Evaluation of Automatically Generated Romanian Legal Documents

1 Introduction

This report presents the evaluation of automatically generated Romanian legal documents using a Romanian language model based on GPT-2 (`readerbench/RoGPT2-medium`). The purpose of the evaluation is to assess how coherent, natural, and structurally appropriate the generated documents are compared with real legal documents used during fine-tuning.

Four types of legal documents were analyzed:

- **Proces-verbal (official report)**
- **Procură (power of attorney)**
- **Cerere (request/petition)**
- **Declarație (statement)**

2 Methodology

2.1 Document Generation

For each document type, the model was fine-tuned using a dataset of real Romanian legal documents. After fine-tuning, the model was prompted using structured prompts designed for each category, and multiple documents were generated for evaluation.

2.2 Automatic Evaluation

The generated documents were evaluated automatically using several quantitative metrics. This report focuses on **perplexity**, a standard metric in natural language processing.

2.2.1 Perplexity

Perplexity measures how well a language model predicts a given text. It is computed using the log-likelihood of the sequence, and lower values indicate:

- more fluent text,
- more natural structure,
- fewer inconsistencies or unexpected elements for the model.

For each document category, the average perplexity across all generated documents was computed.

3 Results

Table 1 shows the average perplexity values for all document types:

Document Type	Average Perplexity
Proces-verbal	10.03
Procură	4.61
Cerere	5.62
Declarație	6.45

Table 1: Average perplexity scores for generated legal documents.

4 Interpretation of Results

The numerical results show meaningful differences between categories:

- **Procură** documents have the lowest perplexity (4.61), indicating the highest fluency and most predictable structure for the model.
- **Cerere** and **Declarație** have moderate perplexity values (5.62 and 6.45), suggesting reasonably good structure but more variation.
- **Proces-verbal** documents show the highest perplexity (10.03), indicating that these texts are structurally more complex and harder for the model to generate consistently.

The results suggest that the model learns simple, formulaic structures (such as *procuri*) more easily, while complex procedural or narrative documents (such as *procese-verbale*) remain more challenging.

5 Conclusion

The evaluation demonstrates that the fine-tuned Romanian GPT-2 model is capable of generating coherent legal documents, with varying performance depending on the structural complexity of each category. Future work may include a 'two-hat' approach, where the LLM is both the generator and the 'critic'.