EAS-Bench: Evaluating Selective Evidence Admission in Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-augmented language systems receive evidence incrementally: from retrievers, tools, memories, files, and other agents. Most work asks whether the final answer is correct after retrieval. We ask a local question: when a new evidence item arrives, should the system admit it before answering or writing memory? We formalize *evidence admission safety* as a selective generation problem and introduce EAS-Bench, a 2,800-event benchmark derived from FEVER, HotpotQA, and 2WikiMultihopQA evidence-state perturbations. Each event compares a base answer with an event-conditioned answer and is labeled as benign, corrective, harmful, or unstable. Across held-out events, confidence-only admission is close to accept-all (75.8% safe precision at 98.1% coverage), while answer invariance removes harmful admissions but misses corrective evidence. Learned transition classifiers reach high harmful-admission AUROC (0.95–0.96) but exhibit fragile source transfer, especially across multihop QA traces. Under a risk-controlled threshold selected for zero harmful admissions on development data, a random forest admits 77.1% of held-out events with zero harmful flips. In live retrieval-agent simulations, admission gating reduces bad memory writes on claim-verification conflicts, but a HotpotQA counterexample shows that detection is not enough: repair must be task-sensitive.

1 Introduction

Retrieval-augmented generation (RAG) combines parametric generation with retrieved non-parametric evidence (Lewis et al., 2020). It is often described as a pipeline: retrieve documents, place them in context, generate an answer, and evaluate the answer. This abstraction hides an important decision. Evidence does not merely appear in a prompt; a system *admits* it into an evidence state from which the model answers, updates memory, or commits an external action.

This admission decision now appears in many NLP systems. Modern LLM applications retrieve documents, call tools, read memories, consume intermediate files, and exchange state with other agents. A misleading observation can override a correct answer. A redundant observation can leave the answer unchanged but perturb the reasoning trace. A corrective observation can repair a previously wrong answer. Existing work on knowledge conflict and robust RAG shows that retrieved evidence can be contradictory or noisy (Chen et al., 2022; Hong et al., 2024; Kim et al., 2025; Choi et al., 2025). Selective prediction asks when a model should abstain (Xin et al., 2021). We connect these lines by studying selective *evidence admission*: before answering, should the system trust the newly arrived evidence state?

The task is deliberately local. Instead of scoring only the final response, we consider an evidence-admission event $e = (q, C_0, a_0, z)$, where q is an input, C_0 is the current evidence context, a_0 is a base answer, and z is an incoming evidence item or tool observation. A policy chooses whether to ADMIT the event-conditioned answer, DEFER, or route to REPAIR. This lets us evaluate risk at the moment evidence changes the model's state.

We make three contributions.

- We formalize *evidence admission safety* as a selective generation problem with coverage, accepted-answer accuracy, harmful-admission rate, and missed-correction rate.
- We construct EAS-Bench, a 2,800-event benchmark from FEVER, HotpotQA, and 2WikiMultihopQA perturbation traces, covering redundant, conflicting, removed-support, and replaced-evidence events.
- We evaluate confidence gates, answer-

invariance gates, transition contracts, logistic classifiers, random forests, risk-controlled operating points, feature ablations, source-transfer settings, and downstream agent simulations. The results show that confidence is a weak admission signal, answer invariance is safe but conservative, and learned transition features detect many harmful events but do not solve repair.

2 Evidence Admission Safety

Events. Let q denote a claim or question, C_0 a current evidence state, a_0 the answer produced from C_0, and z an incoming evidence item. Let $C_1 = C_0 \cup \{z\}$ denote the event-conditioned evidence state, and a_1 the answer produced from C_1. An evidence-admission policy observes transition features

$$\phi(e) = \{c_0, c_1, c_1 - c_0, |c_1 - c_0|, \mathbb{1}[a_0 = a_1], d_{\text{trace}}, t_z\}.$$

where c_0, c_1 are confidence estimates, d_{trace} is a trace divergence score, and t_z is the event type. The policy returns

$$\pi(e) \in \{\text{ADMIT}, \text{DEFER}, \text{REPAIR}\}.$$

Labels. Using gold answers, each event is assigned one of six labels: *benign-stable*, *benign-unstable*, *corrective*, *harmful*, *wrong-stable*, or *wrong-unstable*. An event is safe to admit if a_1 is correct: benign events preserve correctness and corrective events repair a wrong base answer. An event is harmful if a_0 is correct but a_1 is wrong. Wrong-stable and wrong-unstable events are unsafe admissions but not harmful flips because the base state was already wrong.

Metrics. For a policy that admits a subset A_π, we report coverage

$$\text{Cov}(\pi) = |A_\pi|/|E|,$$

accepted-answer accuracy

$$\text{Acc}_{\text{admit}}(\pi) = \Pr[a_1 \text{ correct} \mid e \in A_\pi],$$

unsafe-admission rate, harmful-admission rate, and missed-correction rate. For scored classifiers, we also report AUROC and AUPRC for harmful-admission detection. Unsafe and harmful rates are conditional on admitted events: $\Pr[a_1 \text{ incorrect} \mid e \in A_\pi]$ and $\Pr[a_0 \text{ correct}, a_1 \text{ incorrect} \mid e \in A_\pi]$. Missed correction is measured over all events: $\Pr[e \notin A_\pi, a_0 \text{ incorrect}, a_1 \text{ correct}]$.

Admission label	Count	Admit-safe?
Benign-stable	1120	yes
Benign-unstable	622	yes
Corrective	150	yes
Harmful	334	no
Wrong-stable	181	no
Wrong-unstable	393	no

Table 1: EAS-Bench label distribution. Safe admission means the event-conditioned answer is correct. Harmful admission is the subset where the event flips a correct base answer to an incorrect answer.

Dataset	Source	Events	Harmful
FEVER	FEVER traces	1600	105
HotpotQA	Qwen traces	600	139
HotpotQA	GPT traces	300	53
2WikiMultihopQA	Qwen traces	300	37

Table 2: Benchmark sources. Multihop QA has a higher unsafe-admission rate than FEVER because evidence removal/replacement often breaks multi-hop support.

3 EAS-Bench

EAS-Bench is built from existing experimental traces in three NLP settings: FEVER claim verification (Thorne et al., 2018), HotpotQA multihop QA (Yang et al., 2018), and 2WikiMultihopQA (Ho et al., 2020). The benchmark contains 2,800 events. FEVER events append same-label evidence, opposite-label evidence, synthetic conflict, or duplicate gold evidence. HotpotQA and 2WikiMultihopQA events duplicate, remove, or replace evidence in multihop traces. Each event stores the base and event-conditioned correctness, answer equality, confidence shift, trace divergence, dataset, source, event type, and split.

The benchmark is not intended as a replacement for human annotation. Rather, it turns already gold-scored perturbation traces into a local selective-generation task. This makes it possible to study whether observable transition features can predict when incoming evidence should be trusted. We also run deterministic label audits: safe admission is exactly equivalent to event-conditioned correctness, harmful admission is exactly equivalent to a correct base answer flipped to an incorrect event-conditioned answer, and answer-invariant events are never harmful under the benchmark definition.

4 Methods

We evaluate fixed and learned admission policies.

Label	Event	Base	Event answer / outcome
Benign	duplicate	23 Nov. 1982	23 Nov. 1982; stable answer
Corrective	remove	no	yes; event repairs wrong base answer
Harmful	remove	23 Nov. 1982	1961; event breaks correct answer
Wrong stable	duplicate	Harold Young	Harold Young; both wrong

Table 3: Qualitative EAS-Bench examples from multihop QA traces, summarizing how local evidence perturbations change answer states.

Accept all. This is standard RAG behavior: every event-conditioned answer is admitted.

Confidence gates. We select a threshold on the development split for either high event confidence or low confidence shift. These baselines test whether self-reported confidence is sufficient for admission.

Answer invariance. The policy admits only if the normalized answer is unchanged after the evidence event. This is a strong conservative baseline for harmful-flip avoidance, but it cannot admit many corrective events.

Stable-transition contract. The policy admits only if the answer is invariant and both trace divergence and confidence shift are below thresholds selected on development data.

Learned classifiers. We train logistic regression and random-forest classifiers over base confidence, event confidence, confidence shift, trace divergence, event-confidence delta, answer equality, answer-change indicator, and event type. Thresholds are selected on development data and evaluated on held-out test data. We intentionally use shallow models so the features, not model scale, are the object of study.

Risk-controlled admission. In addition to F1-style threshold selection, we evaluate operating points that maximize development coverage subject to a maximum harmful-admission rate. This mirrors deployment settings where the user specifies an admissible risk budget.

5 Held-Out Results

We organize the empirical study around four questions. RQ1 asks whether confidence is sufficient for evidence admission. RQ2 asks whether answer invariance is a reliable safety certificate. RQ3 asks whether transition features generalize across datasets and model sources. RQ4 asks whether admission errors matter downstream when an agent writes memory or commits an action.

Table 4 shows that confidence is a poor evidence-admission criterion. A threshold on event confidence admits 98.1% of test events and improves safe precision by only 0.6 points over accept-all. Confidence shift is similarly weak. Trace divergence is more useful, improving safe precision to 84.6% while reducing harmful admission from 11.9% to 4.9%.

Answer invariance is the strongest fixed safety rule: it admits 77.2% of events with 88.5% safe precision and no harmful flips. The stable-transition contract provides a small additional precision gain but behaves similarly because answer change is the dominant harmful-flip signal. Its weakness is missed correction: 6.7% of all test events are corrective events it rejects. This exposes a real trade-off: a useful evidence policy should not merely defend a base answer; it should also recognize when new evidence repairs the system.

The learned classifiers show a different trade-off. They are better ranked detectors of harmful admission (logistic AUROC 0.951, random forest AUROC 0.956), but their selected thresholds admit more harmful cases than answer invariance. These results indicate that EAS-Bench is not solved by simply training a small classifier on transition features; objective choice and repair routing are central.

6 Cross-Source Transfer

To test generalization, we train learned admission classifiers on all but one source and evaluate on the held-out source. Table 7 reports safe precision and harmful-admission rate. Transfer to FEVER is strong, but transfer among multihop QA sources is weaker. In particular, a logistic model trained without Hotpot-Qwen admits almost all held-out Hotpot-Qwen events and collapses to 38.2% safe precision. The random forest is more conservative but still admits many unsafe events. The transfer result is important for evaluation: evidence admission has task- and source-specific structure that is not captured by a single shallow transition model.

7 Risk-Coverage Behavior

Selective generation systems are often useful because users can choose an operating point. On aggregate EAS-Bench, a logistic admission score

Method	Coverage	Safe precision	Unsafe rate	Harmful rate	Missed correction
Accept all	1.000	0.752	0.248	0.119	0.000
Event confidence	0.981	0.758	0.242	0.113	0.001
Confidence shift	0.986	0.755	0.245	0.115	0.000
Trace divergence	0.876	0.846	0.154	0.049	0.006
Answer invariance	0.772	0.885	0.115	0.000	0.067
Stable-transition contract	0.769	0.887	0.113	0.000	0.067
Logistic classifier	0.914	0.816	0.184	0.077	0.005
Random forest	0.892	0.836	0.164	0.055	0.006

Table 4: Aggregate held-out EAS-Bench results ($n = 1079$). Confidence-only admission performs close to accept-all. Answer invariance and the stable transition contract remove harmful admissions but miss corrections. Learned classifiers recover more coverage but admit more harmful events.

Features	Model	Harm AUROC	Harm AUPRC
Confidence only	Log.	0.688	0.197
Confidence only	RF	0.643	0.206
Event type only	Log.	0.700	0.238
No answer features	Log.	0.907	0.467
Trace + answer	Log.	0.957	0.636
Trace + answer	RF	0.953	0.606
All features	Log.	0.951	0.555
All features	RF	0.956	0.626

Table 5: Feature ablations for harmful-admission detection. Confidence and event type alone are weak; answer transition and trace divergence carry most of the signal.

Dataset	Method	Cov.	Safe	Harmful
FEVER	Accept all	1.000	0.848	0.076
	Answer inv.	0.826	0.924	0.000
	Trace div.	0.990	0.855	0.072
HotpotQA	Accept all	1.000	0.502	0.247
	Answer inv.	0.626	0.752	0.000
	Trace div.	0.667	0.712	0.041
2Wiki	Accept all	1.000	0.383	0.217
	Answer inv.	0.583	0.657	0.000
	Trace div.	0.583	0.657	0.029

Table 6: Within-dataset held-out results. Multihop QA is substantially harder than FEVER: accepting all evidence is unsafe, but conservative invariance also rejects many events.

reaches 94–96% safe precision below 60% coverage and admits zero harmful flips in those operating regions. At 80% coverage, safe precision drops to 88.3% and harmful admissions reappear. At 90% coverage, harmful-admission rate is 6.6%. This curve shows why a single accuracy number is misleading: the same detector can look strong or weak depending on the requested coverage.

Risk-controlled selection gives a more operational view. When thresholds are selected on development data to allow no harmful admissions, the random forest admits 77.1% of held-out events with 88.7% safe precision and zero harmful flips. Allowing a 5% development harmful-admission

Held-out source	Method	Cov.	Safe	Harmful
FEVER	Log.	0.968	0.883	0.046
FEVER	RF	0.966	0.885	0.044
Hotpot-GPT	Log.	0.730	0.676	0.073
Hotpot-GPT	RF	0.697	0.722	0.019
Hotpot-Qwen	Log.	0.992	0.382	0.234
Hotpot-Qwen	RF	0.652	0.552	0.061
2Wiki-Qwen	Log.	0.783	0.536	0.081
2Wiki-Qwen	RF	0.773	0.543	0.047

Table 7: Leave-source-out transfer. Strong ranking metrics do not guarantee safe operating points under source shift.

Target cov.	Actual cov.	Safe	Unsafe	Harmful
0.20	0.209	0.965	0.035	0.000
0.40	0.414	0.955	0.045	0.000
0.60	0.601	0.935	0.065	0.000
0.80	0.800	0.883	0.117	0.020
0.90	0.900	0.827	0.173	0.066
1.00	1.000	0.752	0.248	0.119

Table 8: Risk-coverage points for the aggregate logistic admission score. High-confidence operating regions admit no harmful flips, but harmful admission reappears as coverage approaches standard accept-all RAG.

budget raises coverage to 89.7%, but the held-out harmful rate rises to 6.1%. Thus the benchmark exposes a practical admission frontier rather than a single best classifier.

8 Downstream Agent Simulation

Evidence admission matters because admitted evidence can drive actions and memory writes. We therefore run live API-backed mini-agent simulations on claim verification and HotpotQA. These are not the main benchmark; they test whether EAS-style admission failures become downstream state failures.

Claim verification. In synthetic-conflict FEVER events, a boundary admission contract compares base and event-conditioned traces before allowing the incoming observation into the

Model/risk budget	Cov.	Safe	Harmful	Harmful n
Log., 0.00	0.770	0.887	0.001	1
Log., 0.02	0.857	0.862	0.034	31
RF, 0.00	0.771	0.887	0.000	0
RF, 0.01	0.776	0.883	0.004	3
RF, 0.05	0.897	0.831	0.061	59

Table 9: Risk-controlled admission. Thresholds are chosen on development data to maximize coverage under a harmful-admission budget, then evaluated on held-out test events.

Task/model	Policy	Cov.	Acc.	Unsafe	Bad mem.
FEVER/Gemini	None	1.00	0.910	9	9
FEVER/Gemini	Conf.	0.98	0.929	7	7
FEVER/Gemini	Contract	1.00	0.990	1	1
FEVER/GPT	None	1.00	0.767	14	14
FEVER/GPT	Conf.	0.93	0.804	11	11
FEVER/GPT	Contract	1.00	0.950	3	1
Hotpot/Gemini	None	1.00	0.725	11	11
Hotpot/Gemini	Conf.	0.98	0.744	10	11
Hotpot/Gemini	Contract	0.98	0.641	14	9

Table 10: Downstream live-agent simulations. Admission gating helps claim-verification conflicts but naive fallback repair fails on HotpotQA, separating detection from repair selection.

action path. With Gemini 2.5 Flash Lite on 100 events, the no-interceptor agent produces 9 unsafe actions and 9 bad memories; a confidence gate produces 7 and 7; the admission contract produces 1 and 1. With GPT-4o Mini on 60 events, the same contract reduces unsafe actions from 14 to 3 and bad memory writes from 14 to 1.

HotpotQA counterexample. On 40 HotpotQA misleading-tool events, the same fallback-style repair fails: the boundary contract reaches only 64.1% accepted accuracy, worse than the no-interceptor agent at 72.5%. This failure matters. In multihop QA, the base answer may already be unreliable, so falling back to the base state is not a valid universal repair. Evidence admission should therefore be decomposed into two decisions: detect unsafe admission, then select a task-appropriate repair such as retrieve-more, verifier call, defer, or fallback.

9 Related Work

Selective prediction and abstention. Selective prediction evaluates risk as a function of coverage and lets a model abstain on uncertain examples (Xin et al., 2021). Recent LLM work extends this idea to generation and knowledge-access settings, including contrastive decoding with abstention (Kim et al., 2025). EAS differs by placing selection at the evidence-state transition rather than only at the final output.

RAG conflict and noisy evidence. Prior work shows that retrieved evidence can conflict with parametric knowledge or with other retrieved passages, and that confidence may remain poorly calibrated under such conflicts (Chen et al., 2022). MacNoise studies counterfactual noise in retrieval-augmented models (Hong et al., 2024). Recent conflict-aware RAG methods train or prompt models to identify unreliable contexts (Choi et al., 2025). A contemporaneous preprint studies conflict detection and resolution as a pipeline problem (Wang et al., 2026). Our benchmark is complementary: instead of proposing a new reader architecture, it evaluates whether an incoming evidence transition should be admitted.

Evidence selection. Evidence retrieval methods select useful passages or sentence sets before generation; recent work frames evidence selection as search over evidence combinations (Sun et al., 2025). EAS focuses on a later runtime decision: given a base state and an incoming evidence item, does admission improve or damage the answer state?

Agent memory and tool use. Agent evaluations increasingly study memory, retrieval, and tool-mediated workflows. Our live simulations show the practical consequence for NLP reliability: unsafe evidence admission can become a bad memory write or an unsafe action. However, we intentionally keep the benchmark centered on evidence admission so the contribution remains measurable and comparable.

10 Reproducibility

All benchmark rows are deterministic transformations of stored traces. The analysis scripts select thresholds on development splits, evaluate on held-out test events, and report leave-source-out transfer without retuning on the held out source. Live-agent runs use fixed model identifiers, fixed temperature, and stored per-example traces so memory/action errors can be audited. To support replication after review, we will release anonymized construction scripts, benchmark tables, and run summaries.

11 Conclusion

We introduced evidence admission safety, a selective generation problem for RAG and retrieval-augmented agents. EAS-Bench shows that confidence-only admission is weak, answer invariance is safe but misses useful corrections, and learned transition classifiers face source-transfer and operating-point challenges. Downstream simulations show why the local decision matters: unsafe admission can propagate into actions and memory. Taken together, the results suggest that reliable retrieval-augmented systems should evaluate not only whether an answer is correct after retrieval, but whether each evidence transition deserved to be admitted in the first place.

12 Limitations

EAS-Bench is derived from perturbation traces rather than new human annotation. Its labels depend on gold answer correctness and automatic answer normalization, although deterministic label audits verify the intended base/event-correctness invariants. The event set is broad enough to expose failure modes but not exhaustive: web pages, tables, code repositories, and long-term conversational memories should be evaluated separately. Some transition features, especially trace divergence and event type, assume instrumentation that may not be available in every deployed system; the feature ablations are intended to make this dependence visible rather than hide it. Admission policies also create their own risks: an overly conservative gate can reject corrective evidence and preserve a wrong answer. Our learned baselines use shallow transition features; larger verifier models and NLI-style conflict detectors are important next baselines. Finally, the live-agent experiments are small and should be treated as downstream validation, not as the primary empirical claim.

References

Hung-Ting Chen, Michael J. Q. Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eunseong Choi, June Park, Hyeri Lee, and Jongwuk Lee. 2025. Conflict-aware soft prompting for retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26981–26995, Suzhou, China. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2474–2495, Mexico City, Mexico. Association for Computational Linguistics.

Hyuhng Joon Kim, Youna Kim, Sang-goo Lee, and Taeuk Kim. 2025. When to speak, when to abstain: Contrastive decoding with abstention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9710–9730, Vienna, Austria. Association for Computational Linguistics.

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

Hao Sun, Hengyi Cai, Yuchen Li, Xuanbo Fan, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, and Dawei Yin. 2025. Enhancing retrieval-augmented generation via evidence tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24116–24127, Vienna, Austria. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Chenyu Wang, Yingmin Liu, and Yang Shu. 2026. ConflictRAG: Detecting and resolving knowledge conflicts in retrieval augmented generation. *Preprint*, arXiv:2605.17301.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1040–1051, Online. Association for Computational Linguistics.

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*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.