# Synthetic Romanian Legal Corpus Generation

## 1 Methodology

The proposed approach utilizes a multi-stage pipeline designed to generate synthetic Romanian legal documents (specifically notary powers of attorney and declarations). The methodology circumvents the need for computationally expensive fine-tuning on a small dataset (approx. 27 documents) by leveraging In-Context Learning (ICL) and a Retrieval-Augmented Generation (RAG) inspired workflow. The pipeline consists of four distinct phases:

### 1.1 Data Preparation and Cleaning

The initial dataset consisted of 27 raw, anonymized real-world templates containing placeholders (e.g., "..."). An initial Large Language Model (LLM) pass was employed to "repair" these documents, replacing ellipses with coherent synthetic entities using a probabilistic approach. This resulted in a "Golden Standard" library of clean, complete text files stored in a persistent environment (Google Drive).

### 1.2 Scenario Generation (AI Screenwriter)

To ensure corpus diversity and prevent overfitting to specific templates, a distinct LLM instance acts as a "Screenwriter." Instead of using static procedural generation, this agent generates complex JSON metadata containing diverse legal scenarios (e.g., real estate sales, vehicle registration, travel declarations for minors, succession). This introduces semantic variety and realistic edge cases into the input data.

### 1.3 Context-Aware Generation

The generation phase employs a retrieval mechanism. Based on the metadata generated in the previous step, the system retrieves the most semantically relevant template from the "Golden Standard" library. The LLM is then prompted using a One-Shot strategy: it is provided with the retrieved template as a strict stylistic guide and instructed to rewrite it using the new synthetic metadata. This ensures 100% adherence to legal formatting and terminology while varying the factual content.

### 1.4 Unified Critic (Evaluation)

A final validation step is performed by a "Critic" agent. This agent evaluates the generated output against two criteria simultaneously:

- **Structural Integrity (Syntax):** Verifies the absence of artifacts (e.g., remaining placeholders), valid JSON formatting, and the presence of mandatory legal clauses (GDPR, signature blocks).
- **Semantic Consistency (Logic):** Verifies that the generated legal text aligns with the intent specified in the metadata (e.g., if the purpose is "sale," the text must authorize a sale).

Only documents that pass both criteria are appended to the final JSONL dataset.

## 2 Model Specifications

The core model used for all agents (Cleaner, Screenwriter, Generator, and Critic) is **RoLlama3-8b-Instruct-DPO-2025-04-23**, a Llama 3 derivative fine-tuned for the Romanian language. To facilitate execution on consumer-grade hardware (Google Colab T4 GPU with 16GB VRAM), the model was loaded using 4-bit quantization (NF4) via the `bitsandbytes` library. This configuration allowed for an efficient inference pipeline without significant degradation in linguistic quality.

## 3 Results

The pipeline successfully generated a synthetically augmented corpus that mimics the stylistic nuances of the 27 original source documents. The implementation of the "AI Screenwriter" and template retrieval mechanism solved the issue of mode collapse, ensuring that the output covers a wide taxonomy of legal acts. The "Unified Critic" loop effectively filtered out hallucinations and incomplete generations, resulting in a high-quality JSONL dataset suitable for future downstream tasks such as Named Entity Recognition (NER) or legal text classification.

## 4 Conclusion

This project demonstrates that high-fidelity synthetic legal data can be generated for low-resource languages (like Romanian) without extensive fine-tuning. By separating the workflow into specialized agents—specifically decoupling the scenario invention from the drafting process—and utilizing rigorous self-verification steps, we achieved a robust system capable of scaling a small set of templates into a large, diverse dataset for machine learning applications.