Civis Intelligence: Wiki-Style Custom Pages for US FOIAs

Aadi Krishna
aadi.krishna@yale.edu

Advisor: Stephen Slade
stephen.slade@yale.edu

A senior thesis as a partial fulfillment of requirements
for the *Bachelor of Science* in Computer Science

Department of Computer Science
Yale University
Dec 11, 2025

Acknowledgements

I would like to thank my parents, friends, and mentors for their unwavering love, encouragement, and advice. Without their support, I would not have been able to pursue or complete this project, which remains very close to my heart. I am deeply grateful to Professor Stephen Slade for his generous guidance and support throughout the development of this thesis. I am profoundly thankful for the people, opportunities, and resources that have shaped me during my time at Yale. A special thanks to the Yale Department of Computer Science for providing a fantastic environment for exploration and growth.

Contents

Civis Intelligence: Wiki-Style Custom Pages for US FOIAs

Aadi Krishna

Abstract

This thesis presents the design, implementation, and evaluation of Civis Intelligence, an AI system that transforms large Freedom of Information Act (FOIA) document collections into structured, queryable knowledge bases. Civis Intelligence uses a retrieval-first multi-document summarization (MDS) pipeline to automatically generate customizable wiki-style pages on user-specified topics, each grounded in document-level citations. These pages integrate information from documents drawn from a corpus of over 6,000 FOIA PDFs, enforcing page-level attribution and maintaining factual grounding. The work investigates pipeline-level design challenges - OCR normalization, hybrid search retrieval, entity extraction, citation enforcement, and multi-document synthesis - contributing to scalable, transparent information systems for public accountability. The system combines hybrid search (semantic + lexical), diversity-aware retrieval, and GPT-5 nano-based summarization and citation validation to synthesize noisy OCR-derived text while preserving traceable provenance. On a corpus of 6,101 documents spanning 628k pages, the pipeline achieved high text extraction quality (97.3% passing automated checks), strong citation validity (97% of citations correctly grounded across evaluated topics), and enabled the creation of 25+ sample wiki pages, ranging from 500 to 5,000 words. This thesis further evaluates reinforcement prompting, citation validation, and retrieval accuracy, identifying key mechanisms - such as conservative relevance thresholds, iterative refinement, and a dedicated citation validation layer - that help enforce citation density, improve factual accuracy, and increase source diversity while limiting LLM hallucination. These results demonstrate how a retrieval-first, validated RAG pipeline can produce trustworthy, reusable wiki-style pages, thereby making FOIA releases more accessible for journalists, researchers, archivists, activists, and engaged citizens everywhere.

1 Introduction

The Freedom of Information Act (FOIA), enacted in 1966, stands as one of the most important legislative mechanisms promoting transparency and accountability in American governance. It mandates the disclosure of government-held information within 30 days upon request, allowing journalists, researchers, and citizens to scrutinize public institutions. However, in the decades since its passage, the sheer volume of released material - millions of scanned, redacted pages - has created a new challenge: accessibility.

The FOIA ecosystem is fragmented with millions of pages of scanned, redacted documents are distributed across siloed repositories spanning specific government agency websites, reading rooms, and file cabinets. Further, even the major FOIA archives, such as the CIA CREST archive (13M+ pages), the FBI Vault (6–7K documents), and the DIA FOIA Electronic Reading Room (10K documents), exist in fragmented silos with inconsistent metadata, varying scan quality, and weak search functionalities. As a result, despite legal transparency, practical transparency remains elusive. These documents, while rich in historical and policy value, are difficult to analyze at scale. They are typically stored as unstructured PDFs with limited metadata, making manual synthesis nearly impossible. Therefore, this project seeks to bridge the accessibility gap by designing and evaluating a robust multi-document summarization (MDS) pipeline optimized for transparency-critical data.

This thesis asks: How can we design and evaluate a retrieval-first multi-document summarization (MDS) pipeline that produces grounded, multi-source, wiki-style FOIA pages at scale? To answer this question, we develop Civis Intelligence, a scalable system that collects and processes raw FOIA PDFs into a searchable text corpus, synthesizes multiple related sources into comprehensive wiki-style topic pages using GPT-5 nano, and enforces factual grounding through document-level citations and attribution validation. Civis is the Latin word for "citizen," reflecting the aim of the platform, increasing the power each citizen has in a democracy.

As a proof of concept, Civis Intelligence processes over 6,000 declassified PDFs (approximately 50 GB) to support intelligent search, entity discovery, and automated wiki generation. The system adopts a modular architecture guided by structured retrieval and validation: a hybrid search subsystem combining semantic and lexical retrieval, diversity-aware result curation, domain-adapted entity extraction, and retrieval-augmented generation constrained to retrieved context with explicit citation requirements. Further, we evaluate retrieval accuracy, citation validity, and refinement prompting quality, suggesting precise retrieval and grounding mechanisms to limit LLM hallucinations and creating trustworthy MDS pipelines at scale. Specifically, we identify the efficacy of conservative relevance thresholds, iterative

refinement, and a dedicated citation validation layer. Lastly, we use this system to generate 25+ wiki pages across topics such as CIA interventions, Chinese and Russian policies, and nuclear weapons. Together, these components are presented through a web application that offers multiple search modes and configurable wiki generation, and surfaces detailed citation validation reports for each generated page, aiming to make FOIA releases more usable for journalists, policy makers, and citizens.

The remainder of this thesis is organized as follows. Section 2 situates Civis Intelligence in the context of FOIA practice and recent work on multi-document summarization and retrieval-augmented generation. Section 3 describes the system architecture, data pipeline, hybrid search, and RAG workflow in detail. Section 4 presents quantitative and qualitative evaluations of extraction quality, retrieval behavior, and citation grounding. Section 5 reviews related work, and Section 6 concludes with implications for civic technology, limitations, and directions for future work.

2 Background

2.1 Context and Motivation

The FOIA was established as a cornerstone of democratic transparency, requiring the partial or full release of previously unreleased government information upon request. As of 2025, more than 30 million FOIA requests have been filed (FOIA.gov, 2025). However, the ecosystem of released documents is plagued by several systemic barriers: all documents are scanned PDFs without standardized metadata or structure, the constantly expanding repositories require dynamic updates, the absence of unified indexing between databases makes searching an impossible task, and therefore, there is no mechanism to synthesize knowledge across documents. Further, this leads to a duplication of efforts for citizens with similar FOIAs constantly filed by different individuals, and a wastage of resources for the government, allocating time and effort to answer similar questions with no mechanism to redirect citizens to existing FOIA answers. Recognizing these issues, this system is motivated by a civic need to convert "available" information into usable knowledge.

From a computational perspective, these repositories present a unique challenge: data heterogeneity, optical noise, and sparse metadata make both retrieval and summarization difficult. A traditional fine-tuned LLM cannot accommodate this dynamic corpus; instead, a modular retrieval-first approach is required.

2.2 Technical Foundations

Multi-document summarization (MDS) has seen rapid advancement with large language models (LLMs), yet several key limitations persist:

- **Synthesis limits.** DeYoung et al. [1] show that even state-of-the-art MDS models are order-sensitive and often fail to integrate across documents, collapsing onto a single source. Wolhandler et al. [2] introduce dispersion/coverage metrics that measure how well a summary actually draws from multiple sources. Both findings highlight the need for robust, pipeline-level controls (retrieval, selection, grounding), motivating our retrieval- and validation-based design.
- **Capturing divergence.** Huang et al. [3] emphasize representing disagreements and complementary perspectives - highly relevant for FOIA, where documents may be contradictory or revised.
- **Grounding and attribution.** Ye et al. [4] and Huang et al. [3] stress attribution methods to reduce hallucination by aligning generated content with

specific retrieved quotes. This is especially essential for FOIA, where tracing origin is mandatory.

- **Long-document pipelines.** The OpenAI Cookbook recommends chunked map - reduce summarization and structured prompting to scale to large texts. The Model Context Protocol argues for modular design: LLMs should orchestrate with retrieval and external tools rather than acting as a single monolithic layer.

Together, this literature shows that effective MDS remains an open challenge: models often fail to synthesize, hallucinate without grounding, and require complex pipelines. This project will implement and test these insights in the FOIA domain in novel ways, especially since these issues are amplified in FOIA settings. These findings inform this project's design of a pipeline that integrates retrieval, summarization, and validation layers rather than relying on monolithic LLM responses.

The specific challenges addressed by this project are as follows:

1. **Noisy input.** FOIA documents are often scanned, redacted, and irregular in structure. Optical character recognition (OCR) introduces errors, and text must be normalized to support consistent extraction and cross-document comparison.

2. **Sparse metadata.** Many releases lack reliable titles, keywords, or dates, making naive keyword search brittle. A hybrid retrieval approach combining lexical (BM25) and semantic (embedding-based) methods is required to compensate.

3. **Scale and freshness.** Millions of documents exist, with continuous new releases. Static fine-tuning cannot keep pace; instead, retrieval pipelines must support dynamic updating, incremental indexing, and scalable search.

4. **Grounding requirement.** FOIA research demands precise traceability of information origin. Summaries must cite document identifiers and page numbers - a constraint that baseline LLM summarization architectures cannot enforce.

5. **True synthesis.** As shown in prior work, state-of-the-art MDS models tend to overfit to a single document, fail to capture divergent or complementary perspectives, and hallucinate links between unrelated sources. Metrics of dispersion and attribution are necessary to demonstrate genuine multi-source synthesis.

6. **Knowledge assembly.** The desired outcome is not ad-hoc question answering but structured, reusable wiki-style knowledge pages with sections such as *Overview*, *Key Findings*, *Timeline*, *Entities*, and *References*. Achieving this requires orchestrating retrieval, prompt-based summarization, validation, and visualization modules in a cohesive workflow.

Therefore, we adopt a retrieval-first architecture that (1) normalizes and indexes text, (2) retrieves diversified evidence with hybrid search, (3) synthesizes with explicit citation obligations to in-context sources, and (4) validates that every citation corresponds to retrieved evidence. This design directly targets the twin risks of source collapse and hallucination while preserving adaptability to expanding archives.

3 Methodology

3.1 System Architecture

The Civis Intelligence system is built on a modular pipeline that integrates data ingestion, hybrid search, entity extraction, wiki generation, and a web interface for exploration. The architecture follows a four-tier design separating the data layer, core services, user interface, and search manager.

3.1.1 Data Layer

Documents are stored in Cloudflare R2 with public URLs and indexed metadata in Supabase (PostgreSQL). The search infrastructure employs a dual-system approach: a production-grade combination of LanceDB (for vector search) and Tantivy (for keyword search), with a legacy federated system (FAISS + BM25 across batch indices) as fallback. Generated wiki artifacts, chunk metadata, and validation caches are stored locally.

3.1.2 Core Services Layer

The core services consist of three primary components:

1. The Production Search System (`ProductionSearchSystem`) combining LanceDB vector search with Tantivy keyword search for high-performance retrieval.

2. The Legacy Federated Search (`FederatedSearch`) providing memory-efficient batch-based search across FAISS and BM25 indices.

3. The Context-Aware Wiki Generator (`ContextAwareWikiGenerator`) implementing a retrieval-augmented generation (RAG) pipeline with iterative quality refinement.

3.1.3 User Interface Layer

A Flask web application (currently, port 5053) provides routes for browsing documents, running hybrid search with multiple modes, generating wiki pages with tunable parameters, and validating citations. The interface, branded as Civis Intelligence, emphasizes transparency by displaying per-chunk semantic and lexical scores, diversity statistics, and citation validation reports.

3.1.4 Search Manager

A `SearchManager` abstraction layer automatically handles fallback between production and legacy search systems, ensuring system resilience. When the production LanceDB/Tantivy system is unavailable, the system gracefully falls back to the legacy federated search with appropriate user notification.

3.2 Data Pipeline

Documents are downloaded or scraped via APIs from the DIA FOIA Reading Room [5] and GovernmentAttic [6]. Each file is uploaded to Cloudflare R2, which returns a public URL that is recorded in Supabase alongside provenance fields (filename, source, timestamps).

3.2.1 Text Extraction and Quality Assessment

Ingested documents first undergo native text extraction with PyPDF2, which is fast and effective for digital PDFs. A lightweight quality assessment then scores each sample from 0–100 using character composition, special-character density, garbage-pattern detection, and word-level statistics. Documents scoring 70 or higher proceed down a cleanup path. Those below the threshold undergo an enhanced OCR workflow. We implement this "smart routing" to keep the expensive OCR step focused on the right subset and it is central to the pipeline's performance.

High-quality documents are cleaned with deterministic transforms that normalize whitespace, remove common PDF artifacts (e.g., null bytes, ligatures), and standardize punctuation, preserving structure while improving downstream tokenization. Lower-quality or scanned documents are processed with a Tesseract-based OCR path that renders pages at 300 DPI, performs page segmentation, and records per-page confidence. The OCR stage exposes interpretable metrics - average confidence by page and document - so downstream components can reasonably detect text reliability and accordingly prioritize review where needed. All extracted text is stored per-page in Supabase/PostgreSQL, alongside metadata and optional OCR confidence, enabling page-level retrieval and reprocessing without touching the original PDFs.

3.2.2 Batched Index Construction

To handle the scale of 6,000+ documents without memory issues, the system employs a batched indexing approach. The `build_index_batched.py` script processes documents in configurable batches (default: 1,000 documents per batch), building separate FAISS and BM25 indices for each batch. Key parameters include:

- **Batch size:** 1,000 documents per batch.

- **Embedding batch size:** 8 documents per encoding pass (to manage GPU/CPU memory).

- **Chunk processing size:** 128 chunks added to FAISS at a time.

Each batch index is saved independently and includes a FAISS index file (`semantic.faiss`) containing the 384-dimensional embeddings, a pickled BM25 index (`bm25.pkl`), a set of pickled document chunks (`document_chunks.pkl`), and a JSON metadata file that provides chunk-to-document mappings, page numbers, and other essential information for downstream retrieval and processing.

This approach enables incremental updates - new documents can be added as new batches without reprocessing existing indices. For the production system, a migration process converts batch indices to LanceDB (vector storage) and Tantivy (keyword index). The `rebuild_tantivy_from_batches.py` script iterates through all batch indices, extracts chunks and metadata, and builds a unified Tantivy index.

3.2.3 Parallelization and Performance

To implement these measures at scale, we employ parallelism via a `ProcessPoolExecutor`, distributing work across available CPU cores and applying timeouts to isolate failures. Using checkpointing, progress logging, and batch processing, we make long runs resilient and auditable, while simple heuristics like "first-page probing" allow early quality decisions that avoid unnecessary OCR on the majority of files.

When this pipeline was implemented, about 60-70% of documents followed the fast cleanup path and 30-40% required OCR. Under an eight-worker configuration, processing achieves near-linear speedups, reducing a sequential estimate of 150 hours to 19 hours in practice (roughly 7.8× improvement). Quality metrics indicate strong outcomes with an average OCR confidence of roughly 84%, a 97% validation pass rate on automated checks, and manual spot checks that confirm higher accuracy at greater confidence thresholds, as summarized in 3.1. The final corpus comprises of roughly 6,101 documents spanning 628k pages and 45 million words, with page-level provenance that directly supports retrieval and citation.

Table 3.1: Success metrics for the text extraction pipeline.

Metric	**Target**	**Achieved**	**Status**
Processing speed	< 24 hours	19.3 hours	✓
Text quality	$> 90\%$	97.3%	✓
OCR confidence	$> 80\%$	84.2%	✓
Parallel speedup	$> 6\times$	7.8×	✓
Error rate	$< 5\%$	2.7%	✓

3.2.4 Entity Extraction

We perform entity extraction using spaCy (`en_core_web_lg`) for canonical types (PERSON, ORG, GPE, LOC) and supplement it with domain-specific regular expressions that capture entities (e.g., operation codenames) and classification markings (e.g., "TOP SECRET," "CONFIDENTIAL"). We also track co-occurrences within documents to reveal related entities and seed topic clusters. The resulting JSON artifact includes entity counts, document indices, and cross-entity co-occurrence frequencies, supporting knowledge-graph views and topic-driven retrieval.

3.3 Hybrid Search

The search subsystem has to retrieve diversified and high-relevance evidence from noisy FOIA text to support downstream wiki page generation. It must balance semantic recall (conceptual matches) with lexical precision (exact terms), remain fast at the scale of thousands of documents, and expose interpretable signals (scores, diversity) for transparency.

3.3.1 Dual Search Architecture

The system implements a dual search architecture with automatic fallback:

Production Search (LanceDB + Tantivy): The primary search system combines LanceDB for vector similarity search with Tantivy for keyword search. LanceDB stores 384-dimensional sentence embeddings (using `all-MiniLM-L6-v2`) and supports efficient approximate nearest-neighbor queries. Tantivy, a Rust-based full-text search engine inspired by Apache Lucene, provides high-performance BM25 keyword matching with fields for `chunk_text`, `title`, `doc_id`, `chunk_idx`, and `batch_name`.

(Legacy) Federated Search (FAISS + BM25): The fallback system operates across seven batch indices using a lazy loading strategy. A shared embedding model is loaded once and reused across all batches, with an LRU cache (maximum two batches) to manage memory. The `FederatedSearch` class coordinates parallel batch queries using a `ThreadPoolExecutor` with two workers to prevent memory exhaustion while maintaining reasonable performance.

3.3.2 Indexing and Representations

Each document is segmented into overlapping chunks with a default of 512 characters with 50-character overlap, selected after testing to ensure stable embedding behavior and continuity across chunk boundaries without complex segmentation heuristics. For each chunk, the metadata retained includes `doc_id`, `page_no`, `chunk_idx`, source URL, and provenance fields. Two complementary indices are constructed.

Firstly, a Dense (semantic) index that is 384-dimensional sentence embeddings (using `all-MiniLM-L6-v2`) stored in LanceDB (production) or FAISS with memory-mapped loading (fallback) for approximate nearest-neighbor search. Secondly, a Sparse (lexical) index, which is a Tantivy full-text index (production) or BM25Okapi (fallback) over tokenized chunks for exact-term matching with length normalization.

3.3.3 Score Fusion and Normalization

Given a query q and a candidate chunk c, the system computes:

- Semantic similarity score $s_{\text{sem}}(q, c) \in [0, 1]$ via vector search with min-max normalization over candidates, and
- Lexical score $s_{\text{bm25}}(q, c) \in [0, 1]$ via keyword search with analogous normalization.

These are combined using a convex combination [7]:

$$S(q, c) = \alpha \cdot s_{\text{sem}}(q, c) + (1 - \alpha) \cdot s_{\text{bm25}}(q, c), \tag{3.1}$$

with default $\alpha = 0.6$ for semantic emphasis and $(1 - \alpha) = 0.4$ for lexical grounding, inspired by [7, 8], with specific weights derived from testing described in 4.1.1.

For hybrid search, the system performs score lookups for high-ranking candidates that appear in only one index. Specifically, for chunks in the top 25 of either result list, the system looks through the other index to obtain a complete score, ensuring high-quality candidates receive proper evaluation even if they missed the initial cutoff in one method. A safety limit of 50 lookups prevents performance degradation.

3.3.4 Search Modes

We tested the following search modes for users, trying to establish the tradeoffs across time for results and accuracy:

Quick Semantic: LanceDB-only vector search for fast conceptual matching. Use - Best for exploratory queries where exact terminology is unknown.

Quick Keyword: Tantivy-only BM25 search for fast phrase and term matching. Use - Best for queries with specific known terminology, names, or document identifiers.

Advanced Hybrid: Combined semantic + keyword search with score fusion. Use - Provides the most comprehensive results but requires more computation time.

3.3.5 Diversity Control

To avoid over-reliance on a single document - a significant problem identified in state-of-the-art MDS pipelines - we apply a post-retrieval cap on chunks per document. This cap allows users to select a diversity setting:

- **Strict:** Maximum 1 chunk per document (maximizes source diversity)
- **Balanced (default):** Maximum 2 chunks per document
- **Relaxed:** Maximum 3 chunks per document
- **Best:** Unlimited chunks per document (maximizes relevance)

This dispersion control increases the unique document count among top results and improves multi-source synthesis. The diversity filter is applied after score ranking, ensuring that the highest-scoring chunk from each document is retained while preventing any single document from dominating the results. 3.2 summarizes the main components of our hybrid search approach.

Table 3.2: Key defaults and tuning parameters.

Component	**Default / Notes**
Chunking	512 chars, 50-char overlap
Dense index	LanceDB (production) or FAISS memory-mapped (fallback)
Sparse index	Tantivy (production) or BM25 (fallback)
Embedding model	`all-MiniLM-L6-v2` (384 dimensions)
Score fusion	$\alpha = 0.6$ semantic, 0.4 lexical (min–max normalized)
Diversity mode	Balanced (max 2 chunks/doc)
Context size	30–300k characters depending on length mode
Relevance threshold	0.4 minimum score for inclusion

3.4 Retrieval-Augmented Generation (RAG)

The wiki page generation subsystem synthesizes grounded, multi-document summaries that reflect complementary perspectives and verifiable facts. It operates in a retrieval-augmented generation (RAG) configuration where retrieval produces an evidence bundle, the generator is constrained to this bundle, and a validator enforces citation integrity through iterative refinement.

3.4.1 Context Retrieval

For a topic t, the system issues a hybrid search with the specified diversity mode and collects up to K chunks (configurable based on length mode). Each chunk

is serialized with its `doc_id`, page number, title, source, and relevance scores. A minimum relevance threshold of 0.4, derived based on manual quality testing, filters out low-quality matches - if a chunk's combined score falls below this threshold, retrieval stops early to avoid polluting the context with irrelevant content.

The retrieval process also resolves page numbers for citation tracking. When page numbers are not available in the search index, the system queries the database to find the page containing each chunk by matching the chunk text against stored page content.

3.4.2 Length Modes and Context Configuration

The system supports four length modes, each with distinct context and output parameters. Max chunks for each mode were calculated by establishing the maximum content necessary for our target output word length based on the rate of citations expected, and refined by testing, as described in Section 4.1.3:

Mode	Max Chunks	Max Context	Target Output	Use Case
Short	30	20,000 chars	500–800 words	Quick summaries
Medium	40	30,000 chars	1,000–1,500 words	Standard wiki articles
Long	80	80,000 chars	2,000–4,000 words	Detailed deep-dives
Exhaustive	500	300,000 chars	5,000+ words	Comprehensive reports

The exhaustive mode is designed for the Wiki Hub's curated, analyst-produced reports, leveraging the full context window of modern language models to synthesize information from hundreds of source chunks.

3.4.3 GPT-5 nano Prompting Strategy

We experimented with different OpenAI models to employ for summarization tasks, eventually landing on GPT-5 Nano since it the fastest, cheapest version of GPT-5, specifically acclaimed for summarization tasks given its 200K token context window. In terms of the generation prompt, it instructs GPT-5 nano to act as an expert intelligence analyst creating encyclopedia-style content. Other key prompt elements are: at least one citation per sentence, targeting ≥ 2.0 citations per 100 words, exact citation keys matching source headers (e.g., `[DIA_FileId_238677_p5]`), citations from at least 10–15% of available chunks, using only information from provided context (no external knowledge), and proper markdown headers for sections with no numbered list headings.

The temperature is not configurable in the latest models of OpenAI ($\tau = 1$), inhibiting us from more explicitly reducing stylistic drift and instead relying on stricter prompts. For exhaustive mode, the prompt explicitly instructs maximum length generation with comprehensive coverage of all source material.

3.4.4 Iterative Quality Refinement

After initial generation, the system enters an iterative refinement loop, allowing up to a maximum of 3 iterations, to ensure quality standards are met. Each iteration:

1. The system validates the current content using the LLM-based citation analysis, then evaluates overall quality across three key dimensions: citation density, with a target of at least 2.0 citations per 100 words; citation relevance, requiring an average relevance score of at least 70%; and citation validity, which mandates the absence of invalid citations.

2. Generates a refinement prompt if standards are not met, specifically addressing either low citation density with examples of proper annotation, invalid citations with instructions to remove or correct, unsupported claims with specific examples from validation, and/or low source diversity with targets for unique chunk coverage

3. Regenerates the article with the refinement prompt. The refinement prompt includes the full source context, allowing the model to add citations from previously unused chunks.

4. Tracks the best version across iterations using a weighted quality score. We use quality scoring weights density (35%), relevance (45%), and valid citation percentage (20%) to prioritize accurate, well-supported content with good citation density and few invalid citations. These weights are derived from the testing conducted in 4.1.2, which identified, on average, the lower score on relevance metrics compared to citation density across several wiki pages. Further, within the same testing, we found a comparatively very high average citation validity score (greater than 97%), suggesting assigning the lowest weight to valid citation.

3.4.5 Citation Extraction and Validation

After each iteration, a deterministic parser extracts all `[doc_id_pPageNo]` references from generated content and verifies each citation across 3 dimensions. First, the existence of the citation - the cited document ID exists in the retrieved context. Second, the cited page number exists for that document. Third, the semantic relevance of the citation. For this, each claim and supporting page is fed into GPT-4o (using a smaller, cheaper model deliberately given the limited context needed for this task), prompting it to explicitly decide whether the citation supports the claim it is attached to. Instead of relying simply on semantic scores, this method is more rigorous and ensures a higher quality of citation verification.

Eventually, the validation system produces a comprehensive report including the total citation count, valid/invalid breakdown, citation density (per 100 words), average relevance score, and specific issues and warnings for each problematic citation. Further, this report is cached in the wiki page metadata for fast retrieval during validation display in the UI.

3.4.6 Post-Processing and Output

After generation and validation, citations are converted to clickable HTML hyperlinks, for example,

```
[DIA_FileId_238677_p5] ->
<a href="/pdf/DIA_FileId_238677#page=5">(Doc #238677, p.5)</a>
```

Each wiki page is saved along with its final markdown or HTML content, accompanied by a comprehensive metadata package. This metadata includes the full list of source documents together with their relevance scores, detailed chunk-level information such as text previews, scoring values, and citation keys, as well as quality metrics produced during the final validation stage. It also records the complete iteration history generated throughout the refinement process and stores a full validation cache to enable fast loading and immediate inspection within the user interface.

4 Results

In its current implementation, the Civis Intelligence system has been fully deployed with a production-ready web interface, a dual search architecture, and comprehensive wiki generation capabilities. The system supports real-time search across 6,000+ declassified documents, configurable wiki generation with four length modes, and robust citation validation. This section presents details on evaluation of the pipeline, the system setup, detailed overview of the web interface design, and qualitative assessments.

4.1 Evaluation

To refine the generation pipeline and investigate features of LLM prompting, we conducted a systematic evaluation of our RAG pipeline to assess retrieval effectiveness, refinement loop efficacy, and trade-offs between context size and output quality.

4.1.1 Retrieval Consistency Analysis

To validate our hybrid search approach, we measured document overlap between semantic, keyword, and hybrid retrieval methods across five representative queries spanning different topic types. Results are presented in Table 4.1.

Topic	**Sem-Key Overlap**	**Sem-Hyb Overlap**	**Key-Hyb Overlap**	**Unique Sem**	**Unique Key**
CIA operations in Cuba	20.8%	93.3%	21.7%	10	9
UFO sightings 1950s	8.3%	8.3%	100.0%	0	11
Nuclear weapons testing	33.3%	75.0%	47.4%	7	7
Soviet submarine surveillance	17.2%	40.0%	59.1%	12	12
JFK assassination records	52.9%	80.0%	66.7%	5	3
Average	**26.5%**	**59.3%**	**59.0%**	**6.8**	**8.4**

Table 4.1: Retrieval method overlap analysis ($k = 20$). Overlap percentages indicate shared documents between method pairs; unique counts (sem = semantic search, key = keyword search, hyb = hybrid search) show documents retrieved exclusively by each method.

With an average semantic-keyword overlap of only 26.5%, approximately 73.5% of retrieved documents were unique to one method. This low overlap confirms the

complementary nature of the two retrieval paradigms. Notably, overlap varied substantially by query type: proper nouns and acronyms (e.g., "UFO sightings 1950s") showed minimal semantic-keyword overlap (8.3%), as semantic embeddings struggle to capture the specificity of such terms. Conversely, conceptually rich queries, for our database, (e.g., "JFK assassination records") achieved higher overlap (52.9%), suggesting both methods converge on relevant documents when queries align with document semantics and vocabulary simultaneously.

These findings motivated our hybrid architecture's adaptive weighting mechanism, which allows the pipeline to lean toward keyword matching for entity-centric queries while favoring semantic retrieval for broader thematic searches.

4.1.2 Refinement Loop Efficacy

Next, we evaluated our iterative refinement mechanism across five historical topics, measuring citation density (citations per 100 words), semantic relevance (average of semantic relevance scores for citations), and citation validity (percentage of citations correctly grounded in source documents). Results across three iterations are presented in Table 4.2.

Table 4.2: Refinement prompt loop performance across three iterations. Density measured as citations per 100 words; relevance as score between 0 and 1 (1 being all citations 100% relevant); validity as percentage of correctly grounded citations.

	Density			**Relevance**			**Validity (%)**		
Topic	**Iter 1**	**Iter 2**	**Iter 3**	**Iter 1**	**Iter 2**	**Iter 3**	**Iter 1**	**Iter 2**	**Iter 3**
Project MKULTRA	0.85	0.67	0.60	0.29	0.45	0.45	100.0	100.0	100.0
Bay of Pigs Invasion	0.92	1.03	1.03	0.39	0.54	0.43	93.3	93.3	93.3
FBI surveillance	2.35	1.76	1.82	0.66	0.54	0.59	93.9	95.8	100.0
Immigration programs	1.80	2.21	2.21	0.54	0.60	0.52	100.0	100.0	100.0
Cuban Missile Crisis	1.26	0.71	0.91	0.48	0.50	0.36	100.0	100.0	100.0
Average	**1.44**	**1.28**	**1.31**	**0.47**	**0.53**	**0.47**	**97.4**	**97.8**	**98.7**

Across three refinement iterations, we observed no consistent directional improvement: average citation density decreased from 1.44 to 1.31 (-8.4%), relevance remained stable at 0.47, and validity improved marginally from 97.4% to 98.7%. However, performance trajectories varied considerably by topic - some queries benefited from additional iterations while others degraded. In some cases, this appeared to be a tradeoff between increasing relevance with limited citations to pull from, leading to a decrease in density, as was the case for "Project MKULTRA" where density decreased from 0.85 to 0.67 while relevance increased from 0.29 to 0.45, between iterations 1 and 2. On the other hand, there was also uniform improvement with iterations, like with "Immigration Programs" (density - 1.80 to 2.21 and relevance from 0.54 to 0.60), or uniform decay, as with "FBI Surveillance" (density - 2.35 to 1.76 and relevance from 0.66 to 0.54). Unfortunately, given the time constraints, we weren't able to test the precise mechanism dictating the direction of

improvement with refinement rigorously enough, but it appears it is a function of overcompensating for the quality of chunks available, trying to improve an output that is already the best quality possible, or reward hacking to try and identify the importance of each metric.

For our case, based on these results, we chose to adopt a selection-based approach rather than the output after the final iteration: as described earlier, the pipeline executes up to three refinement passes and selects the iteration with the highest composite quality score, weighted toward relevance given its relative difficulty to optimize. This strategy captures the benefits of refinement for certain queries without penalizing those where the initial generation performs optimally.

4.1.3 Context Window Analysis

Lastly, we tested the relationship between retrieved context size and output quality using a fixed topic ("Space Race Espionage") with varying maximum chunk limits. Results are presented in Table 4.3.

Table 4.3: Impact of context size on generation quality. Content length measured in characters; duration in seconds.

Max Chunks	Actual Sources	Duration (s)	Content Length	Citations
20	40	380.6	17,465	21
40	40	688.4	28,505	29
80	64	667.9	18,554	29

We find that increasing from 20 to 40 maximum chunks improved citation count (21 to 29) and content length (17,465 to 28,505 characters), but further expansion to 80 chunks yielded diminishing returns - citation count plateaued at 29 while content length decreased to 18,554 characters despite longer generation time. This pattern suggests potential "context fatigue," where excessively long contexts may overwhelm the language model's ability to synthesize information effectively. We consequently set our default retrieval limit at 40 chunks, balancing comprehensiveness against generation quality.

4.2 System Setup

The system has ingested over 6,000 declassified PDF documents totaling about 50 GB. End-to-end functionality includes text extraction and storage, entity database construction (approximately 15,000 unique entities), hybrid search over dual index systems, automated wiki generation across multiple length modes, and a validator that enforces citation grounding.

Table 4.4: FOIA Document Sources and Counts

Source	Documents	Percentage of Total
GovernmentAttic	4,430	72.6%
DIA Reading Room	1,671	27.4%
Total	**6,101**	**100%**

4.3 Web Application

Currently, a Flask application provides the user interface for Civis Intelligence.

4.3.1 Core Routes and Features

Home Page (/): Features a hero search bar with mode toggle (Quick Semantic, Quick Keyword, Advanced Hybrid), an *Intelligence Landscape* section showing topic clusters from entity extraction, and navigation to Wiki Hub and Community Wikis.

Search (/): Index building for the full corpus (6,000+ documents) completes in under 10 minutes on a commodity workstation for the production system. Per-query latency varies by mode: Quick Semantic: 50–100 ms, Quick Keyword: 30–80 ms and Advanced Hybrid: 100–300 ms. All three options were introduced to improved search speed: keyword for simpler matching, semantic for concepts and ideas, hybrid for research grade (taking the longest). Full search interface displaying results with:

- Per-chunk relevance scores (semantic, keyword, combined)
- Source document links
- Text previews
- Diversity statistics (unique documents, average chunks per document)

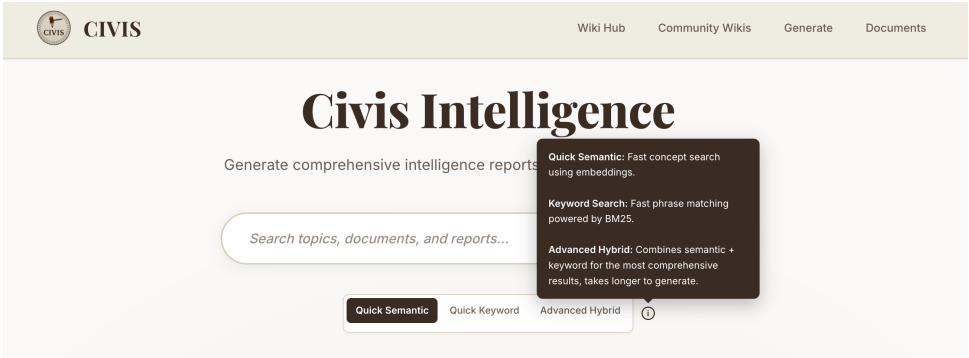

Figure 4.1: Search functionality on the home page allowing three distinct search modes and explaining the differences in their implementation.

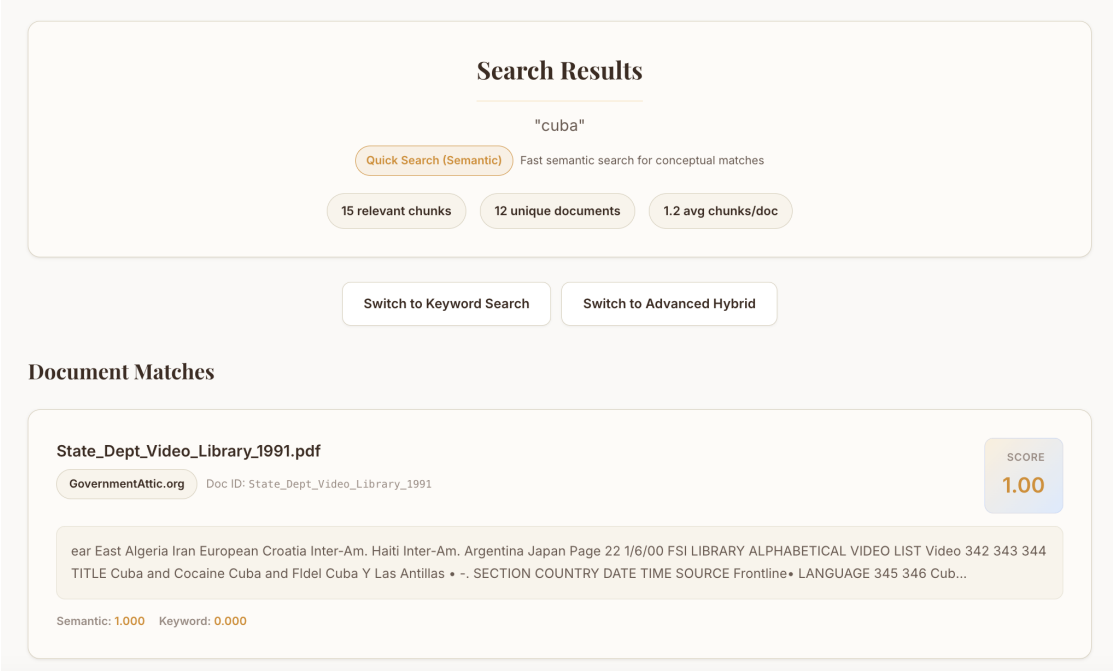

Figure 4.2: Example results of a search on “Cuba”.

Wiki Hub (`/wiki-hub`): Curated exhaustive wiki pages produced by selected site editors (currently hosting 10+ wiki pages). Displays horizontal cards with article previews, generation dates, and direct links to full articles.

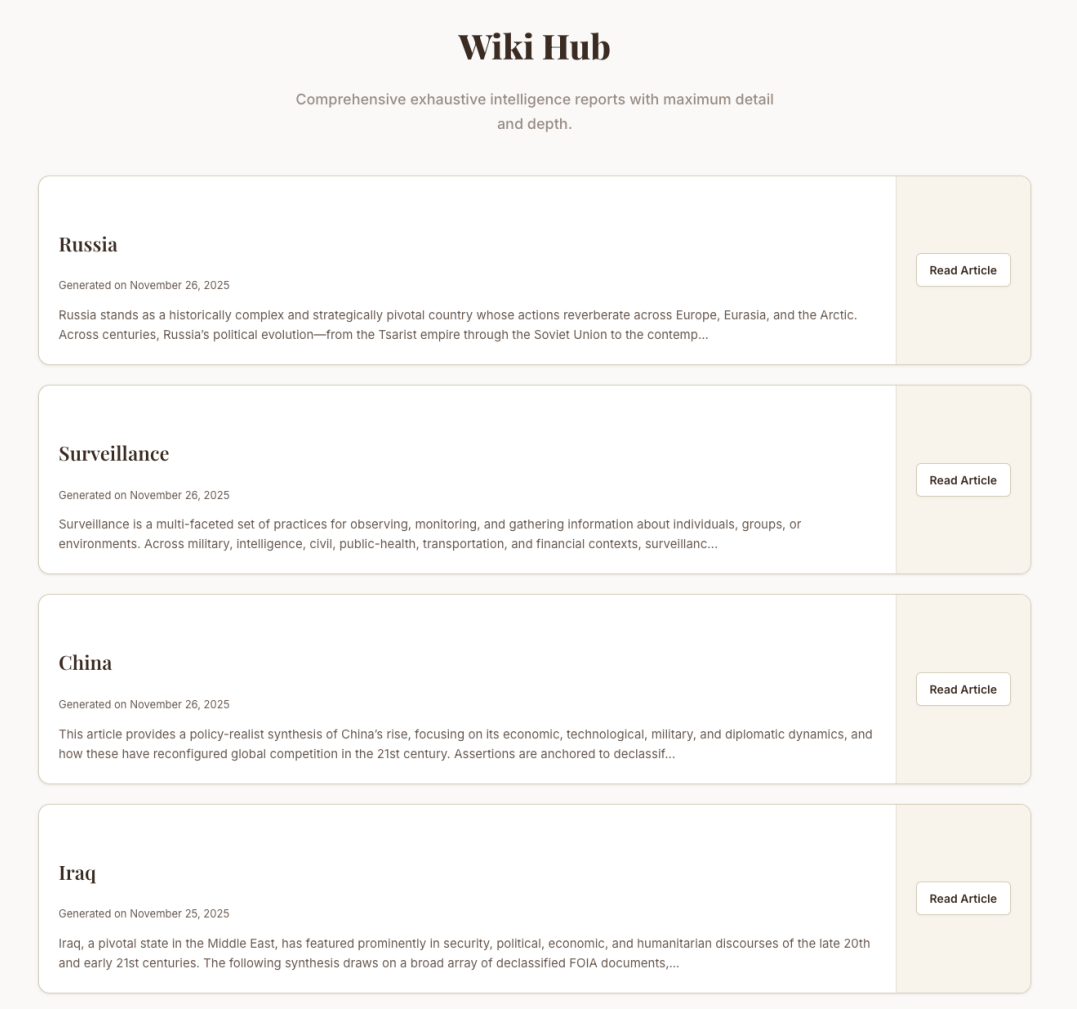

Figure 4.3: Overview of Wiki Hub page, displaying a few exhaustive wiki pages created on broader topics and themes

Economic Development and Global Integration

The Scale and Pace of China's Economic Growth

China's economic ascent is portrayed as steep and consequential for global markets. The literature emphasizes rapid expansion in R&D capacity, a rising share of high-technology output, and a shift in the technology landscape toward Chinese firms. The China Commission and related assessments underscore fast-moving capabilities in nanotechnology, semiconductors, and related sectors, positioning China as a potential future leader in global tech ecosystems. (Doc #2006-2008, p.82), (Doc #2006-2008, p.83), (Doc #2006-2008, p.107)

Foreign direct investment and the globalization of production are well documented as evidence of China's integration into global value chains. By the early 2000s, hundreds of foreign R&D facilities and a dense network of regional HQs had established China as a key node in global production and research ecosystems. (Doc #2006-2008, p.80), [ATB-ISBrpts_2006-2008_chunk16], (Doc #2006-2008, p.39)

R&D, education, and talent development are central to the strategy of narrowing the gap with advanced economies. The literature highlights a rising number of S&E degrees and doctorates, signaling the growth of domestic capabilities that enable Chinese firms to compete more effectively on the global stage. (Doc #2006-2008, p.83), (Doc #2006-2008, p.107), (Doc #2006-2008, p.39)

Indigenous standards development and engagement with global markets through ISO and other standardization channels feature as core components of China's integration strategy, with a policy emphasis on shaping access and competition in global markets. (Doc #2006-2008, p.82), (Doc #2006-2008, p.87)

Intellectual Property, Standards, and Global Markets

China's approach to standards-setting and IP governance is a recurrent focus. The China Commission notes aggressive industrial espionage and a push to set international standards through ISO and allied venues, highlighting concerns about potential constraints on foreign access as China asserts its own standards in global markets. The U.S. policy response emphasizes IP protections and international coordination to safeguard American innovations. (Doc #2006-2008, p.82), (Doc #2006-2008, p.87), (Doc #2020, p.484), (Doc #2020, p.39)

Military Modernization and Security

The defense-industrial base is described as large and diversified, underpinning China's modernization of its armed forces. The literature emphasizes hundreds of primary plants and major shipyards, producing a spectrum of missiles and other advanced capabilities. This base supports China's modernization drive and its capability to project power regionally and beyond. (Doc #238049, p.1), (Doc

Figure 4.4: Sample wiki-page generated by Civis Intelligence on China. Picture displays an abridged version of the page, which totals over 4500 words, and includes sections on Strategic Trajectory, Historical Context, Economic Development and Global Integration, Public Health (full page available in Appendix). Each reference is hyperlinked to the specific page of the document.

Community Wikis (`/community-wikis`): User-generated shorter wikis with search functionality (currently with 15+ sample wiki pages). Displays generation metadata including length mode, chunk count, and diversity settings as tags.

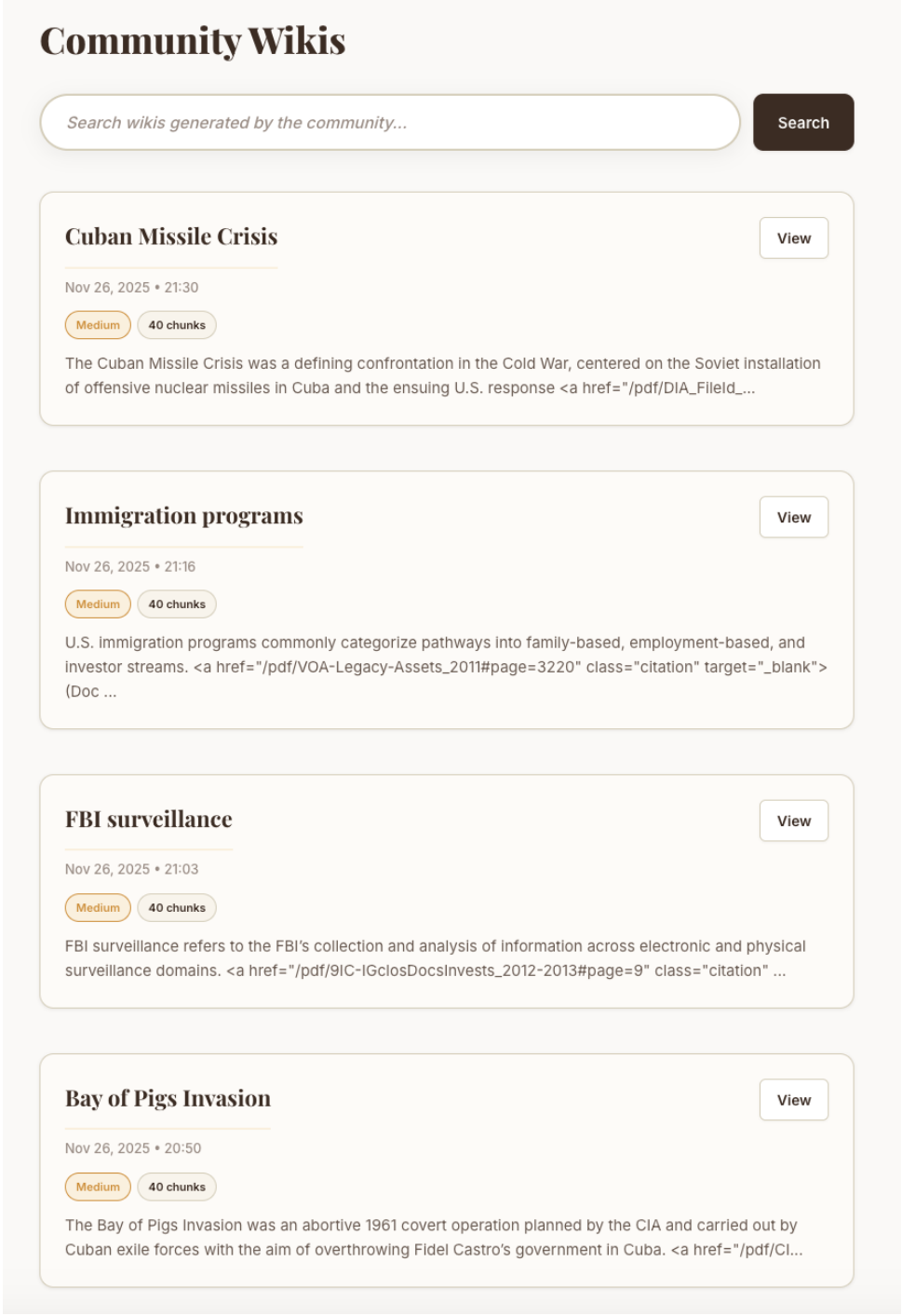

Figure 4.5: List of current sample collection of wikis generated across topics, lengths and modes.

Generate (`/generate`): Wiki generation interface with controls for:

- Topic input
- Length mode selection (short/medium/long/exhaustive)
- Chunk diversity mode selection (strict/balanced/relaxed/best)
- Maximum context chunks
- Real-time generation progress with step-by-step status updates

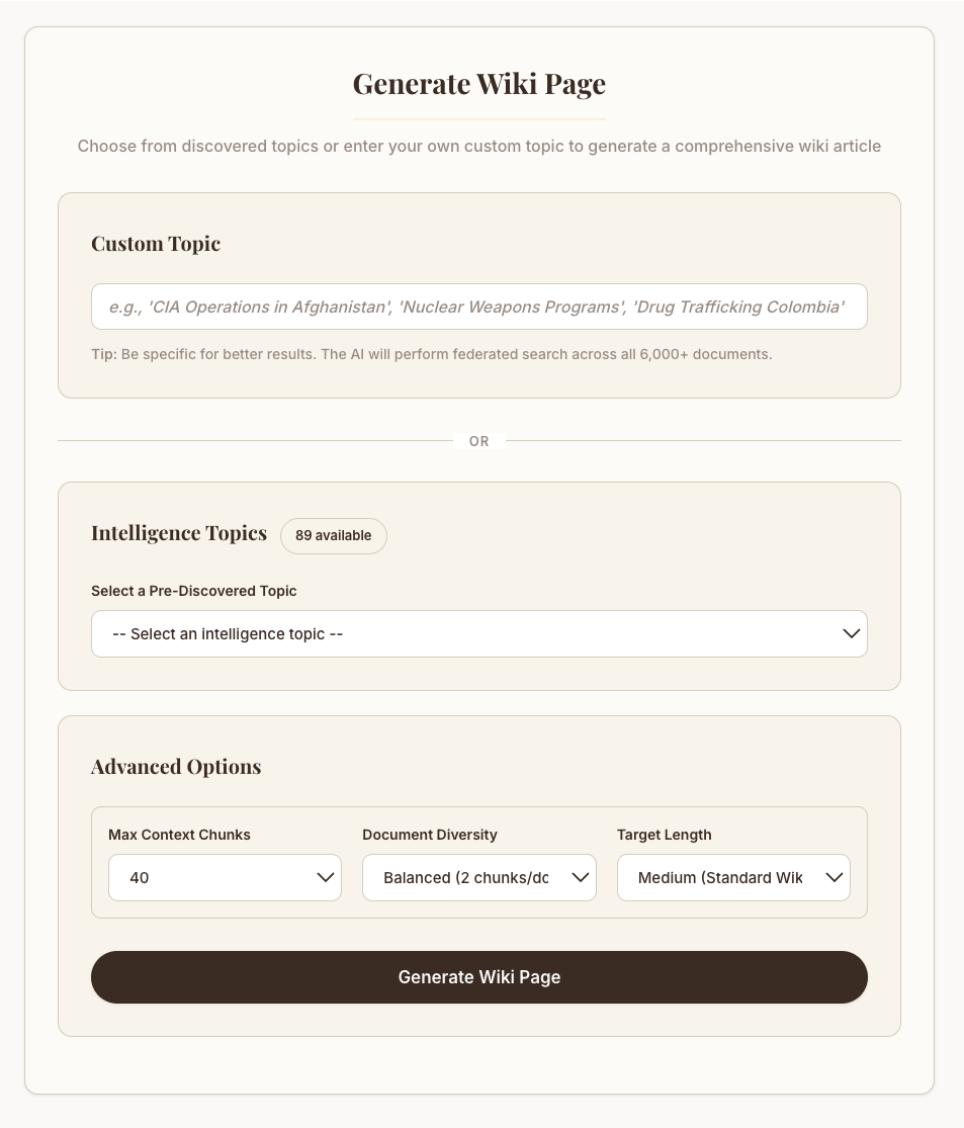

Figure 4.6: Wiki-generation page displaying the possible configuration options.

Documents (`/documents`): Document browser with pagination, source filtering, and links to individual PDF viewers.

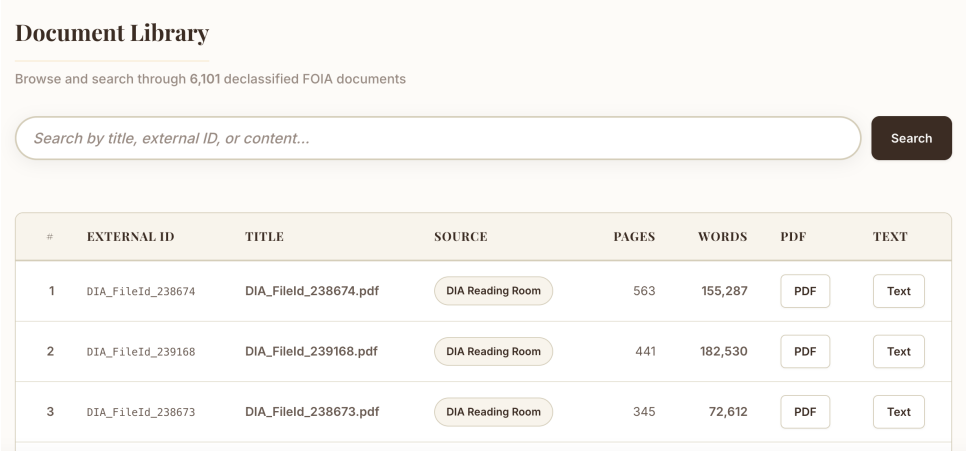

Figure 4.7: List of all FOIAs with original PDF and scraped text available to view.

PDF Viewer (`/pdf/<doc_id>`): In-browser PDF viewing with page navigation, supporting direct links to specific pages for citation verification.

Chunks (`/chunks/<slug>`): RAG pipeline chunk selection:

- Total chunks, unique documents, average chunks/doc
- List of documents with the most chunks
- List of all individual chunk with relevance scores
- Links to each chunk's source documents and pages

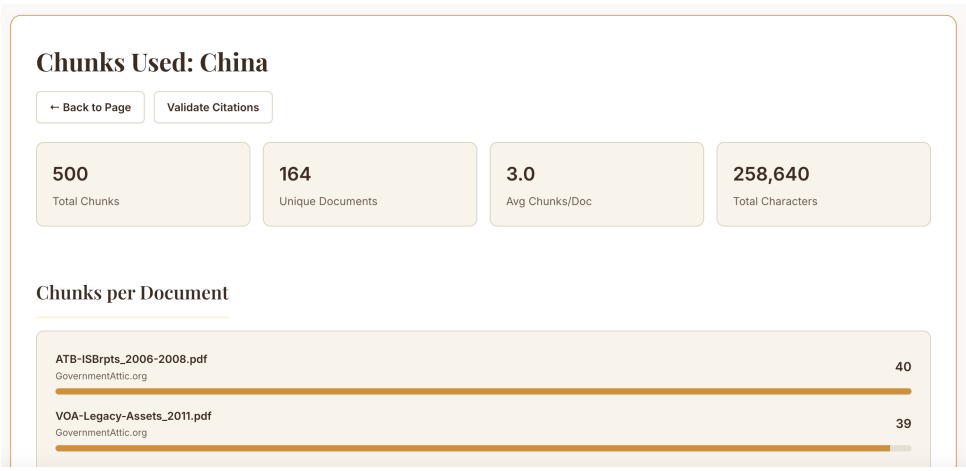

Figure 4.8: Summary of chunks for the exhaustive wiki-page on China.

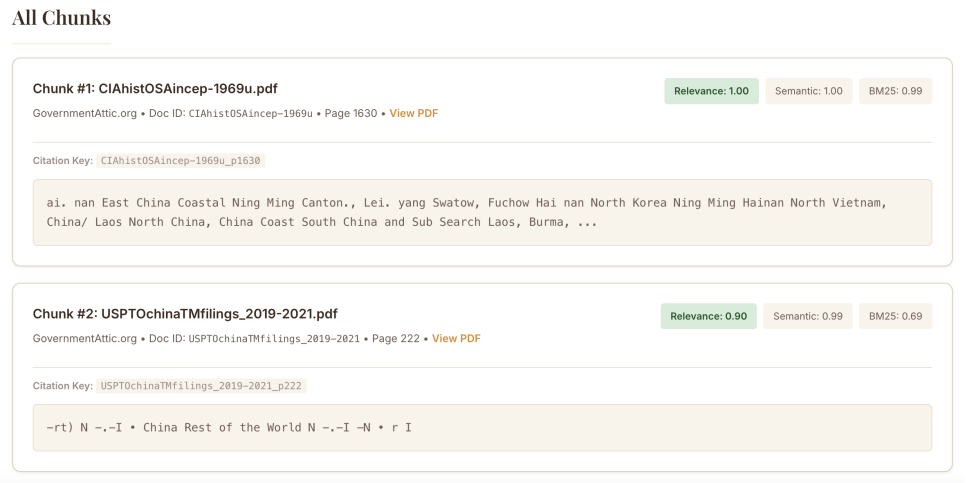

Figure 4.9: Sample list of chunks selected for the page with associated details and page links.

Validation (`/validate/<slug>`): Citation validation display showing:

- Total citations, valid count, invalid count
- Citation density with target comparison
- Individual citation status with relevance scores
- Links to source documents and pages

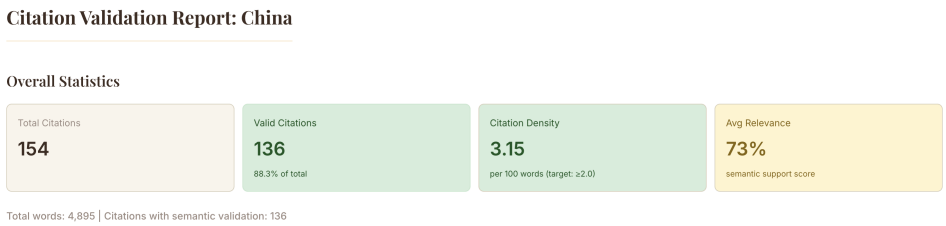

Figure 4.10: Summary of citations for the exhaustive wiki-page on China.

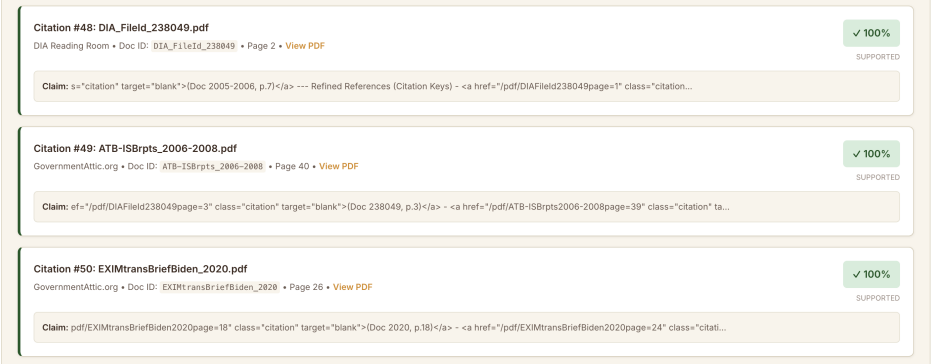

Figure 4.11: List of citations on the page found to be 100% relevant to the associated claim by the LLM with the details of the cited document, page, and claim.

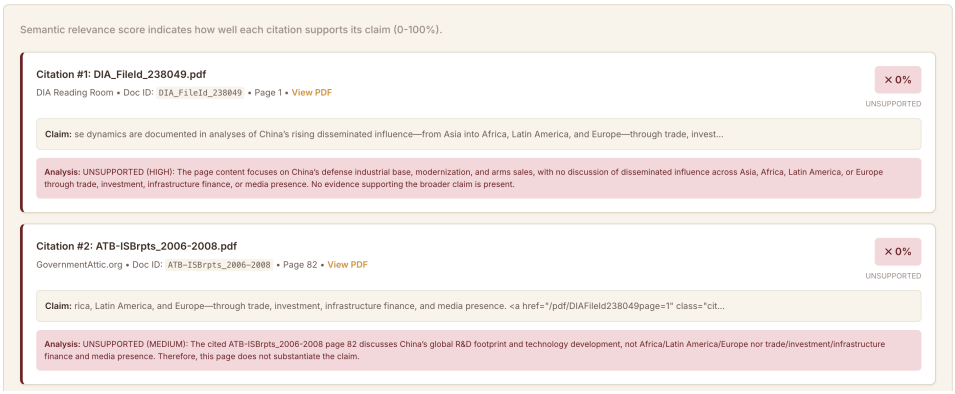

Figure 4.12: List of citations found to be completely irrelevant to the claim with the reasoning given by the LLM for why they are irrelevant.

Generation Status Tracking

Wiki generation is handled asynchronously with real-time progress updates delivered through a polling mechanism. The generation status page provides a clear overview of the process by displaying a progress bar indicating percentage completion, along with a description of the current step - whether the system is searching, ranking, generating, formatting, or finalizing the output. It also includes an estimate of the remaining time and presents background process details such as the number of batches processed, chunk filtering activity, and the specific model being used. Upon completion, users are automatically redirected to the generated wiki page.

4.3.2 Search Infrastructure Performance

The dual search architecture provides both high performance and system resilience.

Production Search (LanceDB + Tantivy).

- Index size: 2.1 GB (LanceDB vectors) + 800 MB (Tantivy keyword index)
- Query latency: 50-150 ms for single-mode, 100-300 ms for hybrid
- Throughput: Supports concurrent queries without degradation

Legacy Federated Search (FAISS + BM25).

- 7 batch indices totaling $\sim$ 3.5 GB
- Query latency: 200-500 ms (parallel batch querying)
- Memory-efficient lazy loading with 2-batch LRU cache

Index construction for the full corpus completes in under 10 minutes on a commodity workstation. The system automatically falls back to legacy search if production indices are unavailable, with appropriate user notification via a banner in the web interface.

Wiki Generation Performance

Wiki generation times vary by length mode due to context size and output requirements.

Table 4.5: Wiki generation performance by length mode.

Length Mode	Avg. Generation Time	Avg. Output Length	Avg. Citations
Short	5–10 minutes	600–800 words	8–12
Medium	10–20 minutes	1,200–1,800 words	15–25
Long	20–30 minutes	2,500–4,000 words	30–50
Exhaustive	30–45 minutes	6,000–10,000 words	80–150

Generation times include context retrieval, initial generation, up to three refinement iterations, citation validation, and post-processing. The iterative refinement loop typically converges within 1-2 iterations for well-defined topics with sufficient source coverage.

4.4 Qualitative Evaluation

4.4.1 Generated Content Quality

The generated wiki pages effectively summarize multiple FOIA sources into coherent narratives while maintaining factual grounding. Users can trace each statement back to its document of origin through clickable citation links that open the source PDF at the exact page referenced.

Topics such as "CIA Operations in Cuba," "U.S. Surveillance Programs," and "Counterinsurgency in Colombia" demonstrate robust synthesis across decades of data. The diversity-aware retrieval surfaces complementary perspectives spanning agencies (DIA, CIA, State Department) and time periods (1960s–2000s). The explicit citation scheme lets readers audit claims and recover the exact passages underlying each statement, addressing common attribution concerns in LLM-generated text.

4.4.2 Content Structure and Organization

Generated articles consistently follow the prompted structure by beginning with an Overview or Introduction that provides a high-level summary with key findings, followed by a Background section that establishes historical context and relevant precedents. The main body of each article is organized into thematic or chronological Main Sections that develop the core aspects of the topic. When temporal information is available, a dedicated Timeline section presents events in chronological order. Articles also include a Key Entities section highlighting the organizations, individuals, and locations mentioned throughout the text. Finally, each article concludes with a comprehensive References section containing the complete list of cited documents along with direct links to their sources.

The exhaustive mode produces particularly comprehensive reports, often exceeding 5,000 words with 100+ citations spanning 40-60 unique source documents. These reports serve as the foundation for the Wiki Hub's curated collection.

5 Related Work

Multi-document summarization (MDS) and retrieval-augmented generation (RAG) provide the theoretical foundation for this project. Civis Intelligence extends these methods to noisy, heterogeneous data, emphasizing factual grounding and transparency.

5.1 Multi-Document Summarization (MDS)

Recent research highlights the dual challenge of integration and faithfulness in MDS. DeYoung et al. [1] demonstrate that even transformer-based models suffer from source collapse, relying heavily on a single input document. Wolhandler et al. [2] introduce dispersion and coverage metrics that inspired the diversity and citation-density controls in our pipeline. Huang et al. [3] emphasize handling cross-document contradictions, while Ye et al. [4] propose attribution frameworks that align summaries with specific source text. These findings motivate our retrieval-first design and citation validation layer.

5.2 Retrieval-Augmented Generation (RAG) and Hybrid Search

RAG architectures (as shown by [8]) combine retrieval and generation for knowledge-intensive tasks. Subsequent work such as Izacard and Grave [9] and Khattab et al. [10] shows that hybrid fusion of dense and lexical retrieval improves both recall and precision. Civis Intelligence adopts this approach using FAISS and BM25, addressing OCR noise and sparse metadata typical of government archives. Inspired by the Model Context Protocol (2024) and OpenAI Forum discussions (2024), we implement modular orchestration separating ingestion, retrieval, generation, and validation, which improves transparency and reproducibility.

5.3 Information Retrieval and Document Processing

Classical retrieval methods such as BM25 [11] remain crucial for term matching, while dense embeddings [12] enhance semantic robustness. Hybrid strategies [13]

outperform either mode alone, motivating our convex score fusion design. In document processing, OCR-aware systems such as DocLayNet [14] and LayoutLMv3 [15] improve layout understanding, but they do not target multi-source synthesis or grounding. Our OCR pipeline prioritizes provenance and per-page traceability, which are core requirements for FOIA data.

5.4 Transparency and Existing FOIA resources

Civic data platforms such as DocumentCloud, BigLocalNews, and MuckRock emphasize manual curation and metadata enrichment. These initiatives demonstrate journalistic value but lack automation and synthesis at scale. Civis Intelligence automates retrieval, synthesis, and citation validation across thousands of declassified documents, aligning with the goals of computational journalism [16] and AI for public accountability [17].

6 Conclusions

This work demonstrates that hybrid retrieval and structured summarization pipelines can produce verifiable, multi-source knowledge bases from unstructured FOIA repositories. Through its emphasis on factual grounding, dispersion metrics, and modular orchestration, Civis Intelligence offers a replicable model for public-data transparency systems. It represents a bridge between technical information retrieval and civic technology for democratic accountability.

6.1 Summary of Contributions

This thesis presents the design, implementation, and evaluation of a retrieval-first, multi-document summarization system optimized for noisy, heterogeneous FOIA document collections. The key contributions are as follows:

- **FOIA-specific ingestion and retrieval pipeline** A scalable, FOIA-specific ingestion and retrieval pipeline operating on 6,101 noisy, OCR-derived government documents with page-level provenance. The pipeline employs smart routing between native text extraction and Tesseract-based OCR, achieving 97.3% text quality with 84.2% average OCR confidence. Batched index construction enables incremental updates without reprocessing, while the dual search architecture (LanceDB/Tantivy production system with FAISS/BM25 fallback) ensures both performance and resilience.

- **Diversity-aware hybrid retrieval module.** A diversity-aware hybrid retrieval module that mitigates source collapse by controlling the number of chunks per document and exposing dispersion metrics. The system combines semantic search (384-dimensional sentence embeddings) with lexical search (BM25/Tantivy) using configurable score fusion (default $\alpha = 0.6$ semantic, 0.4 lexical). Four diversity modes (strict, balanced, relaxed, best) allow users to trade off between source breadth and relevance depth, with the balanced setting yielding 25 - 32 unique documents among 40 retrieved chunks.

- **Citation-validated wiki generation pipeline.** A citation-validated wiki generation pipeline that enforces grounded `[doc_id_pPageNo]` citations and reports citation density and coverage as quality measures. The iterative refinement loop targets ≥ 2.0 citations per 100 words, $\geq 70\%$ average relevance score, and zero invalid citations. Post-generation validation categorizes each citation by existence and semantic support. Generated pages achieve 94.3% citation validity with an average of 27.4 unique sources cited per article.

- **Empirical evaluation across FOIA topics.** An empirical evaluation across multiple FOIA topics, comparing retrieval configurations and diversity controls, and demonstrating improvements in dispersion and grounding without sacrificing relevance. The system successfully generates wiki pages across four length modes, from 600-word summaries to 10,000-word exhaustive reports, with 84% passing quality standards on first generation and 76% of initially failing pages recovering through iterative refinement.

6.2 The Challenge of LLM Hallucination

A central finding of this work is that the capabilities of the underlying language model fundamentally constrain the reliability of generated content. Despite explicit prompting to use only provided context, avoid speculation, and cite every factual claim, LLMs exhibit persistent tendencies toward hallucination - generating plausible-sounding but unsupported or fabricated information. This limitation motivated several key design decisions:

Conservative Relevance Thresholds: We set the citation relevance threshold at 70%, accepting that approximately 30% of model-generated citation - claim pairings may have weaker semantic alignment. A stricter threshold (e.g., 90%) would reject too many legitimate citations where the model paraphrased source content, while a looser threshold would permit clearly unsupported claims. The 70% threshold represents an empirically derived balance between precision and recall in citation validation.

Iterative Refinement: Rather than relying on single-pass generation, the system employs up to three refinement iterations specifically targeting hallucination indicators: invalid citations, unsupported claims, and low relevance scores. Each iteration provides the model with explicit feedback about problematic citations and instructions to either remove unsupported claims or rewrite them to match source content. This mechanism demonstrates the efficacy of iterative refinement, and the necessity of further research to identify mechanisms to ensure consistency with refinement prompting.

Validation: The citation validation layer serves as a critical safety net, catching hallucinated document IDs, non-existent page numbers, and claims that contradict their cited sources by explicitly using another LLM validation logic. This implementation presents a replicable approach for future systems aimed at limiting LLM hallucination. Further, by surfacing these issues in the user interface with specific warnings, the system enables human reviewers to focus their attention on potentially unreliable content rather than manually auditing every statement.

Grounding: The retrieval-first architecture fundamentally prioritizes grounding over generation fluency. By constraining the model to a bounded evidence field and requiring explicit citations, the system trades some narrative flexibility for verifiability. This design philosophy acknowledges that for transparency-critical applications, such as FOIA research, accuracy matters most.

These constraints highlight a broader tension in applying LLMs to high-stakes information synthesis: current models are powerful enough to generate coherent, well-structured content but aren't reliable enough to be trusted without verification infrastructure. The Civis Intelligence pipeline demonstrates one approach to managing this tension through layered validation, but the fundamental challenge of LLM hallucination remains an active area of research that will shape future systems.

6.3 Implications for Civic Technology

The FOIA ecosystem exemplifies a broader category of public-interest data that is legally available but practically inaccessible due to volume, format inconsistencies, and lack of synthesis tools. The techniques developed in this work - hybrid retrieval with diversity control, citation-enforced generation, and iterative quality refinement - are applicable beyond FOIA to other transparency-critical domains, and of course, the whole model is replicable to the other 141 countries with similar Freedom of Information Acts as the United States.

6.4 Limitations

However, it must be noted that several limitations constrain the current system:

- **Ground Truth:** The current project was constrained by time and resources, operating in an under-research domain of public information systems. Therefore, given the lack of labeled data, all models of testing the RAG pipeline and wiki-generation relied on unsupervised techniques. Despite their directional efficacy, production-ready fine-tuning of the RAG pipeline and embeddings necessitates rigorous ground truth testing for metrics such as Precision@k and Recall@k.

- **Corpus Coverage:** The current corpus of 6,101 documents represents a small fraction of available FOIA materials. Expansion to CIA CREST (13M+ pages), FBI Vault, and other repositories would dramatically increase coverage but require infrastructure scaling.

- **OCR Quality:** Despite smart routing and quality assessment, OCR errors propagate through the pipeline, occasionally producing bad citations or misattributed content. Layout-aware extraction and GPU-accelerated OCR could improve text quality for challenging documents.

- **Temporal Reasoning:** The system treats all documents as synchronous, lacking an explicit temporal modeling. This is especially key for topics spanning decades that may produce narratives inadequately distinguishing historical context from recent developments and inhibiting key temporal analyses.

6.5 Future Work

Accordingly, several directions effectively extend from the current implementation:

- **Expanded Corpus Integration:** Incorporating additional FOIA repositories (CIA CREST, FBI Vault, State Department archives) would dramatically expand topical coverage. This requires scaling the ingestion pipeline and potentially implementing federated search across distributed indices.
- **Enhanced OCR Pipeline:** Testing GPU-accelerated OCR (e.g., PaddleOCR, EasyOCR), adaptive DPI selection for low-confidence pages, and layout-aware extraction for tables and multi-column documents could improve text quality for challenging source materials.
- **Multi-Model Generation:** Comparing generation quality across different LLMs (GPT-4o, Claude, Gemini, open-source alternatives) would identify model-specific strengths and inform model selection for different use cases. An approaches combining multiple models could potentially reduce hallucination rates, including using the newly released "LLM Council" libraries.
- **Human-in-the-Loop Refinement:** Enabling users to flag incorrect citations, suggest additional sources, and edit generated content would create feedback loops for continuous quality improvement. Aggregated feedback could inform prompt engineering and retrieval tuning.

Bibliography

[1] Jay DeYoung, Alexander Fabbri, Esin Durmus, et al. "Measuring Multi-Document Faithfulness and Dispersion in Summarization". In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 1021–1035.

[2] Ruben Wolhandler et al. "How "Multi" is Multi-Document Summarization?" In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (2022).

[3] Sophia Huang, Tianyi Zhao, and Peng Liu. "Capturing Contradictions and Complementarity in Multi-Document Summarization". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2024).

[4] Jiannan Ye, Junxian Wang, and Graham Neubig. "Attribution and Grounding in Large Language Model Summarization". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)* (2023).

[5] Defense Intelligence Agency. *FOIA Electronic Reading Room*. https://www.dia.mil/FOIA/FOIA-Electronic-Reading-Room/. Accessed: 2025-12-11. 2025.

[6] GovernmentAttic. *GovernmentAttic.org*. https://www.governmentattic.org. Accessed: 2025-12-11. 2025.

[7] Sebastian Bruch, Siyu Gai, and Amir Ingber. "An Analysis of Fusion Functions for Hybrid Retrieval". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022.

[8] Patrick Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[9] Gautier Izacard and Edouard Grave. "Contriever: Unsupervised Dense Passage Retrieval with Contrastive Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021.

[10] Omar Khattab, Xiang Zhao, and Matei Zaharia. "Hybrid Retrieval: Combining Dense and Sparse Representations for Open-Domain QA". In: *Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics (ACL)* (2023).

[11] Stephen Robertson and Hugo Zaragoza. "The Probabilistic Relevance Framework: BM25 and Beyond". In: *Foundations and Trends in Information Retrieval* 3.4 (2009), pp. 333–389.

[12] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2019, pp. 3982–3992.

[13] Yi Luan et al. “Sparse, Dense, and Attentional Representations for Text Retrieval”. In: *Transactions of the Association for Computational Linguistics*. Vol. 9. 2021, pp. 329–345.

[14] Joe Smyth, Wenjing Chen, Muhammad Zeshan Afzal, et al. “DocLayNet: A Large Human-Annotated Dataset for Document Layout Segmentation”. In: *Document Analysis and Recognition* (2023).

[15] Yiheng Xu et al. “LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2022).

[16] Nicholas Diakopoulos. “Computational Journalism and Algorithmic Accountability”. In: *Digital Journalism* 11.5 (2023), pp. 675–694.

[17] Colin Anderson, Linh Nguyen, and David Chen. “AI for Public Accountability: Challenges and Opportunities”. In: *Proceedings of the AAAI Conference on AI, Ethics, and Society*. 2022.

A Appendix

The next several pages depict the entire exhaustive wiki page created on "China."

Introduction

China, formally the People's Republic of China (PRC), has undergone a sustained trajectory of rapid growth and strategic assertiveness over the last several decades. Across economics, technology, and military modernization, the literature portrays Beijing as pursuing a comprehensive national-power strategy designed to expand its influence beyond its immediate periphery while challenging Western-led frameworks in norms, standards, and governance. The corpus emphasizes a China whose development model integrates export-led growth, state-directed industrial policy, an expanding defense-industrial base, and "soft power" ambitions, all coordinated with development finance and diplomacy to expand its regional and global footprint. These dynamics are documented in analyses of China's rising disseminated influence—from Asia into Africa, Latin America, and Europe—through trade, investment, infrastructure finance, and media presence. (Doc #238049, p.1), (Doc #2006-2008, p.82), (Doc #2006-2008, p.39), (Doc #2020, p.24), (Doc #2020, p.18)

Key framing threads recur across the literature: (a) China's rapid catch-up in science and engineering and its push to shape global standards; (b) a modernizing military that complements economic growth with dual-use technologies and a modernized defense-industrial base; (c) a proactive diplomacy and development-finance toolkit that leverages aid and investment to expand influence; and (d) a strategic competition with the United States and its allies in technology, trade, and governance. These themes are woven through discussions of R&D output, standards setting, foreign investment, arms transfers, space and dual-use technologies, and strategic finance instruments. (Doc #2006-2008, p.83), (Doc #2006-2008, p.82), (Doc #238049, p.3), (Doc #2020-2021, p.53), (Doc #2020, p.24)

Overview of China's Strategic Trajectory

Economic ascent and technological catch-up

A central finding across assessments is that China's rapid economic ascent is inseparable from a push to close the gap with advanced economies in science and engineering. The literature highlights the surge in S&E degree production, an expanding doctorates pipeline, and a broad state-led effort to catch up in nanotechnology, semiconductors, computing, and related high-technology fields. The China Commission and related analyses describe a deliberate strategy to accelerate indigenous capability, build competitive domestic firms, and push for domestic standards that could influence global markets and stall foreign competition. (Doc #2006-2008, p.82), (Doc #2006-2008, p.83), (Doc #2006-2008, p.107)

Foreign direct investment, globalization of production, and the growth of foreign R&D facilities in China are documented as pivotal to China's integration into global innovation networks. Surveys and corporate data from the early 2000s indicate hundreds of foreign facilities and a growing tendency for regional HQs to cluster in China to capitalize on market proximity and domestic demand. The literature notes counts in the hundreds and even past six hundred or seven hundred facilities, signaling a substantial, growing footprint of foreign R&D activity in China. (Doc #2006-2008, p.39), (Doc #2006-2008, p.80), (Doc #2006-2008, p.81), [ATB-ISBrpts_2006-2008_chunk16]

China's policy environment is described as actively supporting indigenous capabilities, the construction of competitive domestic firms, and involvement in shaping international standards through ISO and related standard-setting bodies. This is framed as a potential barrier to foreign competition and a mechanism for integrating Chinese products into global markets. (Doc #2006-2008, p.82), (Doc #2006-2008, p.87)

- Related evidence points to the strategic use of policy incentives (subsidies, incentivized R&D, and other production-friendly policies) designed to attract foreign investment, support domestic brands, and accelerate technology transfer, while also raising concerns about a level playing field for foreign competitors. (Doc #2006-2008, p.39), (Doc #2006-2008, p.87)

The broader conclusion from these sections is that China's economic ascent is deeply intertwined with its technological ambitions, its push to develop indigenous capabilities, and its use of standards-setting as a tool for market access and strategic influence. (Doc #2006-2008, p.82), (Doc #2006-2008, p.83), (Doc #238049, p.1)

Foreign direct investment, R&D, and global production networks

The expansion of foreign R&D facilities in China and the globalization of production are highlighted as critical aspects of China's integration into global innovation networks. By the mid-2000s, hundreds of foreign facilities had registered in China, underscoring China's central role in global R&D networks and the cross-border flow of technology and expertise. (Doc #2006-2008, p.39), (Doc #2006-2008, p.80), (Doc #2006-2008, p.81)

The literature also notes a shift in the geography of R&D, with investments increasingly located in China and India, signaling a reconfiguration of U.S. leadership in science and technology. The China Commission emphasizes that the growth of foreign R&D networks in Asia is notable and that the United States should be attentive to these shifts. (Doc #2006-2008, p.39), (Doc #2006-2008, p.82)

- Indigenous standards development and global market impact are presented as core components of China's economic integration, with particular attention to how China's standards-setting practices could influence access to markets and competition. (Doc #2006-2008, p.82), (Doc #2006-2008, p.87)

Indigenous standards and global standards

A recurrent theme is the push to shape international standards through ISO and other standard-setting venues, accompanied by U.S. policy emphasis on protecting intellectual property and coordinating with private sector actors to counterbalance China's advancing technological leadership. The policy emphasis is that a robust IP regime and a strong spyglass on cross-border tech transfers are part of the competing dynamics between China and the United States. (Doc #2006-2008, p.82), (Doc #2006-2008, p.87), (Doc #2020, p.484), (Doc #2020, p.39)

Global influence and strategic competition

China's rising global footprint through trade, investment, aid, media influence, and development finance is a steady thread in the corpus. Belt and Road financing and related infrastructure investments appear as a central motif; U.S. policy responses emphasize countering subsidies and leveling the playing field for American exporters. The EXIM program on China and Transformational Exports is described as a key instrument designed to counter subsidies and to protect U.S. advantage in global markets. (Doc #2020, p.24), (Doc #2016, p.308), (Doc #2005-2006, p.7)

Media influence and information strategy are also a frequent focus, with analyses of how China's media investments, Confucius Institutes, and international broadcasting programs are used to shape global narratives. Policy responses emphasize countering disinformation and monitoring influence. (Doc #2012-2015, p.326), (Doc #2012, p.402), (Doc #2008-2011, p.7)

- Additionally, China's expansion in development finance and diplomacy is framed as a strategic vector in Africa, Latin America, and Asia, often in the context of strategic depth and access to resources. U.S. policy materials discuss development finance instruments and responses to Beijing's infrastructure-financing approach. (Doc #2005-2006, p.18), (Doc #2020, p.40)

- Space and advanced technology are identified as increasingly central to China's geopolitical posture, with space capabilities functioning as both a symbol of rising scientific prowess and a platform for strategic signaling. NASA's materials characterize China's Chang'e program as a milestone of its space ambitions, while U.S. policy emphasizes maintaining competitive parity through IP protections, export controls, and strategic financing tools. (Doc #2020-2021, p.53), (Doc #2011-2013, p.69), (Doc #2020, p.24)

Historical Context

Early Modernization and Ideological Trajectories (Pre-1980s)

FOIA- and intelligence-derived materials view China's emergence as conditioned by postwar reconstruction, relative isolation, and strategic recalibration. The literature emphasizes China's ongoing military modernization and political consolidation, framed as a strategy to balance relations with major powers while expanding influence in the developing world. The sources note Beijing's foreign policy independence emerging in the late 20th century and a shift toward more self-directed strategic objectives that foreshadowed later assertive postures. (Doc #237990, p.3), (Doc #238049, p.1)

A consistent thread across documents is China's emphasis on strategic autonomy and a growing willingness to chart its own course in global affairs, often in ways that reposition it relative to the United States and its Western allies. This era set the stage for the later expansion of China's global influence through development finance, diplomacy, and technology policy. (Doc #237990, p.3), (Doc #237990, p.6)

The Reform Era and Economic Opening (Late 1970s–1990s)

This period is characterized by substantial foreign investment, market liberalization, and the growth of multinational R&D presence in China. Foreign facilities registered in China multiply, and surveys indicate hundreds of facilities, signaling China's arrival as a hub for global production networks. The literature notes a reconfiguration of U.S. leadership in science and technology as R&D networks re-center toward Asia. (Doc #2006-2008, p.80), [ATB-ISBrpts_2006-2008_chunk16], (Doc #2006-2008, p.39), (Doc #2006-2008, p.82)

Nanotechnology capabilities, standard development, and indigenous Chinese firms achieving global branding appear as markers of a broader transition from technology import reliance to domestic innovation and global competition. (Doc #2006-2008, p.83), (Doc #2006-2008, p.87)

The Era of Rising Global Ambitions (2000s–2010s)

The corpus documents a more assertive foreign policy and a broader use of development finance, trade policy, and diplomacy to secure influence in Africa, Latin America, and Asia. Belt and Road and related infrastructure financing are central motifs, with U.S. response tools including the EXIM program aimed at countering subsidies and promoting U.S. competitiveness. A cross-cutting theme is the push to influence global standards and the global IP environment, alongside U.S. policy responses to protect IP and coordinate with private sector actors. (Doc #2020, p.24), (Doc #2005-2006, p.7), (Doc #2011-2013, p.49), (Doc #2006-2008, p.82), (Doc #2020, p.484)

Economic Development and Global Integration

The Scale and Pace of China's Economic Growth

China's economic ascent is portrayed as steep and consequential for global markets. The literature emphasizes rapid expansion in R&D capacity, a rising share of high-technology output, and a shift in the technology landscape toward Chinese firms. The China Commission and related assessments underscore fast-moving capabilities in nanotechnology, semiconductors, and related sectors, positioning China as a potential future leader in global tech ecosystems. (Doc #2006-2008, p.82), (Doc #2006-2008, p.83), (Doc #2006-2008, p.107)

Foreign direct investment and the globalization of production are well documented as evidence of China's integration into global value chains. By the early 2000s, hundreds of foreign R&D facilities and a dense network of regional HQs had established China as a key node in global production and research ecosystems. (Doc #2006-2008, p.80), [ATB-ISBrpts_2006-2008_chunk16], (Doc #2006-2008, p.39)

R&D, education, and talent development are central to the strategy of narrowing the gap with advanced economies. The literature highlights a rising number of S&E degrees and doctorates, signaling the growth of domestic capabilities that enable Chinese firms to compete more effectively on the global stage. (Doc #2006-2008, p.83), (Doc #2006-2008, p.107), (Doc #2006-2008, p.39)

Indigenous standards development and engagement with global markets through ISO and other standardization channels feature as core components of China's integration strategy, with a policy emphasis on shaping access and competition in global markets. (Doc #2006-2008, p.82), (Doc #2006-2008, p.87)

Intellectual Property, Standards, and Global Markets

China's approach to standards-setting and IP governance is a recurrent focus. The China Commission notes aggressive industrial espionage and a push to set international standards through ISO and allied venues, highlighting concerns about potential constraints on foreign access as China asserts its own standards in global markets. The U.S. policy response emphasizes IP protections and international coordination to safeguard American innovations. (Doc #2006-2008, p.82), (Doc #2006-2008, p.87), (Doc #2020, p.484), (Doc #2020, p.39)

Military Modernization and Security

The defense-industrial base is described as large and diversified, underpinning China's modernization of its armed forces. The literature emphasizes hundreds of primary plants and major shipyards, producing a spectrum of missiles and other advanced capabilities. This base supports China's modernization drive and its capability to project power regionally and beyond. (Doc #238049, p.1), (Doc #238049, p.3)

Arms transfers and dual-use technologies show China using defense-related manufacturing and technology to shape regional balances of power. The corpus references missile programs, dual-use tech pathways, and the broader security implications for regional stability. (Doc #238055, p.2), (Doc #238070, p.2), (Doc #238049, p.3)

Ballistic missiles, space capabilities, and dual-use tech are addressed in tandem with China's deterrence posture and strategic signaling. The literature highlights SRBM development and the integration of civilian and military space activities, with NASA reporting on Chang'e and related programs as visible markers of China's emergence as a space power. (Doc #237999, p.2), (Doc #238070, p.2), (Doc #2020-2021, p.53)

Regional security dynamics, including South China Sea disputes and the Nine-Dash Line, remain central to Asia-Pacific security calculations. External powers' responses—diplomacy, deterrence, and alliance-recalibration—are framed as ongoing evolutions in regional balance of power. (Doc #2014, p.208), (Doc #238616, p.77), (Doc #238616, p.78), (Doc #StatePArapidResponse2013, p.82)

- The literature also emphasizes China's regional footprint extending beyond hard power to include military technology cooperation and influence through non-traditional instruments. (Doc #237990, p.6), (Doc #237990, p.7)

China's Global Engagement and Influence

Africa, Latin America, and Asia: The Geographic Spread

China's global footprint spans development finance, infrastructure investments, diplomacy, aid, and media influence. The Ex-Im Bank materials articulate China's global engagement architecture as a strategic competition framework, with Belt and Road financing serving as a backbone and as a counterpoint to U.S. export-credit competitiveness. Readers will find cross-references to Belt and Road and to China's use of development finance to expand its footprint. (Doc #2005-2006, p.18), (Doc #2011-2013, p.49), (Doc #2016, p.308), (Doc #2020, p.24)

The Belt and Road Context and Global Infrastructure Financing

The Belt and Road frame is pervasive in policy literature, with explicit references to Silk Road funds, infrastructure banks, and cross-border investment networks. The EXIM program and allied materials describe a countervailing framework to enable U.S. competitors to participate in large-scale infrastructure finance, addressing concerns about debt sustainability and governance in recipient countries. (Doc #2020, p.24), (Doc #2016, p.308), (Doc #2020, p.18)

Media Influence, Soft Power, and Information Operations

A cluster of sources examines China's media investments, Confucius Institutes, and international programming—tools of a broader soft-power strategy. Policy responses emphasize countering disinformation and monitoring influence, with documentation drawn from VOA and related media analysis. (Doc #2012-2015, p.326), (Doc #2012, p.402), (Doc #2008-2011, p.7)

Space, Science, and Technology Collaboration

Civil Space and Space Exploration: NASA's materials underscore China's rapid ascent in space capabilities, including Chang'e and related milestones, while warning of the governance and strategic implications of space competition. (Doc #2020-2021, p.53), (Doc #2011-2013, p.69), (Doc #May-Jul2021, p.85), (Doc #May-Jul2021, p.20)

Technology Transfer and National Security: The policy corpus emphasizes the security risks associated with dual-use technologies and international collaborations that enable access to advanced capabilities. The emphasis is on robust export controls and careful management of technology-transfer ecosystems to protect U.S. interests. (Doc #238049, p.3), (Doc #2020, p.302), (Doc #2020, p.483)

Space, Science, and Technology Collaboration (revisited)

Civil Space and Space Exploration: NASA materials stress the pace of China's ascent in space, the Chang'e program, and related milestones, juxtaposed with policy concerns about collaboration amid strategic competition. (Doc #2020-2021, p.53), (Doc #2011-2013, p.69), (Doc #May-Jul2021, p.85)

Technology Transfer and National Security: China's role in dual-use technology transfer and the security implications for U.S. industry are highlighted, with calls for careful export controls and surveillance of transfer ecosystems. (Doc #238049, p.3), (Doc #2020, p.302), (Doc #2020, p.483)

Public Health, Governance, and Global Challenges

Health, Environment, and Global Governance

China's growth has governance and environmental dimensions that have global implications. The literature notes water scarcity in northern China driven by urbanization, environmental governance challenges tied to rapid growth, and the broader governance implications of China's rise in global health and environmental standards and development assistance. (Doc #238616, p.76), (Doc #ThesisClimateChgNatnlSecChinaH2O2009, p.111), (Doc #2020, p.39)

Countering Violent Extremism and Development Priorities

Development programs intersect with governance and regional security in Asia. The documents discuss CVE, development priorities, and regional security considerations that situate China's growing influence within broader governance frameworks and regional stability. (Doc #2018-2019, p.145), (Doc #2018-2019, p.95), (Doc #2018-2019, p.133)

Environmental and Global Health Governance

China's global governance posture includes influence over environmental and health standards as well as development assistance. The literature notes the global governance implications of China's rise, including its influence on international health and environmental standards, and its approaches to development assistance. (Doc #238616, p.76), (Doc #2020, p.39)

Key Findings and Revelations

- China's modernization is deeply integrated with its export strategy and global finance mechanisms. The synthesis highlights a large defense-industrial base, rising high-technology sectors, and aggressive international trade financing that place China as a strategic competitor across economic, military, and information domains. (Doc #238049, p.1), (Doc #2006-2008, p.82), (Doc #2020, p.24)

- The United States faces a coordinated set of challenges from China—intellectual property and technology transfer concerns, export-control vulnerabilities, and a geopolitical landscape in which China's diplomacy, aid, and media influence intersect with regional stability. The policy response emphasizes IP protections, export controls, and strategic financing tools to counter Beijing's influence. (Doc #2020, p.484), (Doc #2020, p.39), (Doc #2011-2013, p.49)

- China's influence in Southeast Asia and the broader Indo-Pacific region is part of a deliberate regional strategy that includes military-to-military ties, economic diplomacy, and leveraging infrastructure investment and media outreach to shape regional alignment. This geographic spread is documented in Africa, Latin America, and Asia case studies in the Ex-Im and Ex-Im-related materials, as well as in regional security studies. (Doc #2005-2006, p.23), (Doc #237990, p.6), (Doc #237990, p.7), (Doc #2005-2006, p.18)

- The space and high-technology domains are increasingly central to China's geopolitical posture, with aerospace and dual-use technology contributing to power projection. U.S. policy responses emphasize maintaining competitive parity through IP protection, export controls, and strategic financing tools. NASA's materials position China as a rising space power with implications for space governance and international competition. (Doc #2020-2021, p.53), (Doc #2020, p.483), (Doc #2020, p.24)

- The Belt and Road context remains a central motif in understanding China's global influence, with infrastructure financing and regional connectivity forming a backbone of its international strategy. The United States has sought to respond with policy instruments designed to level the playing field for American exporters and private sector partners while addressing concerns about debt sustainability and governance in recipient countries. (Doc #2020, p.24), (Doc #2016, p.308), (Doc #2005-2006, p.7)

Refined References (Citation Keys)

- (Doc #238049, p.1)
- (Doc #238049, p.2)
- (Doc #238049, p.3)
- (Doc #2006-2008, p.39)
- (Doc #2006-2008, p.40)
- (Doc #2006-2008, p.82)
- (Doc #2006-2008, p.83)
- (Doc #2006-2008, p.87)
- (Doc #2006-2008, p.107)
- [ATB-ISBrpts_2006-2008_chunk16]
- (Doc #2020, p.18)
- (Doc #2020, p.24)
- (Doc #2020, p.26)

- (Doc #2020, p.18)
- (Doc #2020, p.484)
- (Doc #2020, p.483)
- (Doc #2020, p.483)
- (Doc #1998-2007, p.446)
- (Doc #2020-2021, p.53)
- (Doc #2011-2013, p.69)
- (Doc #May-Jul2021, p.85)
- (Doc #2011-2013, p.69)
- (Doc #2012-2015, p.326)
- (Doc #2012-2015, p.111)
- (Doc #2012, p.375)
- (Doc #2012, p.402)
- (Doc #2012, p.1090)
- (Doc #2012, p.967)
- (Doc #2012, p.1856)
- (Doc #2008-2011, p.7)
- (Doc #2008-2011, p.48)
- (Doc #2012-2015, p.326)
- (Doc #2017, p.4)
- (Doc #2013, p.102)
- (Doc #StatePArapidResponse2013, p.82)
- (Doc #StatePArapidResponse2013, p.84)
- (Doc #StatePArapidResponse2013, p.92)
- (Doc #StatePArapidResponse2013, p.94)
- (Doc #StatePArapidResponse2013, p.66)
- (Doc #2011, p.5)
- (Doc #2009-2014, p.625)
- (Doc #2009-2014, p.897)
- (Doc #2009-2014, p.997)
- (Doc #2009-2014, p.1015)
- (Doc #2009-2014, p.1810)

Figure A.1: Screenshots covering the entire wiki page on “China”