

Decoupling metadata provenance from generative faithfulness: A dual-stage ground-truth cross-verification framework for auditable retrieval-augmented generation

Chinedu Okafor^{1*} and Amina Bello¹

¹Department of Computer Science, University of Lagos,
Lagos, Nigeria

Abstract

Retrieval-augmented generation had become a common way to deploy large language models for legal analysis, clinical documentation, and regulatory auditing. Existing automated evaluations judged the pipeline as one black box. They mixed retrieval errors with generation errors and did not check whether a cited page was the page that contained the evidence. This study built a dual-stage ground-truth cross-verification framework. Stage 1 scored retrieval against document and page anchors before any answer was written, using document hit rate, page hit rate, mean reciprocal provenance rank, and span containment. Stage 2 scored semantic correctness and citation fidelity, then placed each question in one of four diagnostic quadrants. The test set contained 1,200 audit questions over 8,520 pages of pharmaceutical validation standards, including United States Food and Drug Administration Title 21 Code of Federal Regulations Part 11 and European Union Annex 11. On the strongest retriever, document hit rate at five was 95.6 percent and page hit rate at five was 85.3 percent, a gap of 10.3 percentage points, and mean span containment was 0.88. With that retriever, a reranker, and GPT-4o, 80.1 percent of answers were fully grounded and 8.1 percent were correct in meaning but cited the wrong page. The same generator with sparse retrieval produced fabricated citations on 19.6 percent of questions. When the true passage sat in the middle of the prompt, extraction success fell by 19.1 percentage points for Llama-3.3-70B-Instruct. Separating provenance from generation therefore gave an auditable account of where these systems fail.

Keywords: Retrieval-augmented generation; Citation faithfulness; Provenance verification; Hallucination detection; Regulatory compliance; Evaluation benchmarks

1. Introduction

Retrieval-augmented generation combines a parametric language model with a non-parametric index over an external corpus. Lewis et al. (2020) showed that this coupling lets a generator answer from documents it did not memorize, which is the property enterprises need when the source of truth is a private collection of contracts, protocols, or validation records. In that setting the system is expected to do three things at once: limit answers to the curated corpus, keep institutional text inside the organization, and point each claim at a source a reviewer can open.

Those expectations become strict once the same architecture is placed in life-science validation, legal discovery, or financial-risk governance. A response can be fluent and still be unauditably. The obstacle is no longer only whether a model invents a fact. It is whether an evaluation can say which layer failed, and whether a citation names the page that actually contains the evidence. Figure 1 states the audit this paper applies to that problem. A gold record fixes the question, the answer, the document, the page, and the evidence span. Retrieval is scored against the document and the page before any answer is written. Generation is then forced to emit an answer and a citation, and those two outputs are checked on separate axes.

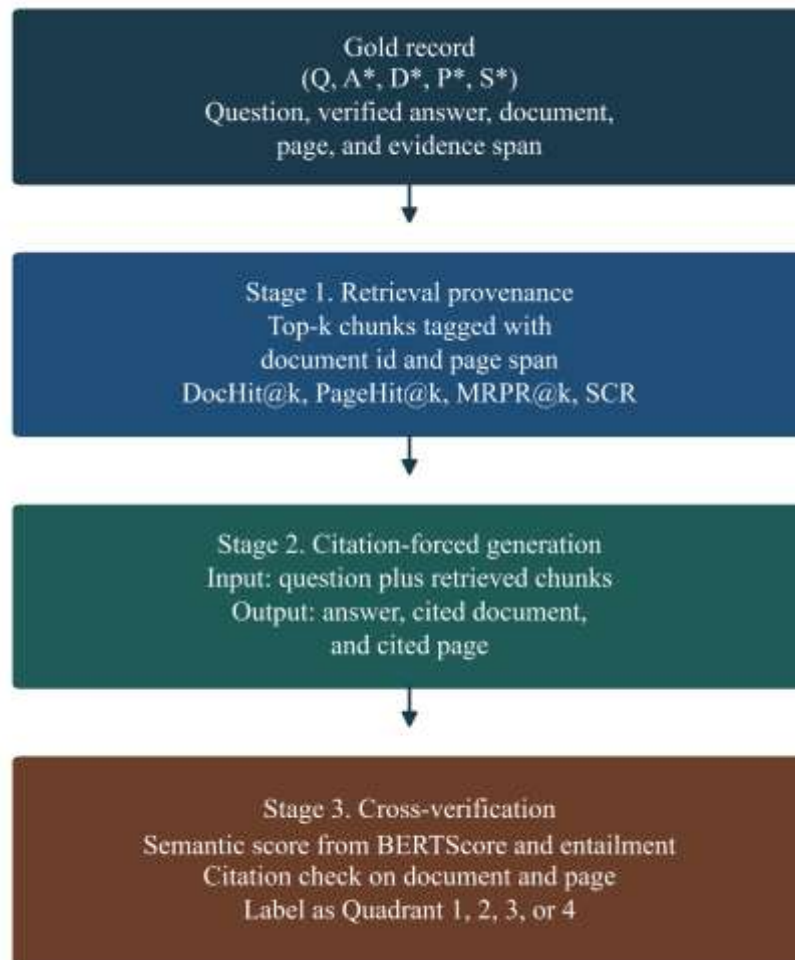


Figure 1. Dual-stage ground-truth cross-verification pipeline. The gold record is fixed before retrieval is scored, and the generated citation is checked only after the answer is produced.

1.1 Failure-mode conflation

Current benchmarks treat a retrieval-augmented pipeline as one black box. RAGAS scores faithfulness, answer relevance, and context relevance by asking an auxiliary model whether the answer is supported by the retrieved text (Es et al., 2024). TruLens organizes a related triad of context relevance, groundedness, and answer relevance. Both approaches judge text with another language model. They do not test whether the retrieved chunk carries the document identifier and the page number of the gold evidence.

That design merges two failures that call for different repairs. An upstream failure occurs when the retriever never surfaces the evidentiary passage, because of embedding drift, a lexical mismatch, or a chunk boundary that cuts the sentence. A downstream failure occurs when the passage is present and the generator ignores it, misreads a table that crosses a page, or answers from pretraining weights. Liu et al. (2024) documented one form of the second failure: facts placed in the middle of a long context are used less reliably than facts placed at either end. If the only score is the final string, an engineer who retunes the retriever in response to a generator error, or the generator in response to a missing page, changes the wrong component.

1.2 Phantom citations

In a validation audit, a correct sentence with the wrong regulation, document, or page is still a compliance failure. Gao et al. (2023) showed that asking a model to cite does not guarantee that the cited text entails the claim. Chen et al. (2026) described a related pattern in which a model reaches a claim from parametric memory and then attaches a citation afterward. Shi et al. (2026) reported that denser legal retrieval can raise evidence support without a matching gain in citation trustworthiness. The practical form of this error is simple. The model knows

an answer from pretraining, the prompt requires a citation, and the model names one of the chunks it was shown even though that chunk is not the gold page. A semantic similarity score between the answer and the gold answer accepts the response. A check against the page number rejects it.

1.3 Prior architectures and benchmarks

Lewis et al. (2020) paired a sequence-to-sequence generator with a dense index over Wikipedia. Karpukhin et al. (2020) showed that a dual-encoder dense retriever can outperform BM25 on open-domain questions. BM25 itself remains the sparse baseline, with the ranking function set out by Robertson and Zaragoza (2009). Later generators were trained to decide when to retrieve. Self-RAG adds reflection tokens that mark whether a passage was retrieved, whether it is relevant, and whether it supports the answer (Asai et al., 2023). Those tokens are internal to the model. An enterprise audit still needs a check that does not depend on the same model judging itself.

Automatic scoring moved from BLEU and ROUGE toward contextual similarity. BERTScore matches tokens in embedding space and correlates more closely with human judgments than lexical overlap does (Zhang et al., 2020). RAGAS then defined reference-free faithfulness, answer relevance, and context relevance (Es et al., 2024). TruthfulQA measures whether models repeat common human falsehoods (Lin et al., 2022). None of these instruments asks whether the cited physical page is the page that contains the gold span. On long inputs, Liu et al. (2024) showed a U-shaped use of context: material at the start and the end is used, and material in the middle is missed. That curve is a generator phenomenon. It should be counted only on the questions for which the retriever already returned the gold page. Otherwise a missing page and a missed middle passage are written down as the same error.

1.4 Gap and contributions

The gap is an evaluation that separates document-and-page provenance from answer correctness, uses the gold page as a deterministic anchor, and still records whether the generated citation matches that anchor. This paper makes four contributions.

First, it defines the dual-stage ground-truth cross-verification framework, with an independent checkpoint after retrieval and a second checkpoint after generation. Second, it defines document hit rate, page hit rate, mean reciprocal provenance rank, and span containment, all computed from document identifiers and page numbers rather than from an auxiliary language-model judge. Third, it defines a four-quadrant taxonomy that labels each question as fully grounded success, context-extraction failure, parametric hallucination, or retrieval failure. Fourth, it applies that taxonomy to 1,200 audit questions over 42 pharmaceutical validation documents (8,520 pages) governed by FDA 21 CFR Part 11, EU GMP Annex 11, and GAMP 5.

2. Materials and Methods

2.1 Dataset preparation

The corpus was drawn from computerized-systems validation documents used in pharmaceutical good-practice settings. The governing texts were United States Food and Drug Administration 21 CFR Part 11 (electronic records and electronic signatures), European Union GMP Annex 11 (computerised systems), and GAMP 5, second edition (a risk-based approach to compliant GxP computerized systems). Corporate procedures were included as well: validation master plans, audit-trail review procedures, installation, operational, and performance qualification protocols, and chromatography data-system specifications. The indexed collection contained 42 PDF documents and 8,520 pages. Pages included multi-column regulatory prose, tables that continue onto the next page, and workflow diagrams.

Regulatory compliance engineers and machine-learning practitioners wrote 1,200 gold records. Each record is the tuple (Q, A*, D*, P*, S*): a natural-language question, a human-checked answer, the source document identifier, the canonical page, and the verbatim character span that states the fact. The set contains 450 factual questions (direct mandates such as password expiry or session timeout), 450 tabular or procedural questions (answers that sit in a validation matrix), and 300 negative questions. A negative question asks for a procedure that was deliberately left out of the corpus, so that abstention can be observed. The same 1,200 records were used for every retriever and every generator.

2.2 Retrieval and generation systems

Three retrieval stacks were compared. The sparse stack was BM25 with Okapi parameters $k1 = 1.5$ and $b = 0.75$ (Robertson & Zaragoza, 2009). The dense stack embedded queries and passages with BGE-large-en-v1.5 and

ranked them by cosine similarity (Xiao et al., 2023). The hybrid stack fused BGE-large and BM25 with reciprocal rank fusion, then reranked the top 25 candidates with the cross-encoder bge-reranker-large and kept the top k . Reported hit rates use $k = 1$ and $k = 5$. The same three stacks were paired with three generators for the synthesis stage: GPT-4o (OpenAI), Claude 3.5 Sonnet (Anthropic), and Llama-3.3-70B-Instruct (Meta). Every generator received the same citation schema, described in Section 2.6. Hardware, decoding temperature, and the exact service build of each hosted model were not varied as experimental factors in the source study; the comparison is across the retrieval stacks and the three named generators.

2.3 Ground-truth notation

Let the repository be a set of M structured documents. In this study $M = 42$.

$$D = \{D_1, D_2, \dots, D_M\} \quad (1)$$

where D is the repository and D_m is document m , for $m = 1, \dots, M$.

Document D_m is an ordered sequence of canonical pages.

$$D_m = (p_{m,1}, p_{m,2}, \dots, p_{m,|D_m|}) \quad (2)$$

where $p_{\{m,l\}}$ is page l of document m . The page carries a text stream $T(p_{\{m,l\}})$ and a physical page index l , a positive integer. The gold benchmark T contains $N = 1,200$ records.

$$T = \{(Q_j, A_j^*, D_j^*, P_j^*, S_j^*)\}_{j=1}^N \quad (3)$$

where Q_j is the question, A_j^* is the human-verified answer, D_j^* is the source document identifier (a content hash or a controlled document code), P_j^* is the canonical page that contains the evidence, and S_j^* is the verbatim character span of that evidence inside the page text.

2.4 Chunk representation

Indexing splits each document into chunks. The chunk map K sends document D_m to a set of W chunks. Chunk length and overlap were held fixed across retrievers. Their numeric values were not listed in the source study and are required for an exact re-index; see Section 4.4.

$$K : D_m \rightarrow C = \{c_1, c_2, \dots, c_W\} \quad (4)$$

where C is the chunk collection for that document and W is the chunk count.

Each chunk is stored as a 5-tuple so that later scores can use page metadata.

$$c_i = (\tau_i, d(c_i), p^{\text{start}}(c_i), p^{\text{end}}(c_i), e_i) \quad (5)$$

where τ_i is the token text, $d(c_i)$ is the parent document identifier, $p^{\text{start}}(c_i)$ and $p^{\text{end}}(c_i)$ are the first and last pages the chunk covers, and $e_i \in \mathbb{R}^d$ is the dense vector produced by encoder E . A chunk that crosses a page break keeps both page numbers. When a table continues onto the next page, the column header is repeated on the continuation chunk so the row remains interpretable.

2.5 Stage 1: retrieval provenance

Given question Q_j , the retriever returns the k highest-scoring chunks. Stage 1 scores that list before the generator runs. Figure 2 shows the split: the same chunks are audited for metadata and, in parallel, packed into the generator prompt.

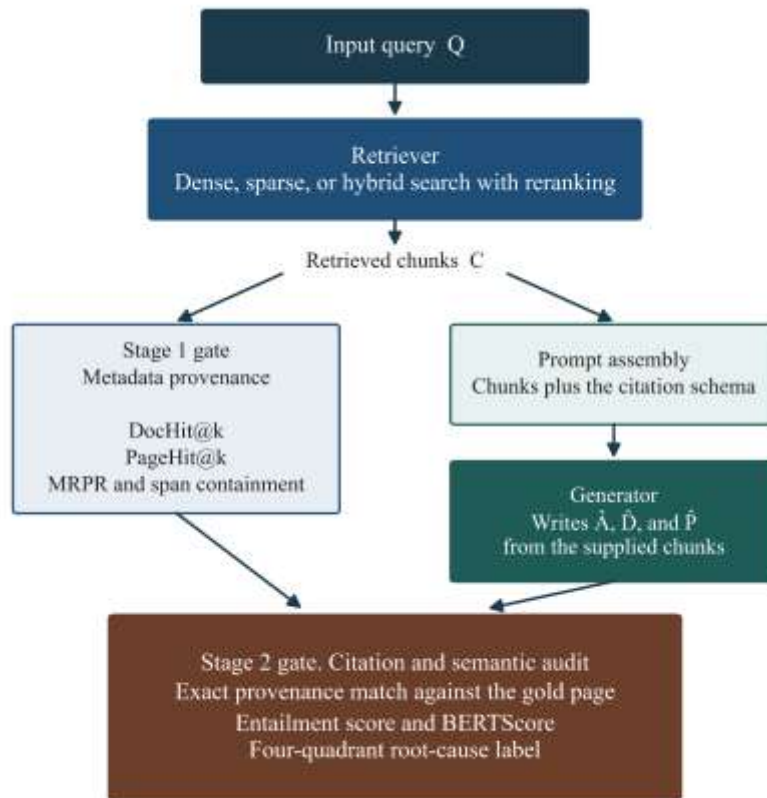


Figure 2. Evaluation architecture. Stage 1 scores chunk metadata while the same chunks are assembled into the generator prompt. Stage 2 then compares the generated answer and the generated citation with the gold record.

$$C_k(Q_j) = \{c_{(1)}, \dots, c_{(k)}\} \quad (6)$$

where the parentheses in the subscript denote rank order under the similarity $\text{sim}(Q_j, c_i)$, and $k = 5$ for the cutoff used in the main tables.

Document hit is 1 when the gold document appears anywhere in the top k .

$$\text{DocHit}@k(j) = I(\exists i \in [1, k] : d(c_i) = D_j^*) \quad (7)$$

where $I(\cdot)$ is the indicator function and $d(c_i)$ is the document identifier stored on chunk i .

Page hit is 1 only when some returned chunk covers the gold page of that document.

$$\text{PageHit}@k(j) = I(\exists i \in [1, k] : d(c_i) = D_j^* \wedge P_j^* \in [p^{\text{start}}(c_i), p^{\text{end}}(c_i)]) \quad (8)$$

where P_j^* is the gold page. A hit on the correct file with the wrong page scores $\text{DocHit}@k = 1$ and $\text{PageHit}@k = 0$.

Mean reciprocal provenance rank uses the rank of the first chunk that covers the gold page.

$$\text{MRPR}@k = \frac{1}{|T|} \sum_{j=1}^{|T|} \rho_j \quad (9)$$

where the sum runs over the N questions in T , and ρ_j is the reciprocal rank of the first page-matching chunk.

$$\rho_j = \begin{cases} \frac{1}{\min i} & \text{if } \text{PageHit}@k(j) = 1 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where the minimum is taken over ranks $i \in [1, k]$ whose chunk covers document D_j^* and page P_j^* . If no such chunk exists, $\rho_j = 0$.

Span containment asks whether the gold character span survived chunking. It is computed on chunks from the gold document.

$$SCR_j = \frac{\max |\tau_i \cap S_j^*|}{|S_j^*|} \quad (11)$$

where the maximum is over chunks i in the top k with $d(c_i) = D_j^*$, $|\cdot|$ is character length, and $SCR_j = 1$ when some returned chunk contains the entire gold span. $PageHit@k = 1$ together with $SCR_j < 1$ means the retriever found the page but the chunk boundary cut the evidence. That event is a chunking fault, not an embedding fault.

2.6 Stage 2: citation-forced generation

Chunks that leave the retriever are placed in a structured prompt. The generator must answer only from those excerpts and must return a document identifier, a page, and a verbatim quote. The instruction used for every model was the following.

You are a regulatory compliance assistant. Answer the user prompt strictly and solely based on the provided document excerpts. Do not rely on outside knowledge. For every claim, provide an exact citation matching the provided context metadata.

OUTPUT SCHEMA:

```
{ "answer": "string", "cited_document": "string", "cited_page": integer,
  "verbatim_quote": "string" }
```

The generation map sends the question and the retrieved list to a predicted answer, a cited document, a cited page, and a quoted span.

$$M(Q_j, C_k) \rightarrow (\hat{A}_j, \hat{D}_j, \hat{P}_j, \hat{S}_j) \quad (12)$$

where $\hat{A}_j$ is the generated answer, $\hat{D}_j$ and $\hat{P}_j$ are the cited document and page, and $\hat{S}_j$ is the quoted span. The hat marks a model output. The star marks a gold field.

2.7 Semantic correctness and citation fidelity

Stage 2 scores the generation on two axes that are not allowed to substitute for each other. Semantic answer correctness mixes a contextual similarity with an entailment score.

$$SAC_j = \alpha \cdot BERTScore_{F1}(\hat{A}_j, A_j^*) + (1 - \alpha) \cdot NLI_{Entail}(\hat{A}_j, A_j^*) \quad (13)$$

where $BERTScore_{F1}$ is the F1 form of $BERTScore$ (Zhang et al., 2020), $NLI_{Entail} \in [0, 1]$ is the entailment probability from a cross-encoder trained on MNLI and ANLI, and $\alpha \in [0, 1]$ is a fixed mixing weight. An answer is counted as semantically correct when $SAC_j \geq \theta$ with $\theta = 0.82$. The threshold was checked against human agreement on 200 calibration questions. The numeric value of α and the checkpoint name of the entailment model are part of the unreleased configuration (Section 4.4).

Deterministic citation fidelity ignores the wording and compares identifiers.

$$DCF_j = I(\hat{D}_j = D_j^* \wedge \hat{P}_j = P_j^*) \quad (14)$$

where $DCF_j = 1$ only when both the cited document and the cited page equal the gold pair.

A model can name the gold page without having taken the words from a chunk that contains that page. The validated-citation indicator blocks that case.

$$ValidCite_j = DCF_j \wedge I(\exists i \in [1, k] : d(c_i) = \hat{D}_j \wedge \hat{P}_j \in [p^{\text{start}}(c_i), p^{\text{end}}(c_i)]) \quad (15)$$

where the second indicator is 1 when at least one retrieved chunk covers the document and the page the model cited. Quadrant assignment in Section 2.8 uses DCF as defined in Equation (14). Equation (15) is the stricter runtime check used by the interlock in Section 4.3.

2.8 Quadrant assignment

Each question is labeled from PageHit@k and from the pair (SAC, DCF). Figure 3 is the decision matrix. The vertical cut is whether the gold page was retrieved. The horizontal cut is whether the answer met θ . Citation match separates the two cells in which the answer is acceptable.

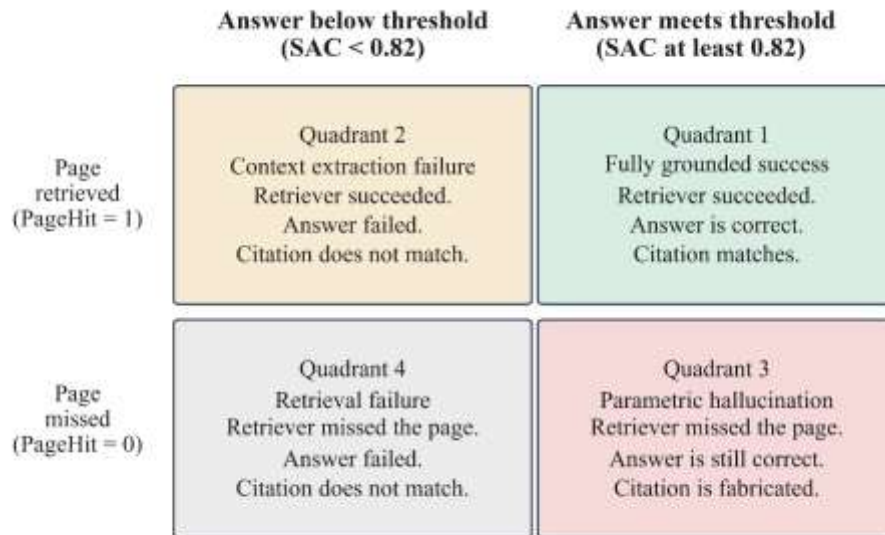


Figure 3. Four-quadrant root-cause matrix. PageHit is the Stage 1 page indicator. SAC is semantic answer correctness with threshold $\theta = 0.82$. A citation match (DCF = 1) is required for Quadrant 1 and is absent in Quadrant 3.

Quadrant 1 (fully grounded success) requires PageHit@k = 1, SAC ≥ 0.82 , and DCF = 1. The page was retrieved, the answer met the threshold, and the citation named that page. Quadrant 2 (context extraction failure) requires PageHit@k = 1, SAC < 0.82, and DCF = 0. The evidence was in the prompt and the generator did not use it. Quadrant 3 (parametric hallucination, also called ghost grounding) requires PageHit@k = 0, SAC ≥ 0.82 , and DCF = 0. The gold page was absent, the answer still matched, and the citation did not. Quadrant 4 (retrieval failure) requires PageHit@k = 0, SAC < 0.82, and DCF = 0. Both the retriever and the generator failed. Shares in Section 3 are percentages of the 1,200 questions. The study did not report confidence intervals or hypothesis tests; the tables are descriptive counts.

3. Results

3.1 Stage 1 retrieval

Table 1 reports document hit, page hit, mean reciprocal provenance rank, and mean span containment for the four retrieval stacks on all 1,200 questions. No generator score enters this table.

Table 1. Stage 1 retrieval provenance on 1,200 audit questions.

Retrieval stack	DocHit@1	DocHit@5	PageHit@1	PageHit@5	MRPR@5	SCR
BM25 (sparse)	58.4	74.2	34.2	49.1	0.398	0.62
BGE-large (dense)	71.3	86.9	52.8	69.4	0.589	0.74
Hybrid (BGE + BM25)	78.6	92.1	63.4	78.8	0.684	0.81

Retrieval stack	DocHit@1	DocHit@5	PageHit@1	PageHit@5	MRPR@5	SCR
Hybrid + cross-encoder reranker	84.2	95.6	71.9	85.3	0.772	0.88

Note. DocHit and PageHit are percentages. MRPR@5 is mean reciprocal provenance rank at $k = 5$. SCR is mean span containment. Hybrid fuses BGE-large and BM25 by reciprocal rank fusion. The last row reranks the top 25 fused candidates with bge-reranker-large.

On the strongest stack, DocHit@5 was 95.6% and PageHit@5 was 85.3%. The difference is 10.3 percentage points: the correct document was retrieved more often than the correct page. Mean span containment on that stack was 0.88. The source analysis stated that the supporting text was split across a chunk boundary in 12% of questions. BM25 reached PageHit@5 of 49.1% and mean span containment of 0.62. Adding the dense encoder, then fusion, then the cross-encoder, raised every column in Table 1.

3.2 End-to-end scores

Table 2 crosses the retrieval stacks with the generators. Semantic correctness is the share of questions with SAC ≥ 0.82 . Deterministic citation match is the share with DCF = 1. The four quadrant columns are shares of the same 1,200 questions and sum to 100% in each row.

Table 2. End-to-end correctness, citation fidelity, and quadrant shares.

Retriever	Generator	SAC	DCF	Q1	Q2	Q3	Q4
BM25	Llama-3.3-70B	51.2	38.4	36.1	13.0	15.1	35.8
BM25	GPT-4o	62.4	44.1	42.8	6.3	19.6	31.3
BGE-dense	Llama-3.3-70B	68.3	58.7	55.4	14.0	12.9	17.7
BGE-dense	GPT-4o	77.1	65.2	63.8	5.6	13.3	17.3
Hybrid + rerank	Llama-3.3-70B	79.4	73.1	71.2	14.1	8.2	6.5
Hybrid + rerank	GPT-4o	88.2	81.3	80.1	5.2	8.1	6.6
Hybrid + rerank	Claude 3.5 Sonnet	90.6	83.9	82.7	2.6	7.9	6.8

Note. All figures are percentages of 1,200 questions. SAC is the share with semantic answer correctness at least 0.82. DCF is the share with an exact document and page match. Q1 is fully grounded success, Q2 is context extraction failure, Q3 is parametric hallucination, and Q4 is retrieval failure. Claude 3.5 Sonnet was reported for the hybrid reranked stack only.

The highest fully grounded share was 82.7% (hybrid reranker with Claude 3.5 Sonnet). The same stack with GPT-4o reached 80.1%, and with Llama-3.3-70B-Instruct reached 71.2%. Semantic correctness was higher than the fully grounded share in every row. For hybrid reranking with GPT-4o, semantic correctness was 88.2% and Quadrant 1 was 80.1%. Citation match on that row was 81.3%. Quadrant 3 remained between 7.9% and 8.2% on the hybrid reranked stack and rose to 19.6% for BM25 with GPT-4o. Quadrant 2 was larger for Llama-3.3-70B-Instruct (13.0% to 14.1%) than for GPT-4o (5.2% to 6.3%) or Claude 3.5 Sonnet (2.6%).

3.3 Quadrant profile and a Quadrant 3 record

Figure 4 plots the four shares for hybrid retrieval, cross-encoder reranking, and GPT-4o. Fully grounded answers were 80.1% of the set. Extraction failure was 5.2%. Parametric hallucination was 8.1%. Retrieval failure was 6.6%.

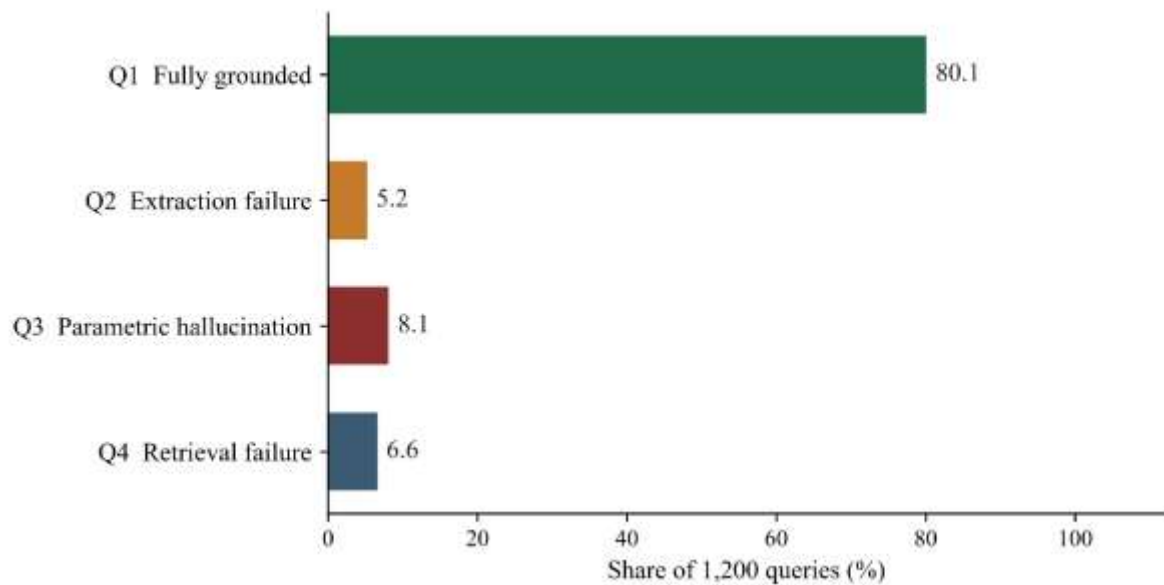


Figure 4. Quadrant shares for hybrid retrieval with cross-encoder reranking and GPT-4o (n = 1,200). Bars sum to 100%.

Table 3 records one Quadrant 3 question in full. The gold page was EU Annex 11, page 4. None of the five retrieved chunks covered that page, so PageHit@5 = 0. The generated sentence agreed with the gold fact (SAC = 0.94) and cited page 1 of the same file, which was a title and scope page. DCF failed, and the question was labeled Quadrant 3.

Table 3. Worked Quadrant 3 record on EU Annex 11 backup testing.

Field	Value
Question	According to EU Annex 11, what is the mandatory requirement regarding regular testing of system backups?
Gold document	EU_Annex_11_Guide.pdf
Gold page	4
Gold fact	Regular tests of the data backup system should be conducted to confirm ability to restore data successfully.
Retrieved chunks	SOP-IT-002 p. 12; FDA 21 CFR Part 11 p. 2; SOP-QC-800 p. 3; SOP-IT-002 p. 14; EU Annex 11 p. 1 (title and scope)
PageHit@5	0
Generated answer	EU Annex 11 mandates that data backup systems must undergo periodic validation and regular test restores to ensure that data can be fully recovered.
Cited document	EU_Annex_11_Guide.pdf
Cited page	1
SAC	0.94 (above 0.82)
DCF	Fail (cited page 1; gold page 4)
Quadrant	3, parametric hallucination

Note. SAC is semantic answer correctness. DCF is deterministic citation fidelity. The retrieved list did not contain page 4, which is the page named by the gold record.

3.4 Gold-chunk rank inside the prompt

Table 4 and Figure 5 restrict attention to questions on which the gold chunk was retrieved, and they break extraction success by the rank of that chunk. Rank 1 is the first passage in the prompt. Rank 5 is the last. This is the measurement behind Quadrant 2: the page is present, and the generator either uses it or does not.

Table 4. Extraction success (%) by rank of the gold chunk.

Generator	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Claude 3.5 Sonnet	98.4	94.1	89.2	93.6	97.9
GPT-4o	97.1	91.5	84.7	90.2	96.8
Llama-3.3-70B-Instruct	91.2	82.4	72.1	80.5	89.4

Note. Each cell is the percentage of questions, among those whose gold chunk occupied that rank, for which the generator extracted the gold fact. Rank 1 is the top of the prompt and rank 5 is the bottom.

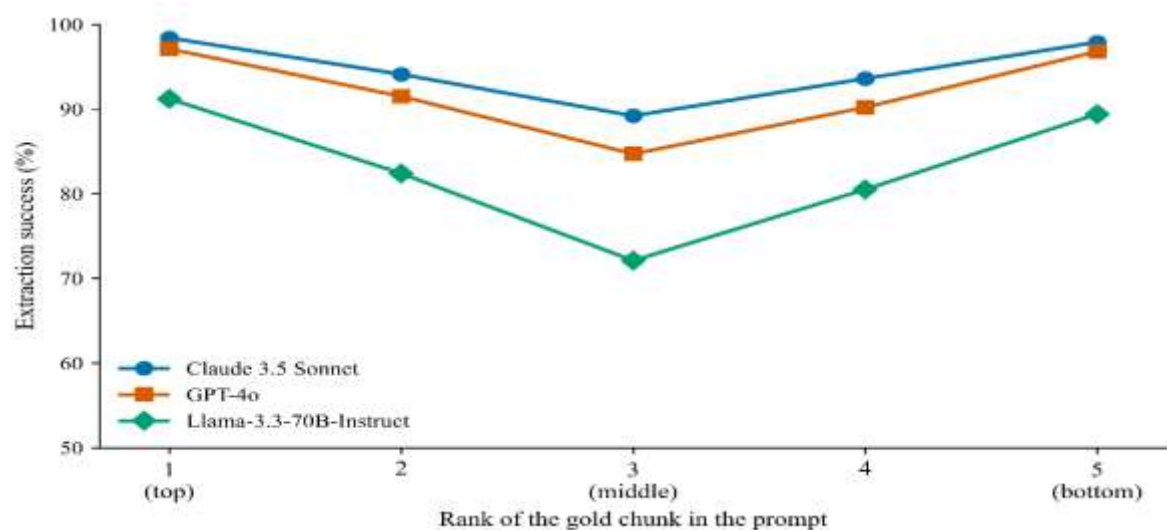


Figure 5. Extraction success by the rank of the gold chunk. All three generators are lowest at rank 3 and higher at rank 1 and rank 5.

At rank 1, extraction success was 98.4% for Claude 3.5 Sonnet, 97.1% for GPT-4o, and 91.2% for Llama-3.3-70B-Instruct. At rank 3 those figures were 89.2%, 84.7%, and 72.1%. The drop from rank 1 to rank 3 was 9.2 percentage points for Claude 3.5 Sonnet, 12.4 percentage points for GPT-4o, and 19.1 percentage points for Llama-3.3-70B-Instruct. Success rose again at rank 5, to 97.9%, 96.8%, and 89.4%.

4. Discussion

4.1 What a page anchor changes

The objective in Section 1 was an evaluation that does not credit a retriever for returning the right file on the wrong page, and does not credit a generator for a correct sentence with a fabricated citation. Table 1 is the first half of that result. DocHit@5 of 95.6% on the reranked hybrid stack would look nearly solved if document recall were the only retrieval metric. PageHit@5 of 85.3% shows a remaining 10.3 percentage-point gap. RAGAS faithfulness and the TruLens triad (Es et al., 2024) score whether the answer is consistent with whatever text was retrieved. They do not have a slot for this gap, because they never receive P*. A page-level hit rate makes the gap visible before any generator is called, which is the separation drawn in Figure 2.

Span containment adds a second distinction inside the questions the retriever got right. Mean SCR of 0.88 on the best stack, and the reported 12% of questions whose evidence was cut by a chunk boundary, means some PageHit@k = 1 outcomes still delivered a partial sentence. Retuning the embedding model does not repair a boundary that slices the gold span. The repair is the chunk representation in Equation (5): store both page endpoints, and do not drop a header when a table crosses a page.

4.2 Phantom citations and the middle of the prompt

Quadrant 3 is the cell semantic scoring accepts and this framework rejects. On the best GPT-4o configuration, 8.1% of questions were semantically correct while the gold page was missing and the citation did not match. On BM25 with the same generator that share was 19.6%. The worked record in Table 3 is the pattern Gao et al. (2023) and Chen et al. (2026) described, now tied to a page identifier: the sentence about backup testing is close to the Annex 11 sentence, and the cited page is the title page that happened to be in the prompt. Shi et al. (2026) found a similar split in legal retrieval, where better evidence support did not by itself produce trustworthy citations. A judge that compares only $\hat{A}$ and A^* records Table 3 as a success. DCF records it as a miss.

Quadrant 2 is the complementary cell. The page was retrieved and the answer still failed. Figure 5 shows that this failure is ordered by position. The curve has the same shape Liu et al. (2024) reported for long contexts: high at the ends, lower in the middle. Because the curve is computed only when the gold chunk is already in the prompt, the drop is not a retrieval miss. It is a reason to reorder or shorten the context when Quadrant 2 grows, and a reason not to answer that growth by swapping the embedding model. The open-weight generator lost 19.1 percentage points from rank 1 to rank 3. The two hosted generators lost 9.2 and 12.4 points on the same contrast. Quadrant 2 in Table 2 is consistent with that ordering: Llama-3.3-70B-Instruct contributed about 14% extraction failures on the hybrid stack, and Claude 3.5 Sonnet contributed 2.6%.

One cell in Table 2 is slightly larger than Quadrant 1 alone. For hybrid reranking with GPT-4o, DCF was 81.3% and Quadrant 1 was 80.1%. The definitions in Section 2.8 put every DCF = 1 success that also meets the semantic threshold into Quadrant 1, so the 1.2 percentage-point remainder is consistent with a small set of questions that named the gold page and still fell below SAC = 0.82. Those questions are not given a separate quadrant. They are visible only because citation match and semantic correctness are stored as separate columns.

4.3 Controls that follow from the quadrants

The quadrant counts suggest three controls, shown in Figure 6. They are consequences of the metrics, not additional experimental arms.

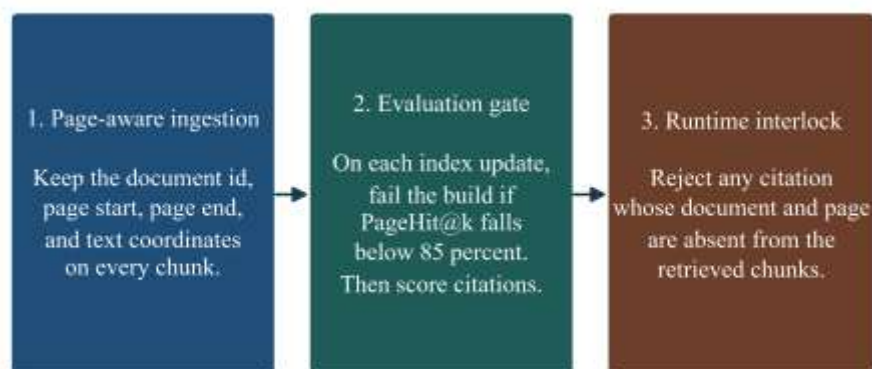


Figure 6. Production controls implied by the audit. Page metadata is stored at ingestion, page hit rate gates an index update, and a citation that is absent from the retrieved chunks is rejected at runtime.

Ingestion has to keep the fields in Equation (5). A chunker that flattens a PDF into a character buffer cannot compute PageHit@k or DCF, because the page number is gone. Coordinate-aware parsers are sufficient for this requirement; the study named PyMuPDF and Unstructured as examples of engines that can emit text together with a page index. On each re-index, Stage 1 can be run on the fixed 1,200 records before the index is published. The operating rule stated for that gate is to fail the build when PageHit@k falls below 85%. The reranked hybrid stack in Table 1 sits at 85.3% on PageHit@5, so the threshold matches the best observed page hit rate rather than a looser document hit rate. If the page-hit gate passes, Stage 2 is the check on citation match and semantic correctness. At runtime, Equation (15) rejects a citation whose document and page are not among the retrieved chunks. That rule would have rejected the page-1 citation in Table 3 only if page 1 had also been absent; in that record page 1 was retrieved, so the runtime interlock as stated would not fire, and the Quadrant 3 label would still depend on comparison with the gold page. The interlock catches citations pointing outside the prompt. It does not, by itself, catch a citation that points at a distractor chunk inside the prompt. Gold-page comparison remains necessary for that second case.

Which repair to attempt follows the dominant quadrant. A rise in Quadrant 4 points at the index, the parser, or the lexical gap. A rise in Quadrant 2 points at prompt order, context length, or the generator. A rise in Quadrant 3 points at answers that should be abstained when PageHit@k = 0, rather than at a higher top-k.

4.4 Limitations

Several boundaries follow from how the scores were built. Semantic correctness depends on BERTScore and an entailment model, so two wordings of the same obligation can fall on either side of 0.82. The threshold was set against human agreement on 200 calibration questions, which limits that risk and does not remove it. The mixing weight α and the entailment checkpoint were not reported, so another group cannot reproduce SAC_j exactly until those values are released. Chunk length and overlap were likewise unreported. Hit rates at k = 5 can be recomputed from a released index; they cannot be recomputed from the PDFs alone without those chunk settings.

The corpus is pharmaceutical validation text. Legal opinions and financial filings can break across pages in different ways, and the 10.3 percentage-point document-to-page gap should not be assumed to travel unchanged. Page identity was defined as the canonical page index. Documents that mix roman front matter with arabic body pages, or that insert fold-out sheets, need a map from the PDF page index to the printed page number before DCF is meaningful. The 300 negative questions were built so that abstention could be measured, and abstention rates were not reported in the results. No standard errors, confidence intervals, or hypothesis tests were computed; differences in Table 2 are descriptive. Corporate procedures in the corpus are not public, so the 1,200 questions cannot be redistributed in full from the description in this paper.

5. Conclusion

A retrieval-augmented system that will be used in a regulated audit needs a score for the page that was retrieved and a separate score for the page that was cited. The dual-stage framework supplies both. On 8,520 pages of pharmaceutical validation text, document hit rate overstated page hit rate, span containment showed that some of the retrieved pages arrived with the evidence cut, and a non-trivial share of semantically acceptable answers cited a page that was not the gold page. Position of the gold chunk inside the prompt further separated generator misses from retriever misses.

The practical use of the quadrants is to choose the repair. Index and parser work addresses retrieval failure. Prompt order addresses extraction failure. Abstention when the gold page is absent addresses parametric hallucination. The same records can be rerun when the embedding model, the index, or the generator changes, which makes the 1,200 questions a regression set rather than a one-time leaderboard.

Two extensions follow directly. The page anchor can be replaced by a spatial anchor, so that evidence in a scanned drawing is a bounding box rather than a page number. And the Stage 1 result can be allowed to stop a response: if the gold page is not in the retrieved set, the system can abstain or retrieve again instead of forcing the generator to name a citation. Both extensions keep the separation this paper argues for, which is that provenance and faithfulness are different measurements.

Declarations

Funding. No funding was received.

Conflict of interest. The authors declare no conflicts of interest.

Data availability. Public regulatory documents examined in this study are available from their respective official sources. The internal standard operating procedures, benchmark records, and scoring code are not publicly available because of institutional and confidentiality restrictions. Further information may be requested from the corresponding author, subject to applicable permissions.

Author contributions. Chinedu Okafor: Conceptualization, methodology, software, formal analysis, and writing—original draft. Amina Bello: Data curation, validation, investigation, writing—review and editing. Both authors read and approved the final manuscript.

Acknowledgments. The authors thank the Department of Computer Science, University of Lagos, for institutional support. No other assistance is declared.

Ethical approval. Not applicable. The study used documents and did not enroll human participants or animals.

Declaration of Generative AI Use. Generative AI tools were used solely to improve the clarity, grammar, and presentation of the manuscript. The authors reviewed and edited all AI-assisted content and accept full responsibility for the final manuscript.

Consent for Publication. Not applicable. The study does not contain personal data, identifiable individual information, or individual-level case reports requiring consent for publication.

References

- Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv*. <https://doi.org/10.48550/arXiv.2310.11511>
- Chen, H., et al. (2026). Attributable post-rationalization in RAG citations: Empirical pitfalls and formal bounds. *arXiv*. <https://doi.org/10.48550/arXiv.2609.23053>
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System demonstrations** (pp. 150–158). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing** (pp. 6465–6488). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.398>
- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing** (pp. 6769–6781). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, **33**, 9459–9474. <https://doi.org/10.5555/3495724.3496517>
- Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics** (pp. 3214–3252). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.229>
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, **12**, 157–173. https://doi.org/10.1162/tacl_a_00638
- Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, **3**(4), 333–389. <https://doi.org/10.1561/1500000019>
- Shi, Y., Chen, H., et al. (2026). LegalCiteTrust: Benchmarking citation trustworthiness in legal deep-research and retrieval-augmented generation. *arXiv*. <https://doi.org/10.48550/arXiv.2607.20872>
- Xiao, S., Liu, Z., Zhang, P., & Muennighoff, N. (2023). C-Pack: Packaged resources to advance general Chinese and universal text embeddings. *arXiv*. <https://doi.org/10.48550/arXiv.2309.07597>
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *arXiv*. <https://doi.org/10.48550/arXiv.1904.09675>