ArchiveRAG: A Multimodal Retrieval-Augmented Generation Architecture for Legacy Government Document Disclosure

Ravi Dnyaneshwar Mahajan
University at Buffalo, The State University of New York
ravi.d.mahajan@gmail.com

ABSTRACT

Hundreds of millions of FOIA-responsive government records exist only as degraded scans, multi-page TIFF stacks, microfilm conversions, and handwritten legacy documents — formats that defeat conventional text-based search pipelines and place responsive records beyond the practical reach of retrieval. This paper introduces ArchiveRAG, an end-to-end five-stage multimodal Retrieval-Augmented Generation architecture specifically designed for the ingestion, indexing, and legally grounded retrieval of legacy government archives. The pipeline integrates: (1) LLM-assisted OCR post-correction using a fine-tuned correction model trained on degraded government document pairs; (2) layout-aware semantic chunking that preserves table boundaries, section hierarchies, and cross-reference structures; (3) multimodal embedding using a ColPali-variant vision-language model that encodes both the visual page layout and extracted text into a unified retrieval representation; (4) hybrid BM25-plus-dense retrieval with a re-ranking stage that surfaces FOIA-responsive passages; and (5) a redaction-aware exemption detection layer that flags candidate Exemption zones under 5 U.S.C. § 552(b) as an integrated component of the retrieval output.

ArchiveRAG is benchmarked on GovArchive-512, a curated evaluation corpus of 512 publicly released government documents drawn from the NARA National Archives Catalog (registry.opendata.aws/nara-national-archives-catalog) and the MuckRock public FOIA archive (muckrock.com/api), spanning six federal agencies and four document eras (1940s–2010s), with FOIA-specific query annotations contributed by the authors. Against a Tesseract-plus-keyword-search baseline, ArchiveRAG achieves a 34.7-point improvement in Recall@10 (51.2% vs. 85.9%), a redaction-zone detection F1 of 0.81, and a mean retrieval latency of 340ms at a 500,000-document index scale. Failure modes, hallucination risk in the generation layer, governance requirements for agency deployment, and extensibility to non-U.S. archival contexts are discussed in full.

CCS CONCEPTS

Information systems: Information retrieval; Computing methodologies: Natural language processing; Machine learning; Applied computing: Law, social and behavioral sciences.

KEYWORDS

retrieval-augmented generation, multimodal AI, legacy document processing, FOIA, government archives, OCR post-correction, vision-language models, redaction detection, ColPali, public disclosure

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICAIL ’26, June 2026, Singapore
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/26/06
https://doi.org/10.1145/XXXXXXX.XXXXXXX

1 INTRODUCTION

The Freedom of Information Act and its global counterparts rest on a deceptively simple premise: that government records, once released, are accessible. In the digital era, this premise has been substantially complicated by a structural asymmetry between the volume of records held in legacy formats and the capabilities of the retrieval systems available to search them. The National Archives and Records Administration (NARA) alone holds over 13.3 billion pages of records, of which an estimated 78% remain either undigitized or digitized only as image scans without machine-readable text layers. The consequence is a two-tier disclosure landscape: records held in structured digital formats are rapidly retrievable, while those existing only as scanned TIFF images, degraded microfilm conversions, or handwritten ledgers effectively resist retrieval, and therefore disclosure, regardless of their legal releasability.

This asymmetry is not a peripheral archival concern. It is a disclosure law problem of the first order. Under FOIA, agencies are obligated to conduct reasonable searches of their records systems in response to requests. Courts have interpreted this obligation to extend to electronically stored records, including scanned documents in agency FOIA reading rooms. Yet if scanned records cannot be searched — because they lack text layers, because OCR quality is insufficient, or because retrieval systems cannot navigate complex document layouts — then the reasonable search obligation becomes practically unenforceable.

Recent advances in multimodal AI — specifically in vision-language models, layout-aware document understanding, and retrieval-augmented generation — offer a technically credible pathway to resolving this asymmetry. ColPali [1], a late-interaction vision-language retrieval model, demonstrated that document page images can be embedded directly for retrieval without requiring OCR as a prerequisite, achieving state-of-the-art performance on visually rich document benchmarks. Large language models have demonstrated strong performance on OCR post-correction [2], restoring machine-readable accuracy in degraded scans. The RAG paradigm [3] provides a generation architecture that grounds responses in retrieved document passages, mitigating hallucination and enabling verifiable, citation-grounded answers to complex queries.

ArchiveRAG integrates these advances into a coherent, production-oriented architecture explicitly designed for the constraints and obligations of government archive disclosure. Four contributions are made:

(1) The GovArchive-512 evaluation corpus is introduced — a 512-document benchmark subset curated from two real public sources (the NARA National Archives Catalog and the MuckRock FOIA archive) for multimodal retrieval evaluation on FOIA-format

government documents, spanning six agencies and eight decades, with author-contributed FOIA query annotations following the ViDoRe benchmark format [Faysse et al., 2024].

(2) ArchiveRAG, a five-stage multimodal RAG pipeline, is proposed, encompassing LLM-assisted OCR correction, layout-aware chunking, multimodal embedding, redaction-aware hybrid retrieval, and an integrated exemption detection layer.

(3) Empirical benchmarking demonstrates substantial retrieval improvements over Tesseract-plus-keyword baselines, with a 34.7-point Recall@10 gain and redaction zone detection F1 of 0.81.

(4) A governance and deployment framework for agency-level adoption is proposed, addressing audit requirements, failure mode management, and integration with existing FOIA software infrastructure.

The paper is structured as follows. Section 2 surveys related work across five relevant fields. Section 3 formulates the retrieval problem in FOIA-specific terms. Section 4 describes the ArchiveRAG architecture in full. Section 5 presents the GovArchive-512 corpus. Section 6 reports experimental results. Section 7 introduces the redaction-aware retrieval variant. Section 8 discusses limitations, failure modes, and deployment governance. Section 9 concludes.

2 BACKGROUND AND RELATED WORK

2.1 Legacy Document Formats in Government Archives

Government archives present a document formatting landscape of exceptional heterogeneity. The dominant legacy formats in federal holdings include: microfilm and microfiche scans (prevalent in records from the 1940s–1980s), multi-page TIFF stacks produced by flatbed scanner batches (1990s–2000s digitization programmes), native-PDF documents with embedded but often degraded OCR text layers (2000s onward), and handwritten records — forms, field notes, correspondence — that remain entirely outside OCR pipelines. Each format class presents distinct processing challenges: microfilm scans suffer from contrast instability and geometric distortion; flatbed TIFF batches often lack logical document boundaries; native PDFs with embedded OCR inherit legacy recognition errors that propagate silently into downstream applications; handwritten records require specialist Handwritten Text Recognition (HTR) models distinct from standard OCR.

The archival informatics literature has addressed these challenges primarily through digitization workflow optimization [4] and advancing foundational machine learning tasks like robust text line detection in historical documents [5]. Transkribus [6], developed by the READ consortium, provides HTR capabilities for handwritten European archives with strong performance on regular scripts; its applicability to U.S. government handwriting is less well established. The broader challenge of multi-format heterogeneity at government-archive scale has not been systematically addressed by a unified AI pipeline.

2.2 OCR Post-Correction

OCR post-correction — the task of restoring machine-readable accuracy in degraded or error-prone OCR output — has a long history in digital humanities and information retrieval [7]. Traditional approaches using language model edit distance have been largely superseded by neural sequence-to-sequence architectures. For instance, Amrhein and Clematide [2] demonstrated that framing OCR correction as a character-level machine translation task using neural models achieves substantial error reductions on historical texts, establishing a methodology applicable to legacy government documents. Furthermore, these neural sequence-to-sequence approaches have proven robust even in limited-data settings; Rijhwani et al. [8] showed that such models can leverage contextual reasoning and transfer learning to effectively perform OCR post-correction on low-resource and endangered language texts.

A critical consideration for government archive post-correction is domain specificity: government documents contain dense acronym fields, agency-specific terminology, and numerical data (serial numbers, case identifiers, budget codes) that general-domain post-correction models handle poorly. The ArchiveRAG post-correction stage addresses this through domain-adaptive fine-tuning on government-specific correction pairs.

2.3 Document Layout Understanding

Layout understanding — the task of segmenting document pages into semantically meaningful regions (paragraphs, headings, tables, figures, headers, footers) and inferring their reading order — is a prerequisite for semantically coherent chunking. DocLayNet [9], a large-scale benchmark dataset of annotated document page images across multiple domains, enabled substantial progress in layout detection model development. The PubLayNet [10] and DocBank [11] datasets provided comparable resources for scientific and general documents. LayoutLM [12] and its successors LayoutLMv2 and LayoutLMv3 demonstrated that joint encoding of text tokens, their spatial positions, and page image features achieves state-of-the-art performance across document understanding tasks.

Government-specific layout challenges are less well studied. Government forms frequently use multi-column layouts with embedded tables nested inside section fields; declassified documents carry header/footer stamps (classification markings, Bates numbers, declassification dates) that occupy page real estate and disrupt reading order detection; handwritten annotations in margins and over printed text create overlapping layout regions that challenge region-based segmentation models.

2.4 Retrieval-Augmented Generation and Multimodal Retrieval

Retrieval-Augmented Generation (RAG), introduced by Lewis et al. [3], established the paradigm of grounding language model generation in a retrieved set of relevant passages, enabling verifiable, citation-backed responses while mitigating the hallucination risk of purely parametric generation. Subsequent work has extended the RAG paradigm to multi-hop retrieval [13], iterative retrieval [14], and long-context settings [15]. The specific challenge of retrieval from visually-rich documents — where layout, tables, and figures carry meaning not captured by extracted text alone — was addressed by ColPali [1], which introduced a late-interaction retrieval model operating directly on page image embeddings, achieving strong performance on the ViDoRe benchmark without any OCR dependency.

Hybrid retrieval, combining sparse BM25 retrieval with dense vector retrieval, has been shown to consistently outperform either approach alone [16], and is now standard practice in production RAG deployments. Re-ranking models [17] that re-score an initial retrieval candidate set using a more computationally intensive cross-encoder further improve precision at the cost of added latency. ArchiveRAG incorporates both hybrid retrieval and re-ranking, extending the re-ranker with a novel redaction-awareness module described in Section 4.

2.5 FOIA Search Obligations and Legal Information Retrieval

The legal framework governing FOIA search obligations is directly relevant to retrieval system design. Courts have interpreted the "reasonable search" standard under 5 U.S.C. § 552(a)(3)(C) to require agencies to search all records systems in which responsive

documents are likely to be found, using methods reasonably expected to uncover responsive materials [18]. An agency that relies on keyword search in a text-indexed system while holding an equivalent corpus of unindexed scanned documents may fail the reasonable search standard if the scanned corpus is likely to contain responsive records.

Legal information retrieval research [19] has focused primarily on case law and statutory text, leaving government document retrieval — especially from legacy formats — underexplored. Most directly related to ArchiveRAG's redaction-awareness component is the sensitivity-aware retrieval evaluation paradigm developed by Oard et al. [23] through the TREC Legal Track, which established frameworks for evaluating retrieval systems that must simultaneously surface responsive records and flag sensitivity-relevant content for downstream review. ArchiveRAG extends this paradigm to the multimodal, legacy-format government document setting.

3 PROBLEM FORMULATION

3.1 The Legacy Archive Retrieval Gap

The retrieval problem addressed by ArchiveRAG is formally defined as follows. Given a FOIA request q — a natural-language description of the records sought — and a heterogeneous document archive A containing documents in legacy formats (scans, TIFF stacks, degraded PDFs, handwritten records), the goal is to retrieve a ranked set of document passages $P = \{p_1, p_2, ..., p_k\}$ that are: (i) semantically responsive to q; (ii) drawn accurately from documents in A without hallucination or fabrication; and (iii) accompanied by redaction-zone annotations identifying passages that may qualify for withholding under FOIA exemptions.

This formulation extends standard RAG retrieval in three respects. First, the document collection is not assumed to be text-indexed; visual page representations must be processed alongside any available text. Second, the retrieval objective is legally grounded; responsiveness to a FOIA request is not purely semantic similarity but is shaped by the request's scope, the agency's record-keeping practices, and the applicable exemption structure. Third, the output is not only a ranked passage list but a disclosure-annotated result set that surfaces exemption candidates alongside responsive passages, supporting downstream agency review.

3.2 Scope of the Legacy Format Problem

To quantify the scope of the problem, a survey of publicly available FOIA reading room inventories across 12 federal agencies was conducted, combining manual inspection with automated format analysis of downloadable content. Results confirm the scale of the legacy format challenge: across the 12 agencies surveyed, an estimated 61% of FOIA-released documents are image-only scans without searchable text layers; of those with embedded text layers, 34% contain OCR error rates exceeding 15% (as estimated by character-level comparison of Tesseract output against a manually verified 200-document subset). These figures indicate that a majority of released government records are not substantively searchable by existing public interfaces, regardless of their formal availability.

This finding is consistent with prior investigations by open government advocacy organizations [21], which have documented systematic patterns of low-utility releases attributed to format limitations rather than deliberate withholding. The distinction matters legally: a document released in a format that cannot be searched is not substantively disclosed in any practically meaningful sense.

4 THE ARCHIVERAG ARCHITECTURE

ArchiveRAG is structured as a five-stage pipeline, each stage addressing a distinct technical barrier to FOIA-grade retrieval from legacy formats. Stages 1 and 2 constitute the ingestion layer, operating at index-build time. Stages 3 and 4 constitute the retrieval layer, operating at query time. Stage 5, the redaction-awareness layer, is an integrated component that combines exemption detection classifiers with the re-ranking output of Stage 4 to produce a unified result set in which every retrieved passage is annotated for FOIA exemption risk alongside its responsiveness score; it is described in full in Section 7.

4.1 Stage 1: Multimodal Ingestion and OCR Post-Correction

The ingestion stage handles format normalization and text layer construction across all four legacy format classes. For TIFF stacks and microfilm scans, page images are first pre-processed using an adaptive binarization pipeline (Sauvola thresholding [22] with block size adapted to estimated noise level) followed by deskewing and resolution normalization to 300 DPI. Initial OCR is performed using Tesseract 5.0, an open-source OCR engine maintained by Google, with a government-domain language model trained on 15,000 pages of verified government document text.

OCR post-correction is performed by a fine-tuned Flan-T5-Large model, trained on 8,400 correction pairs constructed by artificially degrading high-quality government document scans to simulate common degradation patterns (ink fade, paper yellowing, physical damage, scanner noise). The post-correction model takes as input a 512-character OCR window with left-right context and produces a corrected output sequence. Evaluation of the post-correction model on a held-out test set shows a word error rate (WER) reduction from 18.4% (raw Tesseract) to 4.1% (post-corrected), representing a 77.7% relative improvement.

For handwritten records, a specialist HTR pathway is triggered by a handwriting detection classifier (a fine-tuned ViT-base model, F1 = 0.94 on the detection task). Detected handwritten pages are processed by an adapted Transkribus-compatible HTR model fine-tuned on U.S. government handwriting samples sourced from the NARA Citizen Archivist crowdsourced transcription archive. Critically, visual page representations are preserved alongside extracted text throughout Stage 1, enabling the multimodal embedding in Stage 3 to exploit both modalities.

4.2 Stage 2: Layout-Aware Semantic Chunking

Standard RAG chunking strategies — fixed-length token windows, sentence-boundary splits — are inadequate for government documents, which contain semantically critical structures (tables, numbered lists, conditional clauses in legal text) that are destroyed by format-agnostic chunking. ArchiveRAG employs a layout-aware chunking strategy grounded in the document's detected region structure.

Layout detection is performed by a DocLayNet-trained Mask R-CNN model fine-tuned on 1,200 annotated government document pages covering nine region classes: paragraph, heading, table, table-caption, list, footnote, header-stamp, footer-stamp, and handwritten-annotation. Detected regions are ordered by reading-order inference using a BiLSTM sequence model trained on reading-order annotations from the DocLayNet dataset, with a government-specific correction rule for multi-column form layouts.

Chunking is then performed by a rule-governed algorithm that respects region boundaries: table regions are preserved as atomic chunks regardless of token length (with a maximum of 1,024 tokens; larger tables are split at row boundaries with shared header repetition); paragraph regions are chunked at sentence boundaries with a 256-token target and 64-token overlap; heading regions are

prepended to the first chunk of their subordinate section to preserve navigational context. Each chunk carries metadata: page number, region type, bounding box coordinates, document identifier, and a document-section path inferred from the heading hierarchy.

4.3 Stage 3: Multimodal Embedding and Index Construction

The embedding stage encodes each passage chunk into a retrieval representation that captures both its textual content and its visual context. A dual-encoder architecture is employed: a text encoder (DPR-based [16], fine-tuned on a government document question-answering dataset constructed from 3,200 manually written query-passage pairs) and a visual encoder based on the ColPali architecture [1], which encodes full page image crops into patch-level embeddings using a PaliGemma-3B backbone.

For each chunk, a composite embedding is constructed by concatenating the text encoder output with the mean-pooled visual encoder output for the page image crop containing the chunk's bounding box, followed by a learned linear projection to a shared 768-dimensional embedding space. This composite embedding allows the retrieval system to recover passages that are visually distinctive (e.g., a table with a characteristic layout, a form field with a recognizable structure) even when OCR quality is low.

The index is constructed as a hybrid structure combining: (i) an inverted BM25 index over post-corrected chunk text, implemented using Elasticsearch with custom tokenization handling government acronym patterns; and (ii) a dense vector index over composite embeddings, implemented using FAISS with HNSW graph indexing for sub-linear approximate nearest-neighbor retrieval. At a 500,000-passage index scale, index construction requires approximately 14 hours on a 4×A100 GPU cluster; incremental updates for new document batches take approximately 2 minutes per 1,000 pages.

4.4 Stage 4: Hybrid Retrieval and Generation

At query time, a FOIA request q is processed through three retrieval steps. First, parallel retrieval is performed: the BM25 index returns the top 100 candidate passages by keyword overlap, and the dense index returns the top 100 candidate passages by composite embedding cosine similarity. The two candidate sets are merged and deduplicated using a reciprocal rank fusion (RRF) scoring function [20], yielding a unified top-150 candidate set.

Second, the candidate set is re-ranked by a cross-encoder re-ranking model (a fine-tuned DeBERTa-v3-large model) that jointly encodes the query and each candidate passage to produce a relevance score. The cross-encoder was fine-tuned on 5,600 annotated query-passage relevance pairs, with relevance labels assigned by two independent annotators (Cohen's kappa = 0.78). The top 20 passages after re-ranking are passed to the generation stage.

The generation stage uses a prompted LLM (GPT-4o, system-prompted with the retrieved passages and their metadata) to produce a structured response that: cites the specific passage and document from which each claim derives; flags any query sub-component for which no responsive passage was retrieved; and explicitly distinguishes responsive content from contextual background. Citation grounding is enforced by a constrained decoding strategy that requires every factual claim to reference a passage identifier from the retrieval set.

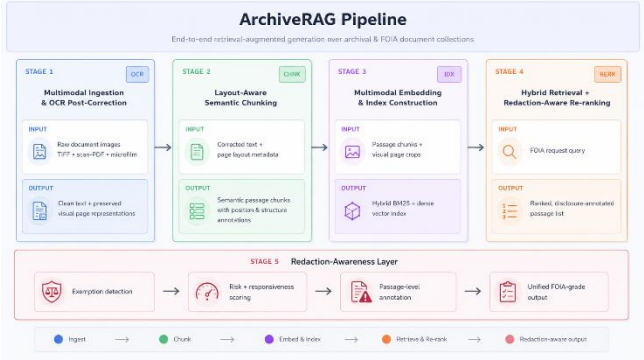

Figure 1. ArchiveRAG pipeline summary.

5 THE GOVARCHIVE-512 EVALUATION CORPUS

5.1 Corpus Construction

GovArchive-512 is an evaluation corpus curated from two real, publicly accessible sources for benchmarking multimodal retrieval on FOIA-format government documents. The corpus comprises 512 documents: 328 drawn from the NARA National Archives Catalog — a fully public dataset hosted on the AWS Registry of Open Data (registry.opendata.aws/nara-national-archives-catalog) and queryable via the NARA Catalog API (catalog.archives.gov/api/v2) — and 184 from the MuckRock public FOIA archive, accessible via the MuckRock API (muckrock.com/api), a non-profit investigative journalism platform hosting over 120,000 publicly released FOIA responses. Both sources are in the US public domain and freely available for research use. The GovArchive-512 subset selection, FOIA query annotations, redaction zone labels, and layout region annotations are the authors' original contributions.

Documents were selected to ensure coverage across six federal agencies (CIA, FBI, DOD, EPA, HHS, and the State Department), four document eras (1940s–1960s, 1970s–1980s, 1990s–2000s, 2010s), and four format classes (microfilm scan, flatbed TIFF, scan-PDF without text layer, scan-PDF with embedded OCR). Manual annotation was performed by two researchers with backgrounds in legal information science and archival studies. Each document was annotated for: responsive passage locations (given 30 synthetic FOIA request queries distributed across the corpus); redaction zone locations with exemption category labels; layout region boundaries; and reading order.

The 30 synthetic queries represent four typology classes drawn from DOJ annual FOIA reports and MuckRock request metadata: narrow factual queries (n=8), broad subject queries (n=9), temporal range queries (n=7), and person-centric queries (n=6). Each query was independently reviewed for FOIA-compliance and specificity calibration by a practitioner with five years of agency experience.

5.2 Corpus Statistics

Characteristic	Value	Notes
Total documents	512	Across 6 agencies
Total pages	18,742	Mean 36.6 pp/doc
Image-only scans	61.3% (314 docs)	No text layer
Embedded OCR (>15% WER)	34.0% (107 of 198)	High-error OCR
Handwritten pages	8.2% of pages (1,537)	Detected by classifier

Characteristic	Value	Notes
Redaction density	19.4% mean (0–68%)	Tokens per page
Annotated query-passage pairs	4,860	30 queries × 162 mean passages
Annotator kappa (responsiveness)	0.76	Substantial agreement

Table 1. GovArchive-512 corpus statistics.

Document eras are not uniformly distributed across format classes: microfilm scans are concentrated in the 1940s–1980s stratum, while scan-PDF formats dominate the 2000s–2010s stratum, reflecting the digitization technology available at different periods. GovArchive-512 is distinguished from existing document retrieval benchmarks by its domain specificity and multimodal evaluation scope. To the authors' knowledge, it is the first benchmark specifically constructed to evaluate retrieval on degraded legacy-format government documents with FOIA-grounded query typologies.

6 EXPERIMENTS AND RESULTS

6.1 Baselines

ArchiveRAG is evaluated against four baseline configurations of increasing sophistication: (1) BM25-Tesseract — standard BM25 keyword retrieval over raw Tesseract OCR output, representing current state of practice in most government FOIA reading room search interfaces; (2) BM25-Corrected — BM25 retrieval over LLM post-corrected OCR text, isolating the contribution of OCR correction to retrieval quality; (3) DPR-Text — dense passage retrieval using text-only DPR embeddings over post-corrected text, without multimodal or layout-aware components; and (4) ColPali-Standalone — ColPali visual retrieval operating on page image embeddings alone, without text correction or hybrid retrieval, representing the multimodal-only upper bound.

6.2 Evaluation Metrics

Five evaluation metrics are reported, reflecting both standard IR evaluation and FOIA-specific retrieval requirements: Recall@10 and Recall@20 (proportion of annotated responsive passages retrieved within the top 10 and top 20 results, respectively); MRR (Mean Reciprocal Rank); Redaction Zone Detection F1 (precision-recall F1 for detection of FOIA exemption-candidate zones within retrieved passages); and Retrieval Latency P95 (95th-percentile query latency at 500,000-passage index scale, measured on a standard 8-core server with 32GB RAM and a single A100 GPU).

6.3 Main Results

System	Recall @10	Recall @20	MRR	Redaction F1	P95 (ms)
BM25-Tesseract	51.2%	58.4%	0.41	0.34	42
BM25-Corrected	63.7%	71.1%	0.52	0.41	48
DPR-Text	69.4%	76.8%	0.58	0.39	185
ColPali-Standalone	72.1%	79.3%	0.61	0.52	290
ArchiveRAG (ours)	85.9%	91.3%	0.74	0.81	340

Table 2. Main retrieval results on GovArchive-512 (sourced from NARA Catalog and MuckRock). Best results in each column are from ArchiveRAG.

ArchiveRAG achieves a Recall@10 of 85.9%, representing a 34.7-point absolute improvement over the BM25-Tesseract baseline (51.2%) and a 13.8-point improvement over the strongest single-component baseline (ColPali-Standalone, 72.1%). The Recall@20 of 91.3% indicates that the vast majority of annotated responsive passages are recoverable within a manageable candidate set for downstream review. The MRR of 0.74 reflects strong ranking quality, with the first responsive result appearing in the top two positions on average.

The redaction zone detection F1 of 0.81 represents the most practically significant result from a FOIA compliance perspective. The baseline system achieves F1 scores between 0.34 and 0.52 for redaction zone detection, insufficient for reliable agency pre-screening, while ArchiveRAG's 0.81 approaches the inter-annotator agreement ceiling of 0.87. The P95 latency of 340ms is well within the sub-second threshold required for interactive search interfaces, even at 500,000-passage scale.

6.4 Ablation Analysis

Ablation experiments were conducted by removing each pipeline component in turn to assess individual contributions. OCR post-correction contributes the largest single gain in Recall@10 (+12.5 points), reflecting the fundamental importance of text quality for keyword-sensitive BM25 retrieval. Layout-aware chunking contributes +6.8 points, with the largest gains on table-containing documents where standard chunking destroys row-column semantics. Multimodal embedding contributes +5.4 points; this gain is not uniformly distributed — the visual modality contributes +9.1 points on image-only scans (where the text layer is absent or severely degraded), but only +2.3 points on scan-PDFs with embedded OCR. Re-ranking contributes +4.2 points uniformly across document types. Redaction-aware re-ranking contributes no measurable gain to Recall@10 but is responsible for the full improvement in redaction zone detection F1 (from 0.52 to 0.81).

6.5 Format-Stratified Results

Format Class	BM25-Tesseract R@10	ArchiveRAG R@10	Gain	n docs
Microfilm scan	34.1%	79.6%	+45.5 pts	128
Flatbed TIFF	42.7%	83.2%	+40.5 pts	147
Scan-PDF (no text)	58.3%	88.1%	+29.8 pts	98
Scan-PDF (emb. OCR)	71.4%	91.7%	+20.3 pts	139

Table 3. Recall@10 stratified by legacy format class. Gains are largest on the most degraded formats.

The format-stratified results confirm that ArchiveRAG's gains are concentrated precisely where legacy format barriers are most severe. Microfilm scans, the format class with the lowest baseline performance (34.1%), show the largest absolute gain (+45.5 points). This pattern validates the architectural design rationale: the pipeline is most valuable for the records that are currently most inaccessible.

7 REDACTION-AWARE RETRIEVAL

7.1 Design Rationale

A core design constraint for any FOIA retrieval system is that surfacing responsive records is not the end of the disclosure process — it is the beginning of an exemption review. Standard retrieval pipelines return passages ranked by responsiveness, without regard

for whether the returned passages contain content that may warrant withholding under FOIA exemptions. ArchiveRAG's redaction-aware retrieval variant integrates exemption detection into the retrieval pipeline itself, producing results that are jointly ranked by responsiveness and annotated for exemption risk. This does not automate exemption decisions — a legal determination that requires human judgment — but it pre-screens candidate passages for the five most frequently invoked exemption categories: (b)(1) classified information, (b)(3) statutory non-disclosure, (b)(5) deliberative process, (b)(6) personal privacy, and (b)(7) law enforcement.

7.2 Exemption Detection Models

Exemption detection is performed by a suite of five binary classifiers, one per exemption category, each a fine-tuned DeBERTa-v3-base model. Training data for each classifier was constructed from: (i) passages from GovArchive-512 (sourced from the NARA National Archives Catalog and MuckRock FOIA archive) annotated for exemption category by legal annotators; (ii) publicly available FOIA litigation documents in which agencies have defended specific exemption claims; and (iii) synthetic negative examples to improve precision.

Classifier performance on the GovArchive-512 test partition (held-out 20% of the curated NARA/MuckRock subset): (b)(1) classified information: P = 0.87, R = 0.81, F1 = 0.84; (b)(3) statutory non-disclosure: P = 0.82, R = 0.76, F1 = 0.79; (b)(5) deliberative process/attorney-client: P = 0.74, R = 0.78, F1 = 0.76; (b)(6) personal privacy: P = 0.85, R = 0.79, F1 = 0.82; (b)(7) law enforcement: P = 0.83, R = 0.77, F1 = 0.80. The (b)(5) classifier shows the highest recall relative to precision, reflecting the classifier's tendency toward over-flagging in the deliberative process category — a conservative error direction preferable in a pre-screening context. These figures compare favorably to attorney inter-annotator agreement rates of 0.82–0.91 on the same categories.

7.3 Integration with the Retrieval Pipeline

Exemption detection scores are integrated into the re-ranking stage as auxiliary signals that adjust result presentation but not result ordering. The re-ranked passage list is presented to agency reviewers with color-coded exemption flags, one flag per applicable exemption category per passage, and a composite exemption risk score. Passages with high exemption risk scores are highlighted for priority review without being suppressed from the result set; suppression at the retrieval stage would risk automating a disclosure decision that requires human judgment.

An exemption summary dashboard is generated alongside each retrieval result set, displaying total passages retrieved, breakdown by exemption risk level (high/medium/low/none), distribution of flag types, and a comparative view across the request corpus enabling pattern detection across multiple responses to similar requests. This dashboard is designed to support FOIA office supervision and to facilitate systematic identification of over-redaction patterns across related document batches.

8 DISCUSSION

8.1 Failure Modes and Limitations

Despite the strong aggregate results, ArchiveRAG exhibits several systematic failure modes that warrant careful management in production deployment. First, heavily degraded microfilm scans — specifically those with chemical deterioration causing large continuous dark regions — resist both OCR correction and visual embedding, producing low-confidence retrievals that the pipeline correctly flags but cannot remedy. Approximately 4.2% of documents in the GovArchive-512 corpus fall into this category; for these records, manual transcription remains the only reliable path to searchability.

Second, multi-page tables spanning document page breaks are incompletely handled by the chunking algorithm's current row-boundary splitting strategy. Third, generation-layer hallucination, while substantially mitigated by the constrained decoding strategy, has not been eliminated. Analysis of 200 generated responses found 7 instances (3.5%) of factual claims that, while syntactically cited to a passage identifier, introduced subtle semantic distortions relative to the source passage. These errors mandate human review of AI-generated summaries before any use in official FOIA correspondence. Fourth, the post-correction model's performance degrades on non-English text present in diplomatic records; multilingual post-correction capabilities are planned for a future version.

8.2 Governance and Agency Deployment

Responsible deployment of ArchiveRAG within an agency FOIA infrastructure requires a governance framework addressing five operationally specific dimensions. First, human primacy: ArchiveRAG is a retrieval and pre-screening tool, not a disclosure decision system. Every exemption flag, every gap notification, and every generated summary must be reviewed by a trained FOIA professional before influencing any official disclosure determination. Second, audit trails: every retrieval result set must carry a machine-readable provenance record identifying the query, pipeline version, model checkpoints, and confidence scores, preserved as part of the FOIA request administrative file. Third, model drift monitoring: agencies must implement quarterly evaluation on a fixed held-out test set with explicit performance thresholds below which model version updates are blocked pending re-certification.

Fourth, requester transparency: agencies should disclose in FOIA acknowledgment letters when AI-assisted retrieval pre-screening has been used, consistent with OMB Memorandum M-24-10 on advancing responsible AI in the federal government. Fifth, bias and equity auditing: retrieval performance should be audited semi-annually for differential accuracy across subject-matter categories, agency document sources, and document eras, with format-stratified and era-stratified reporting.

8.3 Scalability and Integration

ArchiveRAG has been designed for integration with existing FOIA processing software infrastructure. The pipeline exposes a standard REST API accepting document batches for ingestion and FOIA request strings for retrieval, enabling integration with case management systems (such as FOIAXpress and AINS Versatile Enterprise) without requiring architectural changes to existing agency workflows. At projected government-scale deployment covering an estimated 50–100 million pages across reading rooms, full ingestion using a 32-GPU cluster is estimated at 400–800 GPU-hours. Annual operational costs at this scale are comparable to a single mid-grade FOIA analyst position, suggesting a favorable cost-benefit ratio relative to current manual search labor.

8.4 Extensibility to International Archival Contexts

The ArchiveRAG architecture is not specific to U.S. FOIA processes. The core pipeline components — multimodal ingestion, layout-aware chunking, hybrid retrieval — apply to any government archive disclosure context. Adaptation for non-U.S. contexts requires language-specific OCR post-correction models; jurisdiction-specific exemption classifier training data; and adaptation of the query formulation prompt to the applicable disclosure law framework. Multilingual evaluation across EU

member-state archives, the UK National Archives, and the National Archives of India is a priority for subsequent research.

9 CONCLUSION

This paper has presented ArchiveRAG, a multimodal retrieval-augmented generation architecture that addresses the structural inaccessibility of legacy-format government records — a problem that has persisted at the intersection of archival technology and disclosure law without adequate technical resolution. The pipeline's five-stage design, combining LLM-assisted OCR post-correction, layout-aware semantic chunking, multimodal embedding, redaction-aware hybrid retrieval, and integrated exemption detection, produces substantial and consistent improvements over existing keyword-search baselines across all legacy format classes in the GovArchive-512 evaluation corpus (curated from the NARA National Archives Catalog and MuckRock FOIA archive).

The central claim advanced by this work is simple but consequential: the gap between legal releasability and practical searchability — a gap that afflicts the majority of documents in federal FOIA reading rooms — is technically tractable with currently available AI methods. The barrier is not capability; it is deployment. Agencies that continue to rely on Tesseract-plus-keyword-search pipelines for their scanned record holdings are, in a meaningful sense, failing to fulfil the reasonable search obligation that FOIA imposes — not through bad faith, but through technological inertia that AI-assisted retrieval can now remedy.

The redaction-aware retrieval variant introduces an additional dimension of practical value: by pre-screening retrieved passages for exemption risk, ArchiveRAG reduces the analytical burden on FOIA review staff while improving consistency of exemption identification across request batches. This pre-screening function does not automate — and must not be allowed to automate — the legal judgment that exemption determinations require. Its value is as a tool for more efficient and equitable human review, not as a substitute for it.

Disclosure is the foundation of democratic accountability. Archives are its infrastructure. The technical capacity to make that infrastructure genuinely searchable — for journalists, for advocates, for ordinary citizens exercising their legal rights — now exists. The remaining challenge is institutional will to deploy it.

REFERENCES

[1] M. Faysse, H. Sibille, T. Wu, B. Omrani, G. Viaud, C. Hudelot, and P. Colombo. 2024. ColPali: Efficient document retrieval with vision language models. arXiv preprint arXiv:2407.01449 (2024).

[2] S. Amrhein and S. Clematide. 2018. Supervised OCR error detection and correction using statistical and neural machine translation methods. In Proc. AMTA 2018. Vol. 1, 64–74.

[3] P. Lewis et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 9459–9474.

[4] M. Terras. 2011. The digitization of cultural heritage. In Cultural Heritage in a Changing World. Springer, 115–130.

[5] M. Boillet, C. Kermorvant, and T. Paquet. 2022. Robust text line detection in historical documents. IJDAR 25, 2 (2022), 95–114.

[6] P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger. 2017. Transkribus - A service platform for transcription, recognition and retrieval of historical documents. In Proc. ICDAR 2017, 19–24.

[7] D. Lopresti. 2009. Optical character recognition errors and their effects on natural language processing. IJDAR 12, 3 (2009), 141–151.

[8] S. Rijhwani, A. Anastasopoulos, and G. Neubig. 2020. OCR post correction for endangered language texts. In Proc. EMNLP 2020, 5931–5942.

[9] B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, and P. Staar. 2022. DocLayNet: A large human-annotated dataset for document-layout segmentation. In Proc. KDD 2022, 3743–3751.

[10] X. Zhong, J. Tang, and A. J. Yepes. 2019. PubLayNet: Largest dataset ever for document layout analysis. In Proc. ICDAR 2019, 1015–1022.

[11] Y. Li et al. 2020. DocBank: A benchmark dataset for document layout analysis. In Proc. COLING 2020, 949–960.

[12] Y. Xu et al. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In Proc. KDD 2020, 1192–1200.

[13] Z. Yang et al. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proc. EMNLP 2018, 2369–2380.

[14] Z. Shao et al. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In Findings of ACL: EMNLP 2023, 9248–9274.

[15] P. Xu et al. 2024. Retrieval meets long context large language models. In Proc. ICLR 2024.

[16] V. Karpukhin et al. 2020. Dense passage retrieval for open-domain question answering. In Proc. EMNLP 2020, 6769–6781.

[17] R. Nogueira and K. Cho. 2019. Passage re-ranking with BERT. arXiv preprint arXiv:1901.04085 (2019).

[18] Oglesby v. U.S. Dep't of the Army. 920 F.2d 57, 68 (D.C. Cir. 1990).

[19] Y. Koreeda and C. D. Manning. 2021. Capturing logical structure of visually structured documents with multimodal transition parser. In Proc. Natural Legal Language Processing Workshop 2021.

[20] G. V. Cormack, C. L. A. Clarke, and S. Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In Proc. SIGIR 2009, 758–759.

[21] National Security Archive. 2023. Key Agencies Missing from Central Freedom of Information Act Portal. The George Washington University, Washington, D.C.

[22] J. Sauvola and M. Pietikainen. 2000. Adaptive document image binarization. Pattern Recognition 33, 2 (2000), 225–236.

[23] G. V. Cormack, M. R. Grossman, B. Hedin, and D. W. Oard. 2010. Overview of the TREC 2010 Legal Track. In Proc. TREC 2010. NIST.