Backbones and retrieval. The strict structural result reported in the main text uses Claude Haiku 4.5 with dense retrieval on three-hop MuSiQue examples and a four-probe budget. The repository also contains BM25, GPT-4o-mini, and local-model conditions, but incomplete or low-failure-count cells are not used for the main comparison. Static controls use BM25 and token-overlap retrieval; their results are retained as appendix controls, not as the central causal experiment.

Structural interventions and depth eligibility. The structural experiment injects empty retrieval, irrelevant retrieval, false-premise evidence, and stale evidence at hops 1, 2, and 3. For each sample and requested depth, all four intervention types apply to the same base trace, then the downstream suffix is re-executed. A depth-specific case is eligible only when the base trace answered correctly and the actual injected hop equals the requested depth. The primary structural summary uses the available records that support these checks. The matrix is not yet a complete factorial sweep across every backbone, retriever, and depth, so the results are reported as evidence about the observed post-hoc signal rather than as a universal method ranking.

Content-corruption conditions. The content-corruption study uses GPT-4o-mini and Claude Haiku 4.5 on HotpotQA and MuSiQue with BM25, $n=40$ base examples per condition, and

Depth	Doctor-RAG	LLM-Judge	Propagation-Aware	n
Hop 1	0.91 [0.81, 0.98]	0.26 [0.12, 0.40]	0.51 [0.37, 0.67]	43
Hop 2	0.00 [0.00, 0.00]	0.25 [0.11, 0.42]	0.25 [0.11, 0.39]	36
Hop 3	0.00 [0.00, 0.00]	0.43 [0.24, 0.67]	0.48 [0.29, 0.67]	21

Table 2: Exact-hop attribution for the strict dense Claude Haiku 4.5 MuSiQue sweep. Entries are accuracy [bootstrap 95% interval]; n is the per-depth count of failed injected traces.

depths 1–3. All five diagnosers are evaluated. Traces from one model family are judged by the other family, reducing self-diagnosis while retaining the same trace-level task. The resulting answer labels use the certified span pair and therefore do not depend on the LLM judge. Recovery leaves 44 failed cases at hop 1, 18 at hop 2, and 3 at hop 3 in the pooled analysis. The main text therefore treats the depth-2 comparison as exploratory and reports depth-3 method estimates only in the appendix.

Evaluation units. For structural conditions, a requested depth and intervention family define an evaluation unit over eligible base traces. For content conditions, the span selection strategy additionally determines whether a fault changes an answer fact, bridge entity, or salient fallback. All diagnosers receive the same post-intervention trace for a unit. Active probes then create their own counterfactual repairs from that trace, so their additional evidence is part of the diagnoser rather than a change to the benchmark label.

7 Results

7.1 Structural Attribution Across Depths

The completed strict structural sweep shows a loss of post-hoc coverage signal after suffix re-execution. Table 2 reports all three requested depths for dense Claude Haiku 4.5 on MuSiQue. The per-cell n values are failed injected trajectories eligible for exact-hop scoring; each estimate includes its nonparametric bootstrap 95% interval.

At hop 1, coverage identifies the deliberately visible structural fault. At later hops, coverage is 0.00 in both cells, while the judge and frozen-hop probe retain partial, overlapping-interval signal. The result does not support a general ordering between active and post-hoc methods. It shows instead that a regenerated suffix can erase the coverage signature of the injected hop while preserving some information that a more expensive probe can exploit.

Diagnoser	Hop 1 ($n=44$)	Hop 2 ($n=18$)
Doctor-RAG (coverage)	1.00 [1.00, 1.00]	0.00 [0.00, 0.00]
LLM-Judge (cross-family)	0.59 [0.43, 0.73]	0.89 [0.72, 1.00] [†]
Propagation-Aware	0.89 [0.80, 0.98]	0.67 [0.44, 0.89]
Suf-Regen	0.91 [0.82, 0.98]	0.11 [0.00, 0.28]

Table 3: Exact-hop attribution on content faults, pooled over GPT-4o-mini and Claude Haiku 4.5 on HotpotQA and MuSiQue with BM25. Column headings give the failed-case n . The hop-2 estimates are exploratory: their intervals quantify uncertainty within this pooled sample, not a definitive head-to-head comparison. The degenerate 0.00 interval has all bootstrap resamples equal to zero. [†]The judge’s deep result includes a positional prior. Depth-3 values ($n=3$) are reported descriptively in Appendix B.

7.2 Counterfactual Probe Performance

The available structural results do not establish a broad exact-hop advantage for frozen-hop repair over post-hoc diagnosis. The content-fault study instead provides a targeted examination of how two counterfactual scopes behave when the evidence remains topically relevant. Table 3 pools the four content conditions. At hop 2, coverage has no correct attributions in the observed sample, while frozen-hop repair reaches 0.67. These estimates are exploratory because the pooled depth-2 denominator is 18 failed cases.

The two counterfactual probes differ in how they treat downstream evidence. At depth 2, Suf-Regen is 0.11 because regenerating a suffix can retrieve clean evidence and remove the content fault under test. Frozen-hop repair prevents that change. In a small bridge-entity slice, Suf-Regen is 1.00 versus 0.73 for Propagation-Aware ($n=11$), which is consistent with the need to regenerate later hops that depend on a corrupted bridge. This slice is descriptive rather than a method-ranking result. The judge’s 0.89 at hop 2 should also be qualified: cross-family judging excludes self-recognition, but the predicted-hop distribution has a strong mid-trace preference. This makes the estimate unsuitable as a standalone measure of semantic localization.

7.3 Answer Outcomes After Content Corruption

Attribution evaluates only injected trajectories that remain failed. The certified span pair also makes it possible to measure what happened to every generated answer, including trajectories that recover. Table 4 reports absorbed, resisted, and derailed outcomes over all content-fault interventions. At

Depth	Absorbed	Resisted	Derailed	n
Hop 1	0.15	0.58	0.26	106
Hop 2	0.09	0.74	0.18	68
Hop 3	0.00	0.85	0.15	20

Table 4: Deterministic answer outcomes against certified corruption spans, pooled over the four content-corruption conditions. Per-backbone hop-1 rates are reported in Appendix C.

hop 1, 0.15 of cases are absorbed verbatim, 0.58 are resisted, and 0.26 are derailed. The later-depth rows contain all injected cases at their respective depths, not only the failed cases used in the attribution table.

Absorption is strongly answer-shaped. Of 22 absorbed cases, 20 are answer-fact corruptions; salient-entity corruptions are absorbed in 0 of 40 cases. Thus a corrupted value that resembles a direct answer is often repeated, whereas a corrupted chain link more often prevents a coherent answer. Query contamination is only 3 of 129 cases. The changed values therefore usually affect the reasoning that reaches an answer rather than being copied into a later sub-query. Monitoring query text alone would consequently miss most observed content-fault propagation.

8 Discussion

Post-hoc signal after suffix re-execution. Coverage gating attributes a failure to the earliest hop with low answer support. This can be effective when an intervention remains locally visible: empty evidence at hop 1 produces low coverage at hop 1 and the agent does not recover. The assumption fails once the agent re-executes a suffix. A later hop can mask the original fault by retrieving compensating evidence, or it can propagate the fault by issuing a query already biased by the corrupted context. Both outcomes replace the local signature at the injection point with evidence about the later trajectory.

The observed collapse is evidence of an identifiability limitation for the coverage signal available in these interventions. A final trace may show that an answer is unsupported, yet contain no reliable feature that distinguishes an error injected at one hop from consequences that emerge later. Improving a coverage threshold cannot restore information that the regenerated suffix has overwritten. The result does not establish an impossibility theorem for all agentic RAG systems; it identifies a failure mode that any post-hoc localizer must ad-

dress.

This interpretation does not require every later hop to be wrong. A later retrieval can be topically relevant and still be a downstream consequence of a faulty earlier premise. It can also partially repair an earlier failure without restoring an identifiable trace signature. The relevant distinction is between the information an agent needs to answer and the information a post-hoc diagnoser needs to identify the intervention hop. The former can be recovered by compensating evidence even when the latter has been lost.

Counterfactual repair under downstream dependence. Frozen-hop repair tests a useful causal question: does repairing candidate hop h change the answer when all later observed hops are held fixed? It can mislocalize when those later hops were generated from the original fault. Repairing the true early cause may not change the final answer because a later hop still carries the corrupted dependency. Conversely, repairing a later answer-bearing hop can restore correctness and receive credit despite being a downstream consequence. This is why broad stage identification can be more favorable than exact injected-hop attribution.

Suffix regeneration addresses the first problem by rebuilding downstream hops after repairing a candidate. It is therefore suited to bridge-entity corruption, where later queries and retrievals depend on the corrupted bridge. The same operation can obscure deep answer-fact corruption: regenerating a suffix retrieves clean evidence and removes the very fault under test. In that setting, frozen-hop repair preserves the corruption at the other hops and is the more diagnostic probe. The observed contrast between 0.11 for suffix regeneration and 0.67 for frozen-hop repair at depth 2, together with the bridge-entity slice, is consistent with this difference in counterfactual scope. The depth-2 and bridge-entity samples are small, so this is a mechanism-oriented interpretation rather than a definitive comparative result.

The two probes should therefore not be read as interchangeable versions of the same repair. Frozen-hop repair asks whether the observed downstream trajectory would support a correct answer after changing one local input. Suffix regeneration asks whether a repaired prefix can lead the agent to a different downstream trajectory. Each question is useful, but each can award credit to

a different position in a causal chain. Exact-hop evaluation exposes this difference directly, whereas an answer-only evaluation would collapse both repairs into a single correctness outcome.

Recovery as a separate outcome. Live interventions reveal a second property that static trace edits cannot measure: the agent can sometimes heal a corrupted trajectory through a later retrieval. This is not a failure of the intervention. It identifies a case in which the selected fault did not reach the final answer, so no root cause should be credited for an end-to-end failure. The distinction is particularly important for content corruption, where the executed trajectory contains a changed span but later retrieval can access clean corpus evidence. A persistent corpus-level corruption would test a different mechanism and remains future work rather than a feature of the present evaluation.

Implications for diagnosis. The appropriate repair probe depends on the path by which information flows through the trajectory. A single universal localizer must choose between preserving downstream evidence and regenerating it, and either choice can hide a different fault family. Future path-aware or cascaded diagnosis can select a probe based on suspected fault type, then report uncertainty when multiple hops remain causally entangled. Such a design targets the dependence structure that the present exact-hop evaluation exposes.

9 Limitations

The strict dense Claude Haiku 4.5 MuSiQue sweep has adequate failed-case counts at all three requested depths (43, 36, and 21). A matching strict dense HotpotQA sweep has only 9 and 2 failed trajectories at its valid depths because the agent recovers nearly all injected faults; it is therefore descriptive and excluded from the main comparison. The paper does not claim a complete backbone-by-dataset factorial evaluation, and it excludes CRAG from headline results because CRAG is normalized here as a single-turn benchmark.

The content intervention corrupts a trajectory copy while the retrieval corpus remains clean, allowing re-retrieval to heal some faults. This leaves 18 failed cases at depth 2 and 3 at depth 3 in the pooled content study; the depth-2 comparison is exploratory and no depth-3 method comparison is

made. A final content study should pre-specify a larger failed-case target at each depth and include persistent corpus corruption to distinguish recovery from a clean index from recovery caused by later reasoning. Exact-hop accuracy is intentionally strict, and stage-level scoring can credit a diagnoser that finds the failure type but not its injected hop. LLM judges are sensitive to prompting and token cost, and their apparent depth robustness can include a positional prior. Corpus construction also affects intervention and recovery behavior.

10 Conclusion

AgenticRAG-FP evaluates causal failure attribution by injecting a known fault and allowing the agent to generate its downstream response. In the completed strict dense Claude-MuSiQue sweep, coverage-based post-hoc attribution is 0.91 at hop 1 and 0.00 at hops 2 and 3, with 43, 36, and 21 failed trajectories respectively. The smaller content-fault study produces the same coverage pattern at depth 2, where frozen-hop repair reaches 0.67 in an exploratory pooled comparison. Its deterministic labels further show that 15% of hop-1 corruptions are absorbed verbatim. Broader benchmark and backbone coverage is needed before making a universal claim about agentic RAG diagnosis. Future work can evaluate persistent corpus corruption and path-aware diagnosis that chooses a probe by fault type.

References

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. [Self-RAG: Learning to retrieve, generate, and critique through self-reflection](#). *arXiv preprint arXiv:2310.11511*.
- Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [Active retrieval augmented generation](#). *arXiv preprint arXiv:2305.06983*.
- Shuguang Jiao, Chengkai Huang, Shuhan Qi, Xuan Wang, Yifan Li, and Lina Yao. 2026. [Doctor-RAG: Failure-aware repair for agentic retrieval-augmented generation](#). *arXiv preprint arXiv:2604.00865*.
- Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2024. [Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation](#). *arXiv preprint arXiv:2409.12941*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*.
- Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2 edition. Cambridge University Press.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2021. [MuSiQue: Multi-hop questions via single-hop question composition](#). *arXiv preprint arXiv:2108.00573*.
- Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, and 8 others. 2024. [CRAG: Comprehensive RAG benchmark](#). *arXiv preprint arXiv:2406.04744*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). In *International Conference on Learning Representations*.

A Diagnoser details

Rule detects empty retrieval, missing actions, empty answers, grounding overlap, and incorrect answers from the final trace. Doctor-RAG selects the earliest hop with insufficient gold-answer coverage and otherwise attributes the error to answer generation. LLM-Judge reads the hop-level trace and predicts a failing stage and hop. Propagation-Aware repairs a candidate hop, freezes the remaining hops, and tests the answer. Suf-Regen repairs a candidate hop and regenerates the suffix. All active probes test candidates in causal order and return the earliest repair that restores a correct answer.

B Descriptive depth-3 content-fault estimates

Only three content-corruption trajectories at depth 3 remain failed after filtering. The point estimates below are included for completeness, but no comparison or mechanism claim is based on them.

Diagnoser	Hop 3 ($n=3$)
Doctor-RAG (coverage)	0.00
LLM-Judge (cross-family)	0.33
Propagation-Aware	0.00
Suf-Regen	0.00

Table 5: Descriptive exact-hop estimates for the pooled depth-3 content-corruption cases. The denominator is too small for an inferential comparison.

C Additional content-fault outcome details

The hop-1 absorbed rates are 0.18 / 0.19 / 0.11 / 0.11 across the four conditions. Content-fault recovery ranges from 0.54–0.85, compared with 0.00–0.71 for structural faults. The content curves in Figure 1 show 0.75 for Propagation-Aware at hop 2 on both HotpotQA backbones, while coverage gating and suffix regeneration are zero; hop 3 is omitted when no injected trace remains failed.

At depth ≥ 2 , the discordant-pair comparison of Propagation-Aware with the best post-hoc envelope contains 1 versus 6 pairs. This count is too small for a McNemar test, so the direct content-fault comparison concerns Propagation-Aware versus coverage gating rather than Propagation-Aware versus all post-hoc methods.

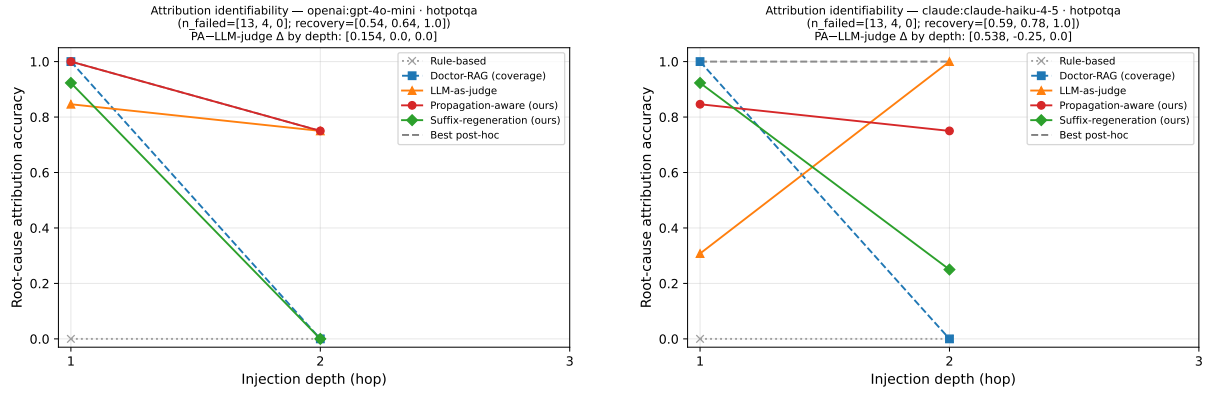


Figure 1: Content-corruption curves on HotpotQA with BM25. Left: GPT-4o-mini agent judged by Claude Haiku 4.5. Right: Claude Haiku 4.5 agent judged by GPT-4o-mini.