Research Design

This study adopts a mixed-method approach that integrates automated text analysis using a large language model (LLM) with traditional manual evaluation methods. The feasibility study is structured to assess the LLM's ability to evaluate administrative acts by comparing its outputs against those generated by human experts. The overall framework comprises data collection, checklist development, automated querying, and performance benchmarking.

Data Collection Process

Municipal documents, specifically determinations (or "determina") available on the public bulletin (albo pretorio), form the primary data source. For each document, the following steps are undertaken:

- Document Acquisition: Download PDF files of relevant determinations.
- Checklist Selection: Use a CSV file that maps the document names to their associated checklists.
- Text Extraction: Convert each PDF into plain text using Python-based tools, ensuring the content is accessible for further processing.

Checklists Development

To standardize the evaluation process, a series of checklists was compiled based on the administrative control requirements used by public administration entities. These checklists were transformed into a JSON format, allowing the division of each checklist into individual questions. Each checklist point includes specific instructions that guide the evaluation of the act in question.

LLM Usage

The LLM is central to this study. Its role is to analyze the text extracted from municipal determinations and respond to each point of the checklist. The process involves:

- **Prompt Design and Configuration:**

  A structured prompt template was created to guide the LLM. The template includes:

  - Instructions to the model as an expert in administrative law.
  - A clear layout where each checklist point is presented as a question.
  - A predefined answer format (e.g., SI/NO/NON PERTINENTE with a brief explanation if needed).

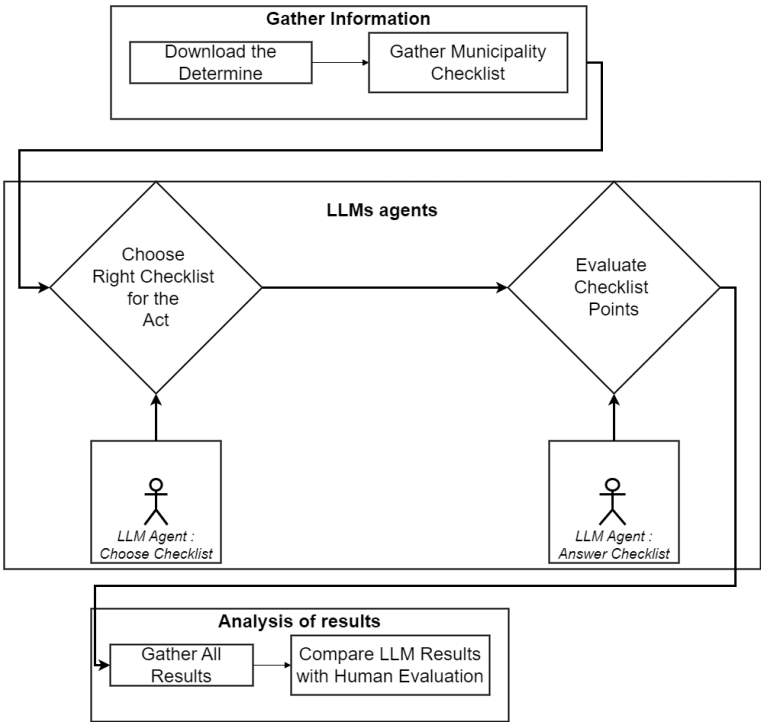
- **Extraction and Analysis Process:**

  The LLM is queried for each checklist point using the custom prompt. The responses, initially in long text form, are processed using regular expressions to extract the essential output: namely, a categorical answer (SI, NO, NON PERTINENTE) and a concise explanation when necessary. A Python script orchestrates this process and aggregates the results into a CSV file.

- **Human Evaluation Procedure for Benchmarking:**

  To benchmark the LLM's performance, a parallel manual evaluation was conducted. A human expert compiled a checklist-based assessment for each determination. The automated results were then compared against these manually obtained values to assess accuracy and reliability.

Workflow



Tools and Performance Metrics

Several tools and performance metrics were used throughout the study:

- **Software Tools:**
  - Python for text extraction, prompt generation, and response processing.
  - Regular expressions for cleaning and standardizing text outputs.
  - OpenAI's API for accessing the LLM.
  - Hugginface Transformers Library to load and use locally llama models

  - CSV and JSON file handling for managing checklists and responses.

- **Performance Metrics:**
  - **Accuracy:** The degree to which the LLM's responses match the human-evaluated answers.
  - **Consistency:** Variability in outputs with changes in the model's temperature and other parameters.
  - **Efficiency:** Processing time and resource usage for automating the evaluation process.
  - **Error Rate:** Frequency and nature of errors encountered during response extraction and analysis.