# iCOR System Design

## 1. Executive Summary

The **iCOR (Intelligent Compliance & Obligation Register)** system is a Legal Tech solution designed to automate the monitoring of UK environmental legislation. By replacing manual legal review with an automated pipeline of web scraping and Generative AI, the project reduces the time required to build legal registers by approximately 80% while ensuring structured, machine-readable outputs.

## 2. System Architecture

The pipeline is divided into three critical stages: Data Extraction, Data Structuring, and AI Synthesis.

### 2.1 Data Extraction (Web Scraping)

Using the BeautifulSoup library, the system targets the UK Legislation database.

- **Targeting:** The script specifically isolates the <title> and the main body text of Statutory Instruments (e.g., *The Hazardous Waste Regulations 2005*).
- **Optimization:** By transitioning from a general soup.find_all('text') to specific tags like class_='legislationBody', the system filters out non-essential HTML elements (headers, footers, and sidebars). This reduces the **Token Count** sent to the LLM, significantly lowering API costs.

### 2.2 Data Structuring

Data is stored in a **Pandas DataFrame** to allow for easy manipulation. The structure includes:

- **Title:** Extracted from the HTML metadata.
- **URL:** Retained for traceability and legal referencing.
- **Text:** The raw legislative prose used as context for the AI.

## 3. Generative AI & Prompt Engineering

The system utilizes the **Azure OpenAI GPT-4** model to perform "Legal Synthesis."

### 3.1 The "Expert Auditor" Prompt

To ensure accuracy and prevent "hallucinations," a specific **System Prompt** was developed:

*"You are an expert in environmental legislation within the UK and building legal registers... Return your responses in a JSON format."*

### 3.2 Structured Output (JSON)

The model is instructed to answer three specific legal queries and return them as a JSON object. This allows the output to be automatically injected into business databases without manual formatting.

- **Question 1 (Summarisation):** Condenses complex law into one concise paragraph.
- **Question 2 (Key Changes):** Identifies specific updates (e.g., changing "special waste" to "hazardous waste").
- **Question 3 (Chronology):** Lists all cited legislations in date order to establish a legal hierarchy.

**4. Business Impact & Cost Management**

- **Token Efficiency:** Since Azure OpenAI charges per token, iCOR implements text truncation and noise filtering. By cleaning the raw HTML before processing, we minimize the "Assistant Context" length.
- **Scalability:** The script is designed to be "loopable," allowing a user to feed a list of 100+ URLs and generate a complete Environmental Legal Register in minutes.

**5. Conclusion**

iCOR demonstrates that AI is not just for chat, but a powerful tool for **Regulatory Technology (RegTech)**. The project successfully proves that unstructured government data can be transformed into a high-value business asset through disciplined scraping and strategic prompt engineering.