A Large Language Model agent based legal assistant for governance applications

Mamalis, Marios Evangelos*
marios.mamalis@uom.edu.gr

Kalampokis, Evangelos*
ekal@uom.edu.gr

Fitsilis, Fotios†
fitsilisf@parliament.gr

Theodorakopoulos, Georgios‡
g.theodorakopoulos@nsk.gr

Tarabanis, Konstantinos*
kat@uom.edu.gr

This is a pre-print version of the following paper: Mamalis, M.E., Kalampokis, E., Fitsilis, F., Theodorakopoulos, G., Tarabanis, K. (2024). A Large Language Model Agent Based Legal Assistant for Governance Applications. In: Janssen, M., et al. Electronic Government. EGOV 2024. Lecture Notes in Computer Science, vol 14841. Springer, Cham. https://doi.org/10.1007/978-3-031-70274-7_18

Abstract

Large Language Models (LLMs) have gained significant traction, primarily due to their potential disruptive influence across industries reliant on natural language processing. Governance stands out as one such sector. Notably, there has been a surge in research activity surrounding the implications of LLMs in deciphering complex legal corpora. This research offers substantial assistance to various stakeholders, including decision-makers, administrators, and citizens. This article focuses on the design and implementation of an LLM-based legal assistant tailored for interacting with legal resources. To achieve this, a real-world scenario has been chosen, incorporating models GPT3.5 and GPT4 as the LLMs, a well-defined legal corpus comprising European Union (EU) legislation and case law concerning the General Data Protection Regulation (GDPR), alongside a series of reference legal queries of varying complexity. Retrieval Augmented Generation (RAG) as well as agent methodologies are employed to seamlessly integrate the LLMs' functionalities with the customized dataset. The results appear to be promising, as the system managed to correctly address the majority of the legal queries, though with variable precision. Expectantly, the complexity of the queries severely impacted the quality of the outcome.

Keywords: Artificial intelligence, legal assistant, large language model, GDPR, policy making, law making, public administration

1 Introduction

As AI continues to advance rapidly, its accompanying tools and services are reaching greater levels of maturity, progressively infiltrating the public sector [30]. Among the array of AI applications are sophisticated tools, for instance, to facilitate law-making, enable oversight mechanisms and enhance political discourse. These are aspects of governance that fall under the EU framework of better regulation [18] that is gradually adopted by individual member states [6]. AI-based technological innovations have the capacity to transform governance institutions fundamentally, reshaping the way public services are generated and delivered [13]. A relevant research and development agenda for the introduction of such tools in the parliamentary workspace has been already outlined and may partially be used for the elaboration of an AI roadmap for governance applications [19]. Accordingly, several of the political, administrative or scientific tasks necessary to modernize governance can be linked to natural language processing (NLP) related technologies. Specifically, over the last few years, large

*Information Systems Laboratory, University of Macedonia, Thessaloniki, Greece
†Department for Scientific Documentation and Supervision, Hellenic Parliament, Athens, Greece
‡Hellenic State Legal Council, Athens, Greece

language models have revolutionised the NLP landscape, having been established as the premier approach for a plethora of related tasks, including chatbots and virtual assistants.

The subset of machine learning models specialized in tasks that concern natural language are referred to as language models. Ever since the inception of the revolutionary attention mechanism [31], transformer based language models have proved to be efficient in understanding and generating natural language, scaling well with the volume of their training data. This breakthrough has lead to the creation of large language models (LLMs), that consistently surpass state of the art performance metrics, and showcase emergent capabilities [33]. In operational contexts such as these, the potential applications of LLMs are readily apparent. Already, numerous use cases and relevant studies emanating from both executive and legislative branches are evident, and public sector bodies are already beginning to deploy generative LLMs to enhance their services; one such case being that of the Greek Government deploying "mAIgov", a chatbot powered by OpenAI's generative LLMs, that has been trained using open data sourced from various public entity websites.

Governance operates within its distinct linguistic domain, primarily characterized by legal language. Given its specialized nature, legal language demands dedicated resources for comprehensive analysis and application. Such resources encompass legal corpora, lexical databases, guidelines governing grammar and style, as well as exhaustive references for acronyms, organizations, and abbreviations. In the EU context, legal resources can be openly accessed via the Publications Office of the EU, which runs the EUR-Lex platform.

The research objective of the article is to explore and evaluate the integration of LLM-based legal assistants in governance applications, aiming to assess their efficacy and potential impact on governance practices. The experimental design includes the utilization of two, state of the art, LLMs (GPT3.5 and GPT4). Employing RAG-based and agent-based methodologies, the LLMs are integrated with a tailored legal corpus focused on EU legislation and case law regarding the General Data Protection Regulation (GDPR). A predefined set of legal queries is posed to the system, and the ensuing results are analysed and discussed.

The article is structured in eight sections. In the introduction (Section 1), the importance of LLMs in governance and the research objective are outlined. The background section (Section 2) provides a thorough examination of foundational concepts such as LLMs, prompt engineering, agent systems, and the RAG methodology, offering a robust theoretical framework for the underlying research. The related work section (Section 3), highlights recent research in the field. The Research approach (Section 4) presents the methodology employed to assess the use of LLMs in governance applications, which is followed by the user requirements (Section 5). The Proof of Concept Implementation (Section 6) details the practical implementation of the LLM system. Ultimately, the results are discussed (Section 7) and summarized (Section 8), while providing key findings, contributions to the field, and directions for future research.

2 Background

2.1 Large language models

Large language models are pre-trained language models comprising hundreds of millions to billions of parameters. They are trained on massive datasets for extended periods of time, acquiring in the process knowledge about language structure, semantics and facts. Their training requires vast amounts of computational resources, but results in state of the art performance in understanding and generating natural language. In the field of NLP, since the invention of the transformer architecture [31] and in recent years, many generative large language models have been created. Amongst these models are the GPT-3.5 and GPT-4 [1] models released by OpenAI, the Llama [28] and Llama2 [29] models released by Meta and Mistral AI's Mistral [11] and Mixtral 8x7B [12]. These foundation LLMs, nowadays widely available, either through application programming interfaces (APIs) or locally hosted solutions, offer new ways to manipulate natural language textual information more quickly and efficiently than humans.

An unexpected byproduct of scaling up LLMs has been a qualitative change in their behavior concerning their abilities. This phenomenon is referred to as "emergence". As stated in [33], "An ability is emergent if it is not present in smaller models but is present in larger models". By scaling up the volume of the models' parameters, the amount of computation performed, or the volume of the data used during training, new functionalities have been observed. Such abilities can range from instruction following to advanced reasoning. The phenomenon of emergence has been detrimental in the adoption of LLMs for complicated tasks that exceed the capacities of smaller models.

Reasoning in generative large language models pertains to their ability to form thought patterns resembling human thought processes. They have been shown to display strong capacity for abstract pattern induction in analogical tasks [32], and can do so in zero-shot scenarios effectively [15]. Reasoning plays a crucial part in planning and following instructions towards a specific and predetermined goal. These instructions can range from simple directions (zero-shot), to example based learning (one-shot and few-shot learning), and result in systems adept at several downstream tasks. However, the level to which these emergent abilities are utilized, largely relies in the way that the model's input is formulated.

2.2 Prompt engineering

Prompt engineering is the process of designing natural language inputs that upon inference will elicit desired responses from the generative LLM. Prompt engineering does not alter the model's weights, but rather relies on the identification of the input that gives the maximum activation of the network towards a specific output.

Individual techniques of prompt engineering include chain of thought reasoning [34], tree of thought prompting [37] and in context learning [25, 2]. These techniques enhance the reasoning capabilities of the LLMs by providing examples of the correct flow of thought that the model should follow, something that allows for more effective access and utilization of its already existing knowledge.

It is of critical importance to note that prompt engineering is not exclusively a process that involves human prompters. Given the highly advanced capacity of LLMs in generating natural language text, it has been shown that even LLM-generated prompts could, under circumstances, rival those of human origin [40]. This further reinforces the belief that self-prompting systems are possible, unlocking in the process various applications that require a higher degree of autonomy.

2.3 Agents

The concept of agents revolves around the notion of self adjusting computational systems performing operations [35, 7]. Harnessing the aforementioned emergent capabilities of LLMs, mainly reasoning through LLM-generated prompt engineering, has shown to lead to advancements in autonomous, or semi-autonomous agents, that can utilize tools and adapt to diverse scenarios more adequately [36].

These agents generate plans towards the actualization of a specific goal, execute the individual steps of the plan with usage of additional tools and capabilities, and have advanced self-correcting capacities enabled by memory modules allowing for flexible and efficient operation.

Autonomous and semi-autonomous agent-based systems have already been used across various domains, such as medicine [14] and the public sector [20], showing great promise in the future.

2.4 Retrieval augmented generation

Even though LLMs hold considerable amounts of world knowledge, thanks to their initial training [24], they still lack the ability to be factually correct in all their responses, which often results in "hallucinations" [39]: factually incorrect responses presented by the LLM as correct ones. In order to tackle this weakness, several methods have been developed [27], mainly utilizing prompt engineering, with perhaps the most prevalent amongst those methods being the supplementation of factually correct context along with the input.

The main proposed architecture supporting the aforementioned method is retrieval augmented generation (RAG) [17]. In RAG, specialized components retrieve information relevant to the original input from corpora and supply it to the model along with it. This has been shown to boost the factual capacities of the models significantly, grounding its responses.

3 Related Work

The importance of LLMs is becoming visible when observing the regulatory activity of major industrial powers. While discussing the regulatory impact on LLMs falls outside of the scope of this study, it is worth mentioning that the EU AI Act that was adopted in March 2024 constitutes an important milestone that will potentially shape the development, training and operation of foundational models within -and maybe beyond- the European space [10]. In the US, an executive order (EO) from October 2023 attempts to harness the potential benefits of AI while also mitigating the associated risks. Generative AI, as represented via LLMs, is encompassed by this EO [23].

Currently, there are several research groups and companies that specialise in legal data using LLMs. Hence, an array of fine-tuned open and close source LLMs is already available to handle and analyze legal corpora [16]. Moreover, dedicated LLMs for the legal domain already started to be developed, signaling the importance of this technology for legal applications [3]. The present study underscores the practical utility and transformative possibilities of such tools in real-world scenarios in all three branches of governance: executive, legislative and judicial. For each of the three branches, a series of indicative yet notable examples are presented below.

In January 2024, the UK government issued a framework for developing generative AI tools and services for governmental organizations [21]. Further analysis suggests that at least 35 to 45 UK government agencies were either piloting or planning to implement generative AI use cases [22]. In Greece, the Hellenic State Legal Council is currently modernizing its IT system, with the perspective of utilizing LLMs for intelligent search for finding and handling related legal cases. Similarly, in Greece, the government has implemented an LLM-based chatbot named "mAigov" (currently in beta) on the Gov.gr platform. This initiative aims to streamline and accelerate citizens' daily interactions with the state.

In the legislative realm, the first intra-parliamentary use of an LLM was recorded in 2021 in the Finnish Eduskunta, when the Committee for the Future interacted with a GPT-3 model on matters related to the UN Agenda 2030 and the opportunities and threats presented by AI technology [5]. In 2023, the US Congress acquired 40 ChatGPT Plus licenses to explore generative AI within its ranks. These licenses were distributed among congressional offices, allowing lawmakers and staff to experiment with the technology internally [9]. At the sub-national level, the State parliament of Berlin operates an open-source AI assistant Germany called Parla[1]. Parla is based on ChatGPT and searches through the parliamentary database, identifies the most relevant documents for the inquiry, and summarizes relevant content into a response. Overall, a rather limited use of LLMs is currently observed in parliamentary environments, a situation that is expected to shift rapidly. Hence, the need rises for a specialized regulatory framework for representative institutions [6].

Though the Judiciary seems to be a preferential site for LLM-based use cases [16], not a single actual LLM use case in the judiciary, i.e. used by its administration and/or the judges, is known to the researchers, a testimony of utmost institutional cautiousness. Nonetheless, LLMs have also found their way into the judicial system - not always for a good reason. In 2023, a lawyer used made up judicial decisions, quotes and citations before court after using ChatGPT, thus opening the discussion about hallucinations in LLMs [4]. In this regard, the Alberta Courts acknowledged that lawyers need to have a human in the loop when utilizing LLMs for support [8].

4 Research approach

This study is designed as a single exploratory case study to examine how LLM-based systems can be used as legal assistants to support policy and decision makers as well as administrators. A single exploratory case study is a useful design to gain insights about phenomena that are thus far understudied or not explored at all, and to construct a new theory or generate propositions about their understanding [38]. In doing so, the research team involved a group of legal professionals, policy making, and technical experts, whose expertise covers every facet of the experimental design, from its inception to its implementation and, finally, its evaluation.

A domain specific legal topic was selected, with relevance to all branches of governance. For this, Regulation (EU) 2016/679 entitled General Data Protection Regulation (GDPR) was considered to be a contemporary and relatively mature legal topic fit for the purpose of this research. The GDPR dictates the responsible handling of personal data by organizations and companies. It is applicable to all EU member states, establishing uniform regulations to safeguard the rights and privacy of both businesses and citizens. Some of its essential contents stipulate transparent data utilization, lawful processing, protection of individual rights, and the obligation to report data breaches promptly [26].

Legal professionals defined a reference set of twelve legal questions related to GDPR, that are considered to be of high value in public administration or policy/law making. The questions were of varying levels of difficulty and were organized accordingly into three groups of four questions each.

A well-defined corpus with legal documents, directly referring to the GDPR was extracted from EUR-Lex[2]. The EUR-Lex repository is a comprehensive database that provides access to EU law. It includes EU treaties, legislative acts, international agreements, and preparatory documents. Additionally, EUR-Lex hosts case law from EU courts, national court decisions related to EU law matters, and the JURE database compiling cases on judicial cooperation. This valuable resource serves legal professionals, researchers, and anyone seeking information on EU law. Using the expert search feature of Eur-Lex, the search for legal texts was limited to the terms "General Data Protection Regulation" and "GDPR", from the period spanning from 1.1.2015 to 15.2.2024. Only texts in English were extracted.

A proof of concept implementation was then created. It utilized the generative large language models GPT3.5 and GPT4, as well as the current state of the art embedding model (text-embedding-3-large) offered by OpenAI. A combination of prompt engineering, retrieval augmented generation architectures and agent-based systems was explored in order to create a system capable of harnessing the advanced reasoning and generative capacities of the LLMs to effectively answer the questions created by the experts.

As an initial evaluation option, a qualitative approach was considered. The set of reference legal questions was posed to the system one by one. The researcher group then engaged in qualitative discussions of the outcome from a legal perspective to evaluate the quality of the responses. This internal approach allows for high flexibility and rapid development across multiple iteration cycles.

5 Reference Legal Questions

A set of 12 questions related to GDPR was defined by legal professionals that have expertise in public administration and policy making. The questions were categorized into three levels of progressive difficulty, namely

[1]https://www.parla.berlin
[2]https://eur-lex.europa.eu

beginner, intermediate, and expert. Each level comprises four questions. It is important to note that the difficulty levels were assessed by human legal experts and do not necessarily reflect the system's understanding of the questions or the quality of its responses. Furthermore, even though legal experts were not expected to have fully delved in the technicalities of an LLM, the language used was clear and unambiguous, avoiding excessive legal jargon and context sensitivity. This ensured that the LLM could perform at its full potential producing useful and accurate responses.

Beginner questions (Q1–Q4) aim at extracting basic information already stated in the pertinent legal framework of GDPR. It is expected that the system generates responses that entail parts or entire provisions, eventually with the respective legal reference (Article/paragraph/etc.). There is no progressive difficulty in this set of questions.

1. What are the primary objective and the EU regulatory framework of the General Data Protection Regulation (GDPR)?
2. How does the GDPR define "consent" in the context of data processing, and what are the requirements for obtaining valid consent?
3. Describe the role, responsibilities, and qualifications of a Data Protection Officer (DPO) under the GDPR.
4. Specify the framework for the protection of childs' rights under GDPR and provide the relevant provisions.

The intermediate set of questions (Q5–Q8)is more demanding and might require multiple levels of processing. There is a progressive difficulty (for a human being) in producing the requested results. The first two questions, Q5 & Q6 ask for numerical results. In the case of Q7, a further analysis is required for identifying court rulings. Similarly, Q8 asks for the identification of preliminary rulings, before classifying them further.

5. How many requests for preliminary rulings have been submitted to the Court during the period of reference?
6. For the requests for preliminary rulings, for which a court ruling has been issued, calculate the amount of time in days between date of submission and date of issuance, and present the average time for Court response.
7. For the entire number of preliminary rulings to the EU Court, identify the cases where a court ruling has been issued.
8. For the period from 2015-2024, identify and classify the requests for preliminary rulings to the EU Court using the following criteria: a. member state courts, b. number of requests per year.

Q9 & Q10 ask more a more advanced processing focusing on the broader context of the legal mater. Q11 goes a step further as asks the system to combine information for identifying potential risks based on the similarity and the number of requests for preliminary rulings. Q12 is a "tricky question" provided that the requested relations is not explicitly mentioned in the GDPR but constitutes basic EU procedural law, according to which whenever national courts have doubts as regards the interpretation of EU legislation or that of domestic legislation as regards its consistency with EU legislation, they may submit question to the EU Court of Justice, which subsequently provides mandatory interpretation for all member states.

9. For the requests for preliminary rulings submitted to the Court by Member State courts, identify overlapping areas as regards the interpretation of specific articles of GDPR, and present the most frequently addressed articles. Classify the findings by article GDPR.
10. From the requests for preliminary rulings submitted to the Court by Member State courts, derive the key principles that govern the processing of personal data.
11. Taking into account the GDPR and the submitted requests for preliminary rulings, identify potential risks factors regarding data protection rights.
12. What is the relation between the EU Court of Justice and the national courts as regards the interpretation of GDPR provisions.

6 Proof of concept implementation

The initial step of the implementation of the proof of concept system pertained to the collection of the data. Upon execution of the expert query, 428 documents were returned, whose metadata were downloaded though the platform, while the individual documents were downloaded, as HTML files, based on their unique CELEX number provided in the metadata file. The metadata that were included were: Title, CELEX number, ELI, Form, EUROVOC descriptor, ECLI identifier, Subject matter, Case-Law directory code (after Lisbon), Type of procedure, Collection of the document, Date of document, Date of publication, Author, Date of effect, Applicant/Appellant, Defendant/Other parties to the proceedings, Number of pages, and Publication Reference.

According to the EUR-Lex classification, this reference legal corpus included EU law and case-law, i.e. EU consolidated legal texts (157), EU case-law (267), as well as national case-law (4). Classified by the type of act, the corpus contained: consolidated legal texts (157), judicial information (139), Opinions of the Advocate General (64), judgments (57), orders (6), decisions by national courts in the field of European Union law (4), and information (1).

In order to answer the aforementioned twelve questions that were posed, two distinct systems were developed. The difference in the approaches used was due to the nature of the initial queries. The questions that belonged in the beginner group (Q1-Q4) required information that could be found only in the contents of the documents in order to be answered effectively. However, questions in the Intermediate and Expert question groups (Q5-Q12) did not only require knowledge extracted from the contents of the documents but also information that pertained to the document type, title, and dates associated with it. Both systems were implemented in Python, partly through the Langchain library, while the LLMs used were models gpt-3.5-turbo-0125 and gpt-4-0125-preview of OpenAI, featuring a context token window size of 16,385 and 128,000 tokens respectively, as well as the text-embedding-3-large embedding model that had an output dimension of 3,072.

6.1 RAG-based system

Beginner questions were deemed to be able to be answered just through examination of the contents of the General Data Protection Regulation (CELEX number: 32016R0679), thus the RAG system only considered this document for the purpose. The document was split in 88 continuous chunks were each chunk corresponded to approximately 2000 tokens, with a 200 token size overlap between chunks, applied to ensure seamless continuation of meaning. Token count was estimated for model gpt-3.5-turbo-0125.

The RAG-based system was comprised of four components. The embedding model, the vector database, the retriever and the generative LLM. For each chunk of the original document, embeddings were calculated using the text-embedding-3-large model. The vector representations of the chunks were stored in a ChromaDB vectorstore. Upon query execution, the query embeddings - created in the same manner - and the stored chunk embeddings were used to retrieve the four most similar chunks to the query based on the cosine similarity of their embedding vectors. All four chunks along with the original query were subsequently passed to the LLM, model gpt-3.5-turbo-0125, in order to retrieve the answer. The prompt was formulated as follows:

"Use the following pieces of context to answer the question at the end.
If you don't know the answer, just say that you don't know, don't try to make up an answer.
Context: {context}
Question: {question}
Answer:"

Where {context} and {question} were populated with the retrieved content and the query respectively. The prompt choice ensured that the model would refrain from hallucinating, and that the answers would be concise and accurate based on the provided context.

The answers generated are quoted below and their numbers correspond to the questions presented in Section 5:

1. The primary objective of the General Data Protection Regulation (GDPR) is to protect the fundamental rights and freedoms of natural persons with regard to the processing of personal data. The EU regulatory framework of the GDPR lays down rules relating to the protection of personal data and the free movement of such data within the Union.
2. The GDPR defines "consent" as a clear indication of the data subject's agreement to the processing of their personal data. In order to obtain valid consent, the controller must be able to demonstrate that the data subject has given consent, the request for consent must be presented in a distinguishable manner, the data subject must have the right to withdraw consent at any time, and consent must be freely given without any imbalance of power between the data subject and the controller. Additionally, the data subject must be informed of the identity of the controller and the purposes of the processing for which the personal data are intended. Consent should not be considered freely given if there is a clear imbalance between the data subject and the controller, or if the performance of a contract is dependent on consent.
3. The Data Protection Officer (DPO) is responsible for informing and advising the controller or processor and employees on their obligations under the GDPR and other data protection laws. They monitor compliance with the GDPR, provide advice on data protection impact assessments, cooperate with the supervisory authority, act as a contact point for the supervisory authority, and ensure that data subjects can contact them regarding their personal data. The DPO must have expert knowledge of data protection

laws and practices, be easily accessible, and have the ability to fulfill their tasks effectively. They must also be involved in all issues related to data protection, have access to resources and personal data, and report directly to the highest management level of the controller or processor. The DPO must maintain confidentiality, avoid conflicts of interest, and have due regard to the risks associated with processing operations. They may be a staff member or work on a service contract basis.

4. The framework for the protection of children's rights under GDPR includes specific protections for children's personal data, especially in relation to marketing, creating profiles, and collecting data through services directly offered to children. The consent of a parent or guardian is not always necessary, particularly in the context of preventive or counseling services offered directly to a child. This is outlined in Article 8 of the GDPR.

6.2 Agent-based system

Due to the complex nature of questions 5 through 12, an agent-based system was implemented. The implementation involved two agent sub-systems, one semi-autonomous and one fully autonomous. The level of autonomy appointed to each sub-system correlated to the flexibility required for each task.

Intermediate and expert level questions required both metadata information about each document as well as information contained within the document itself. For this reason it was necessary to inspect the contents of each document, extract the information -not present in the metadata file- necessary to answer the question posed, augment the metadata file with the extracted and structured information and finally have the augmented metadata file contents analyzed by the LLM.

Initially, each of questions Q5-Q12, along with the column names already present in the metadata file were passed to model gpt-4-0125-preview, that was tasked with identifying and creating sub-questions for the parts of the initial question that could not be answered directly through the metadata file. For each of the sub-questions, the LLM was also asked to provide an associated data type; this ensured to some degree a higher level of homogeneity of the LLM's responses in the next step.

Thereafter, all documents were passed to the LLM, along with orders to generate responses and structure them, for each of the sub-questions of the first step. In this second step model gpt-3.5-turbo-0125 was used, and in cases where the context window size was exceeded, chunking was employed. The structured generated responses were combined with the original metadata file.

In the second stage, a fully autonomous agent based on the gpt-4-0125-preview model was employed to analyze and generate responses for the original questions, based on the augmented metadata file. The answers generated by the autonomous agent are provided below:

5. There have been 209 requests for preliminary rulings submitted to the Court during the period of reference.

6. The average amount of time in days between the date of submission and the date of issuance for requests for preliminary rulings where a court ruling has been issued is approximately 335 days.

7. There are 78 cases where a court ruling has been issued for a request for a preliminary ruling to the EU Court.

8. For the period from 2015-2024, the requests for preliminary rulings to the EU Court from member state courts, classified by member state courts and the number of requests per year, are as follows:
   - In 2016, the Corte suprema di cassazione (Court of Cassation, Italy) made 1 request.
   - In 2017, Latvia, Slovak Republic each made 1 request, and the Supreme Court (Ireland) made 2 requests.
   - In 2018, Belgium made 1 request.
   - In 2024, Austria made 2 requests, Belgium 1 request, Bulgaria 1 request, Germany 6 requests, and Lithuania 1 request.
   (Note: The list provided is a partial summary based on the observation. The full details include requests from various member state courts across the years 2016 to 2024, totaling 99 entries.)

9. Based on the provided data sample, there are no requests for preliminary rulings submitted to the Court by Member State courts, hence no overlapping areas regarding the interpretation of specific articles of GDPR can be identified from this dataset.

10. The key principles governing the processing of personal data derived from requests for preliminary rulings submitted to the Court by Member State courts include, but are not limited to, the following:
    - Protection of personal data, independence of supervisory authorities
    - Purpose limitation, Data minimisation
    - Consent, active consent, separate consent, obligation to fully inform
    - Consent, freely given consent, informed consent, burden of proof

- Protection of natural persons with regard to the processing of personal data, lawfulness of processing, protection of personal data, fundamental rights
- The right to be forgotten, processing of sensitive data, exceptions for journalistic purposes or artistic expression
- Effective and complete protection of data subjects
(Author's note: Output continued in the same way, until the model's output size was exceeded)

11. Potential risk factors regarding data protection rights in the dataset could include ongoing legal challenges or the need for clarifications in the application of data protection laws, as indicated by the 222 documents related to preliminary rulings on data protection. Additionally, the focus on data protection issues and personal data handling practices in 359 and 353 documents, respectively, highlights the importance of these areas in the context of GDPR compliance and potential areas of risk that require careful management and adherence to legal standards.
12. Yes, there is at least one document in the dataset that discusses the relationship between the EU Court of Justice and national courts regarding the interpretation of GDPR provisions.

7 Evaluation

A qualitative evaluation of the results from the LLM system was conducted. Initially, moderate results were anticipated, particularly for beginner-level questions, where the exact response from the question was expected to be found within the legal corpus. However, lower expectations were set for more complex, context-related questions that necessitate a deeper understanding of legal concepts. The evaluation for each separate set of questions is presented below.

Concerning the beginner set of questions, the system seems to understand them and provides reasonable responses that lie within the scope of what was expected. In the intermediate set of questions, the system also seems to understand the question and provide adequate responses. However, the numerical and qualitative nature of the responses is difficult to be immediately verified by human subjects and verification will be provided at a later point in the research. Within the final set of questions, namely the expert set, evaluation of questions took place individually. In Q9, the conditionality the response is based on is wrong, though the conclusion itself might be accurate. In Q10, the response can be considered adequate to what was expected. In Q11, even though the question might be ambiguous regarding the nature of risks (these can be legal, procedural or even social), the system seems to understand it and provides a rather simplistic response. Finally, in Q12, the system seems to evade a direct answer, thus displaying rather human-like behaviors. It might be better, if it had indicated it cannot provide a response based on the available data corpus.

8 Conclusion

LLM-based systems that rely on GPT3.5 and GPT4 models have been successfully demonstrated for a narrowly specified legal corpus concerning the various aspects of GDPR implementation. The two distinct sub-systems created as proof of concept implementations featured a combination of retrieval augmented generation and agent-based methods, highlighting the significance of choice of approach concerning the effective exploitation of the large language models' reasoning capacities in a demanding real-world scenario.

A set of legal questions of progressive difficulty was developed, and tested the LLM-based system. Evaluation of the results reveals encouraging findings for the use of such well-defined LLM systems as legal assistants for governance applications. The system appeared to provide adequate results for all question sets, however the gradual increase of difficulty was evident in the system's responses. In future work, independent evaluation of the system's responses will take place.

As large language models keep advancing in terms of reasoning capacities combined with the improvement of grounding methods that augment the factual reliability of the models, it is expected that systems like the one presented in this paper will become much more capable at performing their tasks effectively in the near future.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*, 2024.

[4] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*, 2024.

[5] Fotios Fitsilis. Artificial intelligence (ai) in parliaments – preliminary analysis of the eduskunta experiment. *The Journal of Legislative Studies*, 27(4):621–633, September 2021.

[6] Fotios Fitsilis and Georgios Theodorakopoulos. Better regulation and its evolution in the hellenic legislative and parliamentary system. *Statute Law Review*, 45(1):hmae003, 2024.

[7] Shaw Green, Leon Hurst, Brenda Nangle, Pádraig Cunningham, Fergal Somers, and Richard Evans. Software agents: A review. *Department of Computer Science, Trinity College Dublin, Tech. Rep. TCS-CS-1997-06*, 1997.

[8] Carolyn Gruske. Alberta courts caution against using unverified citations generated by ai or large language models, Oct 2023.

[9] Marci Harris and Aubrey Wilson. Representative bodies in the ai era: Insights for legislatures.

[10] Natali Helberger and Nicholas Diakopoulos. Chatgpt and the ai act. *Internet Policy Review*, 12(1), February 2023.

[11] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[12] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[13] Evangelos Kalampokis, Nikos Karacapilidis, Dimitris Tsakalidis, and Konstantinos Tarabanis. Understanding the use of emerging technologies in the public sector: A review of horizon 2020 projects. *Digital Government: Research and Practice*, 4(1):1–28, 2023.

[14] Yu He Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. Enhancing diagnostic accuracy through multi-agent conversations: Using large language models to mitigate cognitive bias. *arXiv preprint arXiv:2401.14589*, 2024.

[15] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[16] Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. Large language models in law: A survey. *arXiv preprint arXiv:2312.03718*, 2023.

[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[18] Giulia Listorti, Egle Basyte-Ferrari, Szvetlana Acs, and Paul Smits. Towards an evidence-based and integrated policy cycle in the eu: A review of the debate on the better regulation agenda. *JCMS: Journal of Common Market Studies*, 58(6):1558–1577, 2020.

[19] Jorn von Lucke, Fotios Fitsilis, and Jan Etscheid. Research and development agenda for the use of ai in parliaments. In *Proceedings of the 24th Annual International Conference on Digital Government Research*, pages 423–433, 2023.

[20] Marios Evangelos Mamalis, Evangelos Kalampokis, Areti Karamanou, Petros Brimos, and Konstantinos Tarabanis. Can large language models revolutionalize open government data portals? a case of using chatgpt in statistics. gov. scot. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, pages 53–59, 2023.

[21] Cabinet Office. Generative ai framework for hm government, Jan 2024.

[22] Cabinet Office and Innovation & Technology Department for Science, Mar 2024.

[23] National Archives Office of the Federal Register and Records Administration. 88 fr 75191 - safe, secure, and trustworthy development and use of artificial intelligence. [government]. *Federal Register*, November 2023.

[24] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[26] Daniel J Solove and Paul M Schwartz. *EU Data Protection and the GDPR*. Aspen Publishing, 2023.

[27] SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

[28] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[29] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[30] Colin Van Noordt and Gianluca Misuraca. Exploratory insights on artificial intelligence for government in europe. *Social Science Computer Review*, 40(2):426–444, 2022.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[32] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

[33] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[34] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[35] Michael Wooldridge and Nicholas R Jennings. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152, 1995.

[36] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

[37] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Robert K Yin. *Case study research: Design and methods*, volume 5. sage, 2009.

[39] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[40] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.