

Article

Enhancing E-Government Services through State-of-the-Art, Modular, and Reproducible Architecture over Large Language Models

George Papageorgiou ¹, Vangelis Sarlis ¹, Manolis Maragoudakis ^{2,*} and Christos Tjortjis ¹

¹ School of Science and Technology, International Hellenic University, 57001 Thessaloniki, Greece; gpapageorgiou2@ihu.edu.gr (G.P.); e.sarlis@ihu.edu.gr (V.S.); c.tjortjis@ihu.edu.gr (C.T.)

² Department of Informatics, Ionian University, 49100 Corfu, Greece

* Correspondence: mmarag@ionio.gr

Abstract: Integrating Large Language Models (LLMs) into e-government applications has the potential to improve public service delivery through advanced data processing and automation. This paper explores critical aspects of a modular and reproducible architecture based on Retrieval-Augmented Generation (RAG) for deploying LLM-based assistants within e-government systems. By examining current practices and challenges, we propose a framework ensuring that Artificial Intelligence (AI) systems are modular and reproducible, essential for maintaining scalability, transparency, and ethical standards. Our approach utilizing Haystack demonstrates a complete multi-agent Generative AI (GAI) virtual assistant that facilitates scalability and reproducibility by allowing individual components to be independently scaled. This research focuses on a comprehensive review of the existing literature and presents case study examples to demonstrate how such an architecture can enhance public service operations. This framework provides a valuable case study for researchers, policymakers, and practitioners interested in exploring the integration of advanced computational linguistics and LLMs into e-government services, although it could benefit from further empirical validation.

Keywords: AI Governance; e-government; generative artificial intelligence (GAI); modularity; large language models (LLMs); reproducibility; retrieval-augmented generation (RAG)



Citation: Papageorgiou, G.; Sarlis, V.; Maragoudakis, M.; Tjortjis, C.

Enhancing E-Government Services through State-of-the-Art, Modular, and Reproducible Architecture over Large Language Models. *Appl. Sci.* **2024**, *14*, 8259. <https://doi.org/10.3390/app14188259>

Academic Editor: Luis Javier Garcia Villalba

Received: 6 August 2024

Revised: 9 September 2024

Accepted: 11 September 2024

Published: 13 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction & Background

The rapid advancement of Artificial Intelligence (AI) technologies has created unique opportunities for enhancing public service delivery, business operations, and governance. Large Language Models (LLMs) have emerged as powerful tools capable of transforming a wide range of applications through their advanced data processing and natural language understanding capabilities. This paper explores innovative approaches that leverage LLMs to improve efficiency, accuracy, and accessibility across various domains, with a focus on government services, business applications, and other processes.

The integration of LLMs into e-government services represents a paradigm shift in public administration, leveraging advanced data processing and automation to significantly enhance service delivery. These models hold substantial potential for optimizing operations, increasing efficiency, and improving citizen interactions through more intelligent and responsive systems [1].

The lack of utilization of AI-based systems in e-government can lead to significant losses and missed opportunities for governments. In terms of competitiveness and innovation, governments that fail to adopt AI-based systems risk falling behind their more technologically advanced counterparts, both domestically and globally. Governments also miss opportunities for cost savings by relying on manual processes that require more human resources, leading to higher operational costs. The absence of automation and optimization prevents them from realizing potential cost savings [2]. Furthermore, without

AI-powered data analysis, governments miss out on valuable insights that could inform policy decisions and service improvements. This results in missed opportunities to identify patterns, trends, and areas for optimization within public service delivery [3].

AI-based systems can help tackle challenges in public sector activities, including education, health services, sustainability, public transportation, security services and supervision, waste management, management of electricity and water resources, and the integration of various public services [4]. Embracing the digital transformative potential of AI-powered technologies can enable governments to deliver better public services, make more informed decisions and ultimately, better serve their citizens [5].

The research landscape of LLMs has been marked by significant advancements that have expanded their applicability and efficiency. A notable contribution in this space is the development of an efficient AI model tailored for small- and medium-sized enterprises (SMEs), utilizing the BLOOM approach for multilingual adaptability in English and Vietnamese. By implementing Low-Rank Adaptation (LoRA) and DeepSpeed, this study reduces computational costs and training time, making advanced AI accessible to SMEs with limited resources. The model's domain-specific customization and robust performance validation in smart city applications further demonstrate its versatility and cost-efficiency [6,7].

In the orbit of AI governance, a conceptual framework has been proposed to address the complexities of regulating AI and autonomous systems. This layered approach divides governance into social, legal, ethical, and technical layers, ensuring a comprehensive and adaptable system. The framework emphasizes transparency, accountability, and inclusivity, providing a structured method to navigate the challenges posed by AI technologies. This study advances current methodologies by integrating modularity and layered governance, offering a scalable and flexible model applicable across various contexts [8].

The integration of LLMs with traditional methods to enhance intelligent Q&A systems in the government sector represents another significant advancement. The Large Language Model Help (LLMH) framework leverages the contextual understanding capabilities of LLMs and combines them with traditional retrieval methods to address challenges such as intent recognition and transfer in government-related queries. This integration ensures accurate and contextually relevant answers, significantly improving the performance of existing systems and offering a robust solution for handling complex, multi-round dialogue questions in digital government applications [9].

A transformative approach to enhancing the accessibility and usability of Open Government Data (OGD) through LLMs is demonstrated by employing the GPT-3.5 model to interact with the Scottish open statistics portal. This study shows how LLMs can transform natural language queries into precise SPARQL queries, enabling non-technical users to retrieve accurate statistical information effortlessly. The innovative application of Retrieval Augmented Generation (RAG) ensures factual correctness in responses and proposes a scalable, user-friendly solution for democratizing access to government data [10].

The handling of the massive entry of citizen complaints in e-government systems using advanced data mining algorithms showcases another practical application of LLMs. By comparing various algorithms such as k-Nearest Neighbors, Random Forest, Support Vector Machine, and AdaBoost on a large dataset from the LAKSA app in Tangerang City, Indonesia, the study identifies the most accurate classifier and highlights the importance of continuous supervised training. This research demonstrates the potential of data-driven methods to improve the efficiency and accuracy of complaint management, enhancing citizen satisfaction and offering a scalable solution for smart governance [11].

In the business sector, the development of the Persuasive Message Intelligence (PMI) service introduces advanced Generative AI (GAI) technologies for automated personalized marketing. Utilizing prompt engineering techniques based on persuasive marketing theories and validated through surveys, PMI generates tailored marketing messages that significantly enhance customer engagement. This study addresses the high costs and

complexity of manual message creation by leveraging OpenAI's GPT-4 model, ensuring cost-effective and scalable solutions [12].

Another study presents AsasaraBot, an AI chatbot designed to enhance Content and Language Integrated Learning (CLIL) by teaching cultural content about the Minoan Civilization and facilitating language learning in English and French. This innovative approach leverages AI for interactive and personalized education, supporting human tutors and addressing remote learning challenges. Evaluated in Greek schools, AsasaraBot demonstrates significant advancements in ICT-based education, contributing to improved methodologies in educational technology and language acquisition [13].

Additionally, a novel approach for federal agencies to systematically report on their AI use cases has been introduced, emphasizing transparency, accountability, and effective governance. This draft guidance provides a comprehensive set of criteria and mechanisms for annual AI inventories, focusing on the management of AI-related risks and ethical implications. The framework's innovative approach to safety-impacting and rights-impacting AI use cases sets a new standard for AI governance within the federal sector [14].

The transformation of legal texts into machine-executable computational logic represents another significant advancement. By utilizing Natural Language Processing (NLP) techniques and a structured document inference process, legal provisions are converted into Prolog predicates, making legal rules transparent and explainable. This method addresses the challenges of ambiguity and context dependency in legal texts, providing a foundation for automated legal reasoning and decision-making [15,16].

Finally, the use of LLMs in transforming legal texts into digitized public services via No-Code/Low-Code platforms showcases the potential for reducing manual effort in process digitization. By integrating advanced NLP techniques and formal process modeling standards, this research enhances transparency, efficiency, and accessibility in public service delivery [17,18].

Reproducibility, on the other hand, ensures that LLM deployments can be consistently replicated across different contexts and use cases, maintaining uniformity in performance and reliability. This is essential for upholding transparency, scalability, and ethical standards in AI applications within government systems [1,19].

Other proposed frameworks aim to guide policymakers and practitioners in implementing LLMs effectively, maximizing their benefits while mitigating potential risks. It underscores the importance of collaboration among technical experts, policymakers, and end-users to develop AI systems that are not only efficient but also aligned with public interest and ethical standards [20,21].

In summary, the diverse applications and innovative approaches presented in this paper highlight the transformative potential of LLMs across various sectors. By addressing key challenges and leveraging advanced AI techniques, these studies contribute significantly to the state of the art, offering practical solutions and a detailed roadmap for future research and implementation in public administration, business, and legal contexts. In conclusion, the integration of LLMs into e-government services offers a unique opportunity to enhance public service delivery through advanced AI technologies [21].

By examining best practices and case studies from various sectors, this study proposes a framework that combines modularity and reproducibility to create robust, scalable, and ethical AI solutions for e-government services. By focusing on modularity and reproducibility, this study provides a broad approach to deploying these models, ensuring scalability, transparency, and ethical integrity. The following sections will examine current practices, challenges, and proposed solutions, offering actionable recommendations for future implementations.

1.1. Aims and Objectives

This research aims to advance a modular way to integrate the implications of LLMs [22], RAG [23], and Agents [24] in e-government applications, assessing their potential to enhance public service delivery through advanced data processing and automation. By

highlighting the critical aspects of modularity and reproducibility in the deployment of LLMs and RAG within government systems, this study looks to identify key factors that ensure these qualities. Furthermore, it proposes a framework that ensures public domain and government AI systems are modular, reproducible, scalable, transparent, and ethical, based on a structured approach to implementation.

Starting with the examination of current practices and challenges in deploying LLMs and RAG in government systems through a comprehensive review of the existing literature and case studies, we will provide and highlight best practices for ensuring modularity and reproducibility by identifying successful examples and methodologies from various government sectors. Additionally, this study will provide actionable recommendations for policymakers and practitioners, offering practical strategies for deploying and managing AI systems in e-government applications. Emphasizing the importance of scalability, transparency, and ethical standards in AI deployment, this research aims to ensure that AI systems align with societal values and applied regulations.

1.2. Study Significance

This study's importance lies in its great benefit to the public sector and the e-government implications of using GAI. Integrating LLMs and RAG into e-government applications can significantly enhance public service delivery and productivity, leading to faster processing times, reduced costs, and improved citizen satisfaction in the public sector's applications. GAI technologies enable public organizations responsible for implementing laws, providing public services, and enforcing regulations (such as government agencies) to make quicker, more informed decisions based on comprehensive data analysis, ultimately improving outcomes for the public.

By proposing a framework that ensures that an AI system is modular, meaning that it can be separated into interchangeable components and is reproducible, we aim to support scalable, adaptable, and sustainable solutions. This modularity allows government systems to evolve with changing needs, ensuring they remain responsive and effective over the long term. Ultimately, this study aims to provide a framework that enhances the efficiency and effectiveness of public services and upholds the principles of transparency, scalability, and ethical governance.

2. Technical Architecture and Solution

The proposed technical architecture is within the current state of the art and leverages the latest advances in RAG and agent-based implementations to enhance e-government applications. At its core, the system employs the best-known and established LLMs with memory capabilities and the open-source Haystack LLM orchestration framework [25], enabling more sophisticated data processing and automation. Additionally, the system implements multiple pre-processing and post-processing methods to refine the data in indexing and after retrieval.

Data management is based on a vectorized database, specifically Pinecone (accessed on 1 July) [26], which efficiently manages multiple indexed data sources, ensuring rapid and accurate information retrieval. The primary data sources integrated into our system include Press Corner [27], the Organisation for Economic Cooperation and Development (OECD) library [28], Web Search, and in-memory user documents, all vectorized to facilitate seamless access and processing.

The architecture is based on Haystack's framework [25], a powerful toolset that supports the integration and management of LLM operations. For web search capabilities, we incorporate Serper (accessed on 1 July) [29], an advanced tool that enhances our system's ability to retrieve relevant web-based information. The next sections will detail the methodologies and techniques employed in each component, providing a comprehensive overview of our system's architecture and implementation.

2.1. Architecture Overview

The proposed technical solution is based on the Haystack framework [25] but can be applied straightforwardly with similar frameworks. Each architecture component is considered modular and can be directly modified with alternative options. Additionally, the data sources are selected as examples of e-government applications that could enhance the productivity of public sector employees in their daily tasks or be directly used by citizens based on their interests in various public sectors.

2.2. RAG Integration

The framework is constructed based on two pipelines: the indexing pipeline and the query pipeline. These pipelines enhance their modularity through multiple applications in different contexts. Starting with the indexing pipeline, it is responsible for indexing documents into the vectorized database after the files are acquired. Before the indexing process itself, various preprocessing steps are applied to prepare the files in the correct form. The main components of the indexing pipeline include setting up the vectorized database, preprocessing steps before indexing (such as dividing documents into smaller parts), and indexing using embeddings. We utilized multiple data sources, which will be detailed in the following subsections.

The query pipeline is responsible for generating responsive answers to user queries. We employed RAG to generate answers based on the retrieved documents. Additionally, we introduced different agents, each dedicated to specific indexed sources. We also incorporated conversational memory capabilities into our system. When the passage with the retrieved documents is delivered to the model, memory generated by another LLM is included. Lastly, by utilizing the query pipeline, RAG, conversational memory, and agents, users can access a pipeline tailored to their needs. This setup makes the “steps of thought” of the system visible, including the “Thought” and “Observation” tabs, allowing users to understand the reasoning behind the agents’ actions and the steps before and after providing the definitive answer.

The performance evaluation of the proposed architecture is based thoroughly on the components used. The vectorized database should be configured according to the performance requirements, with advice from the provider documentation. The data must be of sufficient quality to be processed by the modular preprocessor, which should be configured according to its characteristics. The performance of the embeddings and retrieval process depends on the chosen embedding retriever; the better-performing the embedding model, the more precise the retrieval process. Finally, the Q&A and reasoning processes in the conversational and RAG performance depend on the selected GAI models. Each actor should consult the provider’s documentation to select the appropriate model based on the requirements. Therefore, for component selection, the actor should refer to the provider’s performance evaluation documentation and make decisions based on the specific needs of the use case.

2.2.1. Indexing Pipeline

The indexing pipeline presented in Figure 1 is a critical component of our system, and it is responsible for processing and indexing documents into the vectorized database, ensuring that the data are efficiently stored and readily retrievable, enhancing the overall performance and usability of the system.

The first step in the indexing pipeline is setting up the document store. In our implementation, we use a vectorized database to store document embeddings. This database is configured to use cosine similarity, which is particularly effective for comparing the document’s vectorized representations. The setup includes defining the environment and securing access through an API key, ensuring efficiency and security. In addition, to promote the reproducibility and the modularity of our framework, different database solutions can be used (e.g., Astra, Chroma, Elasticsearch, Milvus, MongoDBAtlas, Neo4j, OpenSearch, Pgvector, Qdrant, and Weaviate).

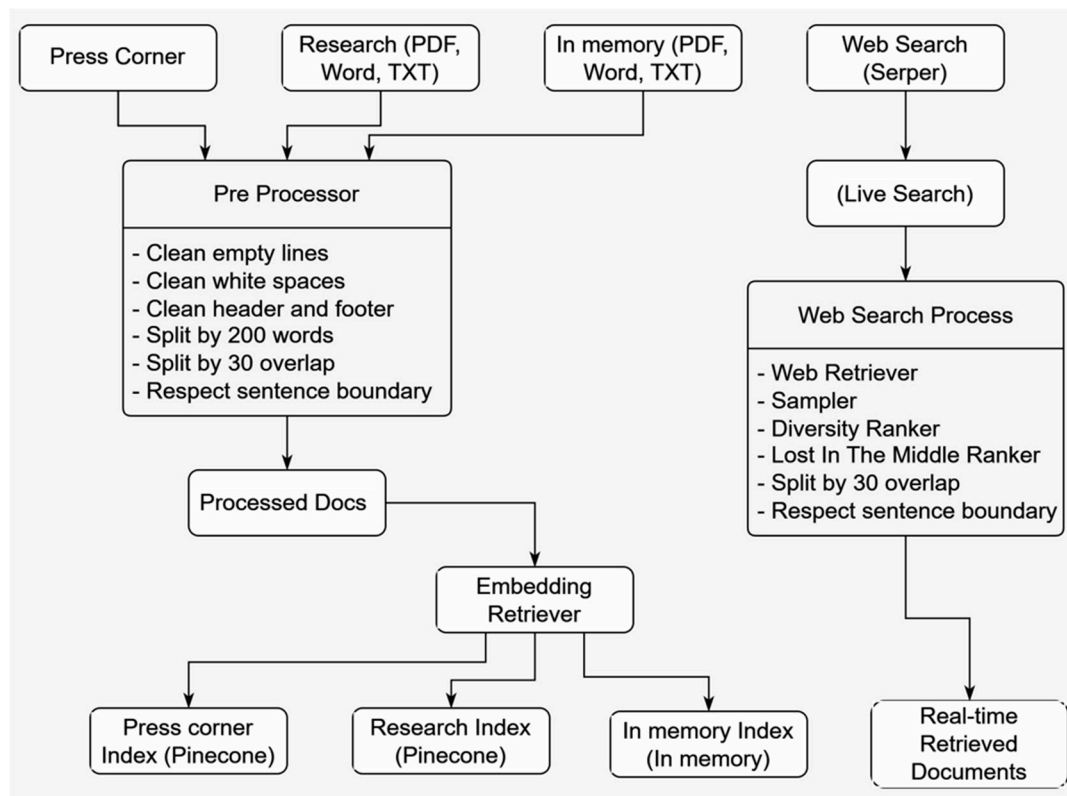


Figure 1. Diagram of Document indexing and Web search processes.

Once the document store is configured, the next step is preprocessing the documents. Preprocessing involves several sub-steps to prepare the documents for indexing. Initially, the documents are cleaned by removing empty lines, unnecessary whitespace, and extraneous headers and footers. After the data cleansing, the documents are split into smaller chunks to facilitate more efficient processing and retrieval. This splitting respects sentence boundaries and includes a sliding window approach to maintain context between chunks.

After preprocessing, the documents are ready for embedding. Each document chunk is transformed into a vectorized representation using an advanced embedding model in this step. For our framework, we used text-embedding-ada-002 [30], but this is again under each preference, using various models imported from Hugging Face, OpenAI, Anthropic, and others. These embeddings capture the semantic meaning of the text, making it possible to perform effective similarity searches later.

The indexing pipeline is designed to be modular, allowing each component to be easily modified or replaced as needed. This flexibility ensures that the system can adapt to different data sources and requirements, maintaining high performance and scalability to support various e-government applications. The main components mentioned above are presented in Figure 1.

2.2.2. Querying Pipeline

The querying pipeline is a critical component of our system, designed to generate responsive answers to user queries. This pipeline employs RAG and integrates GPT-3.5 and GPT-4o models to provide accurate and relevant information from various sources, including research documents from the OECD library (PDFs, word, txt formats), Press Corner releases accessible from Press Corner, and Web search results, retrieved real-time under user query and in-memory documents (PDFs, word, txt formats), which the user can upload real-time and be embedded.

The querying pipeline begins with an embedding retriever that searches the vectorized database for the most relevant documents based on the user's query. For all pipelines except

Web Search, the retriever identifies the most relevant documents, which are then passed to a prompt node. This prompt node, configured with a predefined template, synthesizes a comprehensive answer using the GPT-3.5 [31] or GPT-4o [32] models. The prompt extracts relevant information from the retrieved documents and generates a concise, contextually accurate answer.

For web search results, the querying pipeline incorporates additional components to manage the dynamic nature of web data. It starts with a web retriever that performs a comprehensive search to gather relevant web pages using Serper (modular). The retrieved web pages are then sampled and ranked based on diversity and relevance. We used the Lost in the Middle ranker, which further refines this set based on content length and relevance, as presented in Figure 1. Finally, the prompt node generates a synthesized answer from the top-ranked web snippets. The main components of the abovementioned RAG query and pipeline processes are presented in Figure 2.

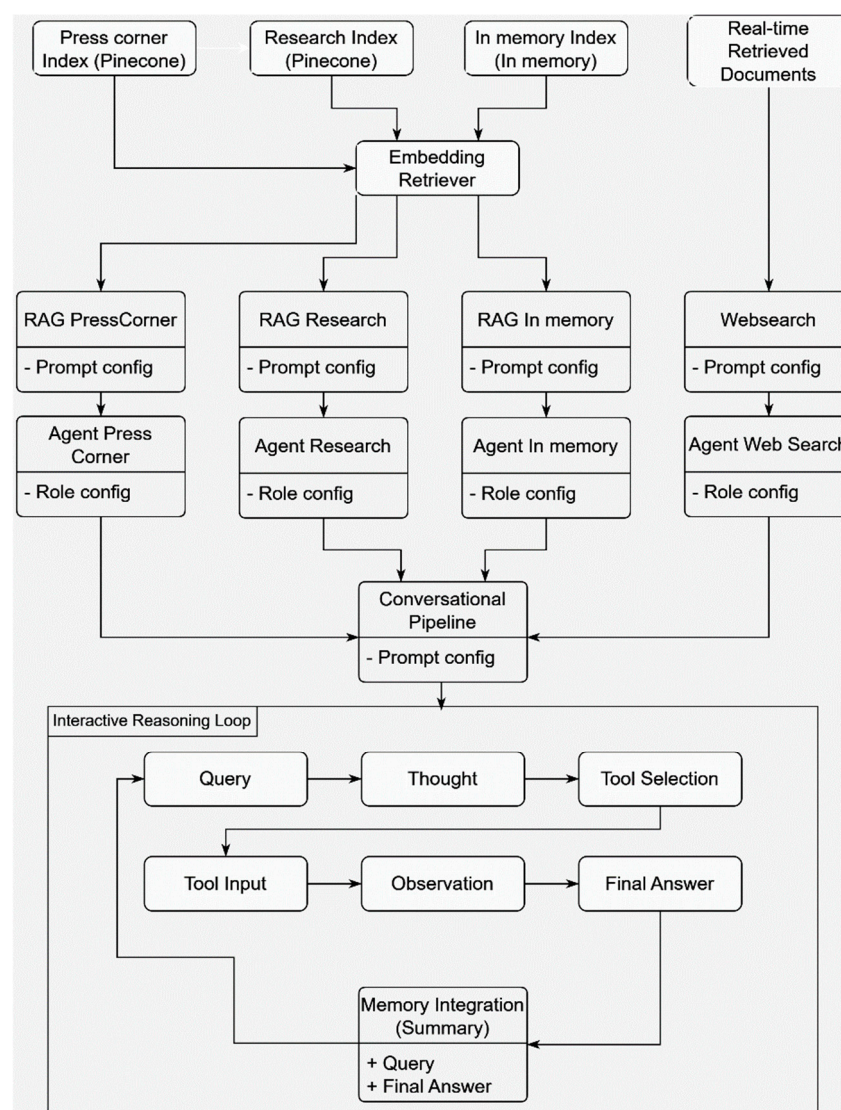


Figure 2. RAG and Conversational pipelines process diagram.

In addition, highlighting the modularity of this system, the models are under each preference due to the available modularity; different open-source models or other models can be used, such as Hugging Face, OpenAI, Anthropic Claude, Llama, Azure ChatGPT (OpenAI), Cohere, and SageMaker.

2.3. Agent Configuration and Deployment

To enhance our framework, we propose a multi-agent system where the RAG pipelines are configured as agent tools, resulting in the implementation of a conversational virtual assistant agent. Each agent has its own configuration, including selecting an applicable LLM (e.g., GPT-4o), setting stop words accordingly (e.g., “Observation:”), and specifying model parameters. Each tool is customarily assigned a role (e.g., “Press Corner” → “useful for when you need to answer questions about the Press Corner website of the European Commission”). Additionally, we incorporated memory using the summary model *philschmid/flan-t5-base-samsum* [33], accessed from Hugging Face [34], for interactions between the user and the virtual assistant, allowing for follow-up and more specific questions based on previous conversations, utilizing any LLM preferred by the user.

Regarding the conversational agent pipeline setup, a new prompt needs to be configured. This includes configuring tool names with descriptions in the prompt, which are set to identify the corresponding agent: memory, which is used to retrieve and incorporate the previous chat conversation; Thought, which represents the AI agent’s reasoning process; Tool input, which is used to transform the input according to the user’s requirements; Observations, which are the tool’s outputs; and the main query and transcript. The corresponding components mentioned above are presented in Figure 2.

2.4. Data Sources

The data sources could vary based on user- or domain-specific area preferences. In the current application, focused on the public sector and e-governance, the primary sources used are Press Corner, the European Commission’s official publicly available informative site, the OECD library, and selected openly available PDFs focused on research in a variety of domains and authors by the European Commission. Additionally, we incorporated web search, for which live documents are retrieved from Google with Serper, so that the user can retrieve the latest information under his preferences, using search features such as normal googling (e.g., for retrieving information from a specific site → “what is the green deal? site: scholar.google.com”). In our Web search configuration, the search engine initially retrieves the top 10 most relevant results based on the user query. These results are subsequently processed and divided into smaller document segments. From these segmented documents, the retriever then identifies and selects the top 50 most relevant sections for further processing. Lastly, we enable live document processing from various sources (e.g., PDF, CSV, and TXT) so the user can upload and use his document for Questions and Answers (Q&A). All of the above can be incorporated into a single search by using the agent’s conversational pipeline, asking for the inclusion of all of the above sources.

3. Results, Case Studies, and Applications

In the domain of the public sector and e-governance, the applicability of the LLMs’ capabilities is extremely important and beneficial. In our framework, we present the advantages of the above by highlighting an advanced practical example of how single or incorporated sources could provide accurate and sufficient information to the user. We will focus on a Q&A on various sources and the inclusion of all of these to provide a comprehensive answer. Each RAG pipeline from each source could be used individually and together, leveraging agents and conversational pipelines.

3.1. Case Study: Public Sector and E-Governance Application

In our case study, we will evaluate the effectiveness of the proposed RAG pipelines and conversational agents. All the results can be reproduced, as presented in the Python notebook example. In the proposed User Interface (UI) framework, the user can select a pipeline and model. For convenience, we will present the results of the conversational pipeline with agents since each tool’s output is shown; however, the reasoning process of the virtual assistant is also examined in detail to provide a definitive answer.

The questions asked for the presented case study are “What is the Green Deal?” and “What is the EU4DigitalUA?”. The results are only based on the retrieved documents that were indexed based on the sources. To leverage documents in the memory, we uploaded a document [35] that was highly correlated specifically with the Green Deal but not with the EU4DigitalUA project. Therefore, the Q&A process with detailed reasoning processes and agents’ usage is presented in Figures 3 and 4.

With the below, a comprehensive reasoning process is overviewed since we can identify the selected tool in each step, the tool input, which can change (see Press Corner compared with the rest of the tools), the observation, which details the answer of each tool in each stage, and the thought process that the conversational virtual assistant follows.

RAG and Conversational AI Interface

What is the Green Deal? Please search in PressCorner, Web search, Research, and In Memory documents and provide a comprehensive answer.

Thought: I will start by searching the Press Corner for information about the Green Deal. If necessary, I will then use the other tools to gather more comprehensive information.

Tool: PressCorner

Tool Input: What is the Green Deal?

Observation: The Green Deal is a plan to build a sustainable future for Europe's citizens. It aims to deliver a stronger economy, future-proof, kinder to nature, zero-pollution, and built around circularity. It emphasizes the importance of a systemic green transition to increase the resilience of socio-economic systems, avoid disasters like floods and droughts, and secure lasting prosperity and strategic autonomy while ensuring fairness and leaving no one behind. It is the only path that ensures a sustainable future for all.

Thought:Thought: I have obtained a basic definition of the Green Deal from Press Corner. To provide a more comprehensive answer, I will now search the Web for additional information.

Tool: Web_Search

Tool Input: What is the Green Deal by the European Commission?

Observation: The European Green Deal is a comprehensive package of policy initiatives introduced by the European Commission with the goal of transitioning the European Union (EU) to a sustainable economic model. It aims to set the EU on a path to becoming the world's first “climate-neutral bloc” by 2050. This means achieving net-zero greenhouse gas emissions while promoting economic growth and environmental sustainability. The Green Deal includes measures to reduce emissions from various sectors, such as transportation, which currently contributes a significant portion of Europe's emissions. Overall, the Green Deal represents the EU's main new growth strategy focused on fostering a green transition and ensuring that economic development is aligned with environmental goals.

Thought:Thought: I have gathered information from both Press Corner and a web search. To ensure a comprehensive answer, I will now search the Research tool for additional details about the Green Deal.

Tool: Research

Tool Input: What is the Green Deal by the European Commission?

Observation: The Green Deal by the European Commission is a transformative agenda focusing on eight areas: climate, energy, agriculture, industry, environment and oceans, transport, finance and regional development, and research and innovation. It aims to achieve carbon neutrality through interconnected regulations, strategies, and action plans. The Green Deal includes investment strategies, such as the NextGenerationEU Recovery Plan and the European Union's Multiannual Financial Framework, to finance green initiatives and support territorial just transition plans, promoting inclusive and sustainable economic growth with a focus on various aspects like fair taxation, social protection, and sustainable agriculture.

Thought:Thought: I have gathered information from Press Corner, a web search, and the Research tool. To ensure completeness, I will now check the In Memory documents for any additional details about the Green Deal.

Tool: In_memory

Tool Input: What is the Green Deal by the European Commission?

Observation: The Green Deal by the European Commission is a comprehensive policy strategy aimed at transforming the EU into a fair and prosperous society with a competitive economy. It outlines ambitions and goals in key areas such as climate ambition, clean energy, circular economy, and sustainable mobility. The Green Deal will involve revising existing regulations, developing new laws, and implementing directives to achieve its objectives.

Thought:Thought: I have gathered comprehensive information from Press Corner, a web search, the Research tool, and In Memory documents. I will now compile this information into a cohesive and detailed answer.

Final Answer:

The Green Deal by the European Commission is a comprehensive policy strategy aimed at transforming the European Union (EU) into a sustainable and competitive economy while ensuring fairness and leaving no one behind. It aims to set the EU on a path to becoming the world's first “climate-neutral bloc” by 2050, achieving net-zero greenhouse gas emissions while promoting economic growth and environmental sustainability.

Key aspects of the Green Deal include:

1. **Climate Ambition**: Achieving carbon neutrality by 2050 through interconnected regulations, strategies, and action plans.
2. **Clean Energy**: Transitioning to renewable energy sources and improving energy efficiency.
3. **Circular Economy**: Promoting a circular economy that emphasizes recycling, reducing waste, and sustainable resource use.
4. **Sustainable Mobility**: Reducing emissions from the transportation sector and promoting sustainable transport options.
5. **Agriculture**: Supporting sustainable agriculture practices.
6. **Industry**: Encouraging green innovation and sustainable industrial practices.
7. **Environment and Oceans**: Protecting biodiversity and promoting sustainable use of natural resources.

Figure 3. Example of the Conversational Q&A process with Agents for the query: “What is the Green Deal?” (“***” denotes intended bold formatting).

RAG and Conversational AI Interface

What is the EU4DigitalUA? Please search in PressCorner, Web search, Research, and in Memory documents and provide a comprehensive answer.

Thought: To provide a comprehensive answer about EU4DigitalUA, I will search in PressCorner, Web search, Research, and In memory documents.

Tool: PressCorner

Tool Input: What is the EU4DigitalUA?

Observation: The EU4DigitalUA is a four-year project called "DigiUni" aimed at developing a high-performance digital platform for Ukrainian universities to strengthen the digital environment for higher education in Ukraine. It will provide educational continuity for students enrolled at Ukrainian higher education institutions in the Ukrainian language and curriculum, as well as offer training in online teaching techniques and adapt learning content for online or virtual delivery.

Thought: I have obtained information from PressCorner about the EU4DigitalUA project, which is a four-year initiative called "DigiUni" aimed at developing a high-performance digital platform for Ukrainian universities. Next, I will search in Web search, Research, and In memory documents to gather more comprehensive details.

Tool: Web_Search

Tool Input: What is the EU4DigitalUA?

Observation: The EU4DigitalUA is a project under the broader EU4Digital Initiative aimed at supporting Ukraine's digital transformation and its integration with the EU Digital Single Market. With a budget of EUR 20.5 million, the project focuses on the development of digital government infrastructure and public e-services. It has facilitated the introduction of 54 e-services to make public services more accessible and efficient. This initiative represents part of the European Union's broader e-government-related support to Ukraine, promoting digital transformation and harmonisation with EU digital standards.

Thought: I have gathered information from PressCorner and Web search about the EU4DigitalUA project. The PressCorner source describes it as a four-year initiative called "DigiUni" aimed at developing a high-performance digital platform for Ukrainian universities. The Web search source indicates that it is part of the broader EU4Digital Initiative, focusing on digital government infrastructure and public e-services in Ukraine. Next, I will search in Research and In memory documents to gather more comprehensive details.

Tool: Research

Tool Input: What is the EU4DigitalUA?

Observation: The EU4DigitalUA is not specifically mentioned in the provided context, so answering this question is not possible given the available information.

Thought: I have gathered information from PressCorner and Web search about the EU4DigitalUA project. The PressCorner source describes it as a four-year initiative called "DigiUni" aimed at developing a high-performance digital platform for Ukrainian universities. The Web search source indicates that it is part of the broader EU4Digital Initiative, focusing on digital government infrastructure and public e-services in Ukraine. The Research tool did not provide additional information. Next, I will search in In memory documents to gather more comprehensive details.

Tool: In_memory

Tool Input: What is the EU4DigitalUA?

Observation: Answering this question is not possible given the available information.

Thought: I have gathered information from PressCorner and Web search about the EU4DigitalUA project. The PressCorner source describes it as a four-year initiative called "DigiUni" aimed at developing a high-performance digital platform for Ukrainian universities. The Web search source indicates that it is part of the broader EU4Digital Initiative, focusing on digital government infrastructure and public e-services in Ukraine. The Research and In memory documents did not provide additional information.

Final Answer:

The EU4DigitalUA is a four-year initiative called "DigiUni" aimed at developing a high-performance digital platform for Ukrainian universities to strengthen the digital environment for higher education in Ukraine. It will provide educational continuity for students enrolled at Ukrainian higher education institutions in the Ukrainian language and curriculum, as well as offer training in online teaching techniques and adapt learning content for online or virtual delivery. Additionally, it is part of the broader EU4Digital Initiative, which focuses on supporting Ukraine's digital transformation and its integration with the EU Digital Single Market, including the development of digital government infrastructure and public e-services.

Figure 4. Example of the Conversational Q&A process with Agents for the query: "What is the EU4DigitalUA?".

Moreover, in the second example related to EU4DigitalUA (as presented in Figure 4), we can observe the Agents' behavior when all data sources do not cover the query topic. This example highlights the agents' ability to comprehend and retrieve relevant information from available sources (Press Corner and Web Search), even when the Research repository and in-memory documents do not contain additional details.

3.2. Comparison of Answers and Agents' Effectiveness

The different sources provided distinct focuses on the Green Deal, contributing different perspectives that enriched the final comprehensive answer. Press Corner Agent, sourced from Press Corner, emphasized the importance of resilience, fairness, and a systemic green transition, stressing circularity and zero pollution as critical elements for a sustainable future. Web Search Agent, sourcing from Google search, highlighted the Green Deal as a policy package aimed at achieving climate neutrality by 2050, connecting economic growth with environmental sustainability and focusing on sector-specific emissions reduction, particularly in transportation. The Research Agent, using open documents from the OECD library, outlined eight specific focus areas, including climate, energy, agriculture, and industry, while also mentioning significant investment strategies such as the NextGenerationEU Recovery Plan. In Memory Agent, retrieving documents from the document [35] that the user uploaded, described the Green Deal as a comprehensive policy strategy with key ambitions in climate ambition, clean energy, circular economy, and sustainable mobility, alongside the need for regulatory revisions.

Each Agent's individual contributions helped to build a well-rounded final answer. Press Corner added a human-centric viewpoint by emphasizing fairness and the necessity of a systemic transition. Web Search contributed a clear goal of climate neutrality by 2050 and highlighted the relationship between economic growth and sustainability. Research Agent enhanced the answer with specific focus areas and investment strategies, emphasizing inclusive and sustainable economic growth. In Memory Agent provided a broad strategic

context, outlining key initiatives and the importance of revising existing regulations. Finally, these contributions created a comprehensive and detailed overview of the Green Deal, integrating various essential aspects to present a holistic view of the initiative.

The different Agents' distinct focuses on the EU4DigitalUA project contributed various perspectives that enriched the final comprehensive answer. The PressCorner Agent emphasized the educational aspect of the project, particularly highlighting the "DigiUni" initiative. This source described EU4DigitalUA as a four-year effort to create a high-performance digital platform for Ukrainian universities, ensuring educational continuity in the Ukrainian language and curriculum while offering training in online teaching techniques and adapting content for virtual learning. The Web Search Agent, sourcing from broader internet resources, offered an expanded scope by placing EU4DigitalUA within the larger EU4Digital Initiative. It highlighted the project's role in supporting Ukraine's digital transformation and integration with the EU Digital Single Market, focusing on developing digital government infrastructure and public e-services.

The Research Agent did not provide additional details, indicating that the already indexed research papers did not include information regarding the EU4DigitalUA. Similarly, the in-memory Agent did not provide further information, indicating the gap in the in-memory stored documents or in this topic, since the in-memory document targeted the Green Deal.

Each Agent's contribution helped build a well-rounded final answer. The Press Corner Agent provided a detailed view of the project's impact on higher education, while the Web Search Agent offered a broader context by connecting the project to EU-wide digital transformation efforts. Despite the lack of new information from the Research and In-Memory Agents, the combined insights from Press Corner and Web Search created a comprehensive understanding of the EU4DigitalUA project, integrating its specific educational goals with the broader objectives of digital government transformation in Ukraine.

3.3. Conversational Pipeline and Agents' thought Process

The thought process of the conversational pipeline, which was enabled by Haystack's framework, applied a systematic and iterative approach to assembling comprehensive information about the Green Deal from multiple sources, adjusting its query to each subsequent source and seeking additional details and different perspectives to ensure a thorough and multi-faceted answer, all of which are initiated from the prompt configuration [36], as presented in Table 1. Initially, as the user suggested, it started with a broad query to Press Corner, obtaining a foundational understanding of the Green Deal and recognizing that the initial information provided a basic definition. After obtaining the initial information, it recognized the need for more comprehensive data and shifted to Web Search, refining the query to gather specific policy details and goals. With foundational and policy information obtained, it moved to Research to acquire detailed descriptions of the various focus areas and investment strategies, enriching the overall understanding. Lastly, it queried In Memory to capture strategic initiatives, objectives, and regulatory aspects, ensuring all essential elements were included.

Regarding the "What is the EU4DigitalUA" example, it adjusted its query to each subsequent source, seeking additional details and different perspectives to ensure, as requested, a thorough and multi-faceted answer initiated from the prompt configuration, as presented in Table 1. Initially, as the user suggested, it started with a query to Press Corner, obtaining foundational and detailed information about the EU4DigitalUA project, specifically highlighting the "DigiUni" initiative to develop a digital platform for Ukrainian universities. Recognizing that this provided an educational perspective, it shifted to Web Search and querying to gather broader context and specific details about the project's role in Ukraine's digital transformation and its integration with the EU Digital Single Market. With these critical aspects covered, the pipeline moved to Research. However, this step did not yield additional information, indicating that the topic was not included in the related index that the Research Agent used. Lastly, it queried In-Memory documents, which did

not provide further details, suggesting that the uploaded document was unrelated to the EU4DigitalUA. Despite these gaps, the process ensured that all available information was included by integrating insights from the Press Corner and web search, resulting in a comprehensive understanding of the EU4DigitalUA project.

Table 1. Conversational Pipeline Prompt config.

Pipeline	Prompt Config
Conversational	<p>In the following conversation, a human user interacts with the AI Agent that has access to the documentation of Press Corner.</p> <p>Press Corner is the official portal of Press material from the European Commission Spokesperson’s Service.</p> <p>The human poses questions and AI Agent should try to find an answer to every question.</p> <p>The final answer to the question should be truthfully based solely on the output of the tool.</p> <p>The AI Agent should ignore its knowledge when answering the questions.</p> <p>The AI Agent has access to this tool:</p> <p>{tool_names_with_descriptions}</p> <p>The following is the previous conversation between a human and The AI Agent:</p> <p>{memory}</p> <p>AI Agent responses must start with one of the following:</p> <p>Thought: [the AI Agent’s reasoning process]</p> <p>Tool: [tool names] (on a new line) Tool Input: [input as a question for the selected tool WITHOUT quotation marks and on a new line] (These must always be provided together and on separate lines.)</p> <p>Observation: [tool’s result]</p> <p>Final Answer: (on a new line) [final answer to the human user’s question]</p> <p>When selecting a tool, the AI Agent must provide both the “Tool:” and “Tool Input:” pair in the same response, but on separate lines.</p> <p>The AI Agent should not ask the human user for additional information, clarification, or context.</p> <p>If the AI Agent cannot find a specific answer after exhausting available tools and approaches, it answers with Final Answer: inconclusive</p> <p>Question: {query}</p> <p>Thought:</p> <p>{transcript}</p>

This methodical process ensured that each query was informed by the previous findings, progressively building a more detailed and comprehensive answer. The conversational pipeline and the agents’ ability to adapt their queries based on the gathered information highlight its thoughtful and strategic approach to information retrieval, resulting in a final answer that integrated diverse perspectives and detailed insights from all agents and sources, providing a well-rounded overview of the Green Deal and the EU4DigitalUA. Moreover, to prevent the system’s hallucination, the GAI models were strictly instructed to respond solely based on the provided documents from retrieval (“The AI Agent should ignore its knowledge when answering the questions”). Additionally, if the retrieved documents did not contain the required information for Q&A, the assistant would answer with “Final Answer: inconclusive”, as configured in the Prompt in Table 1; the example results for this behavior are the Research and In memory Agents’ responses for the EU4DigitalUA topic. However, the system’s behavior is highly correlated with the overall performance of the GAI models. Therefore, system hallucinations should be regularly checked and monitored in accordance with the selected GAI model.

4. Reproducibility and Scalability

The effective deployment of LLMs and RAG systems in e-government applications requires careful consideration of reproducibility and scalability. Reproducibility guarantees that the AI systems can be replicated across different environments and use cases. Scalability addresses the system’s ability to handle increasing data sources and user interactions without compromising performance. It is necessary to accommodate the growing demands of public services and ensure that the GAI virtual assistant’s infrastructure can adapt to

future needs. This section outlines the framework for reproducibility, discusses the considerations for scaling the system, and addresses the challenges and limitations encountered during implementation.

4.1. Reproducibility Framework

Ensuring the reproducibility of our proposed implementation is vital to maintaining consistency and reliability across different applications. Our proposed framework can be easily reproduced by utilizing Haystack, an LLM orchestrator, building production-ready LLM applications. Therefore, we provide a low-code notebook, easily deployed in Google Colab, including configuration files and detailed documentation, and available in a public GitHub repository, including step-by-step guides for setting up the environment, running the code, and reproducing the virtual assistant, even with data as per the user's preference. Detailed specifications of the software and providers, including versions of libraries, dependencies, and alternative providers' options, are provided.

The presented notebook in the GitHub repository (<https://github.com/gpapageorgiouedu/Enhancing-E-Government-Services-through-a-SOTA-Modular-and-Reproducible-Architecture-over-LLMs>, accessed on 5 August 2024) is built in this way to be easy to deploy and demonstrate virtual assistant capabilities. Moreover, with slight modifications and refactoring, the proposed framework can be easily transferred into local or cloud-based environments. The documents used for demonstration purposes in our experiments are publicly accessible, including modular components for preprocessing and indexing. Moreover, all the RAG and agent configurations can be easily reproduced under the user's preferences, adding more pipelines, LLM options, and tools.

4.2. Scalability Considerations

Scaling the proposed implementation to accommodate larger ones involves several considerations and strategies. The technical architecture is modular, allowing individual components to be scaled independently. For instance, the data indexing, retrieval, and processing modules can be distributed across multiple servers and vectorized databases to handle increased loads from more than one data storage. Additionally, each document type from a different source could follow different preprocessing stages, allowing enhanced configuration based on the document's type.

Moreover, utilizing the LLM orchestration framework enables the creation of multiple RAG solutions based on various sources, with modular prompt configuration, accessing each user's needs in e-governance implications. Additionally, using multi-agent solutions, a more detailed search can be done based on needs, creating efficient tools for specific tasks and allowing the conversational multi-agent virtual assistant to identify the tool that should be used based on user queries.

Lastly, the proposed framework could run with local LLMs or with API providers and cloud-based solutions to be used for scaling based on the solution demand. Utilizing cloud services such as AWS, Azure, or Google Cloud provides on-demand scalability, offering tools for automatic scaling, resource management, and monitoring, enabling the system to adapt dynamically to changing demands.

4.3. Challenges and Limitations

While our proposed implementation demonstrates significant potential, it is important to acknowledge and address several challenges and limitations. One of the foremost concerns is ensuring the privacy and security of sensitive government data. The deployment of LLMs in e-government applications requires robust encryption methods, stringent access controls, and strict adherence to relevant legal frameworks to mitigate risks. Moreover, the proposed system does not incorporate users' feedback considerations in the retrieval (whether the retrieved documents are relevant to the query or not) and generator (whether the generated answer is correct and sufficient or not) components. This will be considered in future research.

Another critical challenge is efficiently allocating computational resources, particularly in cloud environments, where costs can rapidly escalate with increased usage. The performance and efficiency of GAI models are closely tied to the size and complexity of the models used. Larger models often deliver higher performance but have significantly higher computational and financial costs. Additionally, using LLMs provided by corporate services may further increase expenses, depending on the specific models and services employed.

To address these challenges, continuous monitoring and fine-tuning of system performance are essential. This approach ensures that deployment remains cost-effective while maintaining optimal performance across various implementations. Moreover, the proposed framework can be implemented with relatively lower costs by leveraging open-source LLMs and Haystack's LLM orchestration framework, based on each use case's specific data sources and requirements. This flexibility allows for a scalable and economically viable solution, making advanced AI technologies more accessible for e-government applications.

4.4. Alignment with the AI Act and Ethical AI Principles

The proposed framework for integrating LLMs and RAG into e-government applications aligns with the key principles of the AI Act and prevailing ethical guidelines for AI deployment in the public sector since it can be incorporated into different deployment models (on-premise or cloud services), preference architectures, and environments based on the public sector's set requirements. Designing the framework with modular components that can be independently scaled and replicated enables public sector agencies to maintain control and auditability over the AI systems powering their e-government services, with an obligation to AI Act policies, as per public sector agencies' needs, and following the information and transparency requirements [37]. This modular approach, with an explanatory reasoning process, facilitates the traceability of decision-making processes and allows for effective monitoring and oversight, which is crucial for upholding the principles of trustworthy AI.

Furthermore, the framework's modular design and its openness to incorporating alternative LLM options and data sources based on user preferences indicate a significant flexibility that can facilitate the tailoring of AI solutions to address potential biases and ensure fairness in public service delivery. This adaptability aligns with the AI Act's emphasis on the need for AI systems to be designed and deployed to respect fundamental rights and avoid discrimination; therefore, each designed application/solution should follow the EU AI Act rules, guidelines, policies, and transparency requirements. Additionally, the framework's modular design and its openness to incorporating alternative LLM options and data sources based on user preferences indicate a significant flexibility that can facilitate the tailoring of AI solutions to address potential biases and ensure fairness in public service delivery. This adaptability aligns with the AI Act's emphasis on the need for AI systems to be designed and deployed in a manner that respects fundamental rights and avoids discrimination [38].

5. Conclusions

This research presents a comprehensive multi-agent virtual assistant framework utilizing Haystack as an LLM orchestrator that is applied to applications in e-governance and the public sector. Our solution features modular, scalable, and reusable components, significantly enhancing the reproducibility and applicability of the framework across various sectors. The core architecture relies on two primary data sources: Press Corner and research publications from the OECD library, while also integrating web search capabilities and in-memory document processing through live uploads. For efficient document indexing, we employed Pinecone, which was selected from a range of vectorized database options to optimize data management. We developed four RAG options for Press Corner, research, web search, and in-memory documents, along with a conversational pipeline incorporating chat memory and introducing specialized tools to utilize these RAG pipelines effectively.

Our framework is designed to be flexible, allowing users to select between GPT-4, GPT-3.5, or other Large Language Models (LLMs), either hosted locally or accessed via an API from various providers, based on user preferences. To facilitate user interaction, we have developed a comprehensive user interface that enables seamless navigation and utilization of the virtual assistant's capabilities. Users can interact with the RAGs, employ conversational pipelines with their chosen LLMs, and efficiently index new documents.

We demonstrated the practical application of our framework through real-world use case scenarios, illustrating the advanced reasoning capabilities of the multi-agent virtual assistant. Those scenarios showcase the system's ability to handle multiple tool inputs, generate diverse observations based on those inputs, and synthesize a robust, coherent, and comprehensive response. This example highlights the potential efficiencies our solution can bring to public sector operations.

The integration of LLMs and RAG techniques into e-government systems, as outlined in this research, shows substantial promise for enhancing public service delivery. By focusing on modularity and reproducibility, our proposed framework ensures that AI deployments are scalable, transparent, customizable, and straightforward to implement. These features hold significant implications for the public sector, where deploying scalable and reproducible AI systems can improve operational efficiency and the accuracy of retrieved information. Enhanced Q&A capabilities enable more informed decision-making, accelerate service delivery, and optimize resource allocation. Furthermore, our commitment to transparency and ethical standards is essential for facilitating public trust and promoting the responsible use of GAI technologies in public service contexts.

Future Work

Future research could explore integrating additional data sources from e-governance and public sector domains into single, purpose-driven tools. Furthermore, future research could focus on integrating the current solution with action bot implications for automated, standardized processes [39]. Moreover, future studies should investigate the user feedback process to improve the agents' efficiency. As LLMs advance, future research should also focus on implementing more sophisticated reasoning processes to enhance the assistant's capabilities [40,41]. Lastly, integrating virtual assistants with other software systems and platforms used in e-governance and the public sector should be explored for compatibility with other used applications for daily tasks.

Author Contributions: Conceptualization: G.P., M.M. and V.S.; methodology: G.P. and M.M.; software: G.P.; validation: M.M., C.T., G.P. and V.S.; formal analysis: V.S. and G.P.; investigation: G.P. and V.S.; resources: M.M., C.T., G.P. and V.S.; data curation: M.M., G.P. and V.S.; writing—original draft preparation: G.P. and V.S.; writing—review and editing: M.M., C.T., G.P. and V.S.; visualization: G.P. and V.S.; supervision: C.T. and M.M.; project administration: M.M., C.T., V.S. and G.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available on GitHub at <https://github.com/gpageorgiouedu/Enhancing-E-Government-Services-through-a-SOTA-Modular-and-Reproducible-Architecture-over-LLMs> (accessed on 5 August 2024).

Conflicts of Interest: The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. This manuscript complies with ethical standards.

References

1. Madyatmadja, E.D.; Sianipar, C.P.M.; Wijaya, C.; Sembiring, D.J.M. Classifying Crowdsourced Citizen Complaints through Data Mining: Accuracy Testing of k-Nearest Neighbors, Random Forest, Support Vector Machine, and AdaBoost. *Informatics* **2023**, *10*, 84. [CrossRef]
2. Vida Fernández, J. Artificial Intelligence in Government: Risks and Challenges of Algorithmic Governance in the Administrative State. *Indiana J. Glob. Leg. Stud.* **2023**, *30*, 65–95. [CrossRef]

3. Chiappetta, A. Navigating the AI Frontier: European Parliamentary Insights on Bias and Regulation, Preceding the AI Act. *Internet Policy Rev.* **2023**, *12*, 1–26. [CrossRef]
4. Desouza, K.C.; Dawson, G.S.; Chenok, D. Designing, Developing, and Deploying Artificial Intelligence Systems: Lessons from and for the Public Sector. *Bus. Horiz.* **2020**, *63*, 205–213. [CrossRef]
5. Klarić, M. Regulation of AI Technology Implementation in Public Administration. In Proceedings of the 2024 47th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia, 20–24 May 2024; pp. 1450–1456.
6. Tuan, N.T.; Moore, P.; Thanh, D.H.V.; Pham, H. Van A Generative Artificial Intelligence Using Multilingual Large Language Models for ChatGPT Applications. *Appl. Sci.* **2024**, *14*, 3036. [CrossRef]
7. Ding, Q.; Ding, D.; Wang, Y.; Guan, C.; Ding, B. Unraveling the Landscape of Large Language Models: A Systematic Review and Future Perspectives. *J. Electron. Bus. Dig. Econ.* **2024**, *3*, 3–19. [CrossRef]
8. Gasser, U.; Almeida, V.A.F. A Layered Model for AI Governance. *IEEE Internet Comput.* **2017**, *21*, 58–62. [CrossRef]
9. Gao, S.; Gao, L.; Li, Q.; Xu, J. Application of Large Language Model in Intelligent Q&A of Digital Government. In Proceedings of the ACM International Conference Proceeding Series, Qinghai, China, 16–18 June 2023; Association for Computing Machinery: New York, NY, USA, 2023; pp. 24–27.
10. Mamalis, M.E.; Kalampokis, E.; Karamanou, A.; Brimos, P.; Tarabanis, K. Can Large Language Models Revolutionize Open Government Data Portals? A Case of Using ChatGPT in Statistics.Gov.Scot. In Proceedings of the ACM International Conference Proceeding Series, Lamia, Greece, 24–26 November 2023; Association for Computing Machinery: New York, NY, USA, 2023; pp. 53–59.
11. Puvača, M.; Kunac, I. Chatbot in Business. In Proceedings of the ICEMIT, Blace, Serbia, 21–23 September 2023; Toplica Academy of Applied Studies, Department of Business Studies Blace: Blace, Serbia, 2023; pp. 219–225.
12. Abdallah, A.M.; Alkaabi, A.; Alameri, G.; Rafique, S.H.; Musa, N.S.; Murugan, T. Cloud Network Anomaly Detection Using Machine and Deep Learning Techniques—Recent Research Advancements. *IEEE Access* **2024**, *12*, 56749–56773. [CrossRef]
13. Mageira, K.; Pittou, D.; Papasalouros, A.; Kotis, K.; Zangogianni, P.; Daradoumis, A. Educational AI Chatbots for Content and Language Integrated Learning. *Appl. Sci.* **2022**, *12*, 3239. [CrossRef]
14. Lee, G.H.; Lee, K.J.; Jeong, B.; Kim, T. Developing Personalized Marketing Service Using Generative AI. *IEEE Access* **2024**, *12*, 22394–22402. [CrossRef]
15. The White House. *DRAFT Guidance for Agency Artificial Intelligence Reporting per EO 14110*; The White House: Washington, DC, USA, 2024. Available online: <https://web.archive.org/web/20240801100051/https://www.whitehouse.gov/wp-content/uploads/2024/03/DRAFT-Guidance-for-Agency-Artificial-Intelligence-Reporting-per-EO14110.pdf> (accessed on 1 July 2024).
16. Kruhlov, V.; Bobos, O.; Hnylianska, O.; Rossikhin, V.; Kolomiets, Y. The Role of Using Artificial Intelligence for Improving the Public Service Provision and Fraud Prevention. *Pak. J. Criminol.* **2024**, *16*, 913–928.
17. Mauch, M.; Bachinger, S.T.; Bornheimer, P.; Breidenbach, S.; Erhardt, D.; Feddoul, L.; Legner, H.; Löffler, F.; Löffler, F.; Raupach, M.; et al. *From Legal Texts to Digitized Services for Public Administrations*; EasyChair Preprint: Manchester, Germany, 2024.
18. Othman, A.; Dhouib, A.; Nasser Al Jabor, A. Fostering Websites Accessibility: A Case Study on the Use of the Large Language Models ChatGPT for Automatic Remediation. In Proceedings of the ACM International Conference Proceeding Series, Corfu, Greece, 5–7 July 2023; Association for Computing Machinery: New York, NY, USA, 2023; pp. 707–713.
19. Deloitte AI Institute. *The Government & Public Services AI Dossier*; Deloitte AI Institute: Costa Mesa, CA, USA, 2024.
20. Yang, J.; Jin, H.; Tang, R.; Han, X.; Feng, Q.; Jiang, H.; Yin, B.; Hu, X. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Trans. Knowl. Discov. Data* **2023**, *18*, 1–32. [CrossRef]
21. Liu, W.; Liu, S.; Gao, D.; Jiao, R.; Huang, Y.; Duan, X. LLM Based Public Message Refinedly Grading Method. In Proceedings of the 2023 3rd International Conference on Digital Society and Intelligent Systems, DSInS 2023, Chengdu, China, 10–12 November 2023; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2023; pp. 41–46.
22. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45. [CrossRef]
23. Chen, J.; Lin, H.; Han, X.; Sun, L. Benchmarking Large Language Models in Retrieval-Augmented Generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 17754–17762. [CrossRef]
24. Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N.V.; Wiest, O.; Zhang, X. Large Language Model Based Multi-Agents: A Survey of Progress and Challenges. *arXiv* **2024**, arXiv:2402.01680.
25. Pietsch, M.; Möller, T.; Kostic, B.; Risch, J.; Pippi, M.; Jobanputra, M.; Zanzottera, S.; Cerza, S.; Blagojevic, V.; Stadelmann, T.; et al. Haystack: The End-to-End NLP Framework for Pragmatic Builders. 2019. Available online: <https://github.com/deepset-ai/haystack> (accessed on 1 July 2024).
26. Pinecone. Pinecone: Vector Database for Machine Learning Applications. Available online: <https://pinecone.io> (accessed on 1 July 2024).
27. European Commission. Press Corner. Available online: <https://ec.europa.eu/commission/presscorner> (accessed on 1 July 2024).
28. Organisation for Economic Cooperation and Development (OECD). OECD iLibrary. Available online: <https://www.oecd-ilibrary.org> (accessed on 1 July 2024).
29. Serper. Serper API. Available online: <https://serper.dev> (accessed on 1 July 2024).

30. OpenAI. Text-Embedding-Ada-002. Available online: <https://platform.openai.com/docs/guides/embeddings> (accessed on 1 July 2024).
31. OpenAI. GPT-3.5. Available online: <https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates> (accessed on 1 July 2024).
32. OpenAI. GPT-4o. Available online: <https://openai.com/index/hello-gpt-4o> (accessed on 1 July 2024).
33. Philschmid. Flan-T5-Base-Samsum. Available online: <https://huggingface.co/philoschmid/flan-t5-base-samsum> (accessed on 1 July 2024).
34. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; Liu, Q., Schlangen, D., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 38–45.
35. Fetting, C. *The European Green Deal*; ESDN Report; European Commission: Brussels, Belgium, December 2020; Volume 2.
36. Deepset. Customizing Agent. Available online: https://haystack.deepset.ai/tutorials/25_customizing_agent (accessed on 1 July 2024).
37. Parliament, E. EU AI Act: First Regulation on Artificial Intelligence: Topics: European Parliament. Available online: www.europa.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed on 1 July 2024).
38. Madiaga, T.; Chahri, S. *EU Legislation in Progress: Artificial Intelligence Act 2024*; European Parliament: Brussels, Belgium, 2024; Available online: [https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI\(2021\)698792](https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792) (accessed on 1 July 2024).
39. Hakimov, S.; Weiser, Y.; Schlangen, D. Evaluating Modular Dialogue System for Form Filling Using Large Language Models. In Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024), St. Julians, Malta, 21–22 March 2024; Graham, Y., Liu, Q., Lampouras, G., Iacobacci, I., Madden, S., Khalid, H., Qureshi, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 36–52.
40. Zhou, Q.; Zhang, Z.; Xiang, X.; Wang, K.; Wu, Y.; Li, Y. Enhancing the General Agent Capabilities of Low-Parameter LLMs through Tuning and Multi-Branch Reasoning. *arXiv* **2024**, arXiv:2403.19962.
41. Shang, J.; Zheng, Z.; Ying, X.; Tao, F.; Team, M. AI-Native Memory: A Pathway from LLMs Towards AGI. *arXiv* **2024**, arXiv:2406.18312.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Reproduced with permission of copyright owner. Further reproduction
prohibited without permission.