ESG Data Extraction System – Design Decisions & Implementation Report

1. Introduction

This document describes the design decisions, implementation steps, challenges encountered, and future improvements for the ESG Data Extraction System. The objective of the project was to build a reliable, auditable pipeline capable of extracting predefined ESG indicators from annual report PDFs.

2. Problem Context

Annual reports are lengthy, heterogeneous documents containing structured ESG disclosures. Traditional keyword-based extraction is unreliable, while fully semantic approaches introduce hallucination risks. The system therefore prioritizes deterministic extraction with explicit handling of missing values.

3. Key Design Decisions

Page-Guided Extraction: Known page ranges are used to locate indicators, avoiding semantic search.
LLM as Parser: Large Language Models are used strictly for structured JSON parsing, not inference.
Batch Processing: Indicators are extracted in a single request per company to reduce cost and latency.
Graceful Failure: Missing indicators are explicitly recorded as 'Not found'.

4. System Architecture

The system consists of configuration, PDF extraction, aggregation, LLM parsing, validation, persistence, and export layers. Each layer has a clearly defined responsibility, ensuring maintainability and auditability.

5. Implementation Steps

Step 1: Define ESG indicators and expected units.
Step 2: Map indicators to page ranges per company.
Step 3: Extract text from relevant PDF pages.
Step 4: Aggregate text and invoke the LLM.
Step 5: Validate results and store them.
Step 6: Export normalized CSV output.

6. Challenges Encountered

Challenges included inconsistent PDF formatting, LLM rate limits, malformed JSON responses, and duplicate data insertion. These were addressed through batching, validation, and strict control over insertion logic.

7. Why Vector Databases Were Not Used

Vector databases were intentionally excluded because page locations were known and indicators were predefined. Page-based extraction provided higher precision and simpler explainability.

8. Limitations

The system does not handle image-only PDFs, relies on text quality, and does not perform semantic summarization. These limitations are acceptable for regulated ESG reporting use cases.

9. Next Steps and Enhancements

Future improvements include table-aware parsing, confidence scoring heuristics, hybrid LLM-table extraction, improved unit normalization, and optional semantic fallback using embeddings.

10. Conclusion

This project demonstrates a pragmatic, production-aligned approach to ESG data extraction. The system balances automation with reliability and is suitable for compliance-driven environments.