Optimizing Legal Information Access: Federated Search and RAG for Secure AI-Powered Legal Solutions

1st Flora Amato
DIETI
University of Naples Federico II
Naples, Italy
flora.amato@unina.it

2nd Egidia Cirillo
DIETI
University of Naples Federico II
Salerno, Italy
egidia.cirillo@unina.it

3rd Mattia Fonisto
DIETI
University of Naples Federico II
Naples, Italy
mattia.fonisto@unina.it

4th Alberto Moccardi
DIETI
University of Naples Federico II
Naples, Italy
alberto.moccardi@unina.it

2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10825815

***Abstract*—Large Language Models (LLMs) have gained increasing importance in the field of Legal Intelligence, enabling the development of applications to assist both legal professionals and ordinary citizens.However, centralized training of these legal LLM models raises concerns about data privacy, as legal data is distributed across various institutions that contain sensitive information about individuals.**

Using Federated Learning (FL), legal LLMs can be locally trained on devices or clients, and their parameters are aggregated and distributed to a central server, ensuring data privacy without directly sharing the raw data. The presence of private information of the parties in legal data calls for new studies based on legal Artificial Intelligence (legal AI) to investigate new decentralized learning methods that can protect the privacy of individuals.

A promising technique in this field is FL, which is useful for enabling multiple participants to collaboratively train a shared model while effectively protecting private sensitive data. However, it has significant limitations, which can be effectively addressed using a technology like FS.

FS is a crucial system in managing distributed information in modern environments and it can play a fundamental role in finding relevant information from diverse heterogeneous sources to generate well-informed answers without requiring specific training on data, which proves to be a less flexible solution.

FS systems aggregate results from different sources, selecting the most appropriate documents to improve the quality of results and align with the user's intent. To achieve this goal, new technologies are being utilized, such as Retrieval-Augmented Generation (RAG) in the case study presented.
In summary, FL therefore allows training AI models on distributed data without centrally sharing them, ensuring protection of sensitive data and collaboration between organizations without compromising data confidentiality.
However, it has important limitations, including computational and communication overhead; in fact, distributed training requires significant computational power, which may not be sustainable for all devices or clients, and it may not be fully efficient, as model optimization in decentralized contexts often doesn't reach the levels of centralized training, especially for complex models like LLMs. FS is a good technology that addresses these limitations, in particularly specific domains, such as the legal field, by focusing on retrieving information from distributed sources rather than on model training. This approach offers several advantages over FL, first of all lower computational requirements; in fact, since it is not necessary to train models on distributed devices, the resources required are significantly lower and subsequently better results for informative tasks in scenarios where it is necessary to generate answers based on heterogeneous information, allowing to retrieve relevant documents from different sources and better integrate the results to generate coherent and informed answers.Furthermore, the major strength concerns flexibility; FS systems can easily adapt to new data or sources without the need for retraining.
This study aims to analyze the limits of FL in the legal field through an excursus on the use of these technologies, supporting the development of new FS methods, especially in the context of RAG pipelines.

***Index Terms*—Large Language Models (LLM), Federated Learning (FL), Federated Search (FS), Retrieval-Augmented Generation (RAG), Artificial Intelligence (AI), Generative-AI, Legal Domain.**

I. Introduction

With the advent of Big Data, the main concern is no longer the management of large amounts of data, but the aspects related to their privacy and security. Data protection is a growing priority for the public, and privacy violations are increasingly serious, so the goal is to improve data security and confidentiality [1], [2].
The General Data Protection Regulation (GDPR) [3], which came into force on May 25, 2018, establishes the new EU data protection regulations, in order to protect the privacy and security of personal information of European citizens. Users must provide their explicit consent before the data is used to train models and operators are required to clearly explain the agreements with users and cannot induce or force consumers

to give up their privacy rights, in addition, users have the right to request the deletion of their personal data.
These laws create new challenges for data processing, leading to the formation of numerous ”data islands”, i.e. collections of data that are isolated and not shared, however, AI relies on the ability to manage and work with large amounts of data, without which model training is impossible.

A solution to this problem is offered by FL, in fact FL offers an alternative, allowing users to collaborate to train distributed models, without having to share their personal data with a central server. In this system, data remains on users' personal devices, while models are trained locally, ensuring privacy protection [4], [5].
This technology opens up new research avenues in the field of artificial intelligence, enabling the development of personalized models without compromising user privacy. With the advancement of computing capabilities of devices, model training can be performed directly on terminal devices, eliminating the need for central servers.
Therefore, FL leverages the computational resources of local devices to train algorithms, thus protecting personal data from possible breaches during transfer and ensuring user privacy by using techniques such as multi-level anonymization and differential privacy, which prevent attackers from accessing the original data.
These techniques ensure that FL does not compromise user privacy, nor violate GDPR or other data protection regulations.

However, in the legal sector, which is an important area of interest, where privacy management and protection are essential, FL may not be entirely sufficient, but new technologies could be used.

Learning, understanding, and correctly using an ever-increasing amount of legal data exceeds the human capabilities of legal scholars [6]. Since most data is textual, there is an “information crisis in law” that is driving research and development of Natural Language Processing (NLP) techniques in the legal sector, in order to provide accessible legal services to both legal professionals and the general public [7]. However, since most of these techniques are based on machine learning, they require training on centralized datasets. This approach raises growing privacy concerns, especially in relation to data protection regulations such as the GDPR [8].

These algorithms, using FL techniques, allow the participants' local devices to collaborate with one or more servers to train a model in a distributed and secure way, ensuring data confidentiality. However, despite the optimistic outlook, FL still faces several challenges, including data variety management [9], protection from privacy attacks [10], and general inefficiency of the system [11].

An important solution is FS a new technology that try to solve the problems of information dispersion and privacy protection, collecting data from different sources and combining them into a single set of results. In the context of FS, one of the main challenges is to ensure that information from distributed sources can be aggregated and returned to the user without compromising the privacy and integrity of the data. In this context, advanced techniques such as RAG, can be seen as an innovative solution that integrates FS principles with intelligent response generation. RAG is a technology that answer complex queries with data from different sources and it was originally developed to optimize the interaction between search and generation, it has features that make it interesting for application in FS scenarios.
In a RAG pipeline, a user's information request is routed to specific resources to obtain the relevant data like in Fig. 1. This data is then aggregated by a LLM, which processes the results to provide a response to the user. Some examples of such pipelines include langchain [12], llamaindex [13], and DSPy [14], [15].

In FS, the crucial steps are:

- **Resource Selection**, which is choosing the sources to which the search request should be sent.
- **Result Fusion**, which is combining the result lists from each source into a unified list that can be used to generate a response.

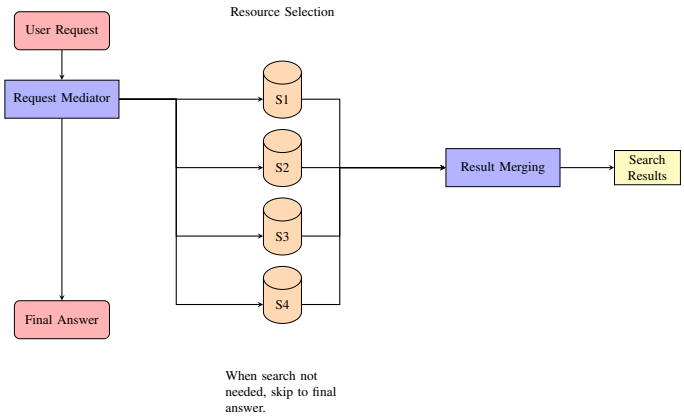

Fig. 1. Diagram of a RAG pipeline demonstrating resource selection, result merging, and final response generation.

Today, Legal LLMs have achieved very promising results, as they are trained on centralized datasets, however they still suffer from the problem related to data privacy, in fact, in a practical scenario, legal data are distributed among different institutions (e.g., courts, prosecutors and legal consultancy firms), containing sensitive data of individuals.

To improve this aspect, the FS comes with the use of technologies such as RAG, which combines two techniques to improve the response of LLM.
First, it searches (retrieval) relevant information from external sources, such as databases or documents, based on the user's question. Then, it uses a language model to generate a response based on both this retrieved information and its own internal knowledge. This approach improves the accuracy of responses, reducing the risk of errors or hallucinations.
With the development and increasing use of Gen-AI, it becomes crucial to develop tools that are accessible and understandable for everyone, especially in sensitive areas such as justice.

The study carried out aims to offer an alternative to the classic static models of RAG, improving them thanks to the use of agents based on LLMs. The goal is to provide a solution generation technology that pays particular attention and protection to privacy and non-dispersion of information of data and documents useful for generating reliable and timely answers.
This will allow less experienced users to navigate the complex processes of legal conflict resolution, providing faster and easier access to justice and at the same time ensuring the reliability of the sources from which legal data are retrieved.
This work is based on the study of numerous research and works that have marked significant progress in the field of Gen-AI, introducing innovative techniques to improve conversational systems. Among the most relevant contributions, there is certainly that of Lewis et al. [16], who in 2020 combined parametric and non-parametric models to create retrieval augmented generators. Another important contribution comes from Gao et al. [17], who in 2024 published an in-depth review of RAG paradigms, highlighting current challenges and suggesting future developments for research. Yao et al. [18] proposed the ReAct framework in 2022, demonstrating its superiority over previous models in terms of interpretability and reliability. ReAct effectively addresses common problems such as data hallucination and error propagation in sequential reasoning models. In 2024, Amatrian [19] presented a comprehensive overview of prompt engineering techniques, delving into approaches such as Reasoning without Observation (ReWOO), Reason and Act (ReAct), and Dialog-Enabled Resolving Agents (DERA). Xi et al. [20] and Hua et al. [21] also introduced LLM Agents and reliable agent frameworks in 2023 and 2024, respectively. An important application of LLMs in the legal field was developed by Cui et al. [22], who in 2024 presented ChatLaw, an open-source legal model capable of addressing the problem of model hallucination. Using a RAG system enhanced by self-attention mechanisms, ChatLaw significantly improves in the management of errors present in the reference legal data. In summary, this work proposes a more human and inclusive approach, capable of responding to users' needs, ensuring data non-dispersion and facilitating access to justice in a transparent and reliable way.

A. Objectives

The objectives of the study regarding the use of FS in the legal context are:

- **Privacy Protection**: Ensuring that all interactions with AI comply with privacy regulations, safeguarding users' sensitive data and ensuring that legal information is handled securely.
- **Data Anonymization**: Implementing data anonymization techniques to protect user identities, ensuring that information used for training and generating responses cannot be traced back to individual users.
- **Access Control**: Establishing strict access control mechanisms to limit access to sensitive data only to authorized personnel, reducing the risk of privacy breaches and ensuring responsible handling of information.
- **Accessibility to Justice**: Facilitating access to legal information for individuals who are not legal experts, providing clear and understandable answers, reducing the need for expensive legal services, and making justice more inclusive.
- **Efficiency in Dispute Resolution**: Automating and speeding up the process of searching and analyzing legal documents, improving the efficiency of conflict resolution through the use of advanced language models capable of generating relevant answers based on extensive sources of knowledge.
- **Trustworthiness and Transparency**: Ensuring that the results provided by AI agents are transparent and verifiable, especially in the legal field where trust is crucial.

II. Federated Learning

FL is a concept introduced by Google to develop machine learning models using data distributed across devices, without having to directly share it, thus reducing the risk of privacy violations [24] [26], this involves the interaction between scattered users and some of the main issues to be addressed are communication costs, non-uniform data distribution and device reliability.
Data is partitioned based on user or device IDs and privacy protection is a key factor in these decentralized collaborative learning situations.
The original concept of FL has since expanded to include all collaborative and decentralized machine learning approaches designed to ensure privacy, this has proven to be particularly useful in enterprise applications where multiple parties need to collaborate without directly sharing their data.
FL is therefore an innovative approach to training machine learning models, characterized by its decentralized architecture and data privacy protection [25].
This paradigm emerged in response to growing concerns about data privacy and security, especially in contexts where personal information is involved.

The fundamental idea of FL is that the data remains on the devices of the users, who participate in the model training by contributing only updates to the model itself [27].
These updates are then aggregated by a central server, which combines information from different devices to improve the model performance.

This approach not only preserves privacy, but also reduces the need to transmit large volumes of data, making the process more bandwidth efficient.

However, FL presents several challenges. One of the main ones is data heterogeneity, as devices can have access to different datasets, with varying distributions and sizes. This leads to generalization problems and can negatively impact the performance of the final model. It is essential to develop robust algorithms that can adapt to these variations and ensure effective training.

Another critical aspect of FL is security. Although data is never transferred, potential attacks could occur during the aggregation phase of model updates. In this regard, techniques such as homomorphic encryption and differential learning mechanisms are proposed to address these issues, ensuring that updates do not reveal sensitive information. In a legal context, where trust is fundamental, the ability to ensure that data remains protected is crucial [33].

Personalization is another point of interest of FL, in fact, thanks to its decentralized nature, FL allows training specific models for each user or group of users, leading to more relevant and performant systems [34].

FL has shown promising results in various fields of application by improving access to information without compromising user privacy and by facilitating the analysis of confidential data without the need for centralization, thus allowing for more fruitful collaboration. However, it shows some important limitations in the legal sector, as the data processed is complex and very sensitive. A major issue is complying with the laws, which vary from country to country and make it difficult to find solutions that are valid for all. Even if FL protects the data by keeping it on devices, the information used to update the models could still be attacked, putting confidentiality at risk. Furthermore, legal data is very diverse, changing depending on the institution or region, and this makes it difficult to create models valid for all contexts. In the legal sector, then, firms and institutions often compete with each other and may be reluctant to collaborate, limiting the potential of FL. Added to this, not everyone has the technical resources or the appropriate devices to participate, making the adoption of this technology slower.

III. Federeted Search

FS is a technology that enables users to search and aggregate information from multiple distributed sources, without the need to physically collect the data in one place, this approach solves the problem of information dispersion, often fragmented across different systems and databases, which puts the privacy associated with that data at risk. The principle behind FS is that, instead of querying a single centralized repository, the system sends search requests to multiple decentralized resources and then combines the results into a single coherent set; this process has numerous applications, but its relevance in the legal field is particularly interesting, especially with regard to privacy protection and secure data management.

In the legal field, crucial information, such as legal precedents, legal documents, laws and regulations, are often stored in multiple repositories, including internal databases, digital libraries, document management systems and external sources such as public or commercial databases.

Access to this information is essential for the work of lawyers, judges and legal advisors, but the difficulty in locating it quickly and efficiently can compromise the timeliness of legal work.

FS helps the work, allowing legal professionals to access dispersed information through a single search interface, furthermore, the adoption of RAG pipelines (Fig. 2.), which combine FS with LLM, further improves the effectiveness of the system.

In these pipelines, the user's search request is sent to different information sources and the obtained results are processed by an LLM, which aggregates and synthesizes the relevant information to generate a complete and coherent answer; this approach not only optimizes information retrieval, but improves the understanding and utility of the retrieved data for the legal professional.

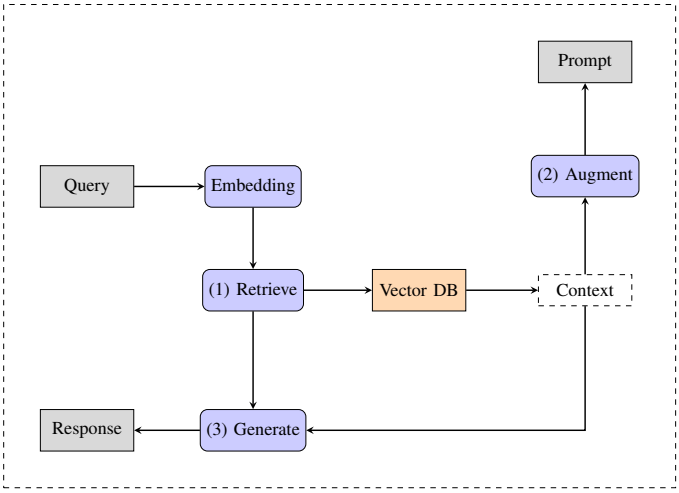

Fig. 2. RAG Workflow.

One of the most important and sensitive aspects of FS is its ability to protect user privacy and ensure data security throughout the search process. Unlike traditional centralized search engines, where data must often be copied and stored on a single server for processing, FS allows data to remain on their original servers, and the generated response is immediately deleted, so the LLM has no way to learn from the solutions provided; only relevant and already selected information is shared, but there is no possibility of self-learning from these responses to avoid issues related to the management of private resources and data. This distributed architecture is particularly advantageous in legal contexts, where confidentiality is essential, in fact many legal information is highly sensitive and the risk of unauthorized access or security breaches could have serious consequences for the parties involved.

RAG further increases this level of security; in fact the system avoids the need to centralize data, since the LLM model processes search results directly from distributed sources and provides an aggregation-based response; this significantly reduces the risk of information leakage and ensures that all data remains safe in the source systems, but most importantly that the centralized model cannot be trained on private data.

In a legal context, where secure information handling is of primary importance, this ability to FS while maintaining data security is a huge advantage, another advantage of FS technology in the legal field is its ability to reduce the risk

of hallucinations, a phenomenon in which AI models generate inaccurate or misleading answers.
This is particularly problematic in the legal field, where an incorrect answer could have significant consequences. However, with careful management of sources and the implementation of more sophisticated result fusion strategies, the risk of inaccurate answers can be mitigated.

Another benefit of FS is the facilitation of collaboration between different clients such as law firms, government agencies and other organizations involved in legal processes, thanks to this technology, it is possible to share information essential for cases without compromising the security or privacy of the data.
This is especially relevant in contexts where it is necessary to comply with strict regulations, such as the GDPR in Europe, which imposes strict constraints on the management of personal data.

By adopting FS, law firms can comply with these regulations without sacrificing operational efficiency or the quality of the search and since the data is not transferred or aggregated centrally, but remains protected in its respective locations, the risk of privacy breaches is drastically reduced.
Additionally, FL techniques can be integrated into the system to ensure that even when training AI models on sensitive legal data, the privacy of individuals is protected.

IV. Methodology

The adopted methodology uses RAG in combination with FS techniques to optimize the retrieval and processing of information in complex contexts. RAG, in particular, exploits LLM to integrate and synthesize information from distributed sources, ensuring a complete and coherent response. FS plays a crucial role by allowing to simultaneously query multiple repositories or information systems, without requiring data centralization. This approach reduces the risk of exposing sensitive information, preserving privacy.
The proposed system comprises two main parts illustrated in Fig. 3.

- **"Solution Explorer"**
  Which leverages language models to find relevant legal documents to provide factual help to user queries.
- **"Digital Journey"**
  Which employs an LLM, along with a framework called Reason plus Action (ReAct), to guide individuals through the necessary reasoning to solve a legal problem.

The proposed system is divided into two main components, designed to improve access and understanding of legal information. The first, called **Solution Explorer**, leverages advanced linguistic models to locate and retrieve relevant legal documents, providing accurate and fact-based answers to user queries. The second component, called **Digital Journey**, combines a linguistic model with the Reason plus Action (ReAct) framework to guide users through the logical steps needed to solve complex legal problems.

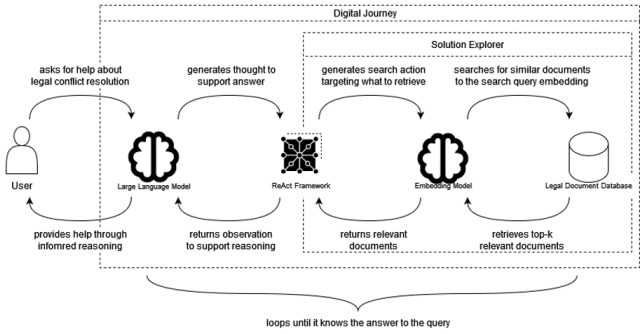

Fig. 3. Diagram depicting the proposed methodology

A. Solution Explorer

This component plays a crucial role in the process of retrieving relevant documents for legal conflict resolution. Its methodology unfolds into the following steps:

1) **Query Analysis**
2) **Document retrieval**

1) Query Analysis: Embedding analysis is essential for any kind of NLP task since it captures the meaning of the text through a numerical representation. This methodology, once the text is converted into this form, allows comparison with other texts to evaluate semantic similarity [28], perform groupings, classifications and other usage scenarios. This family of models extensively exploits transformer architectures to encode sentences derived from legal documents, dividing the text into blocks and then in dense vector representations known as embedding (Fig. 4.). Language models (LMs) which represent the backbone of the embedding models are typically trained on vast amounts of data from various sources to gain a deep understanding of language, although cheaper alternatives exist [29].

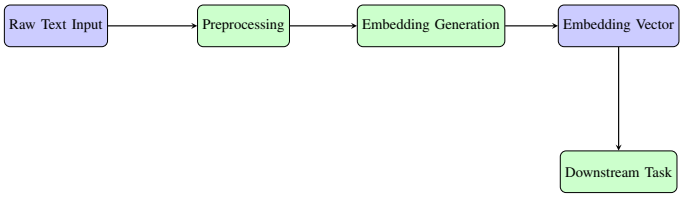

Fig. 4. Text Embedding Workflow

2) Document Retrieval: After analyzing the embedded query, the Solution Explorer compares the semantic vector representation of the query with the pre-computed semantic vector representation of legal documents stored in a vector database like Chroma or V3CTRON which is designed for efficient storage, rapid retrieval of embeddings, and swift lookup of their element. This comparison involves applying semantic similarity measures, in detail cosine similarity which reflects the cosine of the angle between the two embeddings. The retriever calculates a semantic similarity score for each tuple and retrieves the top-ranked section of documents based on the similarity score with the user query (Fig. 5.).

Fig. 5. Text Embeddings 3D example

B. Digital Journey

This component is responsible for guiding model reasoning for the legal conflict resolution proposed by the user query. Its methodology unfolds into the following components:

1) **Large Language Model (LLM)**
2) **ReAct Framework**

1) Large Language Model (LLM): One of the main actors is the Large Language Model which can be seen as the brain behind the intelligent solution, the one in charge of generating text. A sophisticated language model like Llama [30], or a comparable model, possesses extensive pre-existing knowledge of the language. The typical functioning of such language models involves generating text by predicting the next word at each step based on context, demonstrating exceptional proficiency in various tasks, often requiring minimal examples (few shots prompting). Their extensive architecture and training datasets allow them to grasp grammar, semantics, and language syntax comprehensively.

2) ReAct Framework: To augment the reasoning capabilities of the "Digital Journey," the LLM utilizes the ReAct framework. Through a reasoning phase, the model can create, monitor, and revise actions using tools that enable access to external knowledge sources, such as a vector store containing legal content. The Reason plus Action framework, advances the concept of guided prompting by empowering the LLM to engage in reasoning and execute actions by compelling it to produce text following a thought-action-observation structure. Here, the thought serves as an intermediary question that must be addressed to reach the final solution, the action entails retrieving information needed to answer this intermediary question, and the observation represents the output resulting from the action.

V. System Design, Architecture & Results

The architectural backbone of the proposed solution, built using Langchain development framework, combines RAG-based architecture and ReAct presenting a significant advancement in the way in which LLMs agents interact and utilize external sources. The depicted approach intends to grant a more dynamic and context-sensitive retrieval of information with the aim of enhancing the conversational agent capabilities through a human-centric integration, thus not only amplifying the factual correctness and depth of the responses but also improving the system's transparency and interpretability which are crucial for building trust and reliability in AI-driven technologies.

As already mentioned, the proposed ReAct process has been meticulously designed intending to augment the functionality of a legal AI-based assistant [31], mainly crafted to assist individuals unfamiliar with legal jargon in easily accessing justice.

Crucially for this objective is the Prompting schema, inspired by the Chain-of-Thought prompting technique [32] it includes a set of tools that based on similarity search, allows the legal assistant to focus on different parts of the knowledge base.

In the specific context of the proposed solution, just inheritance and divorce legislation of different countries have been considered respectively. The AI agent selects each time which is the most appropriate tool or set of tools and then performs thoughts and actions to satisfy the user query, targeting the RAG process. This agent is appropriately defined in the "Agent definition" through a specific "PREFIX" then governed by a structured protocol: — **when the AI agent determines the necessity of a tool, it follows a sequence of 'Thought', 'Action', 'Action Input', and final 'Observation'** — This structured approach ensures that the AI's interactions are not only contextually relevant but also precise and tailored to the specific legal issue at hand.

The Legal ReAct Agent workflow presents two different phases:

- Through "reasoning" the agent searches different possible useful assets, then, when asset(s) are identified appropriate thought(s) are generated addressing the agent to find that (those) content(s).
- The agent performs action(s) answering the specific questions addressed by the user interrogating a targeted knowledge base as in the classical RAG based application.

The integration of the ReAct framework within the RAG architecture has been evaluated as central to improving the transparency of responses provided by LLM agents.

The example session depicted in Fig. 6.,Fig. 7. illustrates how The ReAct system enhances transparency by breaking down the query resolution process into distinct phases:

1) **Thinking**: The agent identifies the nature of the query and decides on a course of action, demonstrating its initial thought process.

2) **Action**: Based on the thinking step, the agent executes an action—searching a specific legal database for relevant information.
3) **Action Input**: It details the specific query used in the database, showing the direct link between the user's question and the system's search mechanism.
4) **Observation**: The results from the action are presented back to the user, providing a clear and detailed explanation derived from the legal texts enhanced by LLM.

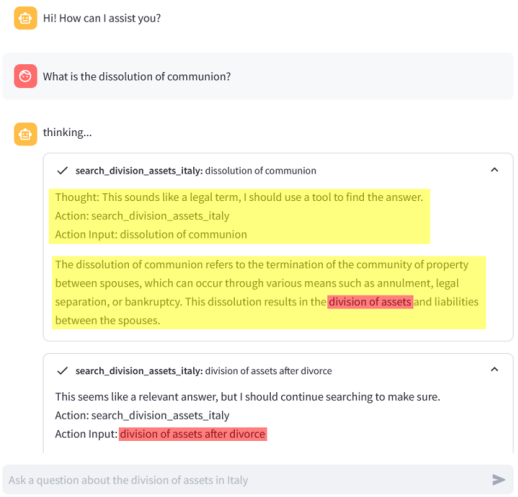

Fig. 6. Legal Inquirying Session

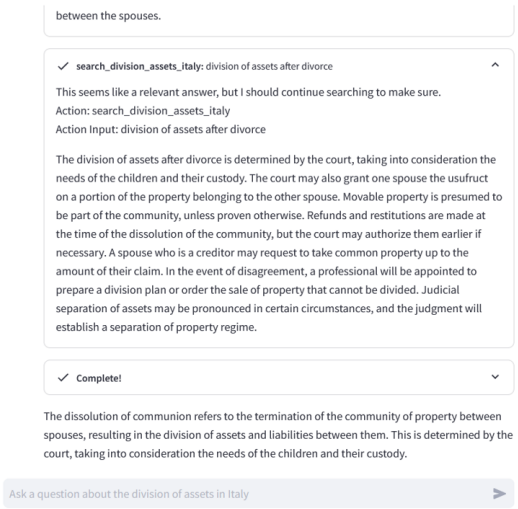

Fig. 7. Legal Inquirying Session

This process ensures that users can follow the agent's reasoning and actions step-by-step, making it easier to understand how the conclusion was reached.

Finally, by incorporating the ReAct framework, the system not only answers legal inquiries but also has the potential to educate the user about the legal reasoning involved while significantly enhancing user trust.

The large-scale deployment of the proposed system within a RAG system could consistently represent a potential advancement toward more transparent, interpretable, and trustworthy legal AI systems that are not only more effective in processing and answering complex legal queries but also more open in their operations, bridging the gaps between advanced AI technologies and users thereby democratizing access to legal information.

VI. Conclusion

In conclusion, FS represents a highly innovative and promising technology, especially for the legal sector, where privacy protection and secure data management are key, offering a more secure approach to information search with its ability to FS from different sources without centralizing data.

Furthermore, by integrating pipelines such as RAG, the system can significantly improve the efficiency of legal searches, reducing the risks of privacy violations and improving the quality of the generated answers.

Thanks to its flexibility, security and regulatory compliance, FS has the potential to transform the way legal information is managed and used. In this study, advanced LLM and RAG-based architectures were integrated into the legal context, highlighting their potential to improve access to justice and efficiency in conflict resolution. Through the adoption of the ReAct framework, greater transparency and traceability in AI-generated responses was ensured, promoting trust in the use of advanced technologies in sensitive areas such as law. The obtained results demonstrate that generative AI can not only automate the search for legal information, but also support users in understanding and navigating complex legal issues.

However, there is still work to be done to further improve the accuracy, standardization and reliability of the generated responses. The adoption of systems such as the one proposed has the potential to transform the interaction between citizens and the legal system, democratizing access to justice and providing a powerful tool for legal professionals.

REFERENCES

[1] C. Zhang, X. Hu, Y. Xie, M. Gong, and B. Yu, A privacy-preserving multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, Frontiers in Neurorobotics, pp. 112, 2020.

[2] K. Bonawitz et al., Towards federated learning at scale: System design, Proceedings of Machine Learning and Systems, 1 (2019), 374-388.

[3] J. P. Albrecht, How the GDPR will change the world, Eur. Data Prot. L. Rev., 2 (2016), pp. 287.

[4] M. Gong, Y. Xie, K. Pan, K. Feng, and A. K. Qin, A survey on differentially private machine learning, IEEE computational intelligence magazine, 15 (2) (2020), 49-64.

[5] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, Federated learning for healthcare informatics, Journal of Healthcare Informatics Research, 5(1) (2021), 1-19.

[6] Marco Gomes, Bruno Oliveira, and Cristóvão Sousa. 2022. Enriching legal knowledge through intelligent information retrieval techniques: A review. In EPIA Conference on Artificial Intelligence, pages 119–130. Springer.

[7] Changlong Sun, Yating Zhang, Xiaozhong Liu, and Fei Wu. 2020a. Legal intelligence: Algorithmic, data, and social challenges. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2464–2467.

[8] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2):1–210.

[9] Suyu Ge, Fangzhao Wu, Chuhan Wu, Tao Qi, Yongfeng Huang, and Xing Xie. 2020. Fedner: Medical named entity recognition with federated learning. arXiv preprint arXiv:2003.09288.

[10] Samyak Gupta, Yangsibo Huang, Zexuan Zhong, Tianyu Gao, Kai Li, and Danqi Chen. 2022. Recovering private text in federated learning of language models. arXiv preprint arXiv:2205.08514

[11] Ruixuan Liu, Fangzhao Wu, Chuhan Wu, Yanlin Wang, Lingjuan Lyu, Hong Chen, and Xing Xie. 2022. No one left behind: Inclusive federated learning over heterogeneous devices. arXiv preprint arXiv:2202.08036.

[12] Harrison Chase. 2022. LangChain. https://github.com/langchain-ai/langchain

[13] Jerry Liu. 2022. LlamaIndex. https://doi.org/10.5281/zenodo.1234

[14] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate Search Predict: Composing Retrieval and Language Models for Knowledge-Intensive NLP. arXiv preprint arXiv:2212.14024 (2022).

[15] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. arXiv preprint arXiv:2310.03714 (2023).

[16] Patrick Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2021. arXiv: 2005.11401 [cs.CL]

[17] Yunfan Gao et al. Retrieval-Augmented Generation for Large Language Models: A Survey. 2024. arXiv: 2312.10997 [cs.CL].

[18] Shunyu Yao et al. "React: Synergizing reasoning and acting in language models". In: arXiv preprint arXiv:2210.03629 (2022)

[19] Xavier Amatriain. Prompt Design and Engineering: Introduction and Advanced Methods. 2024. arXiv: 2401.14423 [cs.SE].

[20] Zhiheng Xi et al. "The rise and potential of large language model based agents: A survey". In: arXiv preprint arXiv:2309.07864 (2023).

[21] Wenyue Hua et al. "TrustAgent: Towards Safe and Trustworthy LLM based Agents through Agent Constitution". In: arXiv preprint arXiv:2402.01586 (2024)

[22] Jiaxi Cui et al. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. 2023. arXiv: 2306.16092 [cs.CL].

[23] Shuai Wang, Ekaterina Khramtsova, Shengyao Zhuang, Guido Zuccon. FeB4RAG: Evaluating Federated Search in the Context of Retrieval Augmented Generation. arXiv:2402.11891 [cs.IR]. https://doi.org/10.48550/arXiv.2402.11891. 2024

[24] K. Bonawitz et al., Practical secure aggregation for privacy-preserving machine learning, presented at the Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017, 2017.

[25] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, Federated learning with non-iid data, arXiv preprint arXiv:1806.00582, 2018.

[26] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, Federated multi-task learning, Advances in neural information processing systems, vol. 30, 2017.

[27] R. C. Geyer, T. Klein, and M. Nabi, Differentially private federated learning: A client level perspective, arXiv preprint arXiv:1712.07557, 2017.

[28] Dhivya Chandrasekaran and Vijay Mago. "Evolution of semantic similarity—a survey". In: ACM Computing Surveys (CSUR) 54.2 (2021), pp. 1– 37.

[29] Nils Reimers and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks". In: arXiv preprint arXiv:1908.10084 (2019).

[30] Hugo Touvron et al. "Llama: Open and efficient foundation language models". In: arXiv preprint arXiv:2302.13971 (2023).

[31] Jason Wei et al. "Chain-of-thought prompting elicits reasoning in large language models". In: Advances in neural information processing systems 35 (2022), pp. 24824–24837.

[32] Jiaxi Cui et al. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. 2023. arXiv: 2306.16092 [cs.CL].

[33] M. Luo et al., Metaselector: Meta-learning for recommendation with user-level adaptive model selection, presented at the Proceedings of The Web Conference 2020, 2020.

[34] L. Li, Y. Fan, M. Tse, and K.-Y. Lin, A review of applications in federated learning, Computers Industrial Engineering, vol. 149, pp. 106854, 2020.

[35] https://www.aplyca.com/en/blog/Semantic-Search-Introduction