RAGGATE: MITIGATING RAG COLLAPSE THROUGH ADMISSION CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) systems can suffer from RAG collapse when their own earlier outputs are retrieved and reused, causing later answers to concentrate on fewer alternatives. We study whether document admission control can mitigate this feedback while preserving adaptation to newly arriving external information. We propose RAGGate, which determines whether an incoming document becomes eligible for retrieval using a score intended to capture information beyond previously admitted evidence. We instantiate this framework in two ways: RAGGate-Geo uses an embedding-based residual, while RAGGate-LLM asks the generator whether a document adds information. In controlled feedback loops designed to isolate repeated self-retrieval, both variants more than halve answer collision probability and improve knowledge renewal relative to admitting all incoming documents. In a complementary real-text experiment, both variants reject most derived documents while still admitting many incoming human-written pages.

1 INTRODUCTION

Retrieval augmented generation (RAG) enables language models to generate answers using retrieved documents (Lewis et al., 2020; Izacard & Grave, 2021). Early work on question answering commonly evaluated systems on static corpora, which represent knowledge at a particular point in time. In practice, however, systems also use evolving corpora that incorporate new documents to answer questions about recent events (Liska et al., 2022). These additions can include the system's own outputs, as in agents that store and retrieve generated reflections (Park et al., 2023). Earlier outputs can then become evidence for later responses, creating a feedback loop.

Druck & Smith (2026) recently reported that retrieving a system's own outputs can concentrate later answers on fewer alternatives, a phenomenon they call *RAG collapse*. Unlike model collapse from repeated training on generated data (Shumailov et al., 2024), this can occur without updating model weights. At the retrieval stage, neural retrievers can favor generated documents over those written by people (Dai et al., 2024), and synthetic content can dominate results even when answer accuracy remains stable (Yu et al., 2026). Keeping original documents in the corpus does not ensure that they remain represented in retrieved evidence. Our reproduction on Multi-News (Fabbri et al., 2019) illustrates this problem: articles generated by the system occupy 97% of retrieved context by round 3, although all original documents remain stored (Figure 1).

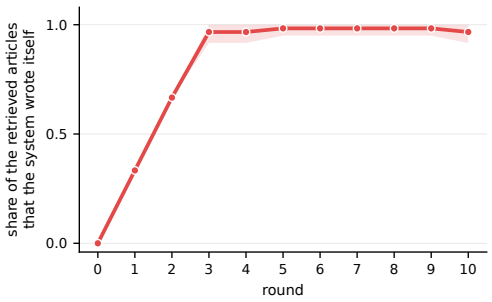


Figure 1: A RAG system's own outputs can dominate retrieval even when all original documents remain stored.

Existing work has not yet shown how to mitigate RAG collapse while still incorporating newly arriving external evidence. Senapati et al. (2026) test several mitigations at the retrieval and generation stages, but none reliably prevents collapse in their experiments. The collapse simulations of Druck

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

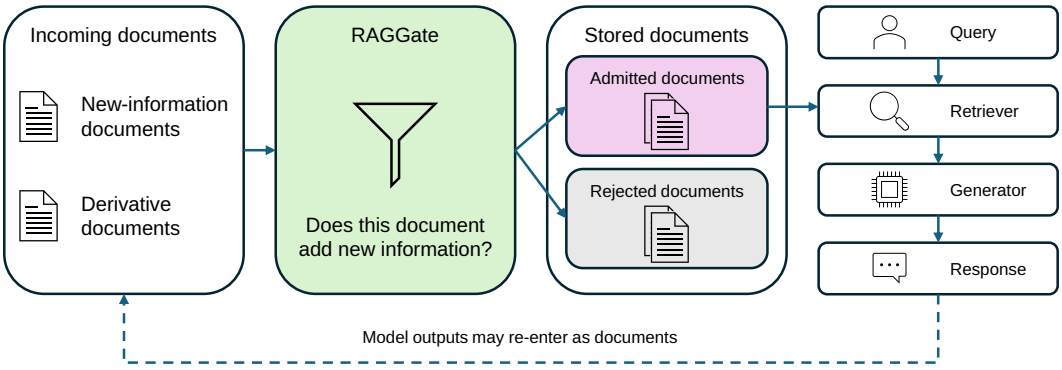


Figure 2: RAGGate separates storage from retrieval eligibility. Incoming documents are tested against admitted evidence before they can enter retrieval. The system's own derived outputs calibrate the threshold. The same admission procedure supports different scores.

& Smith (2026) also omit new external arrivals, leaving adaptation to new evidence untested. We therefore study two objectives jointly: limiting collapse caused by feedback and adapting to newly arriving external evidence, which we call *knowledge renewal*. Freezing the corpus blocks further additions from the system, but also excludes new external evidence. Authorship detectors can filter generated passages from retrieval results (Huang et al., 2026), but authorship does not determine whether a document adds information. A generated report may contain new observations, while a document written by a person may repeat existing material. Textual novelty detection instead asks whether incoming content adds information beyond earlier documents (Allan et al., 2000; Ghosal et al., 2022). This motivates using textual novelty, rather than authorship, as the admission criterion.

> Research question. How can a RAG system mitigate collapse while incorporating new evidence into its answers?

We address this question through document admission control, deciding which incoming documents become available for retrieval. We propose RAGGate, which makes this decision by assessing the information each document adds beyond previously admitted evidence (Figure 2). Only documents that score above a threshold become available for retrieval, although all documents remain stored. RAGGate sets this threshold using examples derived from the system's own outputs. It requires neither the derivation histories of incoming documents nor labeled examples of new information.

RAGGate separates scoring from calibration, so different measures of added information can use the same admission procedure. RAGGate-Geo measures the embedding residual outside the span of nearby admitted documents after removing shared background directions. RAGGate-LLM asks the generator whether a document adds information beyond the references. We also evaluate a nearest neighbour (NN) score within this framework and an independent, natural language inference (NLI) filter with a fixed threshold.

In controlled feedback loops, RAGGate more than halves answer collision probability and improves renewal compared with admitting all incoming documents. Real text experiments similarly show that RAGGate mitigates answer concentration while preserving access to incoming pages written by people.

Our contributions are:

1. We formulate admission control for evolving RAG stores as a joint problem of limiting feedback from repeated evidence and preserving renewal from incoming information (§3).
2. We propose RAGGate, which evaluates incoming documents against admitted evidence and calibrates admission thresholds using the system's own derived outputs. Geometric and model based scores instantiate the same framework (§4).
3. We show in controlled feedback loops and a real-text setting that document admission can reduce answer concentration while still allowing new evidence to affect later answers (§5).

2 Related work

Feedback loops and collapse from generated data. Repeated training on generated data can reduce output quality and diversity (Shumailov et al., 2024; Alemohammad et al., 2024). Gerstgrasser et al. (2024) show that accumulating real and generated data can limit this degradation compared with replacing real data. RAG feedback can change answer distributions without updating model parameters (Wang et al., 2025). Druck & Smith (2026) show that answers can concentrate on fewer alternatives even when original documents remain stored, because documents derived from earlier answers gain disproportionate influence. Senapati et al. (2026) reproduce collapse with open models and find that the preference for documents derived from answers is not fully explained by machine authorship, writing style, or whether the same model generated them. Retaining original documents therefore does not by itself preserve their influence on subsequent answers.

Filtering and diversification in RAG. Neural retrievers can favor generated content, reducing the representation of original sources in retrieved evidence (Dai et al., 2024; Yu et al., 2026). Huang et al. (2026) train authorship detectors on data that simulate evolving corpus contamination and use them to filter retrieved passages. They report reduced performance degradation under repeated contamination. Their filtering target is machine authorship, which does not determine whether a document adds information. Senapati et al. (2026) test detector based reranking, prompt optimization, paraphrasing, and agentic retrieval. None reliably prevents collapse in their experiments. Reranking remains ineffective even with an oracle detector when the retrieved candidate pool contains little original evidence to select.

Retrieval diversification addresses redundancy among selected documents. Maximal marginal relevance balances query relevance with dissimilarity to documents already selected (Carbonell & Goldstein, 1998), and DF-RAG adapts this balance to each query (Khan et al., 2026). These methods select context for a query. RAGGate controls which incoming documents become available for future retrieval, using added information as its admission criterion.

New information in evolving corpora. Prior work has studied whether an incoming document contains information not present in earlier documents. First story detection compares incoming documents with earlier stories to identify new events (Allan et al., 2000), while detection of near duplicates focuses on lexical or structural resemblance (Broder, 2000; Rodier & Carter, 2020). A document may nevertheless recombine existing information without closely resembling any single source. Ghosal et al. (2022) consider semantic redundancy across multiple sources using textual entailment. StreamingQA studies how question answering systems adapt as new knowledge arrives over time (Liska et al., 2022). RAGGate uses document novelty as an admission criterion rather than as an end task: the admission decision determines which documents can influence future retrieval in a feedback loop containing both external evidence and documents derived from the system's own answers.

3 Problem Formulation

We study how a RAG system can limit feedback from its own outputs while incorporating new external evidence.

Evolving RAG system. A document is a retrieval unit, either a complete document or a chunk of a longer document. Let $\mathcal{D}_t$ denote the set of stored documents at time t. For a fixed query q, the system selects retrieved documents $\mathcal{R}_t$ and generates an answer Y_t:

$$\mathcal{R}_t = \mathrm{Retrieve}(q, \mathcal{D}_t), \qquad Y_t \sim \mathrm{Generate}_\theta(\cdot \mid q, \mathcal{R}_t),$$

where $|\mathcal{R}_t| \leq k$ for a retrieval budget k. The generator parameters θ, retrieval procedure, and generation settings remain fixed. We fix q to isolate how document arrivals change its answer distribution. The system may serve other queries as well.

At each time step, the system receives incoming documents $\mathcal{X}_t$, which are added to the store:

$$\mathcal{D}_{t+1} = \mathcal{D}_t \cup \mathcal{X}_t.$$

Incoming documents may derive from past model outputs through copying, rewriting, or synthesis. A model output may produce no incoming document or several descendants. A derived document may also contain new external information, so derivation alone does not establish redundancy. Feedback occurs when past outputs influence later answers through retrieval. All past documents remain stored.

RAG collapse. Retrieving documents derived from past outputs can reinforce the system's earlier answers. Responses to the same open ended query may then concentrate on fewer alternatives, even when other evidence remains stored. Following Druck & Smith (2026), we call this loss of answer breadth due to feedback RAG collapse.

We measure concentration by the probability that two independently generated responses yield the same answer. An answer extractor ϕ maps each response to a discrete representation. For example, it can extract the set of three recommended names, ignoring their order. Two responses have the same answer when $\phi(Y) = \phi(Y')$. For a fixed query and a fixed set of documents eligible for retrieval, let

$$p_t(z \mid q) := \Pr[\phi(Y_t) = z]$$

be the probability of extracted answer z. The eligible set and system settings are implicit in this notation. The collision probability, also known as Simpson's concentration index (Simpson, 1949; Busa-Fekete & Syed, 2025), is

$$\text{Collision}_t(q) := \Pr[\phi(Y_t) = \phi(Y_t')] = \sum_z p_t(z \mid q)^2,$$

where Y_t and Y_t' are independent responses under the same query, eligible set, and system settings. Collision reaches 1 when every response yields the same extracted answer. Evaluation responses do not enter the store. This measure is closely related to the percentage of matching answers of Druck & Smith (2026, §7), with extracted answer equality replacing paraphrase judgments. Section 5.1 and Appendix C.2 specify the empirical measurement and extraction failure convention.

RAGGate: admission control. RAGGate determines which incoming documents become eligible for future retrieval. Let $\mathcal{A}_t \subseteq \mathcal{D}_t$ denote the admitted documents. We assess added information relative to $\mathcal{A}_t$, rather than all stored documents. Starting from $\mathcal{A}_0 = \mathcal{D}_0$, RAGGate selects

$$\mathcal{X}_t^{\text{adm}} = \text{RAGGate}(\mathcal{X}_t; \mathcal{A}_t), \qquad \mathcal{X}_t^{\text{adm}} \subseteq \mathcal{X}_t.$$

The stored and admitted sets evolve as

$$\mathcal{D}_{t+1} = \mathcal{D}_t \cup \mathcal{X}_t,$$
$$\mathcal{A}_{t+1} = \mathcal{A}_t \cup \mathcal{X}_t^{\text{adm}}.$$

Every incoming document is stored, but only admitted documents participate in retrieval. Previously admitted documents remain eligible. Retrieval and generation operate as

$$\mathcal{R}_t = \text{Retrieve}(q, \mathcal{A}_t), \qquad Y_t \sim \text{Generate}_\theta(\cdot \mid q, \mathcal{R}_t).$$

Under admission control, p_t is therefore induced by retrieval from $\mathcal{A}_t$. The gate does not take the evaluation query as input. It compares incoming documents with admitted evidence without access to their derivation histories. The system uses documents derived from its own outputs as calibration examples, without tracing the histories of other incoming documents. Scoring parameters and calibration data are omitted from the notation and specified in Section 4.

Objectives. We evaluate two objectives: limiting answer concentration caused by feedback and enabling knowledge renewal by allowing new external evidence to affect answers. Rejecting all arrivals prevents renewal, so low collision alone is insufficient. The controlled experiment evaluates both objectives in streams containing derived documents and documents with new external information. It supplies valid new facts by construction and measures whether the resulting answers include the newly introduced options (Section 5.1). The real text experiment additionally measures retention of incoming pages written by people, which is distinct from renewal in the answers (Section 5.3).

4 METHOD: RAGGATE

RAGGate scores each incoming document x for information beyond the admitted set $\mathcal{A}_t$. The system's own derived documents supply a threshold τ, and x becomes eligible for retrieval only if its score $s(x)$ exceeds τ. Two scores share this admission rule: RAGGate-Geo uses embeddings, and RAGGate-LLM asks the generator to compare the document with admitted evidence.

4.1 GEOMETRIC SCORE

RAGGate-Geo selects up to M similar admitted documents by cosine similarity to x, using the retriever's frozen encoder without background removal. Let $\mathcal{N}_M(x; \mathcal{A}_t)$ denote this reference set, using all admitted documents when fewer than M are available.

Under an approximately additive embedding model, content recombined from the selected references lies near their span. We use the residual outside this span as a proxy for added information.

Before computing the residual, we remove dominant directions shared across document embeddings. Removing common directions has been studied as an embedding postprocessing step (Mu & Viswanath, 2018; Arora et al., 2017). Here it reduces shared variation that may obscure differences in content. A frozen encoder maps a document to a unit vector $e(x) \in \mathbb{R}^d$. We use the transformed embedding $\tilde{e}(x) = (I - VV^\top)e(x)$, where the columns of V are m orthonormal principal directions estimated once from an unlabeled corpus of originals and system outputs. For $\mathcal{N}_M(x; \mathcal{A}_t) = \{a_1, \ldots, a_n\}$ with $n \le M$, the score is

$$s_{\text{geo}}(x) = \min_{\lambda \in \mathbb{R}^n} \left\| \tilde{e}(x) - \sum_{i=1}^{n} \lambda_i \tilde{e}(a_i) \right\|_2 = \left\| P_{U_t(x)^\perp} \tilde{e}(x) \right\|_2, \qquad U_t(x) = \text{span}\{\tilde{e}(a_1), \ldots, \tilde{e}(a_n)\}. \tag{1}$$

A larger residual indicates that the selected reference embeddings explain less of the incoming embedding. The local span has dimension at most M, regardless of store size. Appendix C.1 gives the transform and its geometric interpretation. Section 5.4 evaluates background removal.

4.2 MODEL SCORE

RAGGate-LLM asks whether an incoming document adds information beyond selected admitted documents. We rank the admitted documents by cosine similarity to x, using the retriever's frozen encoder without background removal. The prompt includes the eight most similar documents, or all admitted documents if fewer than eight are available. This cap bounds the number of references as the store grows.

The generator judges whether x contains an option or factual detail absent from these references. Its next token scores give $s_{\text{LLM}}(x) = \log p(\text{yes}) - \log p(\text{no})$. Scoring requires one generator forward pass, with no sampling or parsing of generated text. Appendix F.3 gives the prompt and token aggregation procedure.

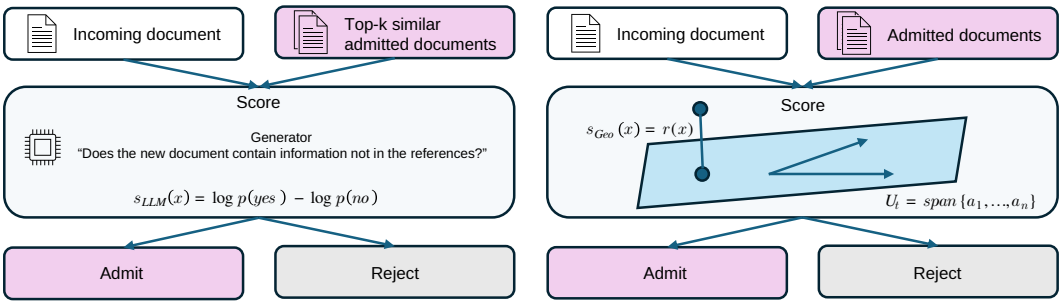
4.3 CALIBRATING τ

RAGGate sets its admission threshold using documents generated by the system from already admitted evidence, without introducing new external observations. These generated documents serve as reference examples of content generated without new external information.

Let $s_1, \ldots, s_{n_{\text{cal}}}$ denote their admission scores, and let $s_{(j)}$ be the j-th smallest score. Given an operational level $\alpha \in (0, 1)$, we set

$$\ell = \lceil (1 - \alpha)(n_{\text{cal}} + 1) \rceil, \qquad \tau = \begin{cases} s_{(\ell)}, & \ell \le n_{\text{cal}}, \\ +\infty, & \ell = n_{\text{cal}} + 1. \end{cases}$$

An incoming document is admitted only when $s(x) > \tau$. Lower values of α therefore make admission more conservative. No labeled examples of new information are required to construct the threshold.



(a) Model-based gate (RAGGate-LLM) (b) Geometric gate (RAGGate-Geo)

Figure 3: RAGGate scoring mechanisms. (a) RAGGate-LLM compares an incoming document with similar admitted documents. The generator's yes/no log odds give the admission score. (b) RAGGate-Geo embeds the incoming document and removes shared directions. Its score is the residual from the span of selected nearby admitted embeddings. The gate admits documents with scores above the calibrated threshold.

The calibration set is initialized with system-generated documents before the feedback loop begins. During the loop, each incoming document is scored against the current admitted set and evaluated using the threshold available before that decision. If the document was produced by the system, its score is then added to the calibration set, regardless of whether it was admitted, and the threshold is recomputed for subsequent decisions. Historical calibration scores are retained rather than recomputed after the admitted set changes. Appendix C.2 gives the initialization and update details.

We use the quantile rule above as an operational way to set the admission threshold from the system's own derived outputs, and evaluate its behavior in the adaptive feedback loop empirically.

5 Experiments

5.1 Experimental Setup

Datasets. The controlled experiment uses twelve authored worlds. Each world has one query and three original documents. Each original presents three options with one trait each, so a world has nine options. The query has one fixed form: "`Which three {topic} would you recommend, and why?`" The original documents were written by the authors in three fixed formats (Appendix F.1). With $k = 3$, the retriever returns all three at round 0. A derived document is the text that the system publishes each round. The system answers the query from the three retrieved documents. The generator rewrites this answer as a text. A new information document is the text that the system injects every five rounds. It introduces an option that no original mentions. Each incoming fact list combines this new option with two existing options and their traits. The generator rewrites all three together, so a document carrying new information also overlaps with the admitted evidence. This construction tests whether a filter can preserve the new option within otherwise familiar content.

The real-text experiment uses 54 distinct questions and their associated web pages, preserving the original twelve questions and adding 42. It tests whether admission control suppresses feedback from the system's own outputs while preserving incoming pages written by people.

Protocol. The controlled loop runs for 20 rounds. New information is injected every five rounds. At rounds $0, 5, \ldots$, we audit each policy by sampling five answers from its documents currently eligible for retrieval. These answers are used only for evaluation and are not fed back into the store. In every round, each policy other than Reject-all and the Oracle also produces one published answer, which is rewritten as a derived document and processed by that policy's admission rule. Each policy maintains its own store, initialized with the same three original documents.

Policy	Rejection rule	Threshold source
Admit-all	Never reject	–
Reject-all	Always reject	–
Oracle	Reject derived documents	Protocol labels
NN	Maximum document cosine similarity $\geq \tau_{\mathrm{NN}}$	Own derived outputs
RAD	Detector classifies document as machine generated	Detector cutoff: 0.5
NLI	Mean entailment coverage ≥ 0.5	Fixed cutoff: 0.5
RAGGate-Geo	Span residual $\leq \tau_{\mathrm{geo}}$	Own derived outputs
RAGGate-LLM	Novelty yes/no log odds $\leq \tau_{\mathrm{LLM}}$	Own derived outputs

Table 1: Admission rules for the baselines, Oracle reference, and two RAGGate variants. All policies use the same retrieval procedure.

Compared policies. We compare RAGGate with five baselines and an Oracle reference (Table 1). All policies retrieve the top three admitted documents by cosine similarity to the query. Admit-all and Reject-all admit every incoming document and none, respectively. NN uses raw embedding similarity, adapting first story detection (Allan et al., 2000) to the same calibration framework as RAGGate. NN and both RAGGate variants calibrate on their own derived outputs at $\alpha = 0.05$. RAD uses the authorship detector of Huang et al. (2026), trained on six separate worlds excluded from evaluation. NLI uses a pretrained entailment classifier to reject documents covered by admitted evidence, with a fixed threshold of 0.5 and no calibration on the feedback stream (Appendix C.3). The Oracle uses protocol labels to admit all new information and never publishes derived documents. Appendix C.2 gives training and calibration details.

Metrics.

- **Answer diversity** is one minus the observed answer agreement rate, the empirical counterpart of collision probability in §3. Among the ten unordered pairs of audit answers, a pair matches only if both yield the same set of three distinct names after normalisation. Appendix C.2 gives the extraction and failure handling rules.
- **Renewal success** is the fraction of options introduced so far that appear in at least one of the five audit answers. It measures whether new information reaches the recommendations.
- **Page admission rate** is the fraction of incoming pages written by people admitted in the real-text experiment. It measures document retention.

We also report rejection of derived documents and the derived share of retrieved context. Appendix C.2 gives parsing details.

Unless stated otherwise, answer metrics are evaluated at round 20 and admission decisions are accumulated over rounds 1 to 20. We pool seeds within each evaluation unit: three seeds per world in the controlled experiment and two seeds per question in the real-text experiment. We bootstrap over worlds or questions with 10,000 resamples and report 95% percentile intervals.

Implementation. The default generator is Qwen3-4B-Instruct-2507. The retriever and gates that use embeddings use bge-small-en-v1.5 as embedder, with cosine retrieval and $k = 3$. Appendix C.2 gives sampling settings and calibration details.

5.2 Controlled Experiment

Figure 4(a) compares answer diversity and renewal success. Tables 4 and 5 report full results and paired comparisons. Appendix A.2 extends the runs to 50 rounds.

At round 20, Admit-all has answer diversity 0.261 and renewal success 0.125, with documents written by the system occupying 55% of the retrieved context. Reject-all retains only the originals: its answer diversity is 0.753, but its renewal success is zero because it admits no new information

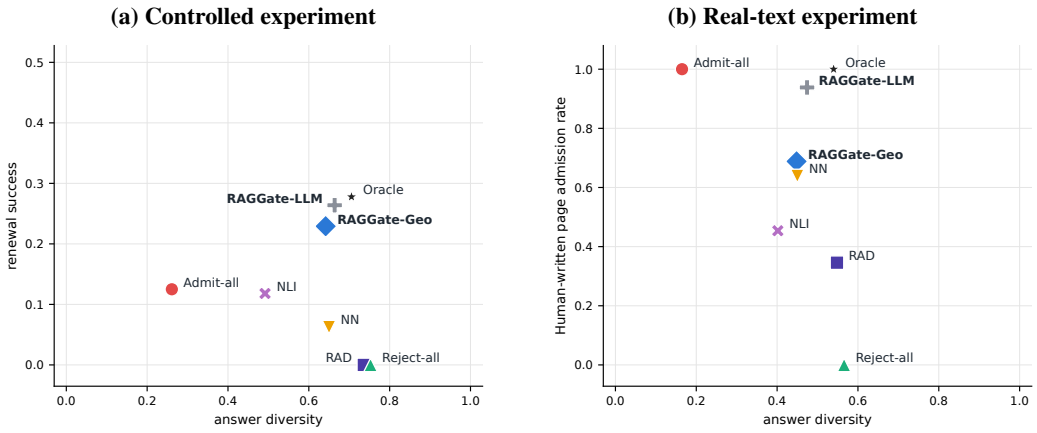


Figure 4: Eight admission policies at round 20: (a) twelve controlled worlds with three seeds and (b) 54 real-text questions with two seeds. Answer diversity is paired with renewal success in (a) and page admission rate in (b). Tables 4 and 7 report full results and confidence intervals.

documents. These endpoints show why diversity alone is insufficient: a policy must also allow incoming information to reach the answers.

RAD rejects all 720 derived documents and all 144 new information documents. It therefore leaves the store unchanged and prevents renewal, as Reject-all does. Both document classes are written by the same generator, so authorship does not distinguish feedback from the system's own outputs from new information. Appendix A.3 examines RAD's scores across decision thresholds. NN instead rejects documents that are too similar to an admitted document. It rejects 702 of 720 derived documents, but also 123 of 144 new information documents. Its answer diversity is 0.650, while renewal success is only 0.062. NLI rejects $669/720$ derived and $106/144$ new information documents, reaching diversity 0.492 and renewal 0.118. Thus, entailment coverage also discards incoming information in this setting.

RAGGate-Geo and RAGGate-LLM achieve renewal success of 0.229 and 0.264, compared with NN's 0.062. Their mean answer diversity is 0.642 and 0.664, compared with NN's 0.650. Both reject most derived documents while admitting more new information documents than NN. Both also achieve higher answer diversity and renewal success than NLI. The Oracle provides a reference at diversity 0.706 and renewal success 0.278. Tables 4 and 5 report the paired comparisons.

5.3 Real-Text Experiment

We expand the real-text evaluation to 54 distinct questions comparing entities spanning travel, food, software, and services, using the web references from Senapati et al. (2026). All eight policies run for 20 rounds with two seeds per question. Each question has three original pages and three reserved pages inserted at rounds 5, 10, and 15. The originals are chosen to cover the extracted entities. Each incoming page names an entity absent from the originals and has limited lexical overlap with them. Both seeds use the same pages, while each policy maintains its own evolving retrieval store. Figure 4(b) pairs answer diversity with page admission rate. Appendix C.4 describes selection without filtering by policy outcomes. Appendix B reports full results and supplementary renewal success measured with extracted entities.

Admit-all admits every incoming page, but its answer diversity falls to 0.165, with derived documents occupying 63% of retrieved context. Reject-all retains diversity 0.566 but admits no incoming pages. NN, RAD, and NLI admit 63.9%, 34.6%, and 45.4% of pages, respectively, with diversity 0.450, 0.548, and 0.402. These filters preserve more diversity than Admit-all but discard substantial fractions of incoming pages.

RAGGate-Geo and RAGGate-LLM admit $223/324$ and $304/324$ incoming pages (68.8% and 93.8%), while rejecting $2101/2160$ and $2083/2160$ derived documents. Their diversity is 0.448

Setting	Derived rejected	New information admitted	Derived share of context	In loop AUC	Answer diversity	Renewal success
Without background removal	708 / 720	19 / 144	0.06	0.824	0.603	0.076
With background removal	689 / 720	93 / 144	0.05	0.971	0.642	0.229
Oracle (reference)	–	144 / 144	0.00	–	0.706	0.278

Table 2: Effect of background removal in the controlled experiment. Appendix A provides metric definitions and confidence intervals (Table 6).

	RAGGate-Geo	RAGGate-LLM
Answer diversity / renewal success	0.642 / 0.229	0.664 / 0.264
Latency per incoming document	5.8 ms	53.9 ms
Generator tokens per incoming document	0	720
Model needed at admission	embedder (33M)	generator (4B)

Table 3: Controlled performance and admission cost of the two RAGGate variants. Latency is the median per document on one RTX 4090. Appendix C.2 gives the measurement procedure.

and 0.474. LLM retains more pages than NN, RAD, and NLI. Table 8 reports paired comparisons for both metrics.

LLM also achieves higher supplementary renewal success than NN, RAD, and NLI (0.205 versus 0.117, 0.050, and 0.081). The Oracle reference has diversity 0.540 and renewal 0.246 while admitting every page. The derived share of retrieved context is 11% for Geo and 12% for LLM.

5.4 Additional Analyses

Background removal. Background removal improves the gate's ability to retain new information (Table 2). It increases new information admission from 19 to 93 of 144 documents and renewal success from 0.076 to 0.229, while both settings reject most derived documents. The admission scores also separate derived and new information documents more clearly, with AUC increasing from 0.824 to 0.971. Appendix A.1 gives the score analysis and the basis for choosing $m = 30$.

Robustness and sensitivity. Across the tested embedder and generator settings, the geometric and LLM implementations achieve higher mean renewal success than the NN baseline. Appendix E reports both answer diversity and renewal success, with confidence intervals and admission counts. Appendix D reports admission and answer metrics for $\alpha \in \{0.01, 0.05, 0.10\}$ in the controlled experiment and the original real-text sample of twelve questions.

Computational cost. Table 3 reports controlled performance and admission cost for the two RAGGate variants. RAGGate-Geo needs one embedding and no generator call, with latency comparable to NN. RAGGate-LLM uses one generator pass over the incoming document and its references, about 720 input tokens per decision. We therefore use RAGGate-Geo as the default implementation and RAGGate-LLM to assess the benefit of using the generator to score documents. Appendix C.2 gives the measurement procedure.

6 Conclusion

RAG collapse shows that retaining original evidence is not enough when a system repeatedly retrieves documents derived from its own outputs. Our experiments indicate that document admission can reduce this feedback while still allowing new external information to affect later answers, by separating storage from retrieval eligibility and admitting documents based on the information they add. More broadly, our findings suggest that admission control is a useful design principle for evolving RAG systems, where it matters not only what is retrieved, but also what is allowed to become retrievable.

AI USE STATEMENT

We used generative AI tools to improve writing, search for relevant literature, assist with code debugging, and generate synthetic data used in the experiments. All research decisions, implementations, analyses, and final claims were reviewed and verified by the authors, who take full responsibility for the content of the paper.

ETHICS STATEMENT

This work involves no human subjects or personal data. The authored worlds use fictional names, and the experiments on real text use public datasets under their existing terms.

REPRODUCIBILITY STATEMENT

The submission includes selected code. Appendix C.2 provides sampling settings and calibration details. Appendix F provides the controlled document templates and prompts. Appendix C.4 specifies page selection and preprocessing for the real text experiment.

REFERENCES

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. In *ICLR*, 2024.

James Allan, Victor Lavrenko, and Hubert Jin. First story detection in TDT is hard. In *CIKM*, 2000.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.

Andrei Z. Broder. Identifying and filtering near-duplicate documents. In *Combinatorial Pattern Matching (CPM 2000)*, volume 1848 of *Lecture Notes in Computer Science*, pp. 1–10. Springer, 2000.

Robert Istvan Busa-Fekete and Umar Syed. Near-optimal algorithms for private estimation and sequential testing of collision probability. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 892–900. PMLR, 2025. URL https://proceedings.mlr.press/v258/busa-fekete25a.html.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. Neural retrievers are biased towards LLM-generated content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 526–537, 2024.

Gregory Druck and Ethan Smith. RAG collapse: LLM responses collapse when retrieved documents are self-authored. *arXiv preprint arXiv:2608.22118*, 2026. v1, August 22, 2026.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *ACL*, 2019.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024.

Tirthankar Ghosal, Tanik Saikh, Tameesh Biswas, Asif Ekbal, and Pushpak Bhattacharyya. Novelty detection: A perspective from natural language processing. *Computational Linguistics*, 48(1): 77–117, 2022.

Zhaoheng Huang, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. LLM-generated text may harm your retrieval! a robust detection strategy for retrieval-augmented generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31973–31988, 2026.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

Saadat Hasan Khan, Spencer Hong, Jingyu Wu, Kevin Lybarger, Youbing Yin, Erin Babinsky, and Daben Liu. DF-RAG: Query-aware diversity for retrieval-augmented generation. In *Findings of EACL*, 2026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://arxiv.org/abs/2005.11401.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13604–13622. PMLR, 2022. URL https://proceedings.mlr.press/v162/liska22a.html.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *ICLR*, 2018.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023. URL https://arxiv.org/abs/2304.03442.

Simon Rodier and Dave Carter. Online near-duplicate detection of news articles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pp. 1242–1249, 2020.

Riddhimaan Senapati, Gregory Druck, Fabeha Fatima, Ozel Yilmazel, Rati Rastogi, Chang Zeng, and Ethan Smith. Does retrieval augmented generation collapse when it retrieves AI-generated documents? OpenReview, https://openreview.net/forum?id=JHwIkYmtxb, 2026.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

Edward H. Simpson. Measurement of diversity. *Nature*, 163:688, 1949. doi: 10.1038/163688a0.

Tianyu Wang, Lingyou Pang, Akira Horiguchi, and Carey E. Priebe. LLM web dynamics: Tracing model collapse in a network of LLMs. *arXiv preprint arXiv:2506.15690*, 2025.

Hongyeon Yu, Dongchan Kim, and Young-Bum Kim. Retrieval collapses when AI pollutes the web. In *Proceedings of the ACM Web Conference 2026*, pp. 8745–8748, 2026.

A CONTROLLED EXPERIMENT: FULL RESULTS

Table 4 reports the results at round 20 for the controlled experiment in Section 5.2 with $\alpha = 0.05$ for NN and both RAGGate variants, and a fixed cutoff of 0.5 for NLI. We pool the three seeds within each of the twelve worlds. We report marginal 95% percentile confidence intervals from 10,000 bootstrap resamples over worlds. Paired differences use the same world for both policies. Appendix C gives the experimental details.

Policy	Answer diversity	Renewal success	Derived rejected	New information rejected	Δ diversity vs. Oracle	Δ renewal vs. Oracle
Admit-all	0.261 [0.169, 0.361]	0.125 [0.062, 0.201]	0 / 720	0 / 144	−0.445 [−0.539, −0.343]	−0.153 [−0.208, −0.090]
Reject-all	0.753 [0.708, 0.797]	0.000	–	–	+0.047 [+0.008, +0.086]	−0.278 [−0.340, −0.208]
Oracle (reference)	0.706 [0.672, 0.742]	0.278 [0.208, 0.340]	–	0 / 144	–	–
NN	0.650 [0.539, 0.733]	0.062 [0.014, 0.118]	702 / 720	123 / 144	−0.056 [−0.169, +0.029]	−0.216 [−0.278, −0.146]
RAD	0.736 [0.694, 0.778]	0.000	720 / 720	144 / 144	+0.031 [−0.011, +0.072]	−0.278 [−0.340, −0.208]
NLI	0.492 [0.408, 0.575]	0.118 [0.062, 0.181]	669 / 720	106 / 144	−0.214 [−0.311, −0.122]	−0.160 [−0.208, −0.111]
RAGGate-Geo	0.642 [0.572, 0.706]	0.229 [0.160, 0.299]	689 / 720	51 / 144	−0.064 [−0.136, −0.003]	−0.049 [−0.090, −0.007]
RAGGate-LLM	0.664 [0.603, 0.719]	0.264 [0.194, 0.333]	708 / 720	24 / 144	−0.042 [−0.106, +0.014]	−0.014 [−0.042, +0.021]

Table 4: Controlled results with 95% bootstrap intervals over worlds. The last two columns show paired differences from the Oracle. Reject-all keeps a frozen store without producing publications or processing arrivals. Dashes denote unprocessed documents or reference comparisons.

Table 5 reports comparisons with NN and NLI and between the two endpoint policies. Differences are computed within each world before pooling across worlds. Table 4 gives the comparisons with the Oracle.

Comparison	Δ diversity	Δ renewal
RAGGate-Geo minus NN	−0.008 [−0.089, +0.086]	+0.167 [+0.104, +0.222]
RAGGate-LLM minus NN	+0.014 [−0.097, +0.164]	+0.201 [+0.118, +0.292]
RAGGate-Geo minus NLI	+0.150 [+0.019, +0.278]	+0.111 [+0.049, +0.174]
RAGGate-LLM minus NLI	+0.172 [+0.064, +0.278]	+0.146 [+0.090, +0.208]
Reject-all minus Admit-all	+0.492 [+0.397, +0.578]	–

Table 5: Paired comparisons in the controlled experiment, with 95% bootstrap intervals over worlds. Positive values favor the first policy. A dash denotes an unreported comparison.

A.1 BACKGROUND REMOVAL

Table 6 gives the confidence intervals for the answer diversity and renewal success reported in Table 2. Both settings use the span residual and the same calibration procedure, with a cold start of 20 documents. They differ only in whether background directions are removed ($m = 0$ without removal, $m = 30$ with removal). In Table 2, "Derived share of context" is the mean fraction of the three retrieved documents that the system wrote itself. "In loop AUC" measures separation of derived and new information documents using the scores recorded at admission time.

Background removal raises the AUC of the scores recorded during admission from 0.824 to 0.971. Without removal, a threshold that rejects most derived documents also rejects most new information documents. We choose $m = 30$ from a static probe: increasing m from 10 to 30 raises the AUC on worlds used for fitting from 0.80 to 0.98 and the AUC when excluding the evaluation world from fitting from 0.72 to 0.83. Increasing m to 60 adds 0.01 to each.

Setting	Answer diversity	Renewal success
Without background removal ($m = 0$)	0.603 [0.503, 0.689]	0.076 [0.028, 0.132]
With background removal ($m = 30$)	0.642 [0.572, 0.706]	0.229 [0.160, 0.299]
Oracle (reference)	0.706 [0.672, 0.742]	0.278 [0.208, 0.340]

Table 6: Background removal results corresponding to Table 2, with 95% bootstrap intervals over worlds.

A.2 Long Horizon Evaluation

Figure 5 extends the original controlled experiment with seven policies from round 20 to round 50, using the same twelve worlds, three seeds, $\alpha = 0.05$, and cold start of 20 documents. One new information document arrives per world every five rounds through round 45, giving nine new options per world. Audits occur every five rounds. The RAGGate variants retain higher mean answer diversity and renewal success than Admit-all at round 50. At round 50, renewal success is 0.130 for Geo, 0.157 for LLM, and 0.164 for the Oracle. At each audit, renewal measures coverage of all options introduced so far, with the denominator growing at each injection.

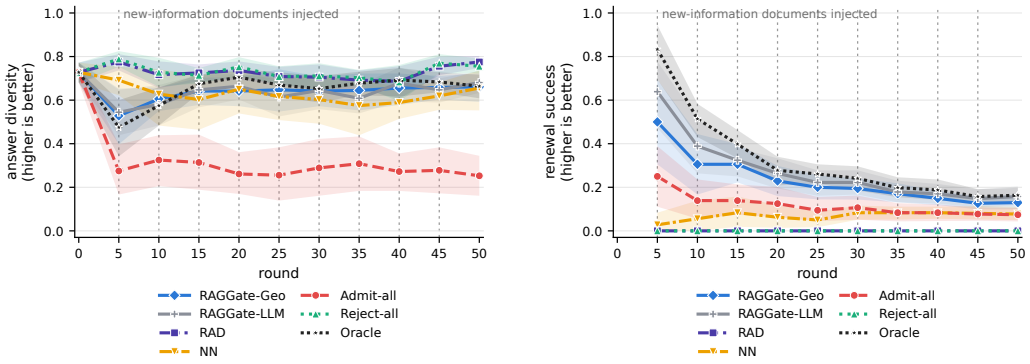


Figure 5: Controlled experiment over 50 rounds. Left: answer diversity. Right: renewal success. Shaded bands give 95% bootstrap intervals over worlds. Dotted lines mark new information arrivals.

A.3 RAD Decision Threshold

RAD's recorded scores provide limited separation between derived and new information documents in the controlled experiment (Figure 6). Across twelve worlds and three seeds, the 1,800 derived and 324 new information documents in its 50-round runs all receive probabilities of machine authorship above 0.998, so the default threshold of 0.5 rejects both classes entirely. Sweeping the threshold yields an AUC of 0.658 for distinguishing derived from new information documents. Increasing the threshold admits more new information documents, but also more derived documents. This is an offline threshold sweep of the recorded scores.

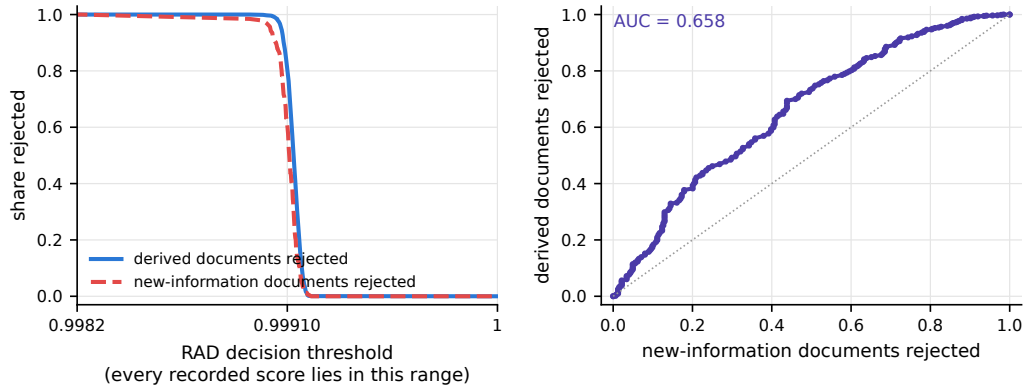


Figure 6: Offline RAD threshold sweep in the controlled experiment. Left: rejection rates by threshold. Right: rejection rates for the two document classes at the same thresholds. The dotted diagonal marks equal rejection rates.

B Real Text Experiment: Full Results

Table 7 reports results at round 20 on 54 questions, pooling seeds 17 and 43 within each question. NN and both RAGGate variants use $\alpha = 0.05$. NLI uses its fixed 0.5 cutoff. We report marginal 95% percentile intervals from 10,000 bootstrap resamples over questions. Appendix C.4 describes the expanded sample and protocol.

Each publishing policy processes 324 arrivals of pages written by people and 2,160 derived documents across the two seeds. Of the 324 pages, NN, RAD, NLI, Geo, and LLM reject 117, 212, 177, 101, and 20, respectively. Admit-all and the Oracle admit every page. Reject-all keeps a frozen store and does not generate publications or process arrivals. Its page admission is zero. Dashes denote documents not generated. The derived share of context is the mean fraction of retrieved documents at round 20 that are derived documents.

Policy	Answer diversity	Page admission	Renewal success	Derived rejected	Derived share of context
Admit-all	0.165 [0.104, 0.231]	1.000 [1.000, 1.000]	0.059 [0.028, 0.095]	0 / 2160	0.63
Reject-all	0.566 [0.474, 0.656]	0.000 [0.000, 0.000]	0.003 [0.000, 0.008]	–	0.00
Oracle (reference)	0.540 [0.456, 0.621]	1.000 [1.000, 1.000]	0.246 [0.190, 0.303]	–	0.00
NN	0.450 [0.374, 0.526]	0.639 [0.559, 0.722]	0.117 [0.077, 0.161]	2103 / 2160	0.11
RAD	0.548 [0.464, 0.630]	0.346 [0.259, 0.438]	0.050 [0.022, 0.082]	2160 / 2160	0.00
NLI	0.402 [0.329, 0.479]	0.454 [0.364, 0.546]	0.081 [0.045, 0.122]	2057 / 2160	0.19
RAGGate-Geo	0.448 [0.370, 0.525]	0.688 [0.608, 0.765]	0.098 [0.057, 0.143]	2101 / 2160	0.11
RAGGate-LLM	0.474 [0.391, 0.557]	0.938 [0.889, 0.978]	0.205 [0.158, 0.253]	2083 / 2160	0.12

Table 7: Expanded real text results with 95% bootstrap intervals over 54 questions. Renewal success is supplementary and depends on extracted entities.

Renewal success measures the share of extracted entities newly introduced by incoming pages that appear in the audit answers, using the parsing procedure in Appendix C.4.

Table 8 reports differences in diversity, page admission, and supplementary renewal, paired within each question.

Comparison	Δ diversity	Δ page admission	Δ renewal
RAGGate-Geo minus Admit-all	$+0.283$ $[+0.195, +0.368]$	-0.312 $[-0.392, -0.235]$	$+0.039$ $[-0.017, +0.096]$
RAGGate-Geo minus NN	-0.002 $[-0.078, +0.074]$	$+0.049$ $[+0.009, +0.093]$	-0.019 $[-0.063, +0.024]$
RAGGate-Geo minus RAD	-0.100 $[-0.178, -0.023]$	$+0.343$ $[+0.250, +0.438]$	$+0.048$ $[+0.001, +0.099]$
RAGGate-Geo minus NLI	$+0.046$ $[-0.031, +0.122]$	$+0.235$ $[+0.117, +0.349]$	$+0.017$ $[-0.022, +0.054]$
RAGGate-Geo minus Oracle	-0.092 $[-0.183, +0.002]$	-0.312 $[-0.392, -0.235]$	-0.148 $[-0.196, -0.102]$
RAGGate-LLM minus Admit-all	$+0.309$ $[+0.222, +0.394]$	-0.062 $[-0.111, -0.022]$	$+0.146$ $[+0.106, +0.188]$
RAGGate-LLM minus NN	$+0.024$ $[-0.066, +0.114]$	$+0.299$ $[+0.216, +0.383]$	$+0.088$ $[+0.047, +0.132]$
RAGGate-LLM minus RAD	-0.074 $[-0.167, +0.020]$	$+0.593$ $[+0.494, +0.688]$	$+0.155$ $[+0.102, +0.210]$
RAGGate-LLM minus NLI	$+0.072$ $[-0.025, +0.164]$	$+0.485$ $[+0.376, +0.586]$	$+0.124$ $[+0.077, +0.173]$
RAGGate-LLM minus Oracle	-0.066 $[-0.112, -0.020]$	-0.062 $[-0.111, -0.022]$	-0.041 $[-0.070, -0.015]$

Table 8: Paired differences on the expanded real text sample, with 95% bootstrap intervals over questions. Positive values favor the first policy.

C Experimental Details

C.1 Geometric Score Details

Background transform. The main experiments use $d = 384$ and remove $m = 30$ directions. To estimate V, we subtract the mean of the fitting embeddings and compute their leading principal directions. When applying the transform, we project out these directions without subtracting the mean or normalizing again. Appendix A.1 reports the analysis of the removal rank.

Reference span. We set the reference cap to $M = 64$. For each incoming document, $a_1, \ldots, a_n$ are the selected references in $\mathcal{N}_M(x; \mathcal{A}_t)$, with $n \leq M$. The same selection rule applies to calibration documents. The coefficients in Equation 1 are unrestricted. They can be negative and need not sum to one. The residual also equals the distance from the original embedding $e(x)$ to

$$\operatorname{span}\{e(a_1), \ldots, e(a_n), v_1, \ldots, v_m\},$$

where $v_1, \ldots, v_m$ are the columns of V. This follows because projecting each selected reference embedding away from V leaves the same augmented span, with the background directions orthogonal to the transformed embeddings. Background directions therefore act as additional reference vectors. Since $\dim U_t(x) \leq M < d - m$, the local span cannot fill the entire transformed embedding space. We retain the mean when applying the transform because translation does not preserve unrestricted linear combinations: subtracting a vector μ changes the reconstruction difference by $(\sum_i \lambda_i - 1)\mu$.

C.2 Controlled Experiment

New information documents. Fact list j contains one new option and its trait, plus existing options j and $j + 1$ and their traits, with existing indices taken modulo nine. The generator rewrites this list using the prompt in Appendix F.

RAD training. RAD uses the RoBERTa-base detector of Huang et al. (2026). We train it on six worlds excluded from evaluation, using their augmentation procedure with four modes and two rounds. It reads the query and document together.

Calibration. NN and both RAGGate variants each set their threshold from their own system's outputs. Before round 0, the system writes twenty derived documents per world from the three originals. It uses the loop's prompts and sampling settings. These documents are scored against the originals but never enter the store. Their scores give the initial threshold at $\alpha = 0.05$ under the rule in §4. Quantile resolution requires at least 19 calibration documents for a finite threshold at this level. Appendix D evaluates $\alpha \in \{0.01, 0.05, 0.10\}$ and gives the corresponding initial sample sizes.

Before round 0, RAGGate-Geo fits its background directions once on the pooled originals and initial generated documents from the twelve worlds. The initial generated documents thus enter both background fitting and calibration. In each later round, the published derived document is scored against the current admitted set and evaluated with the existing threshold. Its score then joins the calibration sample, regardless of admission, and the next threshold is recomputed. Historical scores are retained without rescoring against the updated admitted set. The transform uses no labels or separate worlds.

Admission cost. Table 3 reports median admission latency on one otherwise idle RTX 4090. We replayed the 288 incoming documents from a RAGGate-Geo run through round 20 (seed 17, $\alpha = 0.05$: 240 derived and 48 new information documents) through NN, RAD, and the two RAGGate gates, one document at a time. These gates used the same admitted store, with 5.4 documents per world on average. Reference embeddings were cached, and each incoming document was encoded afresh. RAGGate-LLM used one forward pass of the 4B generator, with a median prompt length of 720 tokens. NN and RAD had median latencies of 5.8 and 5.9 ms. For these gates, means were within 0.3 ms and 90th percentiles within 9 ms of the medians. Over 50 rounds, the store grew to 13.8 admitted documents per world and reached the cap of eight references. The same replay then gave RAGGate-LLM 67.7 ms and 957 input tokens per document, compared with 6.8 ms for RAGGate-Geo.

Answer parsing. We extract the bolded recommendation names and normalise each to an unordered set of lowercase alphanumeric tokens, without using a known option list. A pair matches only when both responses yield the same set of three distinct normalised names. Extraction failures count as nonmatches and remain in the denominator. With successful extraction, this agreement rate estimates collision probability. Matching against known options changes diversity by at most 0.02 in the controlled experiment. Renewal success matches names against options introduced so far. The real text version matches names to extracted entities.

Generation settings and lengths. Sampling uses temperature 0.8, top-p 0.95, and top-k 50. Generation allows at most 160 new tokens per answer and 192 per document. Storage truncates documents at 160 tokens. The embedder receives complete documents without chunking.

C.3 NLI Baseline with a Fixed Threshold

NLI is an independent entailment coverage filter. It uses the pretrained DeBERTa-v3-large classifier released by Moritz Laurer, trained on MNLI, FEVER-NLI, ANLI, LingNLI, and WANLI.[1] We keep its weights fixed. For each incoming document, we select up to eight admitted references using raw BGE cosine similarity. We split the incoming text into sentences with `pysbd` and split long sentences into windows of at most 160 model tokens. References are windowed to fit a 512-token pair of premise and hypothesis, including special tokens; both window types use an overlap of 32 tokens. All resulting pairs are evaluated without truncation.

For each incoming hypothesis window, we take the maximum entailment probability across the reference windows, then average these maxima over the incoming windows. NLI rejects the document when this mean is at least 0.5. An empty document is rejected. The classifier, reference cap, and

[1] Checkpoint: DeBERTa-v3-large NLI, revision `b3546ea6b034`.

threshold are identical in both experiments and were fixed before the runs. NLI uses neither RAGGate's background transform nor calibration on its own outputs, and receives no source or authorship labels.

For the controlled comparison, we retain the existing seven policy runs and add NLI using the same archived incoming documents, worlds, seeds, and generation settings. All 144 incoming texts match the archived inputs. A repeated Reject-all control produces exactly the same 900 audit answers. NLI is evaluated in the default 20-round comparisons.

C.4 Real Text Experiment

Data. We use the dataset of questions comparing entities from Senapati et al. (2026). The expanded sample preserves the original twelve questions and their pages, then adds 42 distinct questions in source order from the cached pool of questions with cached entity extractions. We require at least six usable web references. Three originals are chosen by greedy coverage of extracted entities. Each of three pages reserved for later insertion must name an entity absent from the originals and have longest common subsequence overlap measured over words below 0.5 with every original. We exclude additional questions with overlapping extracted option names across selected questions. These rules were fixed without inspecting policy outcomes. The questions cover travel, food, software, and services.

The loop's generator extracts and clusters entities using the dataset prompts. Rules filter activity headings and name variants. The original sample's manual removal of one journalist's byline is preserved. Each page is clipped to its first $1{,}500$ characters, about 200 words. The reserved pages are not individually verified for factual novelty. We use page admission as the primary retention measure and renewal measured with extracted entities as a supplementary measure.

Generation and evaluation. We rerun all eight policies on the 54 questions for 20 rounds using seeds 17 and 43. Three reserved pages arrive at rounds 5, 10, and 15, giving 162 reserved pages and 324 page arrivals across seeds. Each publishing policy produces 2,160 derived documents. Audits at rounds 0, 5, 10, 15, and 20 sample five answers per question and policy. Models, retrieval of the top three documents, and sampling parameters match the controlled experiment. Prompts omit references to a fictional setting and request web articles of 150 to 250 words. The generation limits are 200 tokens per answer and 400 per article, with storage capped at 1,000 tokens. NN and both RAGGate variants use $\alpha = 0.05$ and twenty cold start documents per question. Geo fits its background directions on the pooled originals and cold start documents from all 54 questions, separately for each seed. NLI uses the fixed rule above.

Answer diversity uses names as written, without the extracted option list. Extraction failures remain nonmatches. At round 20, failures range from 1 to 12 of 540 answers per policy. We average the two seeds within each question before bootstrapping over questions. The original calibration sweep with twelve questions and three seeds remains a separate analysis in Appendix D.2.

RAD checkpoint. The expanded real text experiment uses RAD retrained on the same archived 268 training and validation examples, with a 215/53 split, using RoBERTa-base, four epochs, learning rate 2×10^{-5}, batch size 16, and seed 0 set before initialization. No expanded evaluation examples enter detector training or validation. The controlled results retain the original detector. The expanded real text results use this reconstructed checkpoint.

D Sensitivity to the Operational Level

D.1 Controlled Experiment

Table 9 varies α while keeping worlds, seeds, and m fixed. Rejection and admission rates are computed over 720 derived and 144 new information documents per setting. Cold starts contain 12, 20, and 100 documents per world at $\alpha = 0.10$, 0.05, and 0.01, respectively. A finite threshold requires $\lceil (1-\alpha)(n+1) \rceil \leq n$. At $\alpha = 0.01$, RAGGate-LLM admits 65% of new information documents, RAGGate-Geo 29%, and NN 6%.

Gate	α	Derived rejected	New information admitted	Answer diversity	Renewal success
RAGGate-Geo	0.10	0.929	0.78	0.672 [0.619, 0.722]	0.222 [0.160, 0.278]
	0.05 (main)	0.957	0.65	0.642 [0.572, 0.706]	0.229 [0.160, 0.299]
	0.01	0.988	0.29	0.703 [0.661, 0.747]	0.146 [0.076, 0.222]
RAGGate-LLM	0.10	0.953	0.96	0.658 [0.589, 0.728]	0.271 [0.188, 0.347]
	0.05 (main)	0.983	0.83	0.664 [0.603, 0.719]	0.264 [0.194, 0.333]
	0.01	0.994	0.65	0.644 [0.564, 0.714]	0.236 [0.167, 0.299]
NN	0.10	0.957	0.22	0.617 [0.531, 0.697]	0.069 [0.035, 0.104]
	0.05	0.975	0.15	0.650 [0.539, 0.733]	0.062 [0.014, 0.118]
	0.01	0.996	0.06	0.744 [0.694, 0.794]	0.021 [0.000, 0.056]
Oracle	–	–	1.00	0.706 [0.672, 0.742]	0.278 [0.208, 0.340]

Table 9: Operational level in the controlled experiment. Brackets give 95% bootstrap intervals over worlds.

D.2 Real Text Experiment

Table 10 reports the sweep over $\alpha \in \{0.01, 0.05, 0.10\}$ on the original real text sample of twelve questions with three seeds. Each setting includes 720 derived documents and 108 pages written by people. Cold starts use 12, 20, and 100 documents per question at $\alpha = 0.10$, 0.05, and 0.01, respectively, as in the controlled experiment. At $\alpha = 0.01$, RAGGate-LLM admits 99% of pages written by people, RAGGate-Geo 86%, and NN 68%. At this level, the paired diversity differences from the Oracle are -0.006 $[-0.19, +0.21]$ for Geo and -0.039 $[-0.13, +0.04]$ for LLM.

Gate	α	Derived rejected	Pages written by people admitted	Derived share of context	Answer diversity	Renewal success
RAGGate-Geo	0.10	0.91	0.99	0.24	0.383 [0.233, 0.544]	0.181 [0.093, 0.278]
	0.05	0.94	0.96	0.19	0.400 [0.283, 0.510]	0.162 [0.075, 0.250]
	0.01	0.97	0.86	0.08	0.472 [0.311, 0.639]	0.148 [0.062, 0.243]
RAGGate-LLM	0.10	0.92	1.00	0.22	0.394 [0.233, 0.567]	0.186 [0.108, 0.272]
	0.05	0.96	1.00	0.13	0.414 [0.253, 0.581]	0.212 [0.122, 0.303]
	0.01	0.98	0.99	0.11	0.439 [0.268, 0.606]	0.222 [0.133, 0.314]
NN	0.10	0.96	0.75	0.18	0.344 [0.194, 0.500]	0.102 [0.028, 0.176]
	0.05	0.98	0.75	0.11	0.369 [0.208, 0.539]	0.139 [0.061, 0.222]
	0.01	0.99	0.68	0.06	0.486 [0.310, 0.661]	0.125 [0.043, 0.213]
Oracle	–	–	1.00	0.00	0.478 [0.317, 0.625]	0.268 [0.175, 0.347]

Table 10: Operational level on the original real text sample of twelve questions (three seeds). Brackets give 95% bootstrap intervals over questions.

E Robustness to Model Changes

We change one model at a time from the main controlled experiment. Sampling settings, worlds, seeds, prompts, $\alpha = 0.05$ and the cold start of 20 documents remain fixed. Tables 12, 13 and 14 report means and 95% bootstrap intervals over twelve worlds, with three seeds pooled within each world. Their final columns give paired differences from the Oracle.

Both RAGGate variants retain higher mean renewal success than NN across the tested model changes. Table 11 gives paired differences computed within each world.

Model change	Gate	Δ diversity	Δ renewal
MiniLM	RAGGate-Geo	-0.067 $[-0.181, +0.019]$	$+0.104$ $[+0.049, +0.160]$
MiniLM	RAGGate-LLM	-0.056 $[-0.136, +0.022]$	$+0.167$ $[+0.104, +0.222]$
Phi	RAGGate-Geo	-0.206 $[-0.389, -0.014]$	$+0.208$ $[+0.125, +0.299]$
Phi	RAGGate-LLM	-0.069 $[-0.197, +0.067]$	$+0.056$ $[+0.000, +0.118]$
Gemma	RAGGate-Geo	-0.053 $[-0.111, +0.003]$	$+0.194$ $[+0.146, +0.243]$
Gemma	RAGGate-LLM	$+0.039$ $[-0.019, +0.106]$	$+0.111$ $[+0.042, +0.194]$

Table 11: Paired differences from NN under model changes, with 95% bootstrap intervals over worlds. MiniLM replaces the embedder. Phi and Gemma replace the generator and RAGGate-LLM judge.

E.1 Alternative Embedder

We replace bge-small-en-v1.5 with all-MiniLM-L6-v2 (mean pooling, $d = 384$) for retrieval and every gate that uses embeddings. The background transform is refitted on the new embedder's cold start documents.

Policy	Answer diversity	Renewal success	Derived rejected	New information rejected	Δ diversity vs. Oracle	Δ renewal vs. Oracle
Admit-all	0.444 [0.331, 0.544]	0.076 [0.028, 0.125]	0 / 720	0 / 144	–	–
Reject-all	0.739 [0.689, 0.792]	0.000	720 / 720	144 / 144	+0.089 [+0.025, +0.158]	−0.201 [−0.243, −0.153]
Oracle (reference)	0.650 [0.564, 0.728]	0.201 [0.153, 0.243]	–	0 / 144	–	–
NN	0.700 [0.617, 0.775]	0.028 [0.007, 0.049]	705 / 720	114 / 144	–	–
RAGGate-Geo	0.633 [0.542, 0.714]	0.132 [0.076, 0.188]	691 / 720	51 / 144	−0.017 [−0.131, +0.081]	−0.069 [−0.118, −0.028]
RAGGate-LLM	0.644 [0.561, 0.728]	0.194 [0.146, 0.243]	702 / 720	20 / 144	−0.006 [−0.078, +0.061]	−0.007 [−0.028, +0.014]

Table 12: Controlled experiment with all-MiniLM-L6-v2 embeddings. Brackets give 95% bootstrap intervals over worlds.

E.2 Alternative Generators

We replace Qwen3-4B-Instruct-2507 with Phi-3.5-mini-instruct or Gemma-2-9B-it for answers, derived documents, new information documents and the RAGGate-LLM judge. RAD keeps the detector trained on Qwen outputs from worlds excluded from evaluation. Gemma's chat template has no system role, so its system prompt is prepended to the first user turn.

With Phi and Gemma, derived documents occupy 82% and 70% of Admit-all's retrieved context. RAGGate-Geo reduces these shares to 12% and 9%. Tables 13 and 14 report answer diversity, renewal success, admission counts, and paired differences from the Oracle for both gates.

Policy	Answer diversity	Renewal success	Derived rejected	New information rejected	Δ diversity vs. Oracle	Δ renewal vs. Oracle
Admit-all	0.181 [0.083, 0.297]	0.062 [0.028, 0.104]	0 / 720	0 / 144	–	–
Reject-all	0.850 [0.800, 0.894]	0.000	720 / 720	144 / 144	+0.183 [+0.081, +0.283]	−0.396 [−0.465, −0.326]
Oracle (reference)	0.667 [0.575, 0.750]	0.396 [0.326, 0.465]	–	0 / 144	–	–
NN	0.722 [0.619, 0.819]	0.049 [0.021, 0.076]	703 / 720	133 / 144	–	–
RAD	0.856 [0.786, 0.908]	0.000	720 / 720	144 / 144	–	–
RAGGate-Geo	0.517 [0.383, 0.650]	0.257 [0.181, 0.340]	688 / 720	63 / 144	−0.150 [−0.283, −0.008]	−0.139 [−0.201, −0.083]
RAGGate-LLM	0.653 [0.533, 0.764]	0.104 [0.049, 0.160]	687 / 720	106 / 144	−0.014 [−0.139, +0.108]	−0.292 [−0.354, −0.236]

Table 13: Controlled experiment with Phi-3.5-mini-instruct as generator and RAGGate-LLM judge. Brackets give 95% bootstrap intervals over worlds.

Policy	Answer diversity	Renewal success	Derived rejected	New information rejected	Δ diversity vs. Oracle	Δ renewal vs. Oracle
Admit-all	0.233 [0.131, 0.344]	0.104 [0.035, 0.188]	0 / 720	0 / 144	–	–
Reject-all	0.758 [0.619, 0.867]	0.000	720 / 720	144 / 144	−0.031 [−0.167, +0.069]	−0.319 [−0.396, −0.250]
Oracle (reference)	0.789 [0.744, 0.831]	0.319 [0.250, 0.396]	–	0 / 144	–	–
NN	0.728 [0.639, 0.806]	0.035 [0.014, 0.056]	701 / 720	129 / 144	–	–
RAD	0.733 [0.619, 0.828]	0.000	720 / 720	144 / 144	–	–
RAGGate-Geo	0.675 [0.608, 0.742]	0.229 [0.181, 0.278]	679 / 720	35 / 144	−0.114 [−0.164, −0.069]	−0.090 [−0.153, −0.035]
RAGGate-LLM	0.767 [0.703, 0.828]	0.146 [0.076, 0.222]	705 / 720	95 / 144	−0.022 [−0.094, +0.056]	−0.174 [−0.236, −0.111]

Table 14: Controlled experiment with Gemma-2-9B-it as generator and RAGGate-LLM judge. Brackets give 95% bootstrap intervals over worlds.

F Document Templates and Prompts

The templates and prompts below are used in the controlled experiment. Real text adaptations are described in Appendix C.4.

F.1 Original Document Templates

Every world uses the same three templates, one per original. We reproduce the templates verbatim.

Template 1: Blog Post

Last month I finally tried three of the `{topic}` people keep mentioning. `{A}` was the surprise: `{trait A}`, and nobody made me feel like a beginner. A week later I went to `{B}`, which has `{trait B}`, and then to `{C}`, which has `{trait C}`. All three were cheaper than I expected and easy to find, and I would happily go back to any of them. Which one you prefer probably depends on your mood.

Template 2: Community Bulletin

Community bulletin, new listings under `{topic}`. `{A}`: `{trait A}`. Open to all, modest prices, no prior experience required. `{B}`: `{trait B}`. Also open to all, modest prices, clearly organized for first-time visitors. `{C}`: `{trait C}`. Open to all as well, with the same modest prices and a relaxed pace. The bulletin lists entries alphabetically and does not rank them. Contact the organizers directly for seasonal hours.

Template 3: Review

Four stars for all three, honestly. I was looking for `{topic}` and ended up at `{A}` first: `{trait A}`, welcoming, and nobody rushed us. Then `{B}` the following weekend, which has `{trait B}` and cost about the same. `{C}` came last, with `{trait C}` and the same easy welcome. I could not pick a favourite and I do not think you need to. Go to whichever is closer.

F.2 Generation Prompts

All prompts are in English, and we reproduce system messages verbatim. The `{register}` slot has three possible values: (i) a first-person blog post about your own visits, (ii) a community bulletin entry, listing the places without ranking them, or (iii) a short review, as on a review site. A hash of the seed, world id, and round selects the register.

Prompt 1: Answer (audit and publication)

System message. Answer the question using only the supplied reference notes about a fictional setting. Recommend exactly three distinct options. Use three bullet points, each starting with the option's exact name in bold, followed by one brief sentence explaining the recommendation. Do not include an introduction, a conclusion, or additional options.

User message. The user message lists the three retrieved documents as `[1]`, `[2]`, `[3]` in an order determined by the seed, then the query.

Prompt 2: Derived document (from a publication answer)

System message. Turn the supplied answer into a standalone text of 70 to 100 words for readers interested in the question, written as `{register}`. Preserve its recommended options and their stated reasons. Do not add new options or factual details. Do not mention the question itself. Use ordinary prose rather than a bullet list.

User message. The user message gives the query and the answer.

Prompt 3: New information document (from a fact list)

System message. Turn the supplied facts into a standalone text of 70 to 100 words for readers interested in the question, written as `{register}`. Preserve the named options and their stated details. Do not add new options or factual details. Do not mention the question itself or the fact list. Use ordinary prose rather than a bullet list.

User message. The user message gives the query and the fact list.

F.3 Admission Prompt

Reference selection uses the incoming document as the similarity target, independently of the query used for answer generation. The same selection rule is applied when scoring calibration documents.

The judge reads the next token log probabilities without generating an answer. For each of `yes` and `no`, the implementation takes the first token ID of four encodings: lowercase and initial capital, each with and without a leading space. It aggregates the selected log probabilities with the log of the sum of their exponentials and subtracts the aggregate for `no` from that for `yes`.

Prompt 4: RAGGate-LLM judge

System message. You compare a new note against reference notes about a fictional setting. Answer with a single word, yes or no.

User message. The user message lists the selected admitted documents in decreasing cosine similarity to the incoming document, numbered from `[1]`. The incoming document and the following question come after these references:
Does the new note contain any option or factual detail that is not in the reference notes? Answer yes or no.