Experiment Report: Field Extraction & Confidence Scoring

Context & Methodology

Models Used: Gemini 2.5 Flash and GPT-4o. **Workflow:** 1. **Classification:** Documents are classified into tiers **A, B, and C** with decreasing priority. Tier C is discarded; only **A and B** are retained. 2. **Extraction:** LLMs process documents as images. 3. **Sampling:** Content is extracted from **5 pages** (the first 3 and the last 2) to capture the main information for field extraction.

Experiments & Results

1a. Using Logprobs (Gemini)

- **Model Constraint:** Only supports **Gemini 2.0 Flash** (no newer models available).
- **Mechanism:** Logprobs returns the probability of the tokens *following* the field label.
  - *Example:* For the field `tax_year`, it calculates the average probability of the subsequent tokens (e.g., `2023`).
- **Observation:** Since this is an extraction task, probabilities tend to be naturally high, resulting in very high confidence scores across the board.
- **Result:** View JSON Blob

1b. Using Logprobs (ChatGPT)

- **Model Constraint:** Similar to Gemini, this is supported on **GPT-4o**.
- **Mechanism:** Follows the same logic as the Gemini implementation.
- **Result:** View JSON Blob

2. LLM as a "Judge"

- **Mechanism:** A second LLM is used to evaluate the extracted content.
- **Observation:** This effectively acts as a verification step where the LLM assesses the validity of the content it (or another instance) extracted.
- **Result:** View JSON Blob

3. Multi-Model Consensus (Ensemble)

- **Mechanism:** Multiple distinct models are called (both Gemini and ChatGPT). In this experiment, **4 models** were used.
- **Logic:** The final result is derived by aggregating the outputs from all 4 models and calculating an average or consensus for each field.
- **Drawback:** Significantly higher cost due to multiple API calls.
- **Result:** View JSON Blob

4. LLM Self-Evaluation

- **Mechanism:** The LLM is instructed to evaluate its own output during the extraction process, determining if the field is relevant to the document content.
- **Observation:** Low objectivity and reliability.
- **Result:** View JSON Blob