Improving the accessibility of EU laws: the Chat-EUR-Lex project

Manola Cherubini[1,*,†], Francesco Romano[1,*,†], Andrea Bolioli[2,*,†], Lorenzo De Mattei[2,*,†], and Mattia Sangermano[2,*,†]

[1]Institute of Legal Informatics and Judicial Systems (IGSG-CNR), via dei Barucci 20, Florence, 50127, Italy
[2]Aptus.AI, Largo Padre Renzo Spadoni 1, 56126 Pisa, Italy

Abstract

In this article we describe the results of an ongoing research project on the use of Chat-Based Large Language Models (Chat LLMs) and Retrieval Augmented Generation (RAG) for the access to legal repositories. We are integrating Chat LLMs and RAG to access a dataset of legal acts in English and Italian (a subset of EUR-Lex collection), and interact through a chatbot. We present the state of the art, the objectives, the use cases, the methodology used in the project, and then we discuss the preliminary results.

Keywords

Legal Informatics, Large Language Models (LLMs), Retrieval Augmented Generation (RAG)

1. Introduction

In this article we describe the partial results of an ongoing research project on the use of Chat-Based Large Language Models (Chat LLMs) and Retrieval Augmented Generation (RAG) for the access to normative repositories.

In the project, we are integrating Chat LLMs and RAG to access a dataset of legal documents (European legal acts taken from EUR-Lex repository) and to allow the user to interact through a chatbot.

In the next sections, we will present the state of the art (2. Related works), the objectives and the methodology used (3. Chat-EUR-Lex methodology), the results of a research survey (4. Research survey), the system architecture (5. System architecture), and then we discuss the results presented in the previous sections (6. Discussion and conclusions).

2. Related works

As stated in [1] and many other sources, "Legal professionals rely on accurate and up-to-date information to make informed decisions, interpret laws, and provide legal counsel". The phenomenon of hallucination and nonsensical outputs of systems based on LLMs is obviously not acceptable in the legal context. To the best of our knowledge, the first survey on the challenges faced by LLMs in the legal domain was presented in [2], but mainly for Chinese language. While in other domains, such the financial one, a few LLMs have already been developed [3]. Large language models are also used in healthcare where LLMs are useful for processing and understanding medical text data, providing valuable insights, and supporting clinical decision-making [4].

LLMs are posing interesting challenges to those who are experimenting with these technologies in the legal field, where the "complexities of legal language, nuanced interpretations, and the ever-evolving nature of legislation present unique challenges that

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy

[*] Corresponding author.

[†] These authors contributed equally.

🖂 manola.cherubini@igsg.cnr.it; francesco.romano@igsg.cnr.it; andrea.bolioli@aptus.ai; lorenzo@aptus.ai; mattia.sangermano@aptus.ai

0000-0002-0242-6633 (Manola Cherubini); 0000-0001-5250-7733 (Francesco Romano); 0000-0003-1681-9435 (Andrea Bolioli)

require tailored solutions" [1]. There are many questions and fears about the actual use of these artificial intelligence tools, e.g., their opacity [5] and the possibility of hallucinations, but also "legal problems concerning intellectual property, data privacy, and bias and discrimination" [6]. For this reason, in the European Union it has been decided to regulate the use of artificial intelligence in specific sectors, but also to adopt a regulation that provides for a regulatory framework of reference only for high-risk AI systems [7]. Some experiments conducted on legal datasets show that LLMs can improve the performance of document page classification [8][9], the annotation of legal texts [10], the summarization of legal texts [11] [12], the legal rule classification [13], the legal statute identification from facts [14] and the mining of legal argument [15]. Other trials explore the ability of LLMs "to explain legal concepts from statutory provisions to legal professionals" [16] and to create "a register of obligations from various types of legislative and regulatory material" [17].

Other uses can also be mentioned, such as LLMs as legal tutors in the context of legal training [18] and in one of the most basic tasks required of lawyers, the so-called "statutory reasoning" [19]. Recently, legislative drafting experiments have been carried out with ChatGPT, particularly for "the comparison of legislation among jurisdictions and the synthesis of the best possible policy for the country based on this comparison" [20].

As it is known, generative AI models have been found to hallucinate, i.e., they can generate false or nonsensical statements [21] [22]., two strategies for reducing this problem are Fine-tuning and Retrieval-Augmented Generation (RAG) [23]. In both cases, we try to provide the LLM with the relevant information (according to a domain or a specific query). Fine-tuning involves additional training on a specific dataset, tailoring the model to certain tasks or domains [24] [25]. This improves accuracy but limits the model to knowledge up to the last fine-tuning. RAG merges a pre-trained model with a retrieval system, accessing current data for accurate responses on recent or specific topics. Its success hinges on the quality of retrieved information and requires maintaining a large, updated database. Both methods enhance model performance in specific areas, balancing current information and resource needs.

3. Chat-EUR-Lex metodology

In this section, we present the main problems faced and the methodologies used in the ongoing Chat-EUR-Lex project. Our objective is to create an AI-powered conversational interface that deals with complex legal texts (the regulations in English and Italian published in the EUR-Lex repository), can provide simplified explanations, and allows the user to conduct context-specific interaction. To present the methodologies used, we describe the activities performed:

- Legal and Ethics risk assessment. We performed a legal and ethics assessment. The prototype will be compliant to GDPR and EU AI Act, i.e., we will comply with the rules set by EU regulations.
- UX Research and Survey. We collected data and information from a sample of potential users through a questionnaire, both in Italian and in English. Objective of the survey: understand the needs of people using digital legal resources, and their level of satisfaction; identify users' needs and desires regarding chatbot interaction; know the fears related to the use of generative AI. User experience (UX) research involves studying how users interact with the current EUR-Lex system and identifying pain points and challenges.
- Data collection. Chat-EUR-Lex dataset comprises a selection of in force legal acts in English and Italian sourced from EUR-Lex, covering the period from January 1, 2014, to December 31, 2023. Specifically, it includes all historical texts preserved in Celex 3 sector that remain unaltered over time, along with the most recent consolidated versions in Celex 0 sector for acts that have undergone amendments. Corrigenda are omitted from this dataset. Additionally, the EUR-Lex documents that are not provided with XML or HTML data are excluded from the selection. Number of documents in English: 19062; documents in Italian: 18164.
- Semantic search engine setup. Semantic search must allow users to find relevant legal information even if they don't use precise legal terminology. This involves using Natural Language Processing (NLP) techniques, particularly neural embedding, such as the one introduced by (Lai et al. 2023).
- RAG-based Chat system development. RAG combines retrieval-based methods with generative language models to provide accurate and contextually relevant responses to user queries. The user can read

both the generated answer and the relevant sources, i.e. the portions of regulations used to generate the answer.

- First version release (June 2024). The first version of the prototype is released to a selected group of users. This version should provide basic functionality and serve as a starting point for further improvements.
- Feedback collection and tuning. User feedback is actively collected and analysed. This feedback is used to identify areas for improvement and fine-tune both the chat system and the user interface. This iterative process continues to enhance the system's effectiveness and user satisfaction.

4. Research survey

In this section we present the results of the questionnaire distributed from December 28, 2023, to March 31, 2024, aimed at legal professionals, law researchers, public officials in the legal sector, compliance specialists, and other people interested in the use of Generative AI in the legal domain, in Italy and other European countries. The objectives of the questionnaire were to understand the needs of people using digital legal resources (EUR-Lex in particular) and their level of satisfaction; identify users' needs and desires regarding chatbot interaction; know the fears related to the use of generative AI. The questionnaire was anonymous; the languages used were Italian and English. We distributed it online on websites and with targeted e-mail activity.

The questionnaire contained 22 questions: 4 questions for demographic information (age, gender, education, profession); 9 multiple choice questions; 6 open-ended questions; 2 yes/no questions, and 1 rating question. Regarding the topic of the use of LLMs for accessing European laws, the most important questions are: “7) To search for legal documents, regulations and rulings, do you mainly use the EUR-Lex search engine, or do you use Google Search or something else?”. “17) In the legal domain, could a generative AI chatbot help search and interaction?”. “18) What kind of requests would you make to the chatbot? Write one or more example requests.”. “19) Do you know what generative AI is and/or are you a user of generative AI tools?”. “21) Do you have any concerns about the use of generative AI in the legal field?”.

The following table (Table 1) presents the numbers of Submissions, number of people that did not complete the questionnaire (Starts), number of people that viewed the questionnaire (Views) in Italian and English, as of March 30, 2024.

Table 1
Questionnaire results in Italy (language: Italian), other EU countries (language: English), and total results

LANG	VIEWS	STARTS	SUBMISSIONS
Italian	769	315	192
English	530	184	105
Total	1299	499	297

We report here some statistics on the Italian responses: 54% of the respondents are legal experts (law researchers, jurists, lawyers, compliance specialists, etc.), while 46% are not legal experts. 66% say that they consulted the EUR-Lex repository at least once. When asked which tool they mainly use to search for legal documents and regulations, 48% answer mainly Google search, 37% mainly EUR-Lex search engine, 15% mainly other tools (we do not report here the answers on the other legal sources). 60.4% say that a generative AI chatbot could help search and interaction, 33.3% don’t know, 6.2% say No. The question “Do you know what generative AI is and/or are you a user of generative AI tools?” is answered: 51% “Little”, 27% say “Yes, I use them regularly”, 22% “Not at all” (remember that these percentages concern responses in Italy). Finally, 87% think generative AI must be regulated, 8% don’t know, 5% answer No. For reasons of space, we do not report here the answers to the question “What kind of requests would you make to the chatbot?”.

In summary, these responses allow us to assess the level of knowledge of legal experts in generative artificial intelligence, to see if there are differences between legal experts and other people, to know their fears on these issues, and, above all, to collect the needs and requirements of potential users of the chatbot.

A detailed report containing the complete questionnaire, the aggregate results and a detailed analysis will be published in May 2024 on the GitHub project repository.

5. System architecture

The pipeline of Chat-EUR-Lex prototype is divided into main parts:

- An asynchronous batched pipeline which collects and indexes the documents from EUR-Lex into a search engine.
- A synchronous pipeline that gets the users' queries, retrieves relevant contextual information and provides a response to the users.

The asynchronous batched pipeline comprises three main components:

1. A crawler that collects the data from EUR-Lex.
2. A chunker who chunks the documents into smaller segments.
3. An embedding model that transforms the segments into dense vectors to be indexed in the vector DB.

The synchronous pipeline comprises two main components:

- A retriever that transforms the query into a vector using the same embedding models used by the asynchronous pipeline and looks into the vector DB for similar contents.
- An LLM that gets both the query and the context inserted in a prompt template and produces a response to be provided to the users

Each time the user does a new query, the whole chat history is passed to the LLM until the maximum prompt length is reached; in that case, older chat parts are truncated.

This process involves several parameters to be selected, such as:

- Chunking techniques and size.
- Embedding models.
- K-nearest neighbor search techniques.
- Prompt templates.
- LLM and its parameters.

In summary, the RAG approach is a blend of two key components: a retrieval system and a generator. The retrieval system scans through a database of documents to fetch the most relevant ones in response to a user query. The most recent solutions for retrieval systems employed in RAGs rely on semantic search utilizing embeddings.

The generator, on the other hand, uses these retrieved documents to generate a well-informed answer. This process ensures that the system provides responses that are both informative and contextually accurate. In the current project setup (April 2024), the gpt-4 model powers the generation of responses. For the creation of embeddings, we utilize text-embedding-ada-002 (https://platform.openai.com/docs/models/embeddings).

While our dataset contains about 37000 legal acts, the need for partitioning these laws for a granular retrieval process amplifies the total count of retrievable documents into about 371000 texts ("chunks"). This extensive partitioning provides a more detailed context for the RAG system, allowing for more accurate answer generation. On the other side the increased number of documents naturally presents a challenge for our retrieval process.

6. Discussion and conclusions

In this project, we are trying different combinations of the mentioned parameters using both open-source and closed-source models to investigate the readiness of LLMs to build a system for legislative research.

We are performing two evaluation steps to compare different models and parameters:

1. Search engine evaluation: we are comparing different Embedding models, chunking strategies and k-nearest neighbors search techniques to select the best combinations to retrieve good-quality results.
2. Response generator evaluation: having fixed the best combination for contextual information retrieval thanks to step 1 evaluation, we will compare the quality of the generated response using different prompt templates, LLMs and LLMs parameters.

For step 1 evaluation, we are creating a gold dataset using expert annotators and use standard search engine evaluation metrics such as the Mean Reciprocal Rank. For step 2, evaluation of different settings will be proposed to experts who will ask the same questions and attribute scores to each response. Preliminary results have shown that when the context provided to gpt-4 by the retrieval system is consistent with the question asked, the generated answer is concise, comprehensible, and accurate. This consistency significantly minimizes the problem of hallucination, wherein the model might generate false or nonsensical information.

The large number of chunks that can contribute to the generation of the answer naturally presents a challenge for our retrieval process. In this context, we are actively exploring strategies to improve the efficiency of this crucial component. One of the promising directions we are considering involves leveraging not only the semantic content of the normative sources but also the boundary information, such as metadata.

The inclusion of metadata in our retrieval process could potentially imbue our system with the ability to hone in on the most relevant documents, thereby optimizing the retrieval process and improving the overall performance of the chat system. On the other hand, the utilization of a specific embedding model built on legal data could be beneficial, as opposed to a generic embedder. Indeed, this model could provide a more nuanced understanding of the legal texts, thus enhancing the retrieval process.

The evaluation of the preliminary results is promising: on simple tasks, i.e. simple queries, the system perform on par with human experts, as attested in similar researches [26].

We must highlight the need for legal experts to make qualitative assessments, as well as quantitative evaluations, both due to the complexity of the legal domain and because the evaluation can be done in different ways depending on the use case, aim and user target. It does not seem possible to create a unique dataset to evaluate the quality of the results in an automatic way for QA tasks in the legal domain. Multiple points of view may be equally valid, and a unique 'ground truth' may not exist, as discussed for other tasks in Perspectivist Approaches to NLP [27]

A general problem for experiments with LLMs is "non-repeatability": the experiments are not exactly reproducible because the systems are not deterministic; moreover, in proprietary commercial systems, we do not know the LLM's parameters and the dataset used for training; new versions and new LLMs are released quickly.

Here are some additional critical issues that we are addressing:

- text length limitations: currently, there's a limit on the length of text the LLM can handle effectively. This necessitates breaking down longer texts, which can be cumbersome;
- impact of short queries: the chatbot's accuracy and precision suffer when responding to very short or poorly defined queries. More detailed user queries lead to better results;
- single vs. multiple documents: the chatbot performs best when responding to queries that target information from a single document, rather than synthesizing information from multiple sources.

Acknowledgements

Chat-EUR-Lex project is funded within the framework of the NGI Search project under grant agreement No 101069364. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] J. Cui, Z. Li, Y. Yan, B. Chen, L. Yuan, Chatlaw: Open-source legal large language model with integrated external knowledge bases, arXiv preprint arXiv:2306.16092, (2023).

[2] J. Lai, W. Gan, J. Wu, Z. Qi, P.S. Yu, Large Language Models in Law: A Survey. arXiv preprint arXiv:2312.03718. (2023).

[3] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, G. Mann, BloombergGPT: A Large Language Model for Finance, arXiv preprint arXiv:2303.17564v2 (2023).

[4] S. Reddy, Evaluating large language models for use in healthcare: A framework for translational value assessment, in volume 41 of Informatics in Medicine Unlocked, (2023). https://doi.org/10.1016/j.imu.2023.101304

[5] A. Contaldo, F. Campara, Intelligenza artificiale e Diritto. Dai sistemi esperti "classici" ai sistemi esperti "evoluti": tecnologia e implementazione giuridica, in: G. Taddei Elmi, A. Contaldo (Eds.), Intelligenza artificiale. Algoritmi giuridici. Ius condendum o "fantadiritto", Pacini editore, Pisa, 2020, p. 24.

[6] Z. Sun, A short survey of viewing large language models in legal aspect, arXiv preprint arXiv:2303.09136 (2023).

[7] G. Finocchiaro, Artificial intelligence. What are the rules? Il Mulino, Bologna, 2024.

[8] P. Fragkogiannis, M. Forster, G.E. Lee, D. Zhang, Context-Aware Classification of Legal Document Pages. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), Association for Computing Machinery, NY, pp 3285-3289, 2023, doi: 10.1145/3539618.3591839

[9] D. Trautmann, Large Language Model Prompt Chaining for Long Legal Document Classification, arXiv preprint arXiv:2308.04138.(2023).

[10] J. Savelka, K.D. Ashley, The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts, Frontiers Artificial Intelligence, 6: 1-14, 2023, doi: 10.3389/frai.2023.1279794.

[11] M. Cherubini, F. Romano, A. Bolioli, N. De Francesco, I. Benedetto, The summarization of legal texts: an experiment with GPT-3. Rivista italiana di informatica e diritto, 5, 1: 191-204 (2023) https://doi.org/10.32091/RIID0103

[12] D. Datta, S. Soni, R. Mukherjee, S. Ghosh, MILDSum: A Novel Benchmark Dataset for Multilingual Summarization of Indian Legal Case Judgments, arXiv preprint arXiv:2310.18600v1, (2023) https://doi.org/10.48550/arXiv.2310.18600

[13] D. Liga, L. Robaldo, Fine-tuning GPT-3 for legal rule classification, in volume 51 of Computer Law & Security Review, (2023) doi: 10.1016/j.clsr.2023.105864.

[14] S. Paul, A. Mandal, P. Goyal, S. Ghosh, Pre-trained language models for the legal domain: a case study on Indian law. In: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, (2023), pp 187-196.

[15] A. Al Zubaer, M. Granitzer, J. Mitrović, Performance analysis of large language models in the domain of legal argument mining", Frontiers Artificial Intelligence, 6, (2023), doi: 10.3389/frai.2023.1278796.

[16] J. Savelka, K.D. Ashley, M.A. Gray, H. Westermann, H. Xu, Explaining Legal Concepts with Augmented Large Language Models (GPT-4), (2023) arXiv preprint arXiv:2306.09525v2.

[17] J. Ioannidis, J. Harper, M.S. Quah, D. Hunter, Gracenote.ai: Legal Generative AI for Regulatory Compliance. In: Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023) co-located with (ICAIL 2023), CEUR-WS.org, Elsevier, pp. 20-31.

[18] D. Charlotin, Large Language Models and the Future of Law, SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4548258, 2023, Accessed 04 april, 2024

[19] A. Blair-Stanek, N. Holzenberger, B. Van Durme, Can GPT-3 Perform Statutory Reasoning? In: The Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), ACM, NY, USA, pp. 22-31, 2023, doi: 10.1145/3594536.3595163.

[20] G. Hill, The emerging artificial intelligence (AI) and national uniform legislation, in volume 97.5 of Australian Law Journal, (2023), pp. 303-306.

[21] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232 (2023).

[22] S. M. Tonmoy, S.M. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, A. Das, A comprehensive survey of hallucination mitigation techniques in large language models. arXiv preprint arXiv:2401.01313 (2024).

[23] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goya, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances, in volume 33 of Neural Information Processing Systems (2020), pp. 9459-9474.

[24] L. Xu, H. Xie, S.Z.J. Qin, X. Tao, F.L. Wang, Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. arXiv preprint arXiv:2312.12148 (2023).

[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021).

[26] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, T.B. Hashimoto, Benchmarking large language models for news summarization, in volume 12 of Transactions of the Association for Computational Linguistics, (2024), pp. 39-57

[27] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, No. 6, 2023, pp. 6860-6868.

A. Online Resources

The GitHub project repository can be consulted at https://github.com/Aptus-AI/chat-eur-lex. The dataset has been published on https://huggingface.co/datasets/AptusAI/chat-eur-lex .