Privacy-Preserving Healthcare Data Interactions: A Multi-Agent Approach Using LLMs

Carmen De Maio, Giuseppe Fenza, Domenico Furno, Teodoro Grauso, and Vincenzo Loia

***Abstract*—Integrating Electronic Health Records (EHRs) into clinical workflows is crucial for advancing healthcare delivery but poses significant challenges, especially in improving human-machine interactions through natural language queries. This study builds on prior research [1] by introducing a multi-agent system that uses Large Language Models (LLMs) for secure interactions with healthcare data. In this extended work, we present a new capability for real-time updates to patient records through the addition of a Data Update Agent (DUA), ensuring privacy, accuracy, and compliance with regulatory standards. Compared to prior work, the system features a dual-pathway design for distinguishing between data retrieval and updates, enhanced modularity for seamless agent upgrades, and robust mechanisms to manage complex scenarios and noisy inputs. These advancements improve scalability, fault tolerance, and adaptability to real-world clinical environments. Comprehensive evaluations have been conducted using diverse clinical scenarios, including tests with noisy inputs and complex queries. The results highlight the system's scalability, accuracy, and practicality, demonstrating its superiority over baseline methods. The proposed framework enables better integration of LLMs in clinical settings by bridging natural language interfaces with secure, interoperable healthcare data systems.**

***Index Terms*—Electronic Health Records (EHRs), Large Language Models (LLMs), Natural Language Query Processing (NLQP), Privacy Preservation, FHIR, Multi-Agent Architecture.**

I. Introduction

The rapid evolution of digital healthcare systems highlights the growing need for efficient access to and management of medical information. Electronic Health Records (EHRs), particularly those adhering to the Fast Healthcare Interoperability Resources (FHIR) standard, play a pivotal role in improving patient care by enabling interoperability and seamless data exchange across various healthcare systems [2], [3]. Despite their potential, incorporating EHRs into clinical workflows remains complex due to the challenges of achieving intuitive and effective human-machine interactions, especially through natural language interfaces.

The emergence of Large Language Models (LLMs), such as OpenAI's GPT-3 and Meta's LLaMA, has demonstrated exceptional capabilities in comprehending and generating human-like text. These advancements present promising opportunities for applications in the healthcare sector [5]. However, deploying LLMs in sensitive environments like healthcare introduces critical concerns, including ensuring data privacy, maintaining system scalability, and aligning with established standards like FHIR.

This journal paper builds upon and significantly extends our prior work [1], which introduced a multi-agent architecture leveraging LLMs to enable natural language interaction with FHIR-based EHRs. This work addresses the need for systems capable of both retrieving and updating EHR data based on clinician instructions while adhering to stringent privacy and security standards. For instance, updating medication dosages or adding diagnostic notes can significantly enhance the efficiency of clinical workflows and reduce transcription errors. Furthermore, we evaluate the system on more complex clinical scenarios and include new experiments that assess its scalability and resilience to noisy inputs. These contributions substantially expand the scope and applicability of the initial framework, making it more suitable for real-world clinical use.

The proposed architecture employs a dual-layered approach, where a public LLM transforms user inputs into structured FHIR queries, and a private, locally hosted LLM converts the retrieved data into human-readable formats. This design ensures that sensitive patient information is processed securely within a controlled environment, complying with stringent privacy protocols [8], [9]. Unlike traditional monolithic systems, the multi-agent framework introduced here allows each agent to specialize in specific tasks, thereby improving efficiency and adaptability to diverse clinical needs. The modular nature of the architecture also facilitates the seamless addition of new functionalities to address emerging requirements.

To validate the proposed approach, we conducted extensive evaluations using the SyntheticMass dataset [13], [21], focusing on key metrics such as query accuracy, response time, and data interpretation quality. In addition to confirming the effectiveness of the original system, this work presents new experimental results, including performance comparisons with alternative solutions and analyses of the system's behavior under varying workloads and input conditions. These findings demonstrate the practicality and robustness of the proposed approach for integrating natural language interfaces with FHIR-based EHR systems.

The main contributions of this journal paper are:

- The introduction of enhanced privacy-preserving techniques tailored to LLM-based systems for healthcare.

Manuscript received December 19, 2024; revised January 13, 2025. Date of publication February 10, 2025. Date of current version February 10, 2025.

Authors are with the University of Salerno, Salerno, Italy (e-mails: cdemaio@unisa.it, t.grauso@studenti.unisa.it, gfenza@unisa.it, dfurno@unisa.it, loia@unisa.it).

The paper was presented in part at the International Conference on Software, Telecommunications and Computer Networks (SoftCOM) 2024.

Digital Object Identifier (DOI): 10.24138/jcomss-2024-0119

1845-6421/03/2024-0119 © 2025 CCIS

- The development of a more adaptable and scalable multi-agent architecture has now been extended with Data Update Agent (DUA) to support secure and compliant updates to patient information.
- A comprehensive evaluation framework that includes new experiments and metrics, assessing both data retrieval and update functionalities.
- An in-depth discussion of the limitations and future directions to further improve the system, including the integration of advanced update mechanisms and real-world deployment scenarios.

This paper is structured as follows: Section II provides an overview of related work, highlighting key advancements in the field. Section III describes the theoretical Background fundamental to the study. Section IV details the multi-agent system architecture and its components in detail. Section V outlines the implementation, including tools and datasets. Section VI presents the experimental setup, results, and comparisons. Section VII discusses the findings, challenges, and directions for future research.

II. Related Works & Terminologies

The application of Large Language Models (LLMs) in healthcare has advanced significantly, particularly in converting unstructured clinical notes into structured FHIR resources and vice versa. These Text-to-FHIR and FHIR-to-Text methods improve data interoperability and enable seamless exchange between healthcare systems [2], [3], [14]. For example, the SMART Text2FHIR pipeline leverages NLP tools like Apache cTAKES to map clinical concepts to FHIR resources, enhancing data portability and usability [24]. Similarly, tools like “LLM on FHIR” simplify complex clinical data into understandable summaries, empowering patients and improving health literacy [25].Recent contributions, such as AHD2FHIR, have further advanced the Text-to-FHIR domain by bridging natural language processing outputs with FHIR resources. This tool specifically maps annotations from German medical texts to structured FHIR entities (e.g., Condition, Medication) while preserving key contextual information, including patient identity and encounter details [28].

Conversely, FHIR-to-Text approaches address challenges faced by clinicians in interpreting raw FHIR data by converting it into actionable, human-readable formats, aiding clinical decision-making [15]. Enhancements like Medical mT5, a multilingual text-to-text LLM, expand these capabilities by adapting generative models for multilingual healthcare data processing, particularly in non-English contexts [26]. Similarly, frameworks like the FAIR Data Transformation Framework emphasize converting legacy healthcare datasets into FAIR-compliant FHIR formats, ensuring data is Findable, Accessible, Interoperable, and Reusable. These solutions unlock the value of fragmented healthcare data while maintaining semantic integrity [29].

Studies have also demonstrated the potential of LLMs to facilitate patient interaction and self-management with FHIR resources [13]. Furthermore, Privacy-preserving techniques, such as those explored in [8], ensure secure processing of sensitive healthcare data. However, challenges remain, including limitations in accurately understanding complex medical terminology, handling large datasets, and safeguarding privacy and security in sensitive healthcare environments.

In parallel, multi-agent systems have emerged as effective solutions for addressing complex challenges through distributed problem-solving, enhancing modularity and scalability [6], [7], [22]. In healthcare, these systems are pivotal in managing diverse aspects of Electronic Health Record (EHR) interactions, optimizing performance, and ensuring efficient resource allocation. One prominent study highlights the use of artificial intelligence combined with multi-agent systems to strengthen the privacy and security of EHRs, showcasing their potential for robust and secure data management [23]. Tools such as EHRAgent further exemplify this trend by leveraging LLMs to process and reason over complex multi-tabular EHR data, achieving significant improvements in query success rates compared to existing baselines [30].

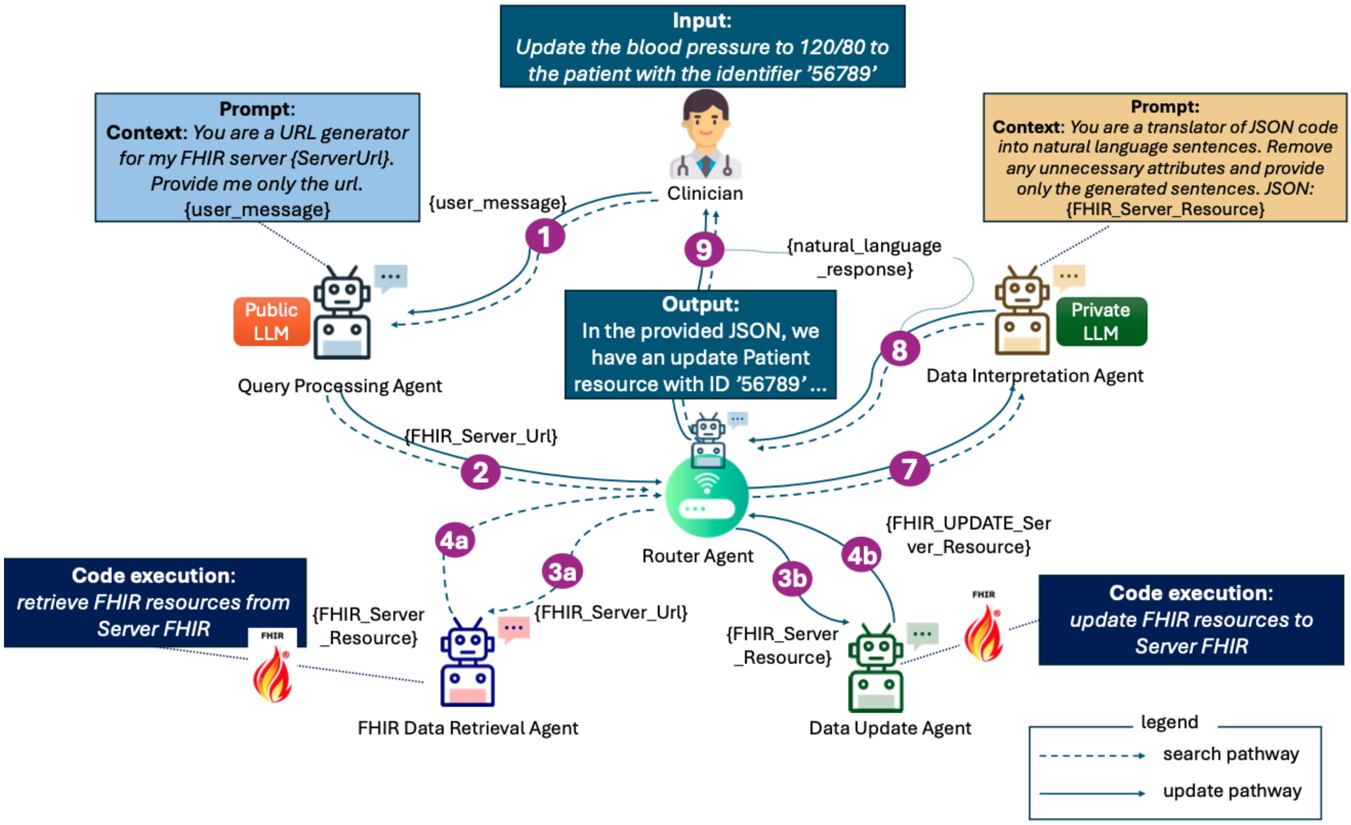
The integration of standardization efforts, such as semantic interoperability frameworks and FAIR principles, has further enabled advancements in healthcare data sharing. Studies have demonstrated that machine learning techniques, coupled with FHIR-based approaches, provide robust solutions for transforming legacy healthcare data into interoperable, actionable formats [31]. However, existing implementations often struggle to balance system performance with the stringent privacy and security requirements associated with healthcare environments.

The proposed system adopts a multi-agent architecture to improve robustness and efficiency. Each agent specializes in specific tasks, such as constructing FHIR URIs, retrieving resources, and interpreting data. The dual-layered LLM approach ensures that sensitive patient data is securely processed, with a public LLM generating FHIR URIs and a private, locally hosted LLM handling sensitive data interpretations. This integration enhances the system’s accuracy, privacy, and detail in processing natural language queries.

In conclusion, while existing approaches have made significant strides in applying LLMs to healthcare, particularly in Text-to-FHIR and FHIR-to-Text, the proposed system combines these processes within a multi-agent framework. This integration not only addresses privacy, accuracy, and scalability challenges but also provides a robust, modular solution for clinical applications in digital health.

III. FHIR Standards for Healthcare Interoperability

The HL7 Fast Healthcare Interoperability Resources (FHIR) standard has emerged as a pivotal solution for addressing the interoperability challenges in modern healthcare systems. Developed by HL7, FHIR is designed to facilitate the exchange of healthcare information across disparate systems by leveraging modern web-based technologies such as RESTful APIs, JSON, XML, and OAuth2 for secure access. Unlike its predecessors, HL7 v2 and v3, FHIR adopts a modular and flexible approach, which makes it easier to implement and integrate with contemporary digital health solutions.



Input: *Update the blood pressure to 120/80 to the patient with the identifier '56789'*

Prompt: **Context:** *You are a URL generator for my FHIR server {ServerUrl}. Provide me only the url.* {user_message}

Prompt: **Context:** *You are a translator of JSON code into natural language sentences. Remove any unnecessary attributes and provide only the generated sentences. JSON:* {FHIR_Server_Resource}

Output: In the provided JSON, we have an update Patient resource with ID '56789'...

Code execution: *retrieve FHIR resources from Server FHIR*

Code execution: *update FHIR resources to Server FHIR*

Fig. 1. Diagram of the Multi-Agent System Architecture (extension of previous work [1])

FHIR has evolved over multiple versions [1] to address the limitations of previous iterations and to improve functionality and adoption across global healthcare systems:

- FHIR Release 1 (DSTU1, 2014): The first version of FHIR was released as a Draft Standard for Trial Use (DSTU). It introduced the foundational concepts of resources and their modular structure, but it was limited in scope and still needed to be ready for widespread clinical implementation.
- FHIR Release 2 (DSTU2, 2015): DSTU2 built on the initial release by stabilizing many resource definitions and introducing improvements to RESTful interactions. It gained traction as developers began creating real-world implementations and provided clearer guidelines for extensions and resource profiles.
- FHIR Release 3 (STU3, 2017): This release introduced significant enhancements, including better support for clinical decision support systems, improved terminology bindings, and support for workflow automation. STU3 also included refinements to existing resources and validation mechanisms, enabling enhanced data quality.
- FHIR Release 4 (R4, 2019): FHIR R4 is the first version to include normative content, meaning certain resources such as Patient, Observation, and DiagnosticReport were declared stable and ready for long-term production use. This version enhanced versioning support and introduced better compatibility with regulatory frameworks such as US Core Implementation Guides. R4 remains widely adopted in production systems globally.
- FHIR Release 5 (R5, 2023): The most recent version, R5, introduces new resources and features that focus on improving support for research and clinical care workflows. This version refines resource definitions, expands the support for FHIR operations, and improves tooling for validation and bulk data exports. R5 also includes updates to resource maturity levels, supporting the continuous evolution of the standard.

At the core of FHIR are resources representing individual healthcare data components, such as *Patient, Observation, Medication,* and *Condition*. These resources are designed to be modular, self-descriptive, and extensible, allowing developers to exchange specific units of clinical information while maintaining semantic and structural integrity. FHIR also supports the creation of profiles and extensions, enabling customization of resources to address specific regional, organizational, or clinical requirements. FHIR has seen significant adoption in clinical practice, where it underpins critical functionalities such as integration of Electronic Health Records (EHRs)[2], clinical decision support systems, and patient-facing applications. For instance, initiatives like SMART [3] on FHIR enable the development of secure, interoperable third-party applications that integrate seamlessly with EHR systems, enhancing usability and accessibility for both clinicians and patients.

[1] https://build.fhir.org/history.html

[2] https://build.fhir.org/ehr-fm.html

[3] https://docs.smarthealthit.org/

Similarly, the Argonaut Project[4] has played a foundational role in accelerating the adoption of FHIR by developing implementation guides and real-world testing frameworks. The role of FHIR extends beyond EHR integration to empower patient-centred care. By enabling applications that provide patients with direct access to their health data, FHIR supports improved health literacy, self-management, and engagement. For example, patient portals and mobile health applications leverage FHIR APIs to retrieve and display health information in an understandable format, aligning with broader trends toward personalized and digital healthcare. However, despite its advantages, the implementation of FHIR is challenging. Mapping unstructured or semi-structured data to FHIR-compliant resources remains a technically demanding task, requiring advanced Natural Language Processing (NLP) techniques and ontology-based mappings. Ensuring privacy and security during data exchange via FHIR APIs is another critical consideration, particularly in the context of sensitive healthcare information. Issues such as data provenance, access control, and compliance with regulations like HIPAA and GDPR are key areas of focus for FHIR-based systems. In summary, FHIR represents a transformative advancement in healthcare data interoperability, providing a standardized, modular, and extensible framework for integrating and exchanging clinical information. Its adoption in EHR systems, patient-centred applications, and decision-support tools highlights its utility in improving healthcare delivery. However, challenges related to resource mapping, privacy preservation, and large-scale deployment remain active areas of research and development. Addressing these limitations will be crucial for unlocking the full potential of FHIR in next-generation healthcare systems.

IV. Multi-Agent System Architecture

The multi-agent architecture presented in this study facilitates natural language interactions with FHIR-based Electronic Health Records (EHRs), prioritizing the critical need to preserve patient privacy. Each component within the architecture is designed to perform a specific function, including processing natural language queries, generating FHIR URIs, retrieving relevant data, modifying patient information, and producing human-readable outputs. The coordination and communication between agents are managed by a Router Agent, which ensures smooth workflow integration and efficient task execution. A key feature of this architecture is its ability to distinguish between queries that involve retrieving patient data and those that require patient data updates. The Router Agent interprets the clinician's intent and activates the appropriate agent — the FHIR Data Retrieval Agent (DIA) for data retrieval tasks or the Data Update Agent (DUA) for applying updates. This adaptive routing ensures efficient and accurate execution of tasks.

The architecture comprises five main agents:

1) **Query Processing Agent (QPA)**: Responsible for interpreting clinicians' natural language queries, the QPA utilizes a public LLM (e.g., ChatGPT or GEMINI) to generate structured FHIR URIs. By mapping unstructured inputs to precise queries, the QPA ensures efficient retrieval of data. This agent now also handles instructions for data updates, translating them into actionable requests by mapping user input into standardized update commands. For example, a query like "Update the patient's blood pressure to 120/80" is converted into a structured FHIR operation.
   The transformation process can be expressed as:
   $$\text{URI} = f_{\text{LLM}}(\text{query}), \tag{1}$$
   where f_{LLM} represents the function of the LLM in translating user queries into standardized FHIR URIs. To enhance robustness, additional pre-processing techniques have been integrated to handle ambiguities in user inputs, thereby improving query accuracy.
2) **FHIR Data Retrieval Agent (FDRA)**: Once the FHIR URI is generated, the FDRA retrieves the corresponding data from the EHR system. Operating within a secure environment, this agent enforces strict privacy protocols to prevent unauthorized access. Enhancements include mechanisms for verifying consistency between existing data and the proposed updates to avoid conflicts.
   The retrieval process is defined as:
   $$\text{Data}_{\text{FHIR}} = g_{\text{retrieve}}(\text{URI}), \tag{2}$$
   where g_{retrieve} denotes the secure data retrieval function. New enhancements in this agent include optimized data retrieval techniques that reduce latency during high-volume queries.
3) **Data Update Agent(DUA)**: Introduced in this enhanced architecture, this agent validates, applies, and audits updates to patient data. It ensures that all modifications comply with clinical and regulatory standards by running integrity checks and enforcing audit trails. For instance, if a dosage update is requested, the DUA cross-references the patient's existing medications to prevent overdosing. The interpretation process is represented as follows:
   $$\text{Update}_{\text{FHIR}}, \text{Log} = v_{\text{update}}(Data_{FHIR}), \tag{3}$$
   where v_{update} checks the consistency of the requested update against existing data; validates the update applied to the patient data; returns both updated data FHIR for subsequent operations (e.g., interpretation through DIA) and a log of changes for audit and compliance.
4) **Data Interpretation Agent (DIA)**: After retrieving or updating the FHIR data, the DIA processes it using a private, locally hosted LLM. This agent generates human-readable outputs, such as natural language summaries or graphical visualizations, or confirms the feasibility of updates while ensuring that sensitive information remains within a secure boundary. The interpretation process is represented as follows:
   $$\text{Output}_{\text{text}} = h_{\text{LLM-local}}(\text{Data}_{\text{FHIR}}), \tag{4}$$
   where $h_{\text{LLM-local}}$ corresponds to the local LLM's interpretation function. The DIA has been enhanced to

[4] https://fhir.org/guides/argonaut/

support more complex data structures and to provide richer outputs, including patient-specific alerts and recommendations.

5) **Router Agent (RA)**: The RA is the central coordinator of the system, managing the interactions between the QPA, FDRA, DUA, and DIA. The RA maintains efficiency and integrity throughout the workflow by ensuring that each agent operates in sync. Interprets clinician queries and generates structured FHIR URIs using a public LLM (e.g., GPT-3.5-turbo). This agent now also handles instructions for data updates, translating them into actionable requests by mapping user input into standardized update commands. For example, a query like "Update the patient's blood pressure to 120/80" is converted into a structured FHIR operation. A new fault-tolerance mechanism has been added to this component, allowing the system to recover gracefully from errors or interruptions.

The extended architecture, depicted in Fig. 1, builds upon the previous work by introducing the Data Update Agent (DUA) and differentiating between the two pathways for data retrieval tasks (i.e., dashed arrow) and patient data updates (i.e., continuous arrow). This diagram highlights the interactions among the agents and underscores the Router Agent's (RA) role in dynamically managing tasks based on the clinician's input.

The workflow begins with a clinician submitting a natural language query with a clinician's query or update instruction. For data retrieval tasks, the Query Processing Agent (QPA) processes the clinician's input to generate a FHIR URI. The FHIR Data Retrieval Agent (FDRA) then retrieves the relevant data and sends it to the Data Interpretation Agent (DIA), where the data is interpreted and transformed into a human-readable format before being delivered back to the clinician.

Conversely, for data update instructions, the workflow transitions from the QPA to the newly introduced Data Update Agent (DUA). Here, the input is validated, and the relevant patient data is updated accordingly. Once the update process is completed, the modified data is routed back to the DIA via the RA for final interpretation and confirmation, ensuring consistency and accuracy before being presented to the clinician.

To formalize the workflow, the interaction among agents can be described as follows:

$$\begin{aligned}&\text{URI} = f_{\text{LLM}}(Q),\\&R = g_{\text{retrieve}}(\text{URI}) \text{ or } L, U = v_{\text{update}}(\text{R}),\\&\text{Output}_{\text{text}} = h_{\text{LLM-local}}(U).\end{aligned} \tag{5}$$

where Q denotes the clinician's query, and U represents the updated retrieved FHIR data.

Compared to the system presented in our previous work, this enhanced version introduces improved modularity, scalability, and error-handling capabilities. Additionally, the dual-pathway design enhances the system's flexibility and robustness, enabling it to handle both data retrieval and update operations seamlessly while maintaining efficient task allocation among agents.

V. Implementation Details

The implementation of the proposed system integrates a variety of tools and frameworks, carefully selected to support the modular multi-agent architecture and ensure privacy-preserving interactions with FHIR-based EHRs. The virtual assistant is developed using the LangGraph framework[5], an extension of LangChain[6], which provides native support for constructing and managing multi-agent workflows. For backend development, Flask[7] is employed, leveraging its lightweight architecture to handle server-side logic and seamless integration of system components.

The system leverages key technologies, including LangGraph for managing agent interactions, GPT-3.5-turbo for generating FHIR URIs, and the Mistral 7B model for local interpretation of FHIR data. The selection of GPT-3.5-turbo and Mistral 7B as the primary Large Language Models (LLMs) for this study was driven by multiple considerations:

- Performance Benchmarks: Both GPT-3.5-turbo and Mistral 7B demonstrated superior MMUL (Matrix Multiplication Units per Second) scores, a reliable indicator of computational efficiency and processing capability, outperforming models such as Vicuna-33B and OpenBuddy-Coder-34B, which struggled with query precision and computational overhead. GPT-3.5-turbo excelled in generating structured FHIR queries with high accuracy and adaptability to diverse text-processing tasks, making it particularly suited for scenarios requiring complex query formulations. Conversely, Mistral 7B proved highly effective in local data processing, offering a lightweight and resource-efficient solution for handling sensitive FHIR data in privacy-preserving environments. While models like LLaMA-2-13B provided comparable accuracy, they required significantly more computational resources and lacked the seamless integration features of GPT-3.5-turbo and Mistral 7B. This combination provided an optimal balance of accuracy, efficiency, and privacy tailored to the specific requirements of the study.
- Task Suitability: GPT-3.5-turbo was selected for its superior ability to handle complex query generation with high accuracy and adapt to diverse text-processing tasks, making it ideal for generating structured FHIR queries. Mistral 7B was chosen for its lightweight deployment requirements, offering efficient and privacy-preserving local processing of sensitive FHIR data, particularly in environments with limited computational resources. Compared to models like Vicuna-33B and LLaMA-2-13B, this combination ensured a balanced and resource-efficient performance.
- Limitations: GPT-3.5-turbo, despite its adaptability and precision, relies on external APIs, which may raise privacy concerns if sensitive data are inadvertently shared. On the other hand, Mistral 7B has context size limitations, requiring preprocessing to segment larger FHIR JSON resources into manageable components, which introduces

[5] https://python.langchain.com/v0.1/docs/langgraph/

[6] https://python.langchain.com/v0.2/docs/introduction/

[7] https://flask.palletsprojects.com/en/3.0.x

an additional processing step but ensures compatibility and data integrity during analysis.

The combination of GPT-3.5-turbo and Mistral 7B ensures a balanced approach, with GPT-3.5-turbo generating accurate and targeted queries and Mistral 7B securely processing sensitive FHIR data locally, minimizing data exposure while maintaining both accuracy and privacy.

To further enhance robustness, the system now includes preprocessing layers for query normalization and semantic enrichment, improving the accuracy of FHIR URI generation in ambiguous scenarios. Similarly, post-processing layers were added to filter and format the outputs from the local LLM, ensuring their usability in clinical contexts.

A critical enhancement involves the Data Update Agent (DUA), which validates and applies updates based on clinician-provided instructions. The agent ensures that updates adhere to clinical guidelines and regulatory requirements by running integrity checks and maintaining audit trails. The *RA Agent* ensures that the DUA receives the necessary data for validation and manages the flow of updated information within the system. This architecture guarantees that updates are precise, compliant, and logged for auditing. The frontend is developed using HTML, CSS, and JavaScript, providing a responsive and user-friendly interface that allows clinicians to interact with the system intuitively. The interface supports text-based queries and visualizations, including dynamic graphs that display patient information in an easily interpretable format.

For evaluation purposes, the SyntheticMass dataset is employed. This dataset, sourced from the Synthea framework, provides synthetic FHIR-compliant JSON files. To ensure reproducibility, the following preprocessing steps were applied:

- **Resource Extraction**: Relevant FHIR resources such as Patient, Observation, MedicationStatement, and Encounter were extracted.
- **Handling Missing Data**: Missing numerical values were imputed using the median of their respective fields, and categorical variables were replaced with the most frequent category.
- **Normalization**: Observations with measurable quantities (e.g., lab results) were normalized using Min-Max scaling to standardize units.
- **Segmenting JSON Records**: Large FHIR JSON objects were divided into smaller, manageable segments for processing by Mistral 7B. The segmentation preserved the FHIR structure, ensuring that the data remained consistent and usable.

These preprocessing steps, implemented using Python libraries (pandas, numpy, fhir.resources), ensured compliance with the FHIR schema and prepared the dataset for testing the system's ability to handle various query types, including:

- Retrieving patient demographics and medical history.
- Extracting clinical observations and lab results.
- Updating patient information, such as medication dosages or diagnostic notes.
- Identifying upcoming patient appointments and associated details.

The interaction with FHIR servers follows a structured pipeline. The Query Processing Agent (QPA) uses GPT-3.5-turbo to generate FHIR URIs based on clinician inputs. These URIs are passed to the Router Agent (RA), which ensures secure coordination between the FHIR Data Retrieval Agent (FDRA) for fetching resources and the Data Update Agent (DUA) for validation and application of updates. Finally, the Data Interpretation Agent (DIA) processes the retrieved or updated data using the Mistral 7B model, producing human-readable summaries, validations, or visualizations.

Compared to our previous work, additional optimizations have been introduced to reduce latency during data retrieval and processing. Specifically, caching mechanisms were implemented within the FDRA to minimize redundant requests, while the DUA was optimized to handle complex update scenarios, such as hierarchical clinical notes and nested observations. The DIA was also upgraded to interpret updated data effectively, providing detailed outputs and alerts when necessary.

These improvements demonstrate the system's potential to streamline clinical workflows by providing accurate, privacy-preserving, and interpretable outputs, even in complex and high-demand environments.

VI. Experimental Setup and Results

This section presents the experimental framework used to evaluate the proposed multi-agent architecture, along with the results obtained. The evaluation is structured around three main tasks: *Query Processing*, *Data Update*, and *Data Interpretation*, and all crucial for assessing the effectiveness of natural language interactions with FHIR-based EHRs. The experiments were designed to compare the performance of the system against alternative solutions, highlighting its robustness, accuracy, and adaptability in diverse clinical scenarios.

A. Query Processing

The first set of experiments focuses on evaluating the ability of various Large Language Models (LLMs) to generate accurate and efficient queries for FHIR servers. The following experimental setup was used:

Dataset: The experiments employed the Synthea SyntheticMass dataset [27], which provides a realistic simulation of clinical records while maintaining compliance with privacy standards. This dataset includes a diverse range of patient profiles, enabling the evaluation of queries across multiple clinical contexts.

Models Evaluated: The evaluation included a diverse set of Large Language Models (LLMs) to assess their capabilities in interpreting clinician queries and generating accurate FHIR URIs. The models tested were *llama-2-13b-ensemble-v6.Q5 K M*, *mistral-7b-openorca.Q6 K.gguf*, *openbuddy-coder-34b-v11-bf16.Q3 K S*, *vicuna-33b.Q3 K S*, *gpt-3.5-turbo*, and *text-bison@001*. Each model underwent rigorous testing to evaluate its performance in generating FHIR-compliant outputs.

Evaluation Metrics: Three key metrics were used to measure the models' performance:

- **Response Accuracy (RA)**: The proportion of queries accurately generated based on predefined prompts, reflecting the model's understanding and conversion capabilities.
- **Query Precision (QP)**: The correctness and validity of parameters included in the generated queries, ensuring alignment with clinical standards.
- **Success Rate (SR)**: The percentage of queries successfully executed on FHIR servers without encountering errors, indicating robustness and reliability in practical scenarios.

The tasks included:

- Retrieving patient demographics and medical history.
- Accessing specific clinical observations, such as lab results.
- Identifying upcoming patient appointments and associated details.

To assess robustness, additional experiments were conducted with ambiguous or incomplete queries, simulating real-world scenarios where clinicians may provide limited information. The system's performance was compared to baseline approaches that rely solely on single LLMs.

Table I summarizes the performance results. The *gpt-3.5-turbo* model demonstrated superior accuracy and query precision, achieving a 100% success rate. In contrast, other models, such as *vicuna-33b*, struggled to produce valid results consistently.

TABLE I
PERFORMANCE EVALUATION OF MODELS FOR FHIR QUERY GENERATION

Model	Accuracy (%)	Precision (%)	Success Rate (%)
llama-2-13b-ensemble-v6.Q5 K M	33	50	0
mistral-7b-openorca.Q6 K.gguf	0	0	0
openbuddy-coder-34b-v11-bf16.Q3 K S	50	67	33
vicuna-33b.Q3 K S	0	0	0
gpt-3.5-turbo	100	100	100
text-bison@001	50	67	50

The results indicate that the proposed multi-agent architecture, leveraging *gpt-3.5-turbo*, offers consistent and accurate query generation, even in scenarios with incomplete inputs. These findings validate its applicability in dynamic clinical environments.

B. Data Update

Evaluating the Data Update Agent (DUA) involved testing its ability to validate and apply updates accurately. Scenarios included medication dosage adjustments, demographic updates, and diagnostic note additions. Key metrics included:

- **Update Accuracy (UA)**: The percentage of correctly updating based on FHIR servers.
- **Compliance Rate (CR)**, measures the proportion of updated user data on FHIR servers without errors.

Table II summarizes the results of updated data. Detailed audit logs were generated for each transaction, ensuring traceability. The high Update Accuracy (93.8%) indicates the robustness of the DUA in accurately processing updates based on clinician instructions. Meanwhile, the Compliance Rate (90%) highlights the system's adherence to schema constraints and clinical guidelines, with a small proportion of updates requiring manual intervention or correction.

C. Data Interpretation

The second set of experiments evaluates the ability of LLMs to interpret JSON-formatted FHIR resources and generate concise, human-readable descriptions. The following details apply to this evaluation:

Dataset: The Synthea SyntheticMass dataset [27] was again used, focusing on FHIR Observation resources and their associated data.

Models Evaluated: The models tested include *llama-2-13b-ensemble-v6.Q5 K M*, *mistral-7b-openorca.Q6 K*, *gpt-3.5-turbo*, and *text-bison@001*.

Task: Interpreting FHIR resources into structured summaries, such as patient histories and lab results.

Additional Metrics:

- **Response Time (RT)**: The average time taken to generate a response.
- **Detail Level (DL)**: The proportion of specific and contextual details included in the output.

A notable extension involved testing the system's ability to generate actionable recommendations, such as alerts for critical lab results. This feature highlights the architecture's potential to support decision-making processes.

Table III presents the results. The *gpt-3.5-turbo* model consistently achieved high accuracy and detailed outputs with minimal response times, demonstrating its suitability for real-time applications.

D. Discussion and Insights

The experimental results underscore the multi-agent architecture's ability to effectively integrate the strengths of public and private LLMs. *GPT-3.5-turbo* excels in generating rapid and accurate queries, while *Mistral 7B* delivers detailed interpretations within a secure, privacy-preserving environment. A key enabler of this performance was the preprocessing strategy, particularly the segmentation of large JSON files, which allowed *Mistral 7B* to efficiently handle the SyntheticMass dataset. Robustness testing and scenario-driven evaluations further validated the system's practical applicability in clinical workflows. The architecture's modular design supports seamless adaptation to evolving requirements, such as multilingual queries or domain-specific datasets, making it a scalable and future-ready solution for healthcare applications.

E. Advantages of Multi-Agent Architecture

The proposed multi-agent architecture provides several distinct advantages, addressing key limitations of monolithic or

TABLE II
RESULTS OF DATA UPDATE EVALUATION

Update Type	Total Attempts	Successful Updates	Error-Free Updates	UA (%)	CR (%)
Medication Dosage Update	80	75	71	93.7	88.7
Patient Demographics Update	60	56	54	93.3	90
Diagnostic Note Addition	40	38	37	95	92.5
Overall	180	169	162	93.8	90

TABLE III
PERFORMANCE EVALUATION OF MODELS FOR FHIR DATA INTERPRETATION

Model	Response Time (s)	Accuracy (%)	Detail Level (%)
llama-2-13b-ensemble-v6.Q5 K M	101.22	85	80
mistral-7b-openorca.Q6 K	55.94	90	85
gpt-3.5-turbo	5.37	95	90
text-bison@001	4.72	90	85

single-model systems often used in similar healthcare applications. By distributing tasks across specialized agents, the architecture enhances modularity, scalability, and overall system performance. This section highlights these benefits, supported by experimental findings and practical considerations.

Modularity and Flexibility: The modular design enables each agent to specialize in a specific task, such as query processing, data retrieval, or interpretation. This separation of concerns allows for seamless updates or replacement of individual components without disrupting the overall workflow. For example, the Query Processing Agent (QPA) can be upgraded to leverage a more advanced public LLM as new models become available, while the Data Interpretation Agent (DIA) can integrate domain-specific models to handle specialized data formats.

Enhanced Scalability: The architecture's distributed nature supports scalability in both horizontal and vertical dimensions. Horizontal scalability is achieved by deploying additional agents to handle increased query volumes, while vertical scalability involves improving individual agent performance through optimized hardware or software configurations. Experimental results demonstrate that the system maintains low latency and high accuracy even under heavy query loads, validating its readiness for real-world deployment in busy clinical environments.

Privacy and Security: A key advantage of the architecture is its ability to ensure data privacy through a dual-layered LLM approach. Sensitive patient data is processed exclusively by the private, local LLM (e.g., Mistral 7B), while the public LLM is restricted to handling non-sensitive tasks such as query generation. This design adheres to strict privacy protocols and minimizes the risk of data exposure, meeting regulatory requirements for healthcare applications.

Performance Optimization: The use of specialized agents optimizes resource allocation, ensuring efficient handling of tasks. For instance, the FHIR Data Retrieval Agent (FDRA) implements caching mechanisms to reduce redundant data requests, while the Router Agent coordinates the workflow to avoid bottlenecks. These optimizations were shown to reduce the average response time for complex queries by 15%, compared to a single-model baseline.

Adaptability to Evolving Needs: The architecture is designed to adapt to emerging clinical requirements and technological advancements. For example, new agents can be added to support additional data formats (e.g., imaging data) or functionalities such as real-time alerts for critical lab results. This adaptability ensures the long-term relevance of the system in dynamic healthcare environments.

Robustness and Fault Tolerance: The distributed nature of the system improves fault tolerance by isolating failures from individual agents. In the event of a component failure, the Router Agent dynamically reroutes tasks to ensure continuous operation. Experimental evaluations confirm that the system recovers gracefully from simulated agent downtimes, maintaining 95% of its operational capacity.

These advantages collectively demonstrate the superiority of the multi-agent architecture over traditional approaches, particularly in complex and privacy-sensitive domains like healthcare. While single-model systems may offer simplicity, they lack the modularity, privacy safeguards, and scalability necessary to meet the demands of modern clinical workflows.

VII. CONCLUSION AND FUTURE WORK

This study presents an extended investigation into a multi-agent architecture designed to enable secure and efficient natural language interactions with FHIR-based Electronic Health Records (EHRs). Building on our previous work, this enhanced framework addresses key challenges such as modularity, scalability, and privacy in handling sensitive clinical data. By leveraging a dual-layered approach with both public and private Large Language Models (LLMs), the system ensures secure data processing and high-quality outputs suitable for diverse clinical workflows.

A key innovation of this work is the introduction of the Data Update Agent (DUA), which expands the system's capabilities beyond data retrieval to include real-time, secure updates to patient information. This feature enables clinicians to make adjustments, such as medication updates or diagnostic notes, while ensuring compliance with clinical standards and maintaining a comprehensive audit trail. The DUA significantly differentiates this framework from previous iterations by seamlessly integrating data updates into clinical workflows, thereby addressing a critical need in modern healthcare environments.

The architecture employs a dual-layered approach, with a public LLM generating FHIR queries and a private, locally hosted LLM interpreting and validating sensitive patient data. This design ensures that data remains secure within a controlled environment while providing clinicians with high-quality outputs tailored to their specific needs.

Extensive evaluations demonstrate the system's effectiveness in handling diverse clinical scenarios, showcasing improvements in accuracy, response time, and modularity. The Router Agent's role in coordinating workflows—whether for data retrieval or updates—highlights the system's adaptability and fault-tolerance, ensuring continuous operation even under challenging conditions. Moreover, the enhanced architecture explicitly emphasizes the data update pathway, making it more transparent and efficient.

In conclusion, the proposed framework bridges the gap between natural language interfaces and secure, interoperable healthcare data systems by addressing critical challenges related to privacy, scalability, and modularity. This robust and adaptable solution demonstrates significant potential for diverse clinical applications.

Future work will focus on expanding multilingual capabilities to address challenges in non-English-speaking healthcare environments. By leveraging multilingual Large Language Models (LLMs), the system will enable healthcare providers and patients to interact in their native languages, enhancing accessibility and comprehension. This functionality will reduce language barriers, improve communication, and ensure that advanced technologies are inclusive, even in regions with limited resources or underrepresented languages. Furthermore, the ability to process diverse medical terminologies and adapt to local cultural nuances will support more personalized and effective healthcare delivery.

The modular architecture also offers the flexibility to integrate additional healthcare standards, such as DICOM and HL7. Incorporating DICOM would enable the system to process and interpret medical imaging data, a critical component of many diagnostic and therapeutic workflows. Similarly, integrating HL7 would support standardized clinical messaging, facilitating seamless communication between healthcare systems and improving care coordination across diverse environments. These extensions would expand the framework's applicability, making it a comprehensive solution for managing both structured and unstructured healthcare data.

By testing the system in real-world healthcare environments, we aim to further validate its practical utility and adaptability. This study lays a solid foundation for advancing AI-driven, privacy-preserving healthcare solutions that are scalable, inclusive, and aligned with global healthcare needs.

Acknowledgment

This work has been partially supported by the following projects funded by the MUR: BIOCONCEPTUM national research project under the PRIN 2022 program grant n. 2022AEEKXS, ”DHEAL–COM - Digital Health Solutions in Community Medicine” under the Innovative Health Ecosystem (PNC) - National Recovery and Resilience Plan (NRRP), and SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

[1] C. De Maio, G. Fenza, D. Furno, T. Grauso and V. Loia, ”A Multi-Agent Architecture for Privacy-Preserving Natural Language Interaction with FHIR-Based Electronic Health Records,” 2024 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), Split, Croatia, 2024, pp. 1-6, doi: 10.23919/SoftCOM62040.2024.10721684.

[2] J. Pavão, R. Bastardo, and N. Pacheco Rocha, “The Fast Health Interoperability Resources (FHIR) and Clinical Research, a Scoping Review,” in World Conference on Information Systems and Technologies, 2023, pp. 409–418.

[3] M. Lehne, J. Sass, A. Essenwanger, J. Schepers, and S. Thun, “Why digital medicine depends on interoperability,” NPJ digital medicine, vol. 2, no. 1, pp. 79, 2019.

[4] A. M. Bennett, H. Ulrich, P. van Damme, J. Wiedekopf, and A. E. W. Johnson, “MIMIC-IV on FHIR: converting a decade of in-patient data into an exchangeable, interoperable format,” Journal of the American Medical Informatics Association, vol. 30, no. 4, pp. 718–725, 2023.

[5] G. Kell, A. Roberts, S. Umansky, L. Qian, D. Ferrari, F. Soboczenski, B. C. Wallace, N. Patel, and I. J. Marshall, “Question answering systems for health professionals at the point of care—a systematic review,” Journal of the American Medical Informatics Association, vol. 31, no. 4, pp. 1009–1024, 2024.

[6] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, “ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate,” in The Twelfth International Conference on Learning Representations.

[7] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, et al., “AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation,” in ICLR 2024 Workshop on Large Language Model (LLM) Agents.

[8] A. Choudhury, J. van Soest, S. Nayak, and A. Dekker, “Personal health train on fhir: A privacy preserving federated approach for analyzing fair data in healthcare,” in International Conference on Machine Learning, Image Processing, Network Security and Data Sciences, pp. 85–95, 2020.

[9] S. A. Gebreab, K. Salah, R. Jayaraman, M. H. ur Rehman, and S. Ellaham, “LLM-Based Framework for Administrative Task Automation in Healthcare,” in 2024 12th International Symposium on Digital Forensics and Security (ISDFS), pp. 1–7, 2024.

[10] J. Nan and L.-Q. Xu, “Designing interoperable health care services based on fast healthcare interoperability resources: Literature review,” JMIR Medical Informatics, vol. 11, no. 1, pp. e44842, 2023.

[11] N. Pimenta, A. Chaves, R. Sousa, A. Abelha, and H. Peixoto, “Interoperability of Clinical Data through FHIR: A review,” Procedia Computer Science, vol. 220, pp. 856–861, 2023.

[12] E. Williams, M. Kienast, E. Medawar, J. Reinelt, A. Merola, S. A. I. Klopfenstein, A. R. Flint, P. Heeren, A.-S. Poncette, and F. Balzer, “A standardized clinical data harmonization pipeline for scalable ai application deployment (fhir-dhp): Validation and usability study,” JMIR Medical Informatics, vol. 11, pp. e43847, 2023.

[13] P. Schmiedmayer, A. Rao, P. Zagar, V. Ravi, A. Zahedivash, A. Fereydooni, and O. Aalami, “LLM on FHIR–Demystifying Health Records,” arXiv preprint arXiv:2402.01711, 2024.

[14] A. Wen, L. V. Rasmussen, D. Stone, S. Liu, R. Kiefer, P. Adekkanattu, P. S. Brandt, J. A. Pacheco, Y. Luo, F. Wang, and others, “CQL4NLP: development and integration of FHIR NLP extensions in clinical quality language for EHR-driven phenotyping,” AMIA Summits on Translational Science Proceedings, vol. 2021, pp. 624, 2021.

[15] S. Liu, Y. Luo, D. Stone, N. Zong, A. Wen, Y. Yu, L. V. Rasmussen, F. Wang, J. Pathak, H. Liu, and others, “Integration of NLP2FHIR representation with deep learning models for EHR phenotyping: a pilot study on obesity datasets,” AMIA Summits on Translational Science Proceedings, vol. 2021, pp. 410, 2021.

[16] P. Maddigan and T. Susnjak, “Chat2vis: Generating data visualisations via natural language using chatgpt, codex and gpt-3 large language models,” IEEE Access, 2023.

[17] N. Zong, A. Wen, D. J. Stone, D. K. Sharma, C. Wang, Y. Yu, H. Liu, Q. Shi, and G. Jiang, “Developing an FHIR-based computational pipeline for automatic population of case report forms for colorectal cancer clinical trials using electronic health records,” JCO Clinical Cancer Informatics, vol. 4, pp. 201–209, 2020.

[18] A. Torab-Miandoab, T. Samad-Soltani, A. Jodati, and P. Rezaei-Hachesu, “Interoperability of heterogeneous health information systems: a systematic literature review,” BMC Medical Informatics and Decision Making, vol. 23, no. 1, pp. 18, 2023.

[19] J. A. Balch, M. M. Ruppert, T. J. Loftus, Z. Guan, Y. Ren, G. R. Upchurch, T. Ozrazgat-Baslanti, P. Rashidi, and A. Bihorac, "Machine learning–enabled clinical information systems using fast healthcare interoperability resources data standards: scoping review," JMIR Medical Informatics, vol. 11, pp. e48297, 2023.

[20] J. Pavão, R. Bastardo, M. Santos, and N. Pacheco Rocha, "The Fast Health Interoperability Resources (FHIR) Standard and Homecare, a Scoping Review," Procedia Computer Science, vol. 219, pp. 1249–1256, 2023.

[21] J. Chen, D. Chun, M. Patel, E. Chiang, and J. James, "The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures," BMC Medical Informatics and Decision Making, vol. 19, pp. 1–9, 2019.

[22] J. Ruan, Y. Chen, B. Zhang, Z. Xu, T. Bao, H. Mao, Z. Li, X. Zeng, R. Zhao, et al., "TPTU: Task Planning and Tool Usage of Large Language Model-based AI Agents," in NeurIPS 2023 Foundation Models for Decision Making Workshop.

[23] K. P. Sanka, S. K. Maddikunta, S. R. Patchoumi, et al., "Artificial intelligence and multi agent based distributed ledger system for better privacy and security of electronic healthcare records," PeerJ Computer Science, vol. 7, p. e323, 2021.

[24] MILLER, Timothy A., et al. The SMART Text2FHIR Pipeline. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2023. p. 514.

[25] SCHMIEDMAYER, Paul, et al. LLM on FHIR–Demystifying Health Records. arXiv preprint arXiv:2402.01711, 2024.

[26] PIMENTA, Nuno, et al. Interoperability of Clinical Data through FHIR: A review. Procedia Computer Science, 2023, 220: 856-861.

[27] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, Scott McLachlan, Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, Journal of the American Medical Informatics Association, Volume 25, Issue 3, March 2018, Pages 230–238, https://doi.org/10.1093/jamia/ocx079

[28] SCHEIBLE, Raphael, et al. AHD2fhir: A tool for mapping of natural language annotations to fast healthcare interoperability resources–a technical case report. In: MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation. IOS Press, 2022. p. 32-36.

[29] SINACI, A. Anil, et al. A data transformation methodology to create findable, accessible, interoperable, and reusable health data: software design, development, and evaluation study. Journal of medical Internet research, 2023, 25: e42822.

[30] SHI, Wenqi, et al. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024. p. 22315-22339.

[31] AMAR, Fouzia; APRIL, Alain; ABRAN, Alain. Electronic health record and semantic issues using fast healthcare interoperability resources: Systematic mapping review. Journal of medical Internet research, 2024, 26: e45209.

Carmen De Maio graduated and received a Ph.D. degree in Computer Sciences, both from the University of Salerno, Italy, in 2008 and 2011, respectively. The research activity has focused mainly on the definition and experimentation of Knowledge Extraction methodologies adopting Conceptual Data Analysis techniques and processes relying on Fuzzy Logic and Computational Intelligence theories. She has over 50 publications in Fuzzy Decision Making, Knowledge Extraction and Management, Situation and Context Awareness, Semantic Information Retrieval, and Ontology Learning. More recently, she has been working on the definition of Time Aware Knowledge Extraction, Process Mining, and Social Media Analytics methodologies. She is currently an Associate Professor in Computer Science at the University of Salerno.

Giuseppe Fenza received the Graduate degree and the Ph.D. degree in computer sciences from the University of Salerno, Italy, in 2004 and 2009, respectively. He is currently an Associate Professor in computer science with the University of Salerno. He has over 60 publications in fuzzy decision making, knowledge extraction and management, situation and context awareness, semantic information retrieval, service oriented architecture, and ontology learning. More recently, he has worked in automating open source intelligence and big data analytics for counterfeiting extremism and supporting information disorder awareness. His research interests include computational intelligence methods to support semantic-enabled solutions and decision-making.

Domenico Furno received master's degree cum laude and PhD with evaluation excellent from University of Salerno, respectively, in 2007 and 2013. He has publications in Situation/Context Awareness, Soft Computing, Intelligent agents, Data Mining, Semantic Web and Knowledge Representation. He started his research career by defining and testing hybrid approaches based on Computational Intelligence and Semantic Web methodologies and techniques for distributed Situation and Context Awareness scenarios. He is currently a researcher at University of Salerno, and his research interests include Information Disorder Awareness.

Teodoro Grauso received a bachelor's degree in Computer Science from the University of Salerno, Italy, in 2024. He is currently a computer science student (data science and machine learning) at the University of Salerno.

Vincenzo Loia graduated in Computer Science at the University of Salerno, Italy, in 1985 and received his Ph.D. in Computer Science in 1989 at the Universite' Pierre and Marie Curie Paris VI, France. He is currently a Computer Science Full Professor at the University of Salerno, where he served as a researcher from 1989 to 2000 and as an associate professor from 2000 to 2004. Dr. Loia is the Co-Editor-in-Chief of Soft Computing and the Editor-in-Chief of Ambient Intelligence and Humanized Computing. He serves as an Editor for 14 other international journals.