Preparing AI for Compliance: Initial Steps of a Framework for Teaching LLMs to Reason About Compliance

Barbara Makovec[1,2], Luis Rei[1,*] and Inna Novalija[1]

[1]Institut "Jožef Stefan", Jamova 39, Ljubljana, Slovenia
[2]Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, Ljubljana, Slovenia

Abstract

The integration of powerful Large Language Models into diverse applications has been rapid, but it faces significant challenges due to the complexity of global regulatory and ethical frameworks, such as those in the GDPR and the AI act. To address the need for AI systems that can navigate these compliance requirements, we propose a tool designed to create a specialized dataset for training AI assistants in regulatory and ethical reasoning and present its initial implementation. Our approach uses a Retrieval-Augmented Generation (RAG) method that preserves the structure of legal texts, ensuring accurate retrieval and interpretation of relevant provisions. This tool automates the generation of compliance reasoning data by selecting and explaining how specific legal and ethical guidelines impact real-world examples of AI technologies. This is to be followed by a refinement process to ensure only the best candidates are presented to the annotators. We aim to facilitate the development of AI-driven compliance assistants that can effectively align with global legal and ethical standards.

Keywords

Large Language Models (LLMs), Regulatory Reasoning, Retrieval-Augmented Generation (RAG), Chain-of-Though, Text mining, AI Governance, Fair Transparent and Trustworthy AI, Artificial Intelligence (AI) Compliance

1. Introduction

In recent years, we have witnessed the disruptive emergence of powerful Large Language Models, which can be utilized as ready-to-deploy AI services with minimal effort. Their rapid adoption spans from small-scale single-developer projects to critical integrations within Fortune 500 companies. Simultaneously, a plethora of legislations, regulations, ethical guidelines, and policy goals have emerged in the technology and data sectors, such as the GDPR [1], the Data Governance Act [2], the Data Act [3], the Artificial Intelligence Act [4]. The rapid technological advancement, coupled with diverse and evolving regulatory landscapes across different countries, presents significant challenges for developers, data scientists, researchers, regulators, and policymakers. We believe that leveraging Large Language Models (LLMs) to explain, review, and assess AI models, datasets, and complete pipelines from the perspective of legislations, regulations, ethical guidelines, and social impact can help address the challenges. For instance, a data scientist developing a new pipeline could ensure compliance with EU and USA regulations by submitting the pipeline description, along with each dataset and model card, to the compliance assistant. By selecting the relevant jurisdictions, potential issues can be identified early in the development process, facilitating faster progress before a more detailed review by the company's compliance experts.

Beyond just understanding the law, any general solution will likely require some form of Retrieval-Augmented Generation (RAG) in which the LLM can reason over the specific set of retrieved compliance requirements that can apply to a single product, service, or company at a given point in time within a certain jurisdiction. The first step towards developing a "compliance assistant" is to build datasets

RuleML+RR'24: Companion Proceedings of the 8th International Joint Conference on Rules and Reasoning, September 16–22, 2024, Bucharest, Romania
*Corresponding author.
makovecbarbara1@gmail.com (B. Makovec); luis.rei@ijs.si (L. Rei); inna.koval@ijs.si (I. Novalija)

[1]https://eur-lex.europa.eu/eli/reg/2016/679/oj
[2]https://digital-strategy.ec.europa.eu/en/policies/data-governance-act
[3]https://digital-strategy.ec.europa.eu/en/policies/data-act
[4]https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

that can be used to teach and evaluate the assistant in this complex task. Annotating and labeling this data demands the expertise of legal professionals to ensure accuracy, making the process both time-consuming and expensive. To address this challenge, we propose a framework that generates high-quality examples for annotation (Figure 1). In this paper, we discuss the details of the first part, the initial generation of examples.

2. Related Work

Our ultimate goal of creating a compliance assistant is not conceptually unique. For example, Gracenote.ai [5] is an AI-driven platform for regulatory compliance. While the legal AI CoCounsel [6] from Thomson Reuters includes contract compliance features. CuratedAI [7] uses RAG approach to answer legal questions about EU laws and regulations. In research, as overall systems, we highlight DISC-LawLLM which includes a retriever with access to a knowledge base of Chinese laws [1] and Chatlaw dynamically builds a case-specific Knowledge Graph within a multi-agent system by various methods and answers using a RAG approach [2]. Several public datasets evaluate LLM assistants' legal reasoning, such as LegalBench [3] and Contract Understanding Atticus Dataset [4]. Our goal is slightly different, as we want to do reasoning on compliance of AI tools with variable provisions. Given an LLM that is instructed to reason only on specific retrieved provisions, the user can select which provisions would be considered by selecting those that can be retrieved, e.g. only laws that apply in the EU, plus provisions that apply to the financial sector, plus the user's ethical guidelines. For generating better responses, Chain-of-Thought Prompting enhances LLM reasoning by generating intermediate steps [5], and LLMs can perform zero-shot reasoning by adding "Let's think step by step" before answers [6]. Self-Consistency improves this by sampling diverse reasoning paths and selecting the most consistent answer [7]. Additionally, LLMs can self-improve by generating and fine-tuning themselves with high-confidence, rationale-augmented answers [8]. The SELF-DISCOVER framework allows LLMs to self-compose reasoning structures using atomic modules [9], and the Self-Instruct framework enhances instruction-following capabilities through self-generated instructions [10]. In ranking and selecting model responses, the use of strong LLMs as judges to evaluate responses to open-ended questions has become one of the most popular options [11]. Building on this, using a Panel of LLM evaluators (PoLL) has been proposed to provide a more diverse and balanced evaluation [12]. The Llama Guard model introduces an LLM-based input-output safeguard for classifying and evaluating responses that can filter out undesirable ones [13]. Self-Refine introduces an iterative feedback mechanism where an LLM generates an initial output, provides feedback on its own output, and then refines itself based on this feedback [14]. The utility of LLM critics is demonstrated in the context of code and mathematics evaluation, where LLMs provide natural language feedback that highlights issues in code [15] or proofs [16].

3. Data and Methods

We focus on the candidate generation phase of our framework (as shown in Figure 1). This process utilizes a RAG approach, starting with the selection of examples from our database, which includes news articles about specific AI technologies or incidents, GitHub README files from AI-related repositories, and Hugging Face model and dataset cards. The next step involves retrieving relevant sections of legal and ethical provisions from our knowledge base, identified through similarity search. These retrieved provisions are then combined, and the language model is prompted to reason and explain how they impact the selected example using a zero-shot Chain of Thought (CoT) prompt [6].

A common limitation of many RAG pipelines is their disregard for the structural integrity of documents, often dividing them into uniform-length chunks. This can lead to critical oversight, especially

[5] https://gracenote.ai/
[6] https://casetext.com/cocounsel/
[7] https://www.curatedai.eu/

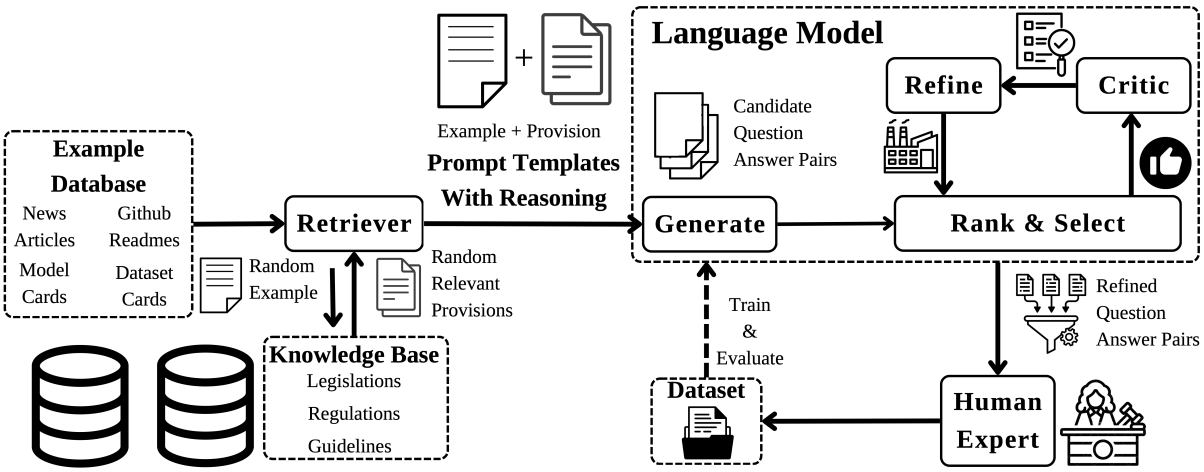

Figure 1: Our framework leveraging RAG and an LLM to generate, judge, criticize, and refine candidate examples.

when dealing with legal documents, which are typically organized into articles and paragraphs. We employ a systematic approach to structuring and querying legal documents for efficient retrieval and compliance analysis, as described in Figure 3. The legal document L is divided into its pre-defined articles and paragraphs as they are structured in the base document. Each paragraph is further segmented into overlapping passages of fixed length s with an overlap o to maintain context across segments. Each passage is then encoded using a dense retrieval embedding model. When querying, we embed the query and compute the dot product similarity between the embeddings of the query and the stored passages. We retrieve the top k passages with the highest scores. We then look up the articles to which these passages belong and generate a prompt using a predefined template and n of these articles. The prompt forms a question asking the LLM to analyze step-by-step [6] the implications of the provided legislative articles with respect to the query.

Algorithm 1 Legal Text Indexing and Retrieval Augmented Generation

Input: Legal document L, query Q, embedding model E, parameters p, o, k, t, n
Output: LLM-generated candidate responses based on Q

Indexing:
Split L into articles $\mathscr{A} = \{A_1, \ldots, A_x\}$
for each A_x in $\mathscr{A}$ **do**
    Split A_x into paragraphs $P_y = \{P_{x1}, \ldots, P_{xy}\}$
    for each P_{xy} in A_x **do**
        Partition P_{xy} into overlapping passages g_{xyz} of length p with overlap o
        Encode g_{xyz} using model E
    end for
end for
Retrieval Augmented Generation:
Encode query Q using model E
Compute similarity scores between the encoded Q and each encoded passage g_{xyz}
Retrieve top k passages $\{g_1, \ldots, g_k\}$ with a similarity score $\geq t$
Get the subset of articles $\mathscr{A}_u$ to which the passages $\{g_1, \ldots, g_k\}$ belong
for each subset of up to n articles in $\mathscr{A}_u$ **do**
    Construct prompt M_Q and obtain LLM response R_Q
end for

In our initial experiments, we used the EU AI Act as our legislative text, and with queries consisting of sentences reporting on AI-related incidents from the news, dataset and model cards, and open-source AI project README files. The retrieval model used was the small BGE [17] model [8] for dense retrieval, while the LLM was GPT-4 [18]. The parameters used were $s = 184$ and $o = 30$, $k = 10$, $t = 0.3$, determined heuristically. We've explored creating queries with both $n = 1$ and $n = k$, the choice influences how many articles are included in a single query. An example prompt template is shown in Listing 1.

Listing 1: Example Prompt for Legal Compliance Analysis

```
Consider the following articles of legislation, provided between triple backticks, and
nothing else:
```\{articles\}```
Under these articles and only these articles and ignoring those that are not applicable,
 as a legal compliance expert, answer: what are the implications of that legislation to
the following \{example type\}, provided between triple backticks:
```\{query\}```
Let's think step by step.
```

4. Conclusions and Future Work

In this work, we introduced the initial phase of a framework and tool designed to prepare datasets for training Large Language Models (LLMs) to perform compliance reasoning in AI applications. Our approach preserves the critical structure and content of legal provisions within a Retrieval-Augmented Generation (RAG) setting, ensuring more accurate and contextually aware reasoning.

Our proposed framework offers significant advantages for companies developing and deploying AI systems across different regulatory landscapes. By integrating a compliance assistant into the AI development process, companies can proactively ensure that their models and data pipelines comply with complex regulations, identify potential legal issues early in the development cycle, and streamline the process by reducing the need for extensive manual reviews by legal experts. As a result, companies can reduce compliance risks, accelerate time-to-market, and maintain high standards of ethical and legal accountability in their AI initiatives.

Looking ahead, our next steps will focus on the implementation of the refinement loop. Additionally, we plan to explore the tool's potential use by the public and policymakers to raise awareness and deepen understanding of AI technologies and the associated regulatory landscape.

5. Acknowledgments

This work was supported by the European Union through enrichMyData EU HORIZON-IA project under grant agreement No 101070284 and ELIAS HORIZON-RIA project under grant agreement No 101120237.

References

[1] S. Yue, W. Chen, S. Wang, B. Li, C. Shen, S. Liu, Y. Zhou, Y. Xiao, S. Yun, W. Lin, X. Huang, Z. Wei, Disc-lawllm: Fine-tuning large language models for intelligent legal services, 2023. arXiv:2309.11325.

[2] J. Cui, M. Ning, Z. Li, B. Chen, Y. Yan, H. Li, B. Ling, Y. Tian, L. Yuan, Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model, 2024. arXiv:2306.16092.

[8] https://huggingface.co/BAAI/bge-small-en-v1.5

[3] N. Guha, et al., Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023, pp. 44123–44279.

[4] D. Hendrycks, C. Burns, A. Chen, S. Ball, CUAD: an expert-annotated NLP dataset for legal contract review, in: J. Vanschoren, S. Yeung (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, 2021.

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022, pp. 24824–24837.

[6] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022, pp. 22199–22213.

[7] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, 2023.

[8] J. Huang, S. Gu, L. Hou, Y. Wu, X. Wang, H. Yu, J. Han, Large language models can self-improve, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 1051–1068. doi:10.18653/V1/2023.EMNLP-MAIN.67.

[9] P. Zhou, et al., Self-discover: Large language models self-compose reasoning structures, 2024. arXiv:2402.03620.

[10] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 13484–13508. doi:10.18653/V1/2023.ACL-LONG.754.

[11] L. Zheng, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023.

[12] P. Verga, S. Hofstätter, S. Althammer, Y. Su, A. Piktus, A. Arkhangorodsky, M. Xu, N. White, P. S. H. Lewis, Replacing judges with juries: Evaluating LLM generations with a panel of diverse models, 2024. arXiv:2404.18796.

[13] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, M. Khabsa, Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. arXiv:2312.06674.

[14] A. Madaan, et al., Self-refine: Iterative refinement with self-feedback, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023, pp. 46534–46594.

[15] N. McAleese, R. M. Pokorny, J. F. C. Uribe, E. Nitishinskaya, M. Trebacz, J. Leike, Llm critics help catch llm bugs, 2024. arXiv:2407.00215.

[16] B. Gao, et al., LLM critics help catch bugs in mathematics: Towards a better mathematical verifier with natural language feedback, 2024. arXiv:2406.14024.

[17] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, C-pack: Packaged resources to advance general chinese embedding, 2023. arXiv:2309.07597.

[18] OpenAI, GPT-4 technical report, 2024. arXiv:2303.08774.