Caprae LeadGen Scraper Report:

Approach

The Caprae LeadGen Scraper is a lightweight Flask-based web app designed to extract contact information (emails, phone numbers, LinkedIn profiles) from any user-supplied URL. It uses HTTP requests to fetch pages, parses HTML using `BeautifulSoup`, and uses regular expressions to extract structured contact data.

Model Selection

This project is non-ML and rule-based. The following tools are used:

- Flask for creating the web interface.
- requests for downloading webpage content.
- BeautifulSoup for parsing and navigating HTML.
- re (regex) for pattern-based data extraction.

This stack was chosen for its reliability, simplicity, and ability to handle a wide range of static web pages.

Data Preprocessing

After extraction, the data passes through several stages:

- Deduplication – Eliminates repeated entries.
- Validation – Ensures correct formats using regex.
- Formatting – Structures the data for clean CSV export

Performance Evaluation

- Accuracy: High precision due to strict regex filtering.
- Completeness: Effective for static websites with publicly visible data.

Rationale

A rule-based system is efficient and easy to maintain for the task at hand. It ensures fast lead extraction without requiring the overhead of training or deploying a machine learning model, making it ideal for early-stage lead generation tools.