Data Science Capstone-DATS 6501

CAPSTONE PROJECT FINAL REPORT

DE-RISKING LAND ACCESS FOR INFRASTRUCTURE

GROUP 36 TEAM MEMBERS:

Samiksha Ramnath Burkul (G46017190)
Shreya Sahay (G36642286)
Tushar Sharma (G24951989)

Under the esteemed guidance of:

Dr. Prof. Abdi Awl
Mercedes Stickler
Beatrice Mora

DATE: April 30, 2025

GEORGE WASHINGTON UNIVERSITY
&
THE WORLD BANK

Table of contents

Glossary of Terms

Abbreviation	Full Form
AI	Artificial Intelligence
ASCII	American Standard Code for Information Interchange
API	Application Programming Interface
CAD	Cadastral
DB	Database
GIS	Geographic Information System
HTML	HyperText Markup Language
LLM	Large Language Model
NLP	Natural Language Processing
NER	Named Entity Recognition
PCA	Principal Component Analysis
PDF	Portable Document Format
P-code	Project Code (Unique World Bank identifier)
QA	Question Answering
RAG	Retrieval-Augmented Generation
RP	Restructuring Paper
URL	Uniform Resource Locator
UX	User Experience
XML	Extensible Markup Language
.txt	Plain Text File Format
KB	Knowledge Base
LDA	Latent Dirichlet Allocation
GPT	Generative Pretrained Transformer
RAKE	Rapid Automatic Keyword Extraction

I. Introduction

A. Background

Weak land administration systems, incomplete or outdated land records, and complex legal frameworks represent some of the most persistent and critical barriers to effective land acquisition in infrastructure development. These issues frequently lead to protracted negotiations, disputed claims, and legal entanglements, which in turn contribute to significant delays, cost overruns, mid-project design changes, or even complete project abandonment. Such challenges are especially pronounced in World Bank-financed projects, which often operate in contexts where institutional capacity and land governance structures are underdeveloped. While restructuring documents offer detailed insights into these setbacks, extracting that information remains a difficult task. The unstructured, narrative-heavy format of these documents makes manual review not only time-consuming and resource-intensive but also prone to inconsistencies and missed signals. As a result, there is a pressing need for scalable, automated approaches to systematically identify and analyze land-related risks embedded in historical project documentation.

B. Problem statement

Efficient land access is crucial for reducing delays and cost overruns in infrastructure projects worldwide. This project will leverage data science techniques to analyze land access challenges in 167 World Bank-financed projects, aiming to quantify impacts and predict risks associated with land tenure insecurity. We will be performing a combination of text and linguistic analytics (natural language processing (NLP) [11] techniques), Large Language Model (LLM)-based quantification of land access impacts, and predictive modeling. This way, we will develop insights to help stakeholders optimize planning and mitigate risks, contributing to more sustainable infrastructure development.

C. Problem elaboration

Land access remains one of the most pressing and complex challenges in global infrastructure development. Despite substantial investments, many projects financed by institutions like the World Bank experience significant delays, cost escalations, or even cancellations due to unresolved land-related issues. These problems stem from weak land governance systems, including outdated or incomplete cadastral records, legal ambiguities, and fragmented land tenure arrangements. Traditional approaches to identifying and assessing land access risks rely heavily on manual review of project documentation, which is inefficient, inconsistent, and difficult to scale across large project portfolios.

In the context of 167 World Bank-financed projects, the need to systematically detect, quantify, and understand the nature of land access challenges is both urgent and unmet. The lack of structured data and standardized reporting formats in restructuring papers further complicates this process, as critical insights are often embedded in unstructured, free-form narratives. As a result, stakeholders face difficulty in gaining early warnings or comparative risk assessments across regions or project types.

D. Motivation

Land-related delays are a persistent and costly bottleneck in infrastructure development, particularly in low- and middle-income countries where land acquisition, legal ambiguities, and community disputes can significantly derail project timelines. While restructuring papers from the World Bank and similar institutions contain valuable historical data on these setbacks, manually reviewing and extracting actionable insights from these documents is laborious, inefficient, and prone to human oversight. The documents are often lengthy, unstructured, and inconsistently formatted, making it difficult to surface recurring patterns, root causes, or geographic hotspots of land-related challenges at scale.

This project addresses that gap by introducing an automated, intelligent system that leverages LLMs and Retrieval-Augmented Generation (RAG) to extract, synthesize, and analyze land access difficulties from unstructured reports. By grouping related documents and enabling precise, contextual querying, the system provides project teams with immediate access to historical challenges and mitigation strategies. The goal is not only to reduce manual effort but also to proactively inform risk assessments, support better planning, and contribute to more resilient infrastructure design. Ultimately, this solution aims to institutionalize learnings from past experiences, enabling the World Bank and its partners to anticipate and mitigate delays more effectively in future projects.

E. Project scope

This project focuses on the end-to-end analysis of restructuring papers from 167 World Bank-financed infrastructure projects to identify, extract, and quantify land access challenges. These documents, published as part of project supervision—offer detailed accounts of delays, scope changes, and implementation hurdles, including those related to land acquisition, compensation disputes, and administrative bottlenecks.

To streamline the data collection process, the project begins with the automated retrieval of restructuring papers directly from the World Bank's online project database using Selenium-based web scraping [10]. This ensures consistent, scalable, and reproducible access to the latest publicly available documents across a wide range of countries and sectors.

To address this, we propose an AI-driven solution that integrates LLM within a RAG framework to automatically extract and summarize land access-related insights from historical project documents. This system enables the identification of recurring patterns, quantifiable risk factors, and common challenges, providing stakeholders with a data-informed foundation for risk-aware planning and mitigation in future infrastructure investments. The goal is to automate the detection of recurring land-related hurdles and support evidence-based decision-making in future project planning. The overall framework not only reduces manual review time but also enhances institutional memory, allowing teams to surface previously underutilized knowledge to inform risk mitigation and improve the long-term success of infrastructure investments.

II. Literature Review

A. Relevant Research

Natural Language Processing (NLP) techniques have shown significant promise in extracting structured insights from unstructured text, particularly in policy-heavy domains such as humanitarian response and infrastructure development.

Liberatore et al. [1] annotated 4,352 quantitative mentions in humanitarian reports using a six-label schema and developed a spaCy-based information extraction pipeline. Their system achieved high inter-annotator agreement (κ ≈ 0.90/0.81) and demonstrated strong performance with strict and fuzzy F1-scores of 0.76 and ~0.60, respectively, outperforming traditional Named Entity Recognition (NER) approaches. The authors have made both the annotated dataset and the code for their extraction pipeline publicly available, fostering further research and development in the field of humanitarian data analysis. This contribution is particularly significant for organizations involved in emergency response, as it enables the rapid extraction of critical quantitative information from vast amounts of textual data, thereby enhancing situational awareness and decision-making processes.
Tamagnone et al. [2] introduced HumBERT, a multilingual, domain-specific language model fine-tuned on over 2 million humanitarian documents from sources like ReliefWeb and UNHCR. Trained on the DEEP and HumSet corpora, HumBERT achieved superior

performance in multi-label classification tasks compared to general-purpose models, both in zero-shot and supervised settings. The authors also implemented counterfactual data augmentation to reduce gender and geographic biases, generating alternate sentence versions to enhance model fairness without sacrificing accuracy—marking a significant advancement in inclusive NLP for humanitarian applications.

Cai [3] conducted a systematic review of NLP applications in urban research, identifying its relevance in extracting insights from land-use documents and highlighting challenges in policy document analysis.

Lewis et al. [4] proposed the RAG framework, which combines a dense retriever and a generative model to enable knowledge-grounded responses. This architecture serves as the technical foundation for our project, allowing for semantic querying of unstructured World Bank documents. To extract keywords from large corpora, unsupervised methods such as RAKE [5] and transformer-based techniques like KeyBERT and KeyLLM [6] have been widely adopted. These tools improve semantic relevance in extracted terms, which is critical for interpreting nuanced policy language. In addition, foundational socio-political literature provides important context for understanding land-related issues. These works guide the interpretation of text-based indicators of tenure insecurity and land governance failures within our dataset. They demonstrate how structured annotation and domain-specific NLP pipelines can uncover latent patterns in unstructured narratives, informing both risk analysis and policy design. By adapting similar methodologies, our project strengthens the extraction of systemic land-related challenges across diverse geographies and institutional contexts.

III. Methodology

A. Data Description

The dataset used in this study is publicly available through the World Bank Documents and Reports portal[1]. Each infrastructure project is uniquely identified by a project code (P-code), which enables efficient retrieval of related documentation. This analysis focuses on *Restructuring Papers* (RPs), which outline approved changes to a project's original design or implementation plan. A total of 167 World Bank–financed projects were examined, each with a varying number of RPs.

B. Data Collection

To build a robust dataset for analysis, we developed an automated extraction pipeline to

[1] https://documents.worldbank.org/en/publication/documents-reports

retrieve restructuring papers for the 167 World Bank-financed infrastructure projects. These projects are identified by unique project codes, known as P-codes. The pipeline uses these codes to navigate the World Bank's Documents and Reports portal, where it searches for and downloads all available restructuring papers in *.txt* format. This process is fully automated using tools like *Selenium*, eliminating the need for manual document collection. From the 167 projects targeted, restructuring papers were successfully retrieved for 117 projects. The remaining 49 projects had no publicly accessible documents at the time of extraction, possibly due to confidentiality, upload delays, or incomplete archival. The overall project workflow is illustrated in Fig. 1 and the first four blocks of the workflow constitute the data extraction pipeline.

All downloaded documents were then stored locally and indexed using *ChromaDB* [9], a vector database that allows for fast similarity-based retrieval. This setup enables seamless integration with the rest of the framework components, allowing users to query the documents efficiently using vector semantic search technique discussed later.

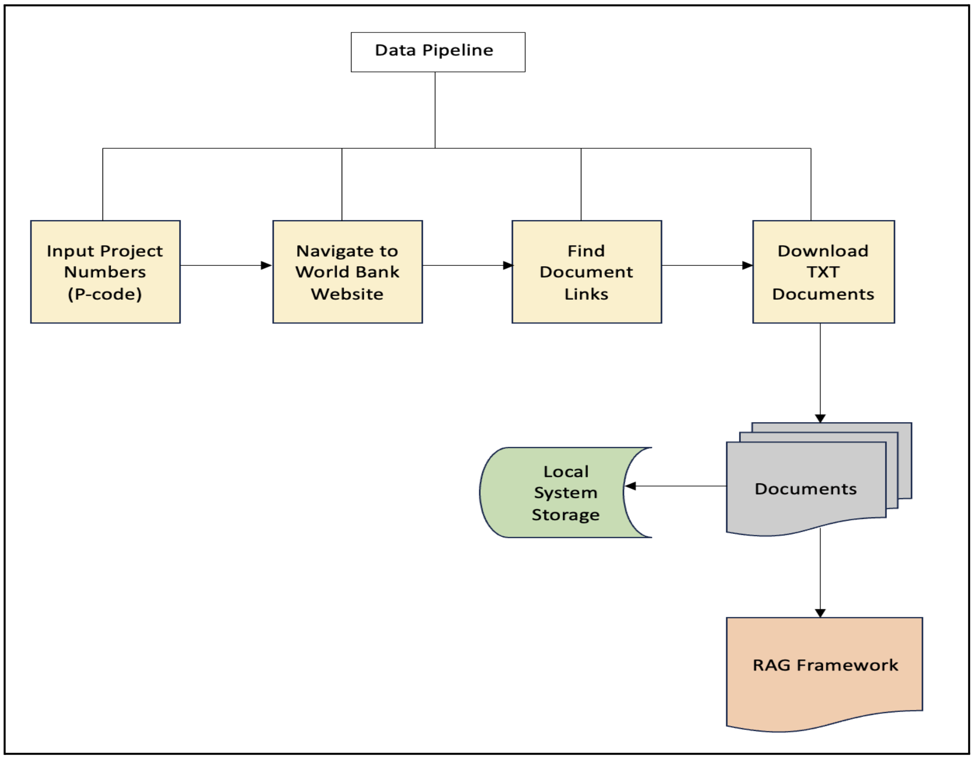


Fig. 1. Project Workflow

C. Data Preprocessing and/or Feature Engineering

The text preprocessing phase employed standard NLP techniques to ensure consistency, cleanliness, and semantic integrity of the input data. All text was first converted to lowercase to minimize case-related variability and reduce dimensionality. Non-ASCII characters were removed to maintain encoding uniformity and prevent downstream processing errors. Redundant whitespace was eliminated to standardize formatting across documents. Additionally, a selective approach to punctuation removal was applied—retaining only those symbols deemed contextually meaningful while discarding others likely to introduce noise. These preprocessing measures were essential for enhancing the quality of subsequent embedding and retrieval tasks.

D. Data Modeling & Visualizations

i) Retrieval Augmented Generation:

Following the data extraction process using the scraping pipeline, the extracted documents are fed to the RAG framework. Fig. 2. shows the architecture of the RAG framework and the components are annotated. The RAG framework is discussed in detail in the subsequent sections.

ii) Architecture:

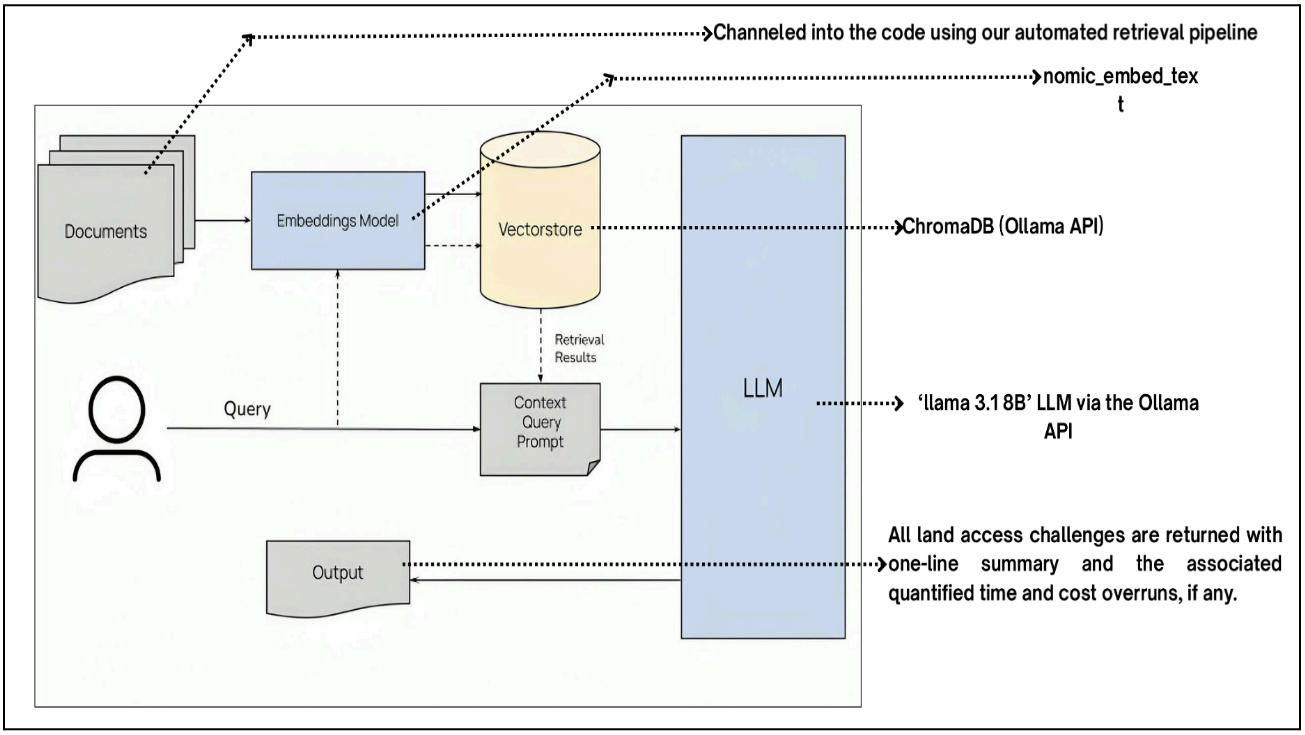

Channeled into the code using our automated retrieval pipeline
nomic_embed_tex
t
Documents
Embeddings Model
Vectorstore
ChromaDB (Ollama API)
Retrieval
Results
Query
Context
Query
Prompt
LLM
'llama 3.1 8B' LLM via the Ollama
API
Output
All land access challenges are returned with one-line summary and the associated quantified time and cost overruns, if any.

Fig. 2. Retrieval Augmented Generation framework

After preprocessing the retrieved restructuring documents, each is broken into smaller text chunks and embedded using the *nomic-embed-text* model, which converts textual data into high-dimensional vector representations that capture semantic meaning. These embeddings are stored in ChromaDB, for similarity-based retrieval.

During inference, when a user submits a query, the query itself is also embedded using the same nomic model. This query embedding is then compared to all stored document vectors using cosine similarity, a metric that measures how closely the meanings of the query and document chunks align. The top-k most relevant chunks—those most semantically similar to the query—are retrieved.

These top-k chunks are then passed to the *Llama 3.1 8B* LLM [8], accessed via the Ollama API. The LLM is prompted to generate a structured and concise summary of the land-related challenges discussed in the retrieved text, with a particular focus on their implications for time delays and cost overruns. This pipeline enables targeted, context-aware analysis of complex development documents, surfacing critical insights with minimal manual effort.

iii) Similarity Search via Cosine Distance:

The system measures cosine similarity between the user's embedded query and all pre-embedded text chunks stored in ChromaDB to identify the most semantically relevant content. To ensure quality and relevance, the retrieved results can be filtered in two ways: by applying a similarity threshold or by using fallback strategy discussed in subsequent sections.

This flexible retrieval logic ensures that users receive the most contextually meaningful results, even in cases where the semantic match is weak or ambiguous. The system also allows for filtering based on metadata, enabling users to focus on specific document types like restructuring papers or project appraisal documents (PAD). For example, code snippets demonstrate how chunk-level metadata—such as document type, project code, or source filename—can be used to refine results and isolate land-access mentions that occur in high-impact contexts. This approach balances precision and recall in information retrieval, ensuring the LLM is grounded in the most appropriate evidence for answering nuanced queries.

iv) Chunking approaches:

Two document chunking strategies were explored in this study:

Physical Chunking: Fixed-size word chunks (e.g., 30 words) offer uniform structure but risk breaking semantic coherence across sentences. This can lead to incomplete or fragmented thoughts within a chunk, making it difficult for embedding models to capture meaningful context and for LLMs to generate coherent responses. As a result, critical information may be split across multiple chunks, reducing retrieval precision and interpretability—particularly problematic in domains like land governance where nuanced language and causal reasoning are essential.

Semantic Chunking:

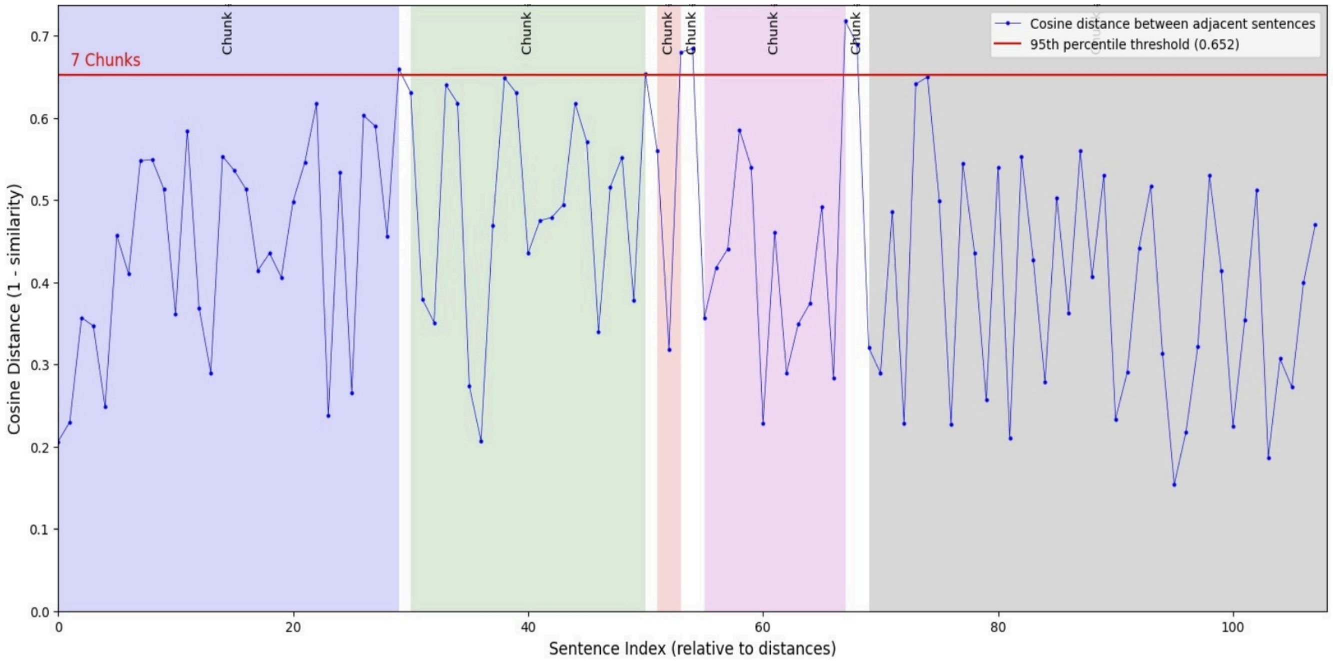


Fig. 3. Document Chunk Breakpoints

Dynamic chunking based on embedding distance between adjacent sentences offers a more semantically aware alternative to fixed-size splitting. The idea behind the particular semantic approach used in this study is taken from Full Stack Retrieval [7]. In this approach, each sentence is embedded individually, and cosine distances are calculated between consecutive sentence pairs. Breakpoints are introduced wherever the cosine distance exceeds the 95 percentile threshold indicating a natural shift in topic or meaning. This ensures that resulting chunks maintain semantic coherence by grouping contextually related sentences together. A cosine distance plot is also generated to visualize and validate these transitions and is shown in Fig. 3. Each obtained chunk is highlighted using a unique colour in the plot. This method preserves linguistic boundaries, reduces fragmentation of key ideas, and improves both retrieval accuracy and LLM response quality by grounding answers in conceptually unified text segments.

v) PCA Visualization

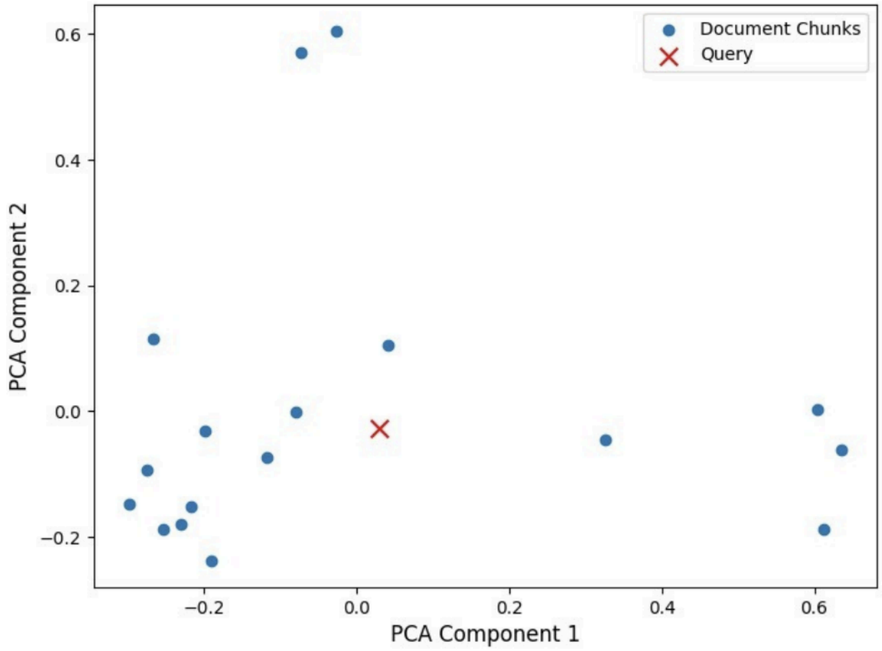


Fig. 4. PCA Visualization of the query and chunks embeddings

To visually interpret the semantic relationship between a user query and retrieved document chunks, Principal Component Analysis (PCA) was applied to project high-dimensional embeddings into a two-dimensional space. This dimensionality reduction technique preserves the most significant variance in the data while enabling intuitive visual comparisons. In the resulting scatterplot, the user query is marked as a red cross, while the document chunks are shown as blue dots. Chunks that appear closer to the query in this 2D space are semantically more relevant, as their embeddings share stronger similarities with the query vector. This visualization not only validates the effectiveness of the retrieval process but also helps diagnose outliers or irrelevant matches, offering a clear, interpretable snapshot of how meaning is distributed across the corpus.

vi) Retrieval strategy:

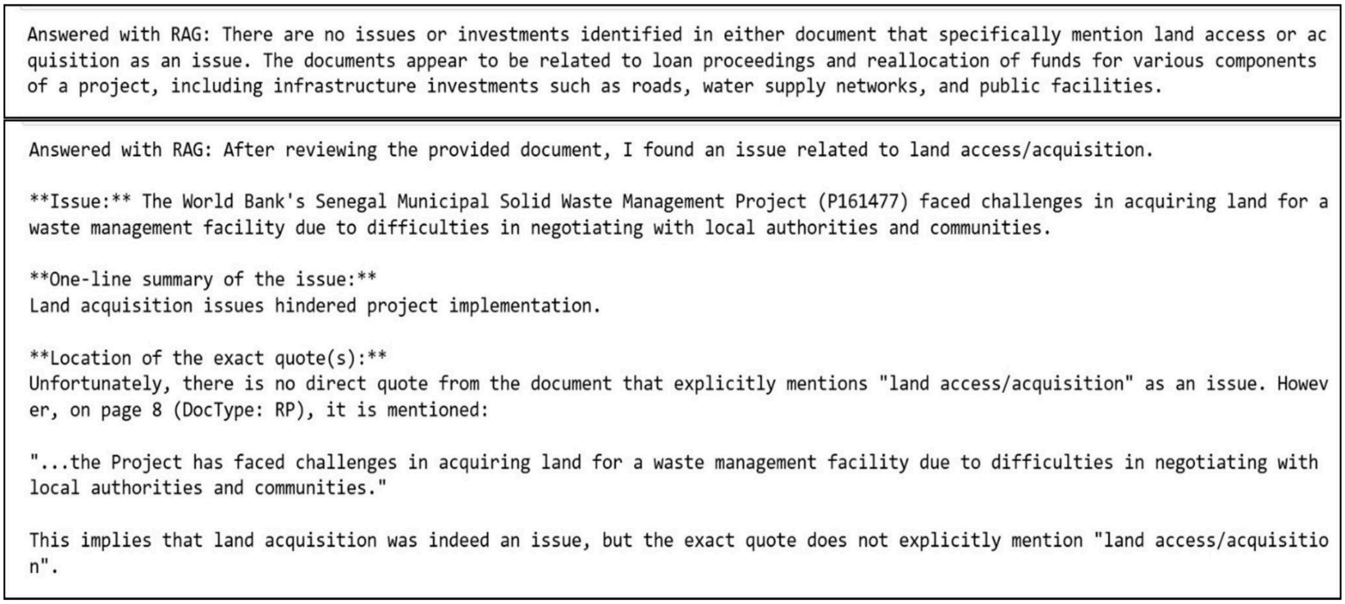
To dynamically extract the most relevant content, the system employs a retrieval strategy based on semantic similarity between the query and the vectorized document chunks.

Methodology - The retrieval process begins by generating an embedding of the user query using the nomic-embed-text model. Candidate chunks are then retrieved from the ChromaDB vector store using cosine similarity. A similarity threshold of 0.6 is applied to select semantically relevant chunks. If no chunk meets the threshold, a fallback mechanism retrieves the top two most similar chunks. The reason for choosing only two chunks is that the land access issues are generally mentioned in fewer paragraphs in the restructuring papers. These chunks are compiled into a context prompt and passed to the LLM within the RAG framework to generate a final output.

Advantages - This retrieval strategy ensures that only highly relevant document segments are selected, thereby increasing semantic alignment with the user query. It guarantees that even low-similarity queries have a fallback context, improving output completeness. Overall, it enhances the accuracy and efficiency of responses generated by the RAG pipeline.

The code for this project is available on our GitHub[2].

IV. Results and Analysis



```
Answered with RAG: There are no issues or investments identified in either document that specifically mention land access or ac
quisition as an issue. The documents appear to be related to loan proceedings and reallocation of funds for various components
of a project, including infrastructure investments such as roads, water supply networks, and public facilities.
```

```
Answered with RAG: After reviewing the provided document, I found an issue related to land access/acquisition.

**Issue:** The World Bank's Senegal Municipal Solid Waste Management Project (P161477) faced challenges in acquiring land for a
waste management facility due to difficulties in negotiating with local authorities and communities.

**One-line summary of the issue:**
Land acquisition issues hindered project implementation.

**Location of the exact quote(s):**
Unfortunately, there is no direct quote from the document that explicitly mentions "land access/acquisition" as an issue. Howev
er, on page 8 (DocType: RP), it is mentioned:

"...the Project has faced challenges in acquiring land for a waste management facility due to difficulties in negotiating with
local authorities and communities."

This implies that land acquisition was indeed an issue, but the exact quote does not explicitly mention "land access/acquisitio
n".
```

Fig. 5. Results based on two different projects

[2] https://github.com/tushar2016sharma/WB36-DATS-6501.git

The RAG output from an example project illustrated in Fig. 5 showcasing the effectiveness and interpretability of the framework when applied to the restructuring paper documents. The system is designed to respond to user queries about land access or acquisition challenges by either retrieving and summarizing relevant content or clearly indicating when no such issue is identified. The first example demonstrates the latter case: the model reviews the input documents and accurately concludes that there are no references—either direct or implied—to land-related delays or issues. This "no-issue" response underscores the system's ability to filter out irrelevant results, thereby reducing false positives and preserving analytical precision.

In contrast, the second example illustrates a positive identification. The system detects a land acquisition challenge in the Senegal Municipal Solid Waste Management Project (P161477), where difficulties in negotiating with local authorities and communities impeded the establishment of a waste management facility. While the term "land acquisition" is not explicitly mentioned, the model successfully interprets the underlying implication from the phrase "challenges in acquiring land," demonstrating its capacity for semantic understanding beyond surface-level keyword matching. The extracted output is formatted into a structured template that includes: (1) a full description of the issue, (2) a concise one-line summary for quick review, and (3) the location and content of the supporting quote, including metadata such as page number and document type.

This structured formatting improves the interpretability and usability of the extracted insights, making them suitable for integration into dashboards, evidence databases, or downstream policy analyses. Overall, these examples highlight the system's dual strengths: (1) high precision in distinguishing between relevant and non-relevant cases, and (2) the ability to surface and clearly present contextually inferred land-related challenges in a user-friendly, standardized format.

The implemented RAG framework was tested on a total of 117 World Bank restructuring papers to identify and extract projects with explicit mentions of land access or acquisition issues. The system successfully identified 49 projects containing land access challenges, while correctly determining no such issues existed in the remaining 68 documents. The example output in Fig. 5 shows clear differentiation in responses—explicitly acknowledging absence or presence of land-related issues. Although the RAG pipeline occasionally produced inaccuracies or "hallucinations"—for instance, incorrectly interpreting unrelated issues as land access problems (false positives)—overall, the framework demonstrated robust and reliable extraction capabilities. These results suggest that the developed RAG system provides substantial value in rapidly pinpointing land acquisition-related issues from lengthy documents, thus enabling faster and evidence-based decision-making.

V. Conclusion

A. Conclusion

This study demonstrates the potential of RAG-based systems in addressing land access challenges for infrastructure development. By significantly reducing manual effort, the system enables faster insight discovery and ensures consistent, evidence-based findings across restructuring documents. It highlights root causes of delays and cost overruns, providing valuable decision support to project stakeholders. The pipeline improves resource efficiency, allowing analysts to focus on solution development rather than document review. Additionally, the framework is scalable and adaptable, capable of re-running analyses as new documents become available. Future improvements could include LLM fine-tuning, transfer learning, or integrating topic modeling for deeper thematic understanding.

B. Project Limitation

While the proposed framework enables automated extraction of land-access challenges, it is not without limitations. The system is computationally expensive and may require significant resources for embedding generation and vector storage. Its performance is also sensitive to the phrasing of user queries, which can influence the relevance of retrieved results. Furthermore, the language model occasionally exhibits hallucinations, producing inaccurate or unverifiable responses.

C. Future Research

Future improvements may include the development of a user-friendly interface such as a chatbot, website, or dashboard to enhance accessibility for non-technical users. Fine-tuning the language model could further improve the accuracy and specificity of generated outputs. Alternate approaches to document chunking could be considered in future such as clustering-based approaches. Additionally, integrating GIS-based analysis would enable spatial mapping of land-access challenges, offering valuable insights into geographically clustered risks and supporting proactive intervention planning.

VI. References

[1] Liberatore, D., Kalimeri, K., Sever, D., & Mejova, Y. (2024, September). Quantitative Information Extraction from Humanitarian Documents. In Proceedings of the 2024 International Conference on Information Technology for Social Good (pp. 240-248).

[2] Tamagnone, N., Fekih, S., Contla, X., Orozco, N., & Rekabsaz, N. (2023). Leveraging domain knowledge for inclusive and bias-aware humanitarian response entry classification. arXiv preprint arXiv:2305.16756.

[3] Cai, M. (2021). Natural language processing for urban research: A systematic review. Heliyon, 7 (3), e06322.

[4] M. Lewis, E. Perez, A. Piktus, V. Karpukhin, N. Goyal, and S. Chakrabarti, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[5] Aydogan, A. F., An, M. K., & Yılmaz, M. (2021, June). A new approach to social engineering with natural language processing: RAKE. In 2021 9th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-7). IEEE.

[6] https://www.maartengrootendorst.com/blog/keyllm/

[7] G. Kamradt, "Full Stack Retrieval," [Online]. Available: https://fullstackretrieval.com/. [Accessed: Apr. 20, 2025].

[8] H. Touvron, et al., "LLaMA: Open and Efficient Foundation Language Models," arXiv preprint, arXiv:2302.13971, 2023.

[9] Chroma Developers, "ChromaDB: Fast, Open-Source Embedding Database," Chroma Documentation, 2024. [Online]. Available: https://docs.trychroma.com. [Accessed: Apr. 20, 2025].

[10] SeleniumHQ, "Selenium WebDriver Documentation," 2025. [Online]. Available: https://www.selenium.dev/documentation. [Accessed: Apr. 20, 2025].

[11] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python, Sebastopol, CA, USA: O'Reilly Media, 2009.