UNIVERSITEIT VAN AMSTERDAM

Data Systems Project

Using Generative AI in the Legal Domain

Group D1
1 November 2024

Background

- Introductions.
- We received mixed instructions from the course, and wanted to quickly run with you the approach.
- First, we expected a well-defined topic that we needed to research.
- After our discussion with Tom it appeared that we have certain freedom on the actual topic of research within the context of the topic of the project of “Using Generative AI in the Legal Domain”, as described in the brief.
- As such, we have looked into the issue of interpretability of the law and intersected it with LLMs evaluation methods.

Problems of interpretability

Main research papers: Doshi-Velez and Kim (2017) **and** Gao et al. (2023)

Trust and Transparency Risks

- Lack of clarity affects confidence in AI-driven legal interpretations which can lower trust.
- Critical in the domain of Law where end-users need to trust and understand AI outputs.

Ambiguity in Legal Language

- AI struggles with nuanced terms, leading to potential misinterpretations.
- Problematic in law, where misinterpretation can lead to incorrect legal advice or decisions.

Inconsistent Interpretations

- Inconsistent interpretations can arise when AI interprets similar legal terms differently across contexts.
- This variability in AI responses can lead to unreliable legal outcomes.

Human Dependency

- Persistent need for expert oversight to verify AI-generated interpretations.
- Current AI systems in law aren’t fully autonomous, which limits efficiency gains AI could offer.

Evaluation methods

Main research papers: Guo, Jin, et al. and Qiguang, Ziyu, et al.

1. Human-Based Evaluation

- Expert Review: Assessments from domain experts (e.g., legal professionals) on coherence, relevance, and applicability of LLM outputs to legal/regulatory documents.
- Tasks & Benchmarks: Datasets like ContractQA and TruthfulQA used for evaluating model explanations and interpretations.
- Use Case: Legal document interpretation, where human judges validate the correctness of law-based generative outputs.
- Important to ask the right question. We really need to have clear the research question we want to answer, so that it can be embedded in the survey/task for the expert.

2. Objective Metrics

- Automated Scoring:
  - Metrics such as ROUGE, BLEU (these metrics may differ depending on the LLM, we could research how to make it more robust and coherent) , and exact match scores measure word overlap, syntax, and accuracy.
  - Sentence Similarity tests using cosine similarity or embedding models to evaluate whether two pieces of text convey the same meaning
- Robustness Tests: Methods like input perturbations (synonym swaps, OCR typos) check if the model's predictions remain stable under slight input modifications

Human-based methods

Main research papers: Guo, Jin, et al. and Qiguang, Ziyu, et al.

- **Expert Review:** Assessments from domain experts (e.g., legal professionals) on coherence, relevance, and applicability of LLM outputs to legal/regulatory documents.

- **Tasks & Benchmarks:** Datasets like ContractQA and TruthfulQA used for evaluating model explanations and interpretations.

- ContractQA and TruthfulQA are datasets designed to test how well language models explain and interpret legal and regulatory content. These benchmarks help consistently assess AI explanations, ensuring they meet standards for accuracy and relevance in law.

- **Use Case:** Legal document interpretation, where human judges validate the correctness of law-based generative outputs.

- Important to define clear questions for these evaluations. Precise questions help experts provide focused feedback, making it easier to evaluate the AI's effectiveness in legal interpretations.

Objective metrics

How to measure word overlap, syntax and accuracy?

BLEU & ROUGE

- Both metrics rely on n-gram matching to evaluate how closely generated text resembles a reference. BLEU focuses on precision, and ROUGE emphasizes recall.
- Both metrics provide complementary insights: BLEU is more stringent and precise, while ROUGE captures overall content recall.
- In the context of legal interpretation, ROUGE can help ensure that all relevant legal terms are included, while BLEU can assess the accuracy of specific legal phrases.

Exact match scores

- Accuracy metric

How can we ensure robustness and coherence?

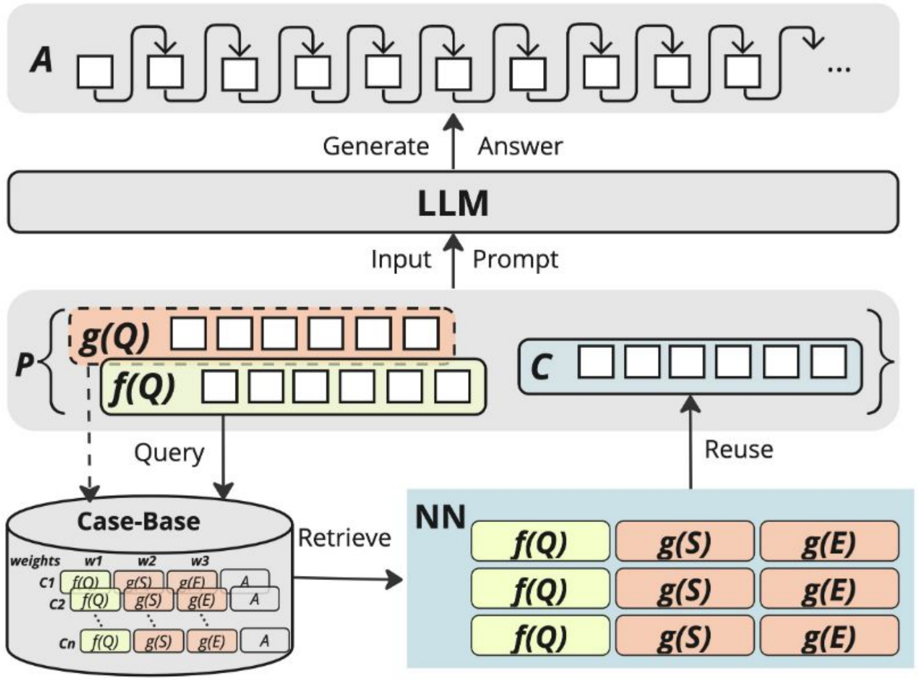
Technology Overview

SOTA systems mostly utilize the RAG pipeline systems that leverage domain-specific databases, including case law, statutes, and legal commentary, using retrieval mechanisms like BM25 and dense retrieval models (BERT architecture):

- DistilBERT (Sanh et al., 2019)
- Legal-BERT (Chalkidis et al., 2020)
- Case-Based Reasoning for RAG (Wiratunga et al., 2024)

Some notable baseline models that we've identified on GitHub:

- https://github.com/rgu-iit-bt/cbr-for-legal-rag (2024)



CBR-RAG (Wiratunga et al., 2024).

Research document

Following our discussions on the legislation that we could use in our case study as a proof of concept, we evaluated a few pieces of legislation and had some considerations:

1. AI Act is a framework from the EU that sets the standard for AI compliance across the EU. Although this could be a pertinent topic to use for our case study, it has been recently enacted, so there is limited literature on interpretability guidelines specific to the Act.

2. GDPR has stronger regulatory precedent which provides more robust evaluation criteria for language models. For instance, the "right to explanation" is a concept coined in the GDPR that has become a well-defined principle and can serve as a foundation for developing interpretability metrics.

3. Language considerations. Although TNO has largely focused on the evaluation based on Dutch pieces of legislation, we believe that results in English should broadly extend to Dutch due to shared legal concepts and model adaptability across languages.

Meeting notes

Main research papers: Guo, Jin, et al. and Qiguang, Ziyu, et al.

1. Human-Based Evaluation

- Important to ask the right question. We really need to have clear the research question we want to answer, so that it can be embedded in the survey/task for the expert.

2. Objective Metrics

Metrics such as ROUGE, BLEU (these metrics may differ depending on the LLM, we could research how to make it more robust and coherent)

BM25 with multilingual BERT models (they focus on dutch) Dragon or Spleet? are hybrid. There are no new retrievers specific to law that are new, so the ones from 2 or 3 years ago seem fine.

Probably we want to use a dataset that has already been out there for a little bit:

Get one of their pipelines as the baseline.

Conclusion & Next Steps

- We would like to explore further some specific evaluation methods in the context of the interpretability of law (around 2 to 4 methods that include both human-based and objective metrics).

- Our goal would be to design one or two evaluation methods that could be used within the context we described.

- For the human-based evaluation methods, we have been able to reach out to some UvA alumni of the MSc in International Law that have agreed to help in the coming months.

- https://gitlab.com/normativesystems/flintfillers/aqa_preconditions

Thank You For Your Attention!

Team D1: Eduard, Jobeal, Gabriela & Filip

Comments from Tom

1. **The main point that he said was about making sure that we ask the right questions. We need to make sure that we have example questions with good answers that we need to first verify with people with legal domain expertise. The next thing is to really focus on the problem of interpretability. that is, we need to "categorise" the ways in which a specific article can be interpreted (e.g., textualism, intentionalism and purposivism). I would personally add examples of this in the next presentation and make it interactive (e.g., given text X, who would say that this means Y or Z, to highlight the problems of interpretability.**

1. **Important: next time let's have a specific title and subtitle to the presentation. I think it might not have been clear that we want to focus on the *evaluation methods* of interpretability, rather than on the interpretability itself.**

1. **Feedback from Tom: "The general idea is okay (using genAI to answer GDPR questions) but not thought about test cases (the questions, model answers, etc.)".**