

Article

A Multimodal Framework Embedding Retrieval-Augmented Generation with MLLMs for Eurobarometer Data

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: This study introduces a multimodal framework integrating retrieval-augmented generation (RAG) with multimodal large language models (MLLMs) to enhance the accessibility, interpretability, and analysis of Eurobarometer survey data. Traditional approaches often struggle with the diverse formats and large-scale nature of these datasets, which include textual and visual elements. The proposed framework leverages multimodal indexing and targeted retrieval to enable focused queries, trend analysis, and visualization, across multiple survey editions. The integration of LLMs facilitates advanced synthesis of insights, providing a more comprehensive understanding of public opinion trends. The proposed framework offers prospective benefits for different types of stakeholders, including policymakers, journalists, nongovernmental organizations (NGOs), researchers, and citizens, while highlighting the need for performance assessment to evaluate its effectiveness based on specific business requirements and practical applications. The framework's modular design supports applications, such as survey studies, comparative analyses, and domain-specific investigations, while its scalability and reproducibility make it suitable for e-governance and public sector deployment. The results indicate potential enhancements in data interpretation and data analysis by providing stakeholders with the capability not only to utilize raw text data for knowledge extraction but also to conduct image analysis based on indexed content, paving the way for informed policymaking and advanced research in the social sciences, while emphasizing the need for performance assessment to validate the framework's output and functionality, based on the selected architectural components. Future research will explore expanded functionalities and real-time applications, ensuring the framework remains adaptable to evolving needs in public opinion analysis and multimodal data integration.

Keywords: Eurobarometer; ethical AI; data analysis; generative artificial intelligence (GAI); image analysis; multimodal large language models (MLLMs); retrieval-augmented generation (RAG)



Academic Editor: Gianni D'Angelo

Received: 23 December 2024

Revised: 11 February 2025

Accepted: 28 February 2025

Published: 3 March 2025

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

A Multimodal Framework Embedding Retrieval-Augmented Generation with MLLMs for Eurobarometer Data. *AI* 2025, 6, 50. <https://doi.org/10.3390/ai6030050>

Copyright: © 2025 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

The Eurobarometer surveys provide an extensive repository of data, capturing public opinion on a wide range of socio-political, sociodemographics, and economic issues across the member states of the European Union (EU). Traditionally, the analysis of such surveys has relied heavily on structured text and statistical data alone, but recent advancements in multimodal machine learning have made it possible to integrate diverse data formats—including textual and visual data—into a single analytical framework. The retrieval-augmented generation (RAG) approach has gained traction for its capacity to improve the

performance of large language models (LLMs) by supplementing them with external data, retrieved in real time, from a range of sources.

Using RAG architecture combined with MLLMs, stakeholders can quickly retrieve highly targeted data, based on specific queries, ensuring that relevant textual and visual information is effectively analyzed. For instance, a query regarding public attitudes towards climate policy could be multifaceted, requiring the analysis of different questions asked to the public and accompanying results' charts. Investigating these results separately may lead to missing important information, which an RAG-powered virtual assistant could efficiently retrieve, based on the given query. Therefore, with RAG, this information can be combined effectively, accompanied by visuals for further analysis, to provide comprehensive results for the stakeholders' specific needs.

With the initiation of multimodal large language models (MLLMs) and the growing application of RAG in complex data analysis, there is an opportunity to transform Eurobarometer survey analysis by integrating nontextual information, such as infographics, charts, and documents. This integration can yield a holistic and more context-rich understanding of public sentiment and trends, enhancing the accuracy and depth of insights gained from these surveys. This manuscript presents a novel framework that utilizes a multimodal RAG approach to incorporate textual, images, and visual data sources, thereby enabling a comprehensive analysis of Eurobarometer data. We aim to explore the potential of this multimodal approach to overcome the limitations of traditional methods, improving both interpretability and the scope of findings in policy and social research.

1.1. Research Questions

The integration of advanced data retrieval methods, such as RAG, and LLMs in the analysis of Eurobarometer survey data represents a novel approach to addressing the complexities associated with mixed-format and large-scale data sources. As Eurobarometer data consist of a variety of input types, including plain text based on the survey's press release, PDF attachments (for example, Summary, Data Annex, and Report), and images, traditional methods of analysis are often limited in their ability to provide comprehensive insights.

The following research questions focus on the potential of a multimodal RAG framework combined with LLMs to address these challenges, enhancing both accessibility and in-depth analysis. Through these questions, we aim to investigate the effectiveness of this approach in retrieving, interpreting, and synthesizing insights from Eurobarometer data, with trend analysis implications, comparative studies, and public opinion studies.

1. How can multimodal integration of RAG and LLMs improve knowledge extraction and interpretability of Eurobarometer data for various data inputs?
2. To what extent does the RAG approach in multimodal data retrieval support synthesized data analysis, information comparison, and data visualization analysis across Eurobarometer surveys?

The first research question explores how multimodal integration of RAG and LLMs can improve access to and interpretation of Eurobarometer data, enabling seamless retrieval and synthesis across mixed data formats.

The second research question evaluates the RAG framework's ability to support survey analyses, including trend analysis, comparative studies, and data visualization. It further examines how combining LLMs with RAG can enhance the extraction and synthesis of insights, offering a comprehensive understanding of public opinion trends and societal dynamics.

1.2. Aims and Objectives

The primary aim of this study is to investigate the potential of a multimodal RAG framework, combined with LLMs, in enhancing data retrieval, analysis, and insight genera-

tion from Eurobarometer survey data. By unifying diverse data formats, including texts and images, within a single RAG framework, this study seeks to improve data accessibility, interpretability, and analytical depth, benefiting researchers, policymakers, and analysts by deriving meaningful insights from large-scale survey data.

This study involves several specific objectives to achieve this aim. First, it aims to develop a multimodal RAG-based framework, designed to support indexing, retrieval, and processing of heterogeneous Eurobarometer data formats, enabling comprehensive access and seamless extraction of relevant information across mixed-format data sources. This framework is also designed to support advanced analytical tasks, such as trend analysis, comparative studies, and data visualization across different editions of Eurobarometer surveys, with the goal of facilitating data-driven decision making in public opinion research and social sciences. Finally, this study seeks to explore the practical applications and potential added value of this multimodal framework, especially in making Eurobarometer data more interpretable for stakeholders in policymaking, research, and academia. By enabling in-depth analysis of complex, mixed-format datasets, this framework aims to demonstrate how multimodal RAG and LLM integration can transform survey data analysis, offering both methodological innovation and practical impact.

1.3. Study Significance

This study holds significant value in advancing methods for managing complex, large-scale survey data, particularly within the Eurobarometer context, which is widely used for public opinion research and policy development across Europe. Traditionally, analyzing Eurobarometer data has posed challenges due to its diverse formats, including textual, tabular, and visual content spread across multiple reports and survey editions, linked one to another. By introducing a multimodal RAG framework combined with LLMs, this study offers a transformative approach to data analysis and interpretation addressing the limitations of traditional methods.

The multimodal RAG approach allows users to retrieve and synthesize information from various data formats seamlessly. It provides valuable insights for stakeholders such as researchers, policymakers, journalists, and NGOs. These stakeholders can analyze trends, conduct investigative reporting, and streamline fact-checking. The framework also supports survey studies and data-driven decision making using Eurobarometer data. Extracting insights from both visual and textual data helps shape policies, create trustworthy reports, combat disinformation, and address social concerns.

This study also explores LLM integration, offering new opportunities for processing large-scale survey data efficiently. It enables the extraction and synthesis of complex insights from extensive datasets. LLMs provide contextual understanding and nuanced interpretation, enhancing the framework's analytical capabilities. This tool is particularly valuable for social sciences and public opinion research. It supports evidence-based policymaking and improves our understanding of evolving social trends.

Overall, this study's contribution lies in its methodological innovation, its ability to address complex data integration challenges, and its practical applications for various sectors. By offering a scalable solution for analyzing and interpreting mixed-format survey data, this research has the potential to reshape the landscape of public opinion analysis, fostering greater accessibility and utility of Eurobarometer and other large-scale survey data for informed decision making across fields.

The structure of the remainder of this paper is as follows: Background Section 2 reviews the literature related to the data use case, applications of LLMs and MLLMs, and technological advancements in these areas. Section 3 details the data sources, outlines the benefits of the proposed framework, and provides technical justifications for the use of RAG and Agents.

Section 4 presents case studies and applications of the proposed framework, detailing the outcomes with examples, followed by discussions on reproducibility, scalability, challenges and limitations, and alignment with AI policies. Finally, Section 5 concludes the paper with a summary of the findings and an outline of potential directions for future research.

2. Background

The Eurobarometer, established in 1974, is a public opinion survey conducted regularly on behalf of the EU Institutions (European Commission (EC), European Parliament (EP), etc.). It covers a range of topics pertinent to EU member states, such as economic policy, social issues, and environmental attitudes. These surveys provide EU policymakers with valuable insights that guide strategic planning and policy development [1]. However, the traditional analysis of Eurobarometer data has been predominantly text-based and numerical, often overlooking valuable contextual information available in related visual data and documents.

The recent rise in MLLMs offers an advanced method for processing and analyzing data across different modalities. MLLMs can combine textual inputs with visual and PDF data, leading to improved analysis and understanding of complex information by leveraging diverse data sources [2,3]. These models utilize pretrained language backbones with additional visual encoders and vision-to-language adapters, enabling them to process images and documents alongside text inputs, significantly enhancing interpretive capabilities [4,5]. The RAG technique further enhances these models by dynamically retrieving relevant data from external repositories, allowing the integration of up-to-date information, which is essential for analyzing Eurobarometer data that spans multiple years and varied themes.

The application of RAG to multimodal data—particularly in the context of integrating textual and visual sources—has shown promising results in fields such as government data analysis and decision-making systems [2,4]. The combination of retrieval-based techniques with multimodal data is anticipated to bridge the gap between structured survey data and the broader, more complex information contained in PDFs and visual media associated with Eurobarometer reports.

In this paper, we outline an innovative multimodal RAG framework tailored for the analysis of Eurobarometer data, setting the groundwork for enhanced public opinion analysis and facilitating more nuanced insights into the EU's policy and strategic priorities.

The integration of multiple data modalities in data analysis has shown substantial potential in enhancing the interpretive capabilities of LLMs. Traditional LLMs have excelled in understanding and generating human language, but the limitations of purely text-based models become evident when interpreting complex, multimodal datasets, like those found in Eurobarometer surveys, which include text, visual data, and document formats such as PDFs. Recent advances in MLLMs, like GPT-4o, Gemini, NVLM, Llama 3.2, PaLM-E, CLIP, etc., and models with RAG highlight the capacity of AI to process and synthesize data from varied sources, thus enabling deeper insights and more accurate analysis.

The development of MLLMs has been driven by the recognition that real-world tasks and environments often require processing information across several sensory domains simultaneously [6,7]. For instance, systems like ChartLlama and AnyGPT have demonstrated how combining visual and textual inputs can aid in specialized tasks, such as chart understanding and responding to multimodal queries, by aligning visual elements with language representations through visual encoders and modality adapters [8,9]. These models leverage various strategies for multimodal alignment, such as training on interleaved text–image data and developing task-specific encoders, which together contribute to a more holistic understanding of complex datasets.

A critical advancement in this field has been the use of RAG techniques. RAG enhances LLMs by retrieving relevant external data, allowing the model to ground its responses in a broader context. This method is particularly valuable for applications like Eurobarometer survey analysis, where real-time or up-to-date information may be necessary to contextualize survey responses or analyze public opinion trends accurately [10,11]. Moreover, models such as WorldGPT and VITA have explored the potential of multimodal world modeling, enabling LLMs to understand temporal and spatial dynamics across multiple data types, further enriching the analytical scope of these models [12,13].

In the evolving landscape of MLLMs, recent advancements underscore the importance of cross-modal integration and efficiency in enhancing data analysis. Models such as GPT-4o and NExT-GPT highlight advancements in seamlessly processing and generating content across multiple modalities, including text, image, video, and audio [14,15]. These models reflect the growing trend of integrating sophisticated multimodal capabilities within large language models to support complex applications, like sentiment analysis, visual question answering, and multilingual user interaction in mobile interfaces.

The integration of RAG with MLLMs represents a significant advancement in addressing the limitations of traditional data processing and analysis frameworks. While LLMs, such as GPT-3.5 and GPT-4, have demonstrated exceptional capabilities in natural language understanding and generation, they often face challenges in handling dynamic and domain-specific knowledge, leading to inaccuracies and hallucinations. RAG addresses these limitations by incorporating external knowledge retrieval during the generative process, ensuring more precise, contextually relevant, and up-to-date outputs [16,17].

Recent studies have explored various applications of RAG frameworks across multiple domains. For instance, Byun et al. demonstrated the effectiveness of RAG in personalized database systems by improving retrieval precision through context-based tagging and NoSQL database integration. Similarly, in [18], the role of RAG in automating systematic literature reviews is highlighted, showcasing its ability to streamline data extraction and synthesis [18]. Furthermore, adaptive RAG frameworks, such as CRP-RAG, have introduced reasoning graphs to enhance logical reasoning, improving the quality and robustness of outputs in complex query scenarios [19].

The application of RAG in multimodal settings, which involves integrating textual, visual, and tabular data, further expands its utility. These studies proposed a framework using reflective tags and step-by-step reasoning to refine retrieval and mitigate hallucinations. These developments underscore the potential of RAG to transform data analysis, particularly in scenarios involving diverse data formats [19,20].

The architectural choices in MLLMs often involve specialized components, such as modality-specific encoders, input projectors, and adapters that bridge visual and textual data within a shared latent space [6]. These components allow MLLMs to process inputs from diverse modalities effectively, addressing tasks that demand cross-modal reasoning and decision-making capabilities. For instance, MobileFlow, a multimodal LLM designed for Graphical User Interfaces (GUIs), employs a hybrid visual encoder and tailored alignment strategies to handle complex mobile GUI tasks, demonstrating the potential of MLLMs in real-world applications [21].

In addition to architectural advancements, the training methodologies such as instruction tuning, retrieval-augmented generation, and modality-specific alignment strategies improve the capabilities of MLLMs to perform diverse multimodal tasks. Those approaches advance the models' contextual understanding and efficiency, making them well suited for applications involving dynamic content generation and interactive AI [22,23].

The ongoing shift from text-only LLMs to multimodal MLLMs, empowered by the RAG framework, marks a transformative step in AI research. This evolution facilitates the

comprehensive integration of structured text, unstructured documents, and visual data, thus offering promising applications in public opinion analysis and other complex social science inquiries. These developments underscore the value of multimodal data processing for capturing a nuanced understanding of public sentiment, policymaking, and other areas relevant to Eurobarometer's extensive dataset.

This paper addresses these gaps by presenting a comprehensive multimodal RAG framework tailored to the Eurobarometer dataset, integrating textual and visual data sources. Hence, this study explores potential improvements in data interpretation and analysis, demonstrating the broader applicability of RAG in socio-political and economic research, while considering the need for evaluation when implemented with specific components, based on their requirements and intended use.

3. Technical Architecture and Solution

Traditional data analysis approaches often struggle with diverse formats and the large-scale nature of similar datasets, which include plain textual, PDF, and visual input, especially on public opinion survey data. The difficulty is in combining vast amounts and different types of information for knowledge extraction. Therefore, for optimizing the knowledge discovery, leveraging MLLMs and RAG architecture, the proposed architecture can provide the user with combined information they are interested in, in the form of questions and answers (Q&A).

This study introduces a framework that combines advanced AI techniques with the analysis of Eurobarometer survey data, a collection of public opinion surveys conducted across EU member states. The proposed implementation leverages innovative AI models, known as MLLMs, and a technique called RAG. These tools collaborate to analyze and extract insights from data in various formats, including text, images, and PDF documents. By doing so, the framework provides policymakers, researchers, and other stakeholders with a powerful tool to understand public opinions and trends more effectively, supporting evidence-based decision making and policy development. This high-level integration of AI into survey analysis aims to bridge gaps in traditional methods, making complex data more accessible and actionable.

The proposed technical architecture, with its overview presented in Figure 1, leverages state-of-the-art advancements in RAG, incorporating simple RAG mechanisms, agent-based conversational stages, and image-based query handling. The basic pillar of the framework integrates well-established image-processing tools, LLMs, and the open-source Haystack LLM orchestration framework [24]. However, similar frameworks can be employed as alternatives. For this reason, each of the framework's components is selected to be modular, allowing easy adjustment and integration of alternative solutions, without significant modifications.

The proposed framework supports plain text, PDF files, and images for indexing. Therefore, for our case study we used Eurobarometer survey results data. Those included press release plain text and images, along with metadata, and, for enriching data input, the PDF report and summary of the selected survey were indexed too.

Also, the architecture included multiple preprocessing and processing methods to transform the data into a format appropriate for efficient indexing and subsequent retrieval for Q&A. Data storage is implemented using an open-source vectorized database solution, FAISS, chosen for its efficiency, flexibility, and straightforward applicability. This approach enabled indexing directly in memory (as implemented in Google Colab in this instance), ensuring accurate indexing and reliable information retrieval.

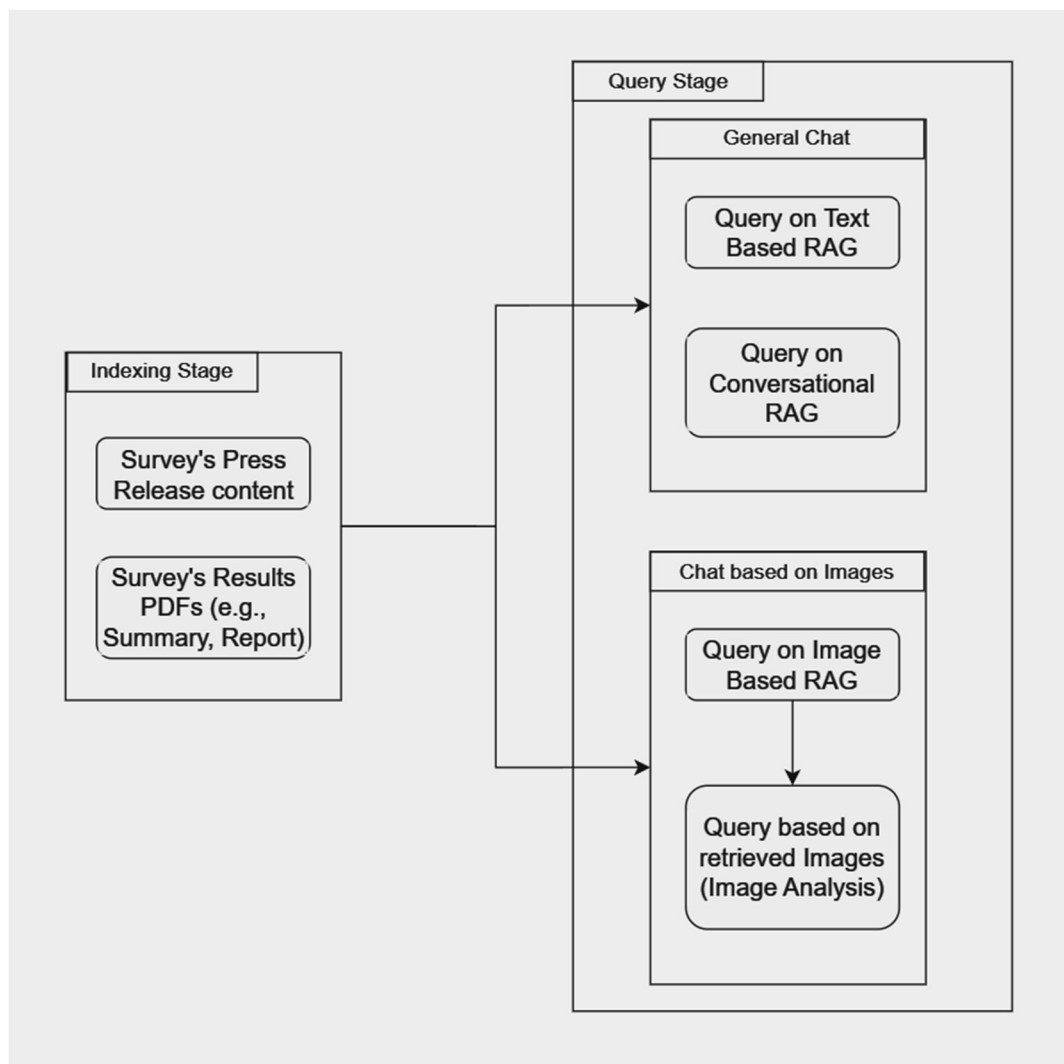


Figure 1. Overview of the proposed framework architecture, integrating RAG and MLLMs for Q&A.

The primary data source, based on user selection, is the Eurobarometer survey press releases and related files. Each survey includes its respective metadata along with associated PDFs, such as Data Annexes, Reports, or Summaries. Additionally, the press release for these surveys includes images in PNG and JPEG format, which present the survey's key findings. During our implementation, these images were incorporated using dedicated preprocessing techniques, leveraging also the alternative text provided by the official EC Eurobarometer website.

The architecture's main component is the Haystack LLM orchestrator [24], which is powerful and applicable to RAG solutions. This open-source component was selected for its modularity, allowing easy modification and customization of each part, including the database, processing models, and LLMs. The following subsections provide a detailed explanation of the methodologies and techniques employed for each component, offering a comprehensive overview of the proposed system.

3.1. Data Sources

This case study focuses on the public sector and the implications of e-governance, with the primary data source, the Eurobarometer surveys. Specifically, the official press releases, publicly available from the EC and the European Parliament, serve as the main content foundation. Each official press release is published online on the Eurobarometer

section of the europa.eu website, providing an overview of the survey results as the main content. Those releases are accompanied by metadata, such as the type of survey conducted (e.g., face-to-face, phone calls, online, etc.). Additionally, the press release includes images that show the key findings from the survey, augmented with alt text (provided in the HTML alt attribute to describe the image and improve accessibility), which is used in our implementation as Supplementary Information.

Each survey also provides publicly available detailed results in PDF files, categorized into several types based on the content, such as Data Annex, Country Factsheets, Infographics, Reports, and Summaries. Lastly, all available types of data sources are employed for the proposed implementation, ensuring a comprehensive retrieval and Q&A process based on press release content, text, images, and text information in PDF files.

3.2. *Benefits to Stakeholders*

The proposed framework is based on Eurobarometer surveys. It illustrates a practical example of an e-government application, targeting to enhance public awareness of such important survey information. It aims to improve public sector employees' productivity in fulfilling their support roles but also provides a valuable tool for multiple stakeholders, such as policymakers, researchers, NGOs, journalists, and data analysts. It enables public sector employees to provide well-informed, data-driven assistance to citizens, while supporting researchers and policymakers in crafting solutions that align with public needs.

Comprehensive data analysis based on user-preferred survey data, in the form of Q&A functionality, empowers multiple stakeholders, such as journalists and NGOs. Hence, it helps to achieve their objectives more effectively by enabling journalists to communicate accurately and provide reliable information while supporting NGOs to address press releases' social topics. Additionally, visual data analysis enriches this capability by allowing the previously mentioned stakeholders to extract and interpret statistical information based on charts, graphs, and other visual figures. Therefore, it facilitates more in-depth knowledge extraction of trends and insights that effectively inform their respective efforts.

3.3. *RAG Integration*

The proposed RAG framework is built based on two stages, each distinct stage based on the type of data being processed. These stages handle the main tasks of data indexing and querying. The former, the indexing stage, is responsible for transforming the relevant content text and files into documents for database indexing through specific preprocessing steps. For images, dedicated processing methods are implemented to handle their unique requirements. Text data processing varies depending on the document type. In our Eurobarometer data implementation, two types of text files were indexed: the native content of press releases and user-selected PDF files related to surveys. Moreover, each press release of those surveys includes images in PNG and JPEG format, presenting some of the survey's key findings. The images are handled separately and integrated with their alt text as supplementary content for indexing to enhance retrieval accuracy.

The latter, called the query stage, is configured to retrieve the most relevant content and create responses based on user questions using indexed data. This stage has three main steps.

The first step is the simple RAG process, which manages basic retrieval and generation tasks for both text and image data. Simple text-based RAG handles queries related to text, while simple image-based RAG processes queries requiring image retrieval and interpretation.

The second step introduces an agent-based Q&A system with two agents configured to manage specific data types, text, and images independently. Each RAG component (text or image) operates as a distinct agent. The agent-based conversational configuration also includes

memory capabilities. This feature enables the LLM to recall previous discussions and adapt to business requirements. To enhance user experience and transparency, the conversational pipeline incorporates “thought” and “observation” steps. These steps help users understand the reasoning behind the agent’s responses.

Lastly, the third Q&A stage for image processing handles the most relevant images retrieved from the Image RAG. This stage processes images to effectively address visual data inquiries. It is supported by models capable of managing native image formats encoded in base64. This feature enhances the framework’s flexibility, allowing it to handle diverse query types efficiently. All the steps work together to create an adaptable system that is both efficient and accurate in retrieving information.

3.3.1. Indexing Stage

The indexing stage, as illustrated in Figure 2, outlines the main components involved. This stage is responsible for preprocessing content based on file types (e.g., native text, PDFs, and images), transforming it into embeddings and documents, and indexing them into the database. In our implementation, with the indexing process handled in the backend, the user can index the survey of their preference, starting with the press release and its metadata, along with his selected images. In a subsequent step, the user can choose to index additional documents retrieved as PDFs. These PDFs undergo processing by the same processor used for press release content, which first converts the files to extract their text content. The processor performs data cleansing, removing empty lines, excess whitespace, and any headers or footers if present. A chunking strategy is applied, splitting the content into segments of 200 words, with a 30-word overlap, while maintaining sentence boundaries.

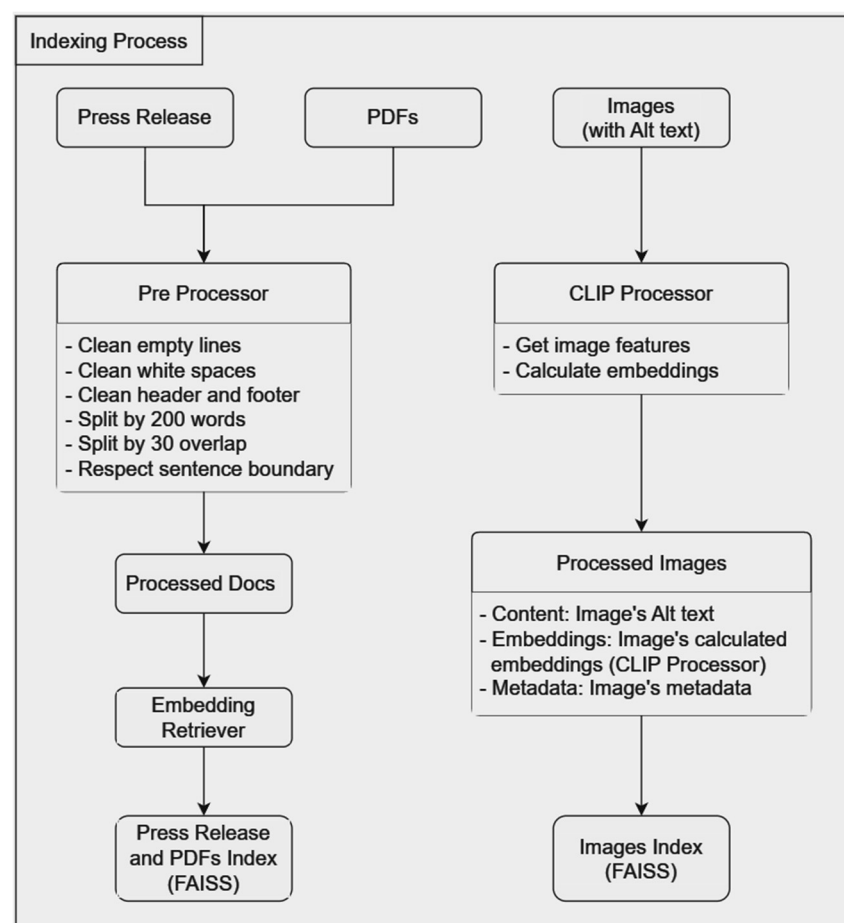


Figure 2. Indexing stage for textual, visual, and PDF data.

This approach ensures that each chunk is small enough to fit within the input limits of modern NLP models, while preserving sufficient context for accurate processing. The 30-word overlap prevents information loss at the boundaries of chunks, enabling the model to handle queries that span across segments. Maintaining sentence boundaries ensures that chunks remain coherent and meaningful, which is critical for downstream tasks like question answering or summarization, where logical continuity enhances performance.

For image data in PNG or JPEG formats, the images are processed using the CLIP processor and model, specifically “openai/clip-vit-base-patch32” [25,26] in our implementation. The preprocessing for images includes extracting their features and generating embeddings, which are then indexed alongside the alt text and metadata of the images.

The next step involved embedding the preprocessed data. Text data are transformed into vectorized representations using the “text-embedding-ada-002” [27] embedding model for each document chunk. Similarly, images are embedded using the CLIP model to extract features, which are indexed alongside their associated metadata and alt text. This two-part approach makes sure that both text and images are seamlessly included in the indexing process.

The indexing pipeline is designed with a multimodal approach to handle and index diverse types of data. Text content, such as the native content of press releases and PDFs, is processed in a consistent manner. Images included in press releases are processed and indexed with their associated alternative text label (i.e., alt), provided by the official EC Eurobarometer website. The goal of this pipeline is to provide a complete, modular indexing solution tailored to specific case study requirements. All models used are modular and easily replaceable, ensuring flexibility and adaptability for varying needs.

The performance evaluation of the proposed framework is closely tied to the configuration and quality of its components. The vectorized database is set up to meet performance requirements, following recommendations outlined in the provider’s documentation. Data quality is critical to ensure proper processing by the modular preprocessor, which should be configured based on the characteristics of the input data. The accuracy of the embeddings and retrieval process depends on the performance of the selected embedding retriever; higher-performing embedding models lead to more precise retrieval outcomes.

Similarly, the Q&A and reasoning processes within the conversational and RAG framework are influenced by the chosen generative AI (GAI) models. The same applies to image analysis, in which models’ performance may vary based on image quality and quantity of information. Detailed investigation should be conducted based on the provider’s evaluation. Stakeholders should consult the provider’s documentation to identify and select models that align with the specific requirements of their use case. In our use case example, the evaluation set for GPT-4o and GPT4o-mini models can be found in [28–31]. Component selection should be guided by the provider’s performance evaluation resources, ensuring decisions are tailored to the specific demands of the application.

3.3.2. Querying Stage

The querying stage, as illustrated in Figure 3, is dedicated to building a versatile environment for Q&A interactions, offering multiple options for users. It employs an RAG framework, which, in this example, uses GPT-4o-mini [32] to provide accurate responses by retrieving the most relevant information through the embedding retriever “text-embedding-ada-002” [27]. The implemented pipelines are divided into three main categories: Simple RAG, Conversational Agents with Memory, and Image-Integrated RAG, each tailored for specific data and query scenarios.

The first category, Simple RAG, includes two implementations: one for text data and another for images. The text-based RAG is designed to manage queries about textual information, while the image-based RAG processes queries requiring visual data. The configuration

for Simple RAG is designed to optimize the performance with a custom prompt template, as presented in Table 1, tailored for GPT-4o-mini [32]. The prompt allowed up to 4096 tokens for detailed answers. The setup included controls like a low temperature of 0.1 for consistent responses and a top- p value of 0.9 to use nucleus sampling for more logical and coherent outputs. These parameters ensured high-quality responses while maintaining a balance between randomness and reliability.

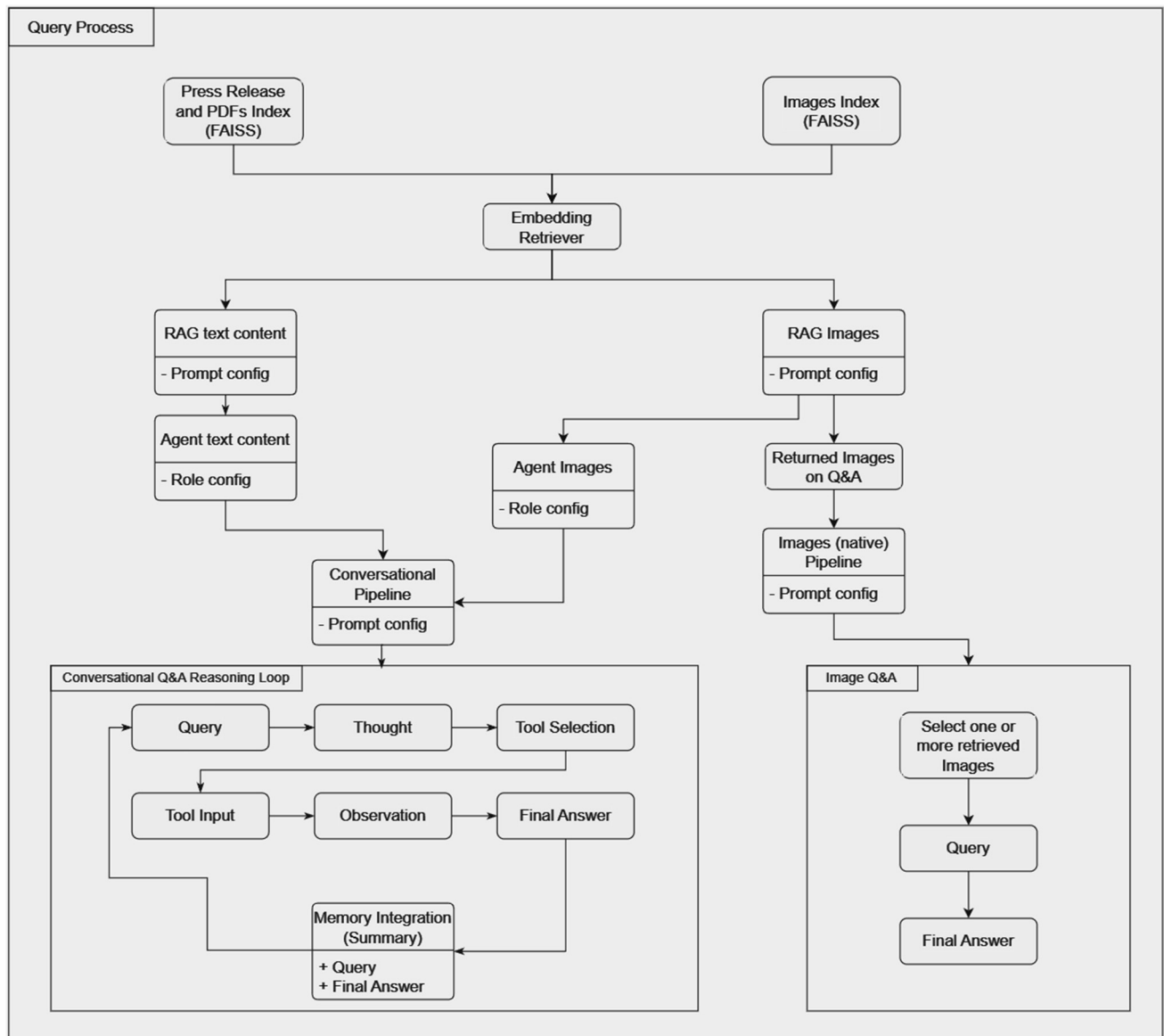


Figure 3. Query stage with RAG pipelines, conversational agents, and Q&A in images.

The second category is conversational agents with memory, built upon the simple RAG implementation and incorporating memory capabilities. This setup configured the RAG pipelines as agents, enabling context-aware interactions, using GPT-4o [33]. Memory is implemented using the “philschmid/flan-t5-base-samsum” model [34], accessed from Hugging Face [35], which excels in summarization tasks. The memory prompt is designed to summarize conversation history and integrate a structured thought process, allowing the system to maintain context across sessions. The agents are configured [36] with a conversational prompt tailored to manage shorter responses, with a maximum length of 256 tokens. Additionally, stop words such as “Observation:” are defined to refine the agent’s response flow, with the

custom prompt presented in Table 2. These agents are employed to be adaptive, making them suitable for dynamic and evolving conversational needs.

The final category is the image-integrated RAG, a dedicated implementation for handling image-based queries. It combines the capabilities of the image RAG pipeline for retrieving the related images based on a standard query with GPT-4o’s [33] image query capabilities. This configuration allows users to query the system using visual references retrieved from the image RAG. Users can select a specific image from the results and initiate follow-up queries based on that image. This process employed a dedicated, prompt configuration for image queries previewed in Table 3, ensuring that responses are contextually accurate and detailed. Using the same embedding retriever ensured consistency across the text and image pipelines. Overall, the querying stage is designed to be flexible, modular, and highly efficient based on user needs. The configuration for each pipeline is tailored to the specific use cases, ensuring seamless integration and adaptability for diverse applications. Advanced prompt configurations and memory integration enhanced the system’s ability to deliver accurate, context-aware, and insightful responses across various data types and query scenarios.

Table 1. Simple RAG prompt configuration.

Pipeline	Prompt Config
For text-related data	<p>In the following conversation, a human user interacts with the AI virtual assistant that has access to the survey data of Eurobarometer.</p> <p>Eurobarometer is a collection of cross-country public opinion surveys conducted regularly on behalf of the EU Institutions since 1974.</p> <p>Based on the survey data, generate a response that answers the question.</p> <p>If the information is not sufficient, say that the answer is not possible from the documents alone.</p> <p>You should ignore your knowledge when answering the questions and be based solely on the below documents.</p> <p>Provide a clear and concise answer, no longer than 200 words.</p> <p>Always include a disclaimer in your answer regarding AI generated content.</p> <p>Disclaimer: This is AI generated content—please use it with caution.</p> <p>\n\n Context: {join(documents)} \n\n Question: {query} \n\n Answer:</p>
For image-related data	<p>In the following conversation, a human user interacts with the AI virtual assistant that has access to the survey images metadata of Eurobarometer.</p> <p>Eurobarometer is a collection of cross-country public opinion surveys conducted regularly on behalf of the EU Institutions since 1974.</p> <p>Based on the image metadata, generate a response that answers the question.</p> <p>If the image information is not sufficient, say that the answer is not possible from the image metadata alone.</p> <p>You should ignore your knowledge when answering the questions and be based solely on the below documents.</p> <p>Provide a clear and concise answer, no longer 200 words.</p> <p>Always include a disclaimer in your answer regarding AI generated content.</p> <p>Disclaimer: This is AI generated content—please use it with caution.</p> <p>\n\n Context: {join(documents)} \n\n Question: {query} \n\n Answer:</p>

Table 2. Conversational agents prompt configuration.

Pipeline	Prompt Config
Conversational	<p>In the following conversation, a human user interacts with the AI Agent that has access to the documentation of Eurobarometer.</p> <p>Eurobarometer is a collection of cross-country public opinion surveys conducted regularly on behalf of the EU Institutions since 1974.</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>You can use each tool only once!</p> <p>The AI Agent has access to this tool: {tool_names_with_descriptions}</p> <p>The following is the previous conversation between a human and The AI Agent: {memory}</p> <p>AI Agent responses must start with one of the following:</p> <p>Thought: [the AI Agent’s reasoning process] 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.) Observation: [tool’s result] Final Answer: (on a new line) [final answer to the human user’s question] 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. If the AI Agent cannot find a specific answer after exhausting available tools and approaches, it answers with Final Answer: inconclusive</p> <p>Always include a disclaimer in your answer regarding AI generated content. Disclaimer: This is AI generated content—please use it with caution</p> <p>Question: {query} Thought: {transcript} """"</p>

Table 3. Image analysis prompt configuration.

Pipeline	Prompt Config
Image analysis	<p>In the following conversation, a human user interacts with the AI virtual assistant that has access to the survey images of Eurobarometer.</p> <p>Eurobarometer is a collection of cross-country public opinion surveys conducted regularly on behalf of the EU Institutions since 1974.</p> <p>Based on the image titled ‘{image_title}’, generate a response that answers the question.</p> <p>If the image information is not sufficient, say that the answer is not possible from the image alone.</p> <p>You should ignore your knowledge when answering the questions and be based solely on the attached image.</p> <p>Always include a disclaimer in your answer regarding AI generated content. Disclaimer: This is AI generated content—please use it with caution</p> <p>Question: {query}</p>

3.4. Agent Configuration and Chat Based on Images

To enhance our framework, we incorporated a multi-agent system comprising simple RAG pipelines, conversational agents, and direct interactions with images relevant to the initial RAG query. Each component is modular and configured independently, as illustrated

in Figure 3. The setup began with the initialization of simple RAG pipelines based on different indices (text content and images). Following this, the conversational pipeline is configured by assigning the RAG pipelines as agents within the tool, defining specific roles for each, and providing detailed instructions in the prompts. Regarding image utilization, the system allows users to select one or more images retrieved through the RAG on the Images pipeline based on their query. Users can then ask further questions about the selected images, enabling detailed information or analysis.

For the conversational agent configurations, in addition to the standard RAG prompts, an extra prompt was required to configure additional tools (agents) with specific names, roles, and descriptions, as shown in Table 4. For example, an agent might be named “Eurobarometer”, with a description such as “Useful for answering questions about the Eurobarometer surveys conducted by the EC and European Parliament”. This prompt incorporated a reasoning loop to provide comprehensive instructions for the model, guiding it on how to act while ensuring transparency in its reasoning process. The reasoning process is broken down into three components:

- Thought, which explains the agent’s reasoning process;
- Tool, which lists and selects the available agents;
- Observations, which represent the tool’s output (e.g., from the simple RAG pipeline), as presented in Table 2.

Table 4. Agents tools’ roles configuration.

Tool Name	Prompt Configuration
Eurobarometer	Useful for when you need to answer questions about the Eurobarometer surveys of the European Commission and European Parliament.
EurobarometerImage	Useful for when you need to answer questions about the Eurobarometer surveys based on Images’ metadata of the European Commission and European Parliament.

In addition to general guidelines, the model is instructed to generate reasoning output alongside the final answer, with the reasoning process depicted in Figure 3.

Furthermore, we integrated an alternative chat functionality centered on image-based interactions. This process utilized the previously configured simple RAG for images. When a user asked a question, the system not only provided an answer but also returned the images from which the answer was derived, constructed based on the alt text associated with these images. The images are stored in the environment’s session memory, initially captured during content retrieval at the start of the process. Alongside the RAG-based answer, the related images are presented to the user. The user can then select one or more images, which are sent in base64-encoded format for further queries. This enables users to delve deeper into specific insights provided by the Eurobarometer survey press releases.

This process required an additional prompt configuration for querying native images, incorporating the alt text of each image to enrich the content available to the model. For this implementation, we used GPT-4o [33], which is renowned for its performance; however, the system is modular and can accommodate other models capable of handling images. This flexibility ensures adaptability to varying requirements and use cases.

4. Results, Case Studies, and Applications

In the domain of digital transformation and e-governance, especially in large-scale survey data, the usage of LLMs could be very helpful and important. In our proposed framework, we used the well-known Eurobarometer survey data as the main source,

which provides valuable information for important matters within the EU. The reason for this selection is to highlight the potential benefits of state-of-the-art LLMs applicability in the e-governance sector, enhancing the digital transformation in the field. In our experiments, we utilized the available text and image data in dedicated Q&A pipelines based on their type and created a complete UI for querying.

To empower our Q&A pipelines, we indexed the press release page of the selected survey from the official Europa website [37], including the plain text and images, along with selected PDF files, which included an overview of the survey results; however, the index is under business preference and can be carried out in the framework's backend. Through the provided UI and virtual assistant tabs, the users can navigate through the general chat, where they can either query the conversational RAG or interact with standard text-based data. Users also have the option to switch to the image-based chat, where they can input their preferred questions, and related images are displayed in a listing format in a sidebar. Finally, users can query the system with a specific image to request additional information. All these features are explained in detail in the following subsections.

4.1. Conversational Virtual Assistant for Eurobarometer Public Opinion Data

In our case study, to demonstrate and evaluate the effectiveness of the proposed framework, we used Eurobarometer survey data results about the EU Humanitarian Aid, a survey with fieldwork dated September–October 2023 and published in January 2024, incorporating both text and image results, utilizing multi-agent RAG pipelines, and MLLMs for image analysis. To do so, we also constructed a user-friendly UI framework in Python 3.11.11, so the user could easily query in different pipelines based on their preference. Accordingly, for convenience, we present the results of the conversational pipeline with the agents (which can also be used separately) with detailed reasoning, the RAG based on images, and the image analysis based on the most relevant retrieved images on the query. However, using Google Colab, all results are reproducible, and the framework can easily be implemented.

For demonstration purposes, we selected the EU Humanitarian Aid [37] press release because it is rich with content in native text, a variety of types of images for analysis, and detailed PDFs. Moreover, while indexing, we selected to index the related Summary [38] and the Report [39] sections of this survey because they provided a detailed overview. Furthermore, all the press release content is indexed automatically upon retrieval, accompanied by the related/attached images.

Different types of questions were asked in the different pipelines. Starting with the conversational pipeline, we asked, “Please provide an overview of EU humanitarian aid based on the Eurobarometer survey and accompanying images”, as demonstrated in Figure 4. This pipeline had two agents configured: one for text-related content and one for image related content, each heading to a dedicated FAISS index. Therefore, with the above query, we instructed the model to investigate the answer in both agents.

As shown in Figure 4, the agents provided a comprehensive answer, illustrating the whole reasoning process. The answer started with detailing the thought process, stating the goal based on the query, and describing the model's next steps. Next, it stated that formulating the final answer would start with the first agent, which headed to the Eurobarometer text-related index, gathering the necessary information. Subsequently, it accessed the image-related RAG to gather additional information, listing additional key insights based on the dedicated image-related RAG. Lastly, after collecting all the necessary information and accessing both agents, the inputs were combined to construct and provide the final answer.

Choose a pipeline: Conversational Agent

Welcome to the chat!

Please provide an overview of EU humanitarian aid based on the Eurobarometer survey and accompanying images.

Tool: Eurobarometer
 Tool Input: Provide an overview of EU humanitarian aid based on the Eurobarometer survey.
 Observation: The Special Eurobarometer survey on EU humanitarian aid, conducted between September and October 2023, reveals key insights into EU citizens' awareness, attitudes, and support for humanitarian aid activities. The survey indicates a general awareness of EU funding for humanitarian aid, with significant variations based on socio-demographic factors such as age, education, and financial stability. Younger individuals and those with higher education levels tend to have more positive perceptions of the EU's role as a humanitarian donor.

Respondents prioritize health assistance, food insecurity, and climate change-related disasters as key areas for EU humanitarian aid. The majority believe that the EU should increase its investment in humanitarian aid, particularly in countries like Greece, Croatia, Malta, Romania, and Portugal. TV is identified as the most trusted source of information regarding EU humanitarian aid, followed by print and digital media.

Overall, the survey highlights a strong support for coordinated EU efforts in humanitarian aid, reflecting a positive sentiment towards the EU's role in addressing global humanitarian challenges.

Disclaimer: This is AI generated content — please use it with caution.

Tool: EurobarometerImage
 Tool Input: Provide metadata and insights from images related to EU humanitarian aid in the Eurobarometer survey.
 Observation: The metadata from the Eurobarometer survey images indicates several key insights regarding EU humanitarian aid.

1. ****Importance of Humanitarian Aid****: A significant 91% of EU citizens believe that EU funding for humanitarian aid activities is very or fairly important.
2. ****Efficiency of Coordination****: About 71% of respondents feel that humanitarian aid is more efficient when coordinated and provided by the EU as a whole.
3. ****EU's Role as a Donor****: The EU is recognized as one of the leading humanitarian aid donors globally, with annual spending ranging from approximately 1.5 to 2 billion euros, translating to about 4 to 5 euros per taxpayer in the EU.
4. ****Public Opinion on Spending Levels****: Regarding future spending, 47% of respondents support maintaining the current level, while 40% advocate for increased investment in humanitarian aid.
5. ****Pride in EU's Contributions****: A strong 83% of respondents express pride, enthusiasm, and satisfaction with the EU's role as a major donor of humanitarian aid.
6. ****Awareness of Funding****: The survey indicates that 68% of EU citizens are aware of the EU's funding for humanitarian aid activities.

Disclaimer: This is AI generated content — please use it with caution.

Final Answer: The Eurobarometer survey on EU humanitarian aid, conducted between September and October 2023, reveals that EU citizens generally support and are aware of the EU's humanitarian aid efforts. Key findings include a high level of awareness (68%) of EU funding for humanitarian aid and a strong belief (91%) in its importance. The survey highlights that younger individuals and those with higher education levels have more positive perceptions of the EU's role as a humanitarian donor. Health assistance, food insecurity, and climate change-related disasters are prioritized areas for aid. A majority of respondents (71%) believe that aid is more efficient when coordinated by the EU, and there is a call for increased investment, particularly in certain EU countries. The EU is recognized as a leading global donor, with annual spending of 1.5 to 2 billion euros. Public opinion is divided on future spending levels, with 47% supporting the current level and 40% advocating for an increase. Additionally, 83% of respondents express pride in the EU's contributions. TV is the most trusted source of information on EU humanitarian aid.

Disclaimer: This is AI generated content — please use it with caution.

Ask a question

Send

Figure 4. Example of the reasoning process in the conversational Q&A pipeline, demonstrating how agents collaborate to provide a comprehensive response to user queries about Eurobarometer data ("**" denotes intended bold formatting).

The Eurobarometer text and image-related agents offered complementary insights based on the initial query about the EU humanitarian aid but focused on various aspects.

The text-related agent highlighted public sentiment, showing that 91% of respondents value EU-funded aid. Moreover, it identified priorities, such as health, food security, and climate-related challenges, as key focus areas for EU humanitarian aid based on the survey's results and explored how people learn about aid, with TV being the most trusted source, and reflected citizens' pride in the EU's role as a leading donor.

The images-related metadata agent took a more detailed approach, detailing financial and operational aspects. It confirmed the 91% importance figure but added that 71% see aid as more effective when co-ordinated by the EU. It highlighted EU spending of EUR 1.5–2 billion annually (about EUR 4–5 per taxpayer) and noted that 47% of respondents want funding maintained, while 40% support an increase. Emotional responses are also quantified, with 83% expressing pride, enthusiasm, and satisfaction with the EU's efforts.

In summary, the text-related agent provides a qualitative view of public opinions and priorities, while the image-related agent offered a quantitative look at funding, efficiency, and future preferences. This is because the text-related agent included documents that provide a comprehensive overview of EU humanitarian aid, since it includes the whole press release text, the summary, and the report of the survey; however, the image-related content is stuck to the highlights provided in images in a press release, giving more specific information. Together, they gave a full picture of EU humanitarian aid, blending public attitudes with measurable outcomes.

4.2. Virtual Assistant Based on Images and Analysis for Eurobarometer Public Opinion Data

Moreover, in our case study, we structured and evaluated the use of combining the RAG pipeline based on image-related data and the follow-up images attached for further analysis. The images' related pipeline and chat involved two stages. The process is initiated by querying in the simple RAG indexed to image-related content. Based on this Q&A, the relevant images are retrieved, as previously indexed. Those images are previewed for a user to select to proceed with the image analysis pipeline with their query and attach the selected image with 64 formats, based on the multimodal GPT-4o [33]. If the user selects no images and no attached images, the query is headed again to simple RAG based on images.

The selected survey about EU humanitarian aid was contextually rich, with different types of visuals, enabling us to test and evaluate its performance. Therefore, as presented in Figure 5, we asked, "Please provide key insights into the awareness of EU humanitarian aid funds and activities". Therefore, the RAG based on image-related data provided a comprehensive answer, synthesizing data from all image contexts. Additionally, with the presented response, the related images are previewed based on the user's query in the right scroll-down toolbar, making them accessible for further queries.

Next, we selected one of the previewed images for further analysis, a heatmap as presented in Figure 6. The prompt configured for this image analysis system also included the question, which is related to the attached image, to provide additional context to the LLM, enhancing the response accuracy. Based on the previewed results, the LLM provided a comprehensive answer, analyzing the heatmap and elaborating further on this. As we can see, even though heatmap visualizations can be difficult to interpret, incorporating LLMs, especially in e-governance applications like our case study, can provide an overview of key insights that might be hard to detect at first glance. Additionally, the current overview of the results offers useful information that may not be immediately obvious, even in a visual representation. By providing context to the model, such as the title of the image, we enable further analysis.

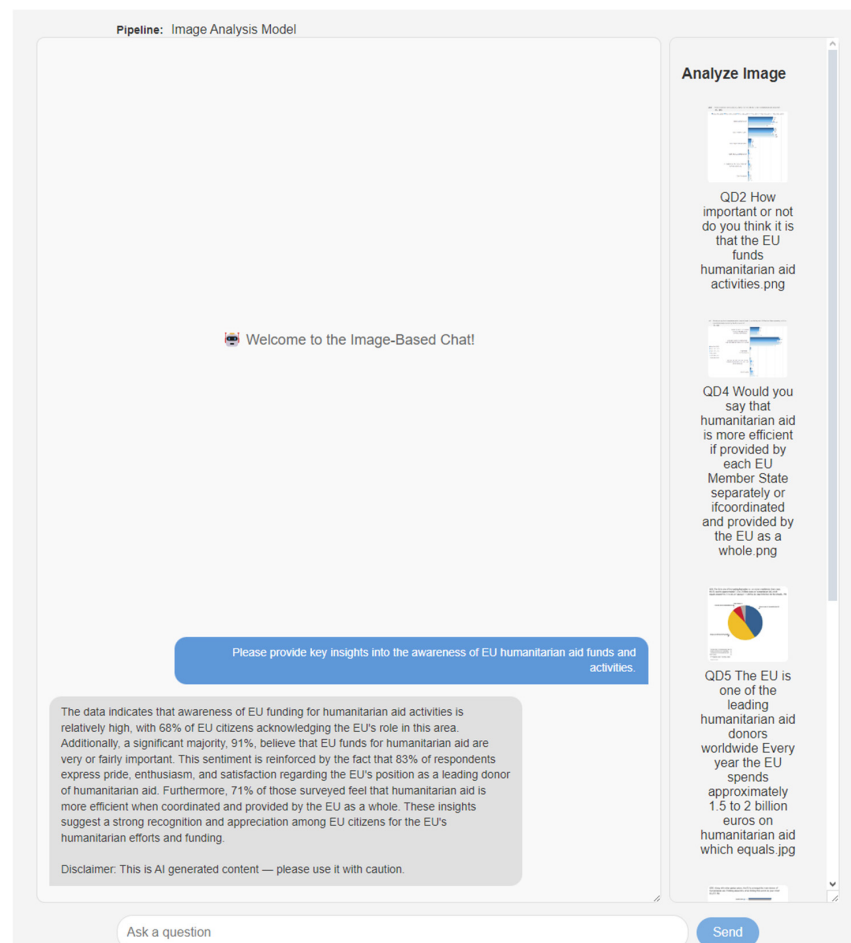


Figure 5. Image content RAG pipeline.

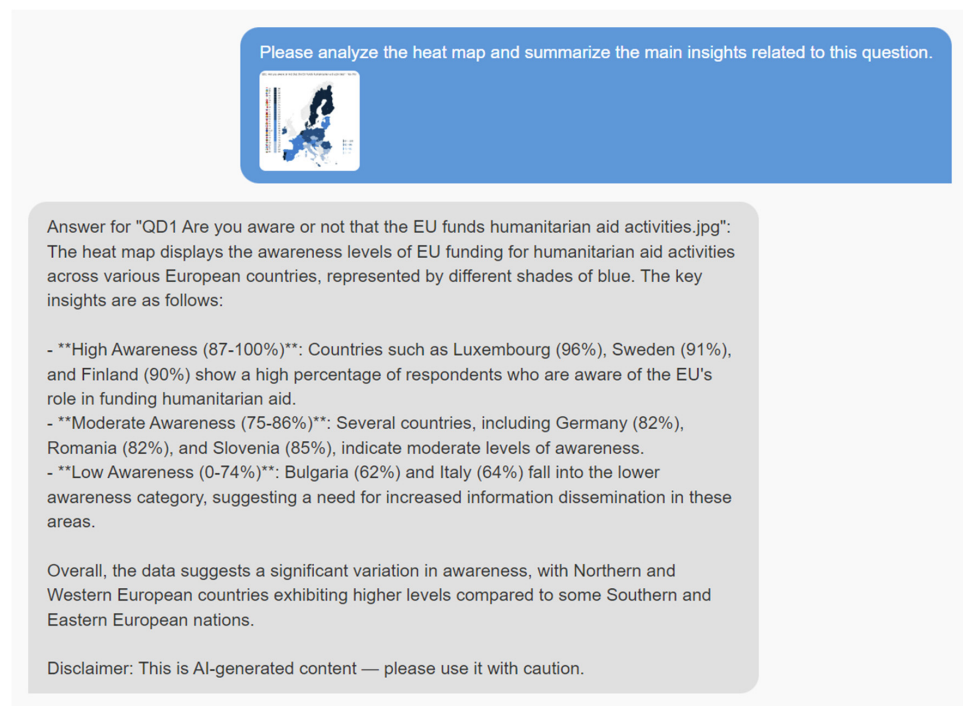


Figure 6. Query in analyzing image content related to QD1: Are you aware or not that the EU funds humanitarian aid activities ("***" denotes intended bold formatting).

The analysis process highlights how requesting additional information through attached images and charts allows for a deeper, more context-specific understanding of the data. By focusing only on the visual content and statistical elements included in the image, the insights are based on the presented evidence, ensuring accuracy and relevance, whilst identifying gaps or trends that require further analysis. Incorporating a focused prompt configuration on Eurobarometer data with an analysis of visual data allowed us to extract useful insights. The findings highlight critical areas for improvement in awareness strategies, ensuring equitable understanding and engagement across all EU member states.

Moreover, we continued by attaching another related image, a pie chart, and asking for more details about this analysis. As presented in Figure 7, the response was more elaborated within the context of the related survey question, since, in the prompt configuration, we instructed the LLM to provide an answer solely based on the attached image context and ignore its pretrained knowledge. The analysis organized the input data in detail, highlighting trends and patterns, ensuring that the response included the main components, and it was structured to facilitate straightforward interpretation of the insights, maintaining relevance to the question's context, which were also visualized in the pie chart.

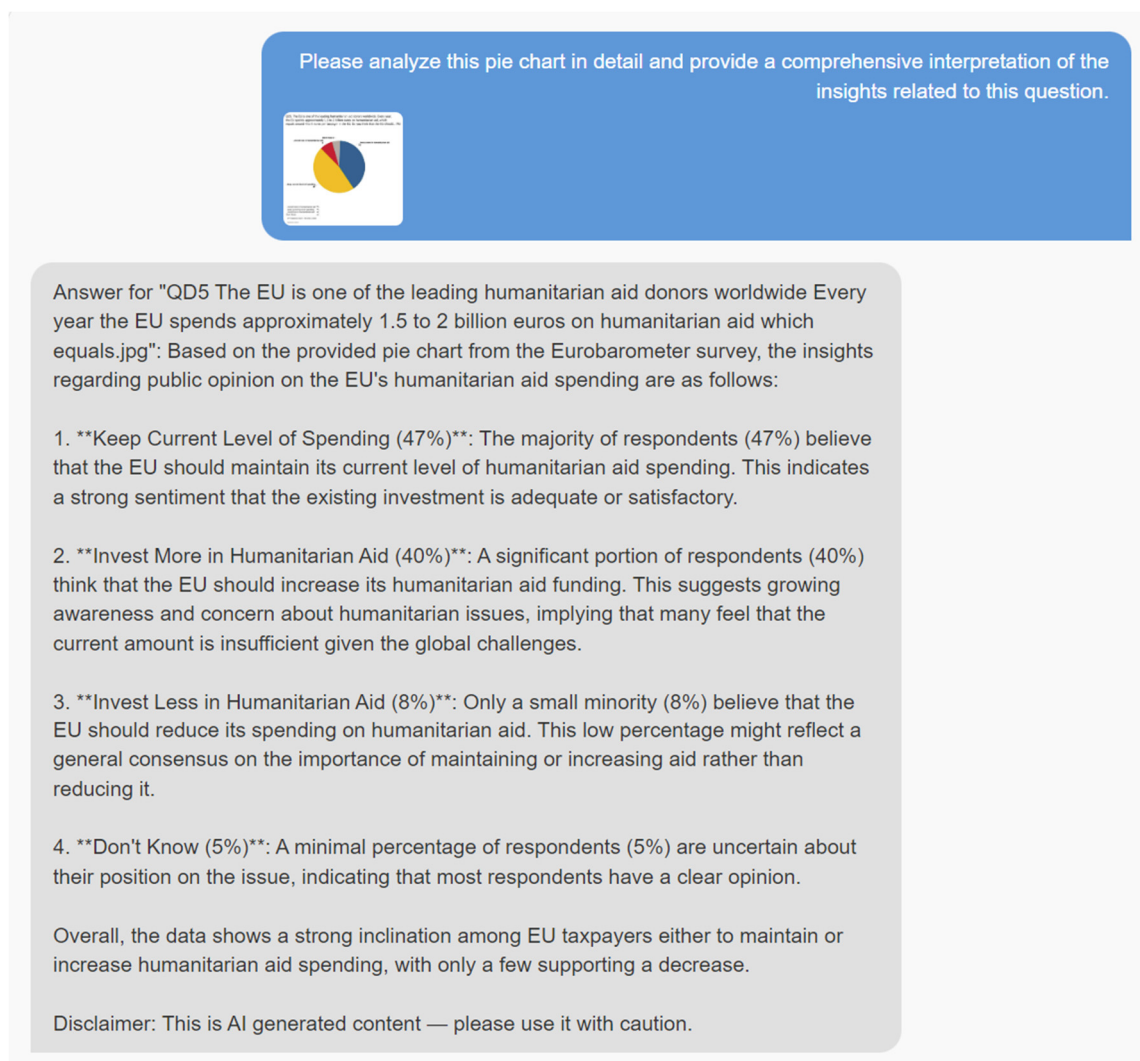


Figure 7. Query in analyzing image related to QD5: The EU is one of the leading humanitarian aid donors worldwide. Every year the EU spends 1.5 to 2 billion euros on humanitarian aid which equals ("**" denotes intended bold formatting).

Lastly, we queried upon a bar chart, asking for analysis thoroughly and providing an overview of the insight for the related question. As previewed in Figure 8, the response included a detailed explanation of what this image is about with reference to the related question, grouping the findings based on the results and providing a complete overview of the percentage results and differences between the years.

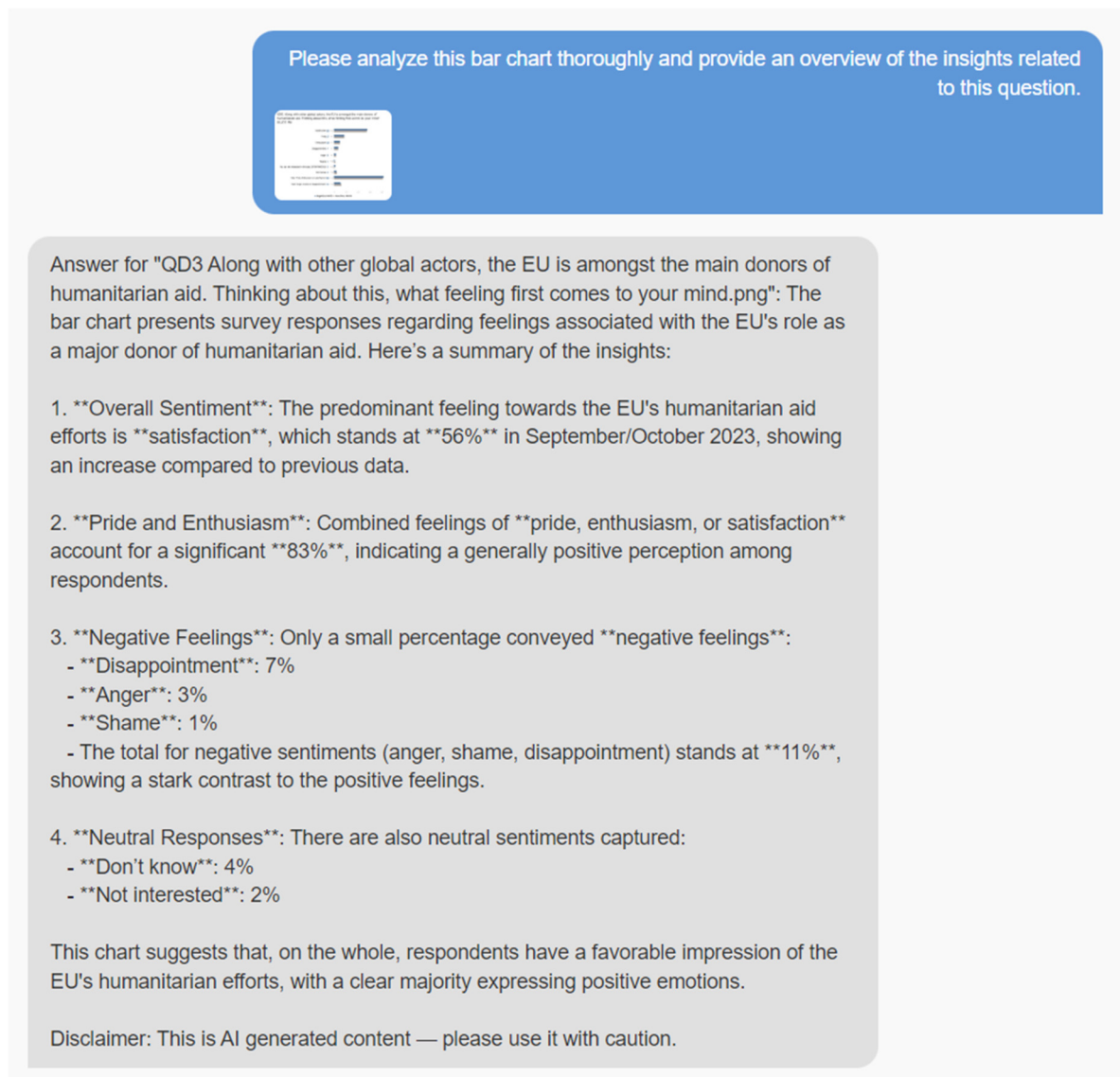


Figure 8. Query in analyzing image related to QD3: Along with other global actors, the EU is amongst the main donors of humanitarian aid. Thinking about this, what feeling first comes to your mind ("**" denotes intended bold formatting).

Overall, the RAG based on image-related content and further analysis of these provides a valuable tool for a detailed understanding of the finding of each piece of survey information, since the user can either generally query on the RAG and then be able to see the related images from his initial query to understand further and in more detail about the results, using visual OCR, but also fast and accurately obtain the insights of each. It is worth mentioning that the number of retrieved relevant images is based on the user query to be able to fast track these and no other unrelated images based on his needs, with all images being publicly available at [37]. Accordingly, if the user does not select any image, they should ask again any question related to their preference and accompany the answer

to obtain the newly related images. Also, the configuration of this is modular, using the multimodal LLM under business needs and the number of the relevant retrieved images but also configuring the prompt as business needs, with our example presented in Table 3.

4.3. Reproducibility and Scalability Framework

To maintain consistency and reliability across different e-governance applications, it is essential to ensure that our proposed implementation can be reproduced. Our presented framework, built in Python with Haystack Fast API and OpenAI's components, is provided in a low-code notebook publicly available at <https://github.com/gpapageorgiouedu/A-Multimodal-Framework-Embedding-Retrieval-Augmented-Generation-with-MLLMs-for-Eurobarometer-Data> (accessed on 23 December 2024), and is easily deployed in Google Colab (Intel Xeon CPU @ 2.20 GHz, 12 GiB RAM, NVIDIA Tesla T4 with 15 GB memory, CUDA 12.2 support, Operating Environment: 64-bit x86_64 architecture with little-endian byte order), including all required files for UI and required libraries. The components for reproducing the chat capabilities are included in the repository, with detailed documents and step-by-step instructions for deployment.

The modular design of the proposed framework makes it adaptable for various domains and data sources beyond Eurobarometer surveys. By leveraging RAG and MLLMs, the framework can manage diverse data formats, such as textual reports, images, and structured datasets. This versatility allows it to be applied in domains such as healthcare (e.g., analyzing electronic health records and medical imaging), education (e.g., integrating multimedia resources for adaptive learning platforms), and finance (e.g., analyzing market trends from reports and charts). Furthermore, the use of scalable indexing and retrieval mechanisms ensures that the framework can accommodate large datasets, making it suitable for applications in research, governance, and industrial sectors. Future iterations of the framework could explore domain-specific enhancements, such as fine-tuned retrieval pipelines or custom multimodal encoders, to further optimize its performance in specialized contexts.

Additionally, with slight modifications, the presented framework can be adjusted based on different business needs. It is built in a modular way and all its components, including RAG configuration, models, and processes, can be changed with alternatives and can be easily transferred to cloud or local hosting environments.

Implementing the proposed framework on a larger scale involves several considerations; however, since the architecture is modular, each component can be scaled individually based on business needs for e-governance implications. Lastly, the proposed framework can run both locally, with locally hosted LLMs, or with API providers and cloud-based solutions, scalable on demand. However, it should be considered that the effectiveness and accuracy of the framework are based on the components selected and, before adjustments, the provider's documentation release notes should be considered.

4.4. Challenges and Limitations

Our proposed framework highlights the advantages of knowledge extraction compared with typical observatory methods; however, it is crucial to address several challenges and limitations. A primary concern that needs to be considered is privacy and security of sensitive government data if it needs to be used. E-government applications necessitate robust encryption techniques, stringent access controls, and strict adherence to legal and regulatory frameworks to mitigate risks effectively.

Another challenge lies in efficiently managing computational resources. In large-scale applications, the performance of GAI models is often tied to their size and complexity; while locally hosted larger models typically provide superior results, they also come with substantially higher computational and financial costs. Since the framework's perfor-

mance depends on the selected models and components, the corresponding performance evaluation should be conducted based on the chosen components and the stakeholders' acceptance criteria. Moreover, costs should be considered based on the efficiency needed, and reliance on corporate LLM services may further increase expenses, depending on the models and services employed.

To overcome these challenges, continuous monitoring and fine-tuning of system performance are essential, ensuring efficient deployment, while maintaining optimal performance across various applications. Additionally, the proposed framework offers a flexible and open-source-compatible architecture, since it could be implemented with open-source LLMs, and Haystack's orchestration capabilities tailored to the specific data sources and requirements of each use case. This flexibility enables the development of scalable, financially viable solutions, facilitating broader access to advanced AI technologies for e-government initiatives. However, based on stakeholders' needs for performance, security, and scalability, a detailed investigation and careful selection of the framework's preferred components should be conducted, along with testing aligned with their acceptance criteria.

4.5. Alignment with the AI Act and Ethical AI Principles

Our presented framework for combining multimodal RAG into e-governance applications aimed to align with the principles of the AI Act under the ethical guidelines for deploying AI in the public sector, aligning with AI Act guidelines for avoiding discrimination, adaptability, and respecting fundamental rights [40]. The proposed multimodal RAG for Eurobarometer survey data is designed to be modular, with the option for using different LLMs based on the needs and requirements, and, with each of the components to be easily modified under preferences, having the option to run both on-premises or cloud services, having as the target to provide assistance to public sector agencies, maintaining control and compliance to AI Act requirements and transparency standards [41]. The modular prompt configuration, agents and tools set up, and their reasoning process provided a clear view of the results in each Q&A and, on query, based on the returned images of the RAG, provided the option for insightful analysis, allowing effective monitoring and oversight, which is crucial for trustworthy AI deployments. Moreover, each user has the option to select which survey they need to Q&A on and analyze, providing flexibility for multiple use cases.

The proposed framework incorporates multiple measures to mitigate bias and ensure transparency, aligning with ethical AI principles and the EU AI Act. To address bias, the framework uses diverse and representative data sources during the indexing phase to minimize potential skewness in retrieved content. Transparency is achieved through the modular design of the implementation, which allows users to trace the origin of retrieved content, supported by detailed metadata for each indexed document or image. The inclusion of reasoning steps in the conversational agents ensures that users can understand how responses are generated, promoting trust and accountability. These features collectively ensure that the framework adheres to ethical standards while providing reliable and unbiased insights to stakeholders.

5. Conclusions

This study proposed a comprehensive multimodal framework integrating RAG with LLMs to address the challenges of analyzing Eurobarometer survey data. The primary aim was to enhance data analysis, interpretability, and insight extraction across diverse formats, including text and images. By combining multimodal indexing with advanced LLM capabilities, the framework enables targeted data retrieval, sophisticated analysis, and trend visualization, providing a holistic understanding of public opinion dynamics.

It is a complete multimodal virtual assistant with agents, using an open-source backend, Haystack, FAISS, and Fast API, applied to e-governance and public sector survey data. The proposed framework encompasses independently modular and scalable components, using GPT-4o [33] and GPT-4o-mini [32] as main LLMs for Q&A and flan-t5-base-samsum for Q&A summary generation and different kinds of embedding models for text and image data.

Key objectives achieved include the development of a scalable and modular framework capable of indexing and retrieving mixed-format data efficiently. The proposed implementation demonstrated its ability to support survey studies, comparative analyses, and real-time querying, showcasing its value in policy-making and social research. The integration of image-based queries and the use of conversational agents further expanded the analytical depth, bridging gaps in traditional methods.

The main source of this research is Eurobarometer data, a collection of cross-country public opinion surveys conducted regularly on behalf of the EU Institutions since 1974. The indexing process was based on the FAISS open-source vectorized database, and embeddings were calculated and indexed from Eurobarometer's press release, accompanied by PDF files (under user selection) and images included in the press release. Moreover, all the mentioned Q&A functionalities are accessible from a custom CSS and HTML UI and deployed to increase reproducibility in Google Colab.

The results underscore the framework's potential to transform Eurobarometer data into actionable insights, supporting evidence-based decision making. By ensuring alignment with ethical AI principles and compliance with the EU AI Act, the framework is positioned as a trustworthy and adaptable solution for e-governance applications.

There were four implemented pipelines: two standard RAG pipelines, one for text-related data, one for image-related data, one conversational pipeline with agents and memory, and, lastly, an image analysis Q&A pipeline based on the related images in the query from the RAG based on images. To implement this framework, we based our experiments on a case study scenario using one of the latest Eurobarometer surveys. We showcased the efficiency the proposed framework could provide in the public sector, highlighting the accessibility and interpretation ability of Eurobarometer insights with a virtual assistant. Additionally, we showcased the flexible and straightforward way of multimodal data retrieval and dedicated image analysis for identifying trends and combining multiple information types of sources. Lastly, with a priority on transparency and respecting the ethical standards for the responsible use of Gen AI capabilities, we demonstrate a detailed solution for Q&A on rich survey data for knowledge extraction, demonstrating the benefits that could be extracted for public sector agencies, with enhanced multimodal Q&A capabilities, enabling more informed and efficient data retrieval, while highlighting the need for performance assessment to evaluate its effectiveness, based on the selected architectural components, specific business requirements, and practical applications. Even though the framework suggests potential improvements in data interpretation, formal evaluations with stakeholders and quantitative metrics are necessary to confirm these benefits.

Future Work

This research establishes a robust foundation for integrating RAG frameworks with LLMs to analyze complex Eurobarometer survey data. However, several fields remain open for future exploration and development. One of the main future directions involves incorporating additional modalities, such as audio input and output [42]. Moreover, the indexing of vast survey data and agents is configured based on the themes of the surveys for domain-specific applications, such as healthcare, education, or environmental

policy, which would underscore its versatility and expand its practical utility to enhance cross-modal reasoning capabilities.

Additionally, further research should be applied to the configuration of the agents for multimodal capabilities, for identifying trends in survey data, and to provide an overview of comparative studies across survey iterations, enhancing cross-modal reasoning capabilities. Moreover, based on the current modular and scalable framework, improvements can be applied to UI adjustments, incorporating advanced data visualization tools to improve further the accessibility and usability of the framework for stakeholders without technical expertise.

The development of the proposed multimodal RAG framework paves the way for collaborations with governmental bodies, research institutes, and private organizations to further enhance its capabilities. Partnerships with governmental institutions could support the integration of the framework into e-governance platforms, enabling real-time public opinion analysis to inform policymaking. Collaborations with academic and research institutes could focus on advancing the technical aspects of the framework, such as fine-tuning retrieval mechanisms, improving cross-modal reasoning, and extending its applicability to new domains like healthcare or education. Looking ahead, the roadmap for future work includes the incorporation of additional data modalities, such as audio and video inputs, and exploring real-time data retrieval for dynamic datasets. Another priority is to enhance user interfaces to make the framework more accessible to nontechnical users, alongside establishing mechanisms to gather user feedback for iterative improvements.

Lastly, integrating mechanisms to collect user feedback on the retrieval part and Q&A are essential for iterative improvement, since insights from survey researchers, policymakers, and other end-users could inform targeted enhancements, ensuring the framework evolves to meet diverse research and policy needs effectively. By focusing on this future research, the proposed framework can advance as a transformative tool for extending multimodal survey data analysis, facilitating innovation in public opinion research and related disciplines for e-governance implications.

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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in this study are openly available on GitHub at <https://github.com/gpapageorgiouedu/A-Multimodal-Framework-Embedding-Retrieval-Augmented-Generation-with-MLLMs-for-Eurobarometer-Data> (accessed on 23 December 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. Nissen, S. The Eurobarometer and the Process of European Integration: Methodological Foundations and Weaknesses of the Largest European Survey. *Qual. Quant.* **2014**, *48*, 713–727. [CrossRef]
2. Febrian, G.F.; Figueredo, G. KemenkeuGPT: Leveraging a Large Language Model on Indonesia's Government Financial Data and Regulations to Enhance Decision Making. *arXiv* **2024**, arXiv:2407.21459.
3. Kagaya, T.; Yuan, T.J.; Lou, Y.; Karlekar, J.; Pranata, S.; Kinose, A.; Oguri, K.; Wick, F.; You, Y. RAP: Retrieval-Augmented Planning with Contextual Memory for Multimodal LLM Agents. *arXiv* **2024**, arXiv:2402.03610.

4. 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. [CrossRef]
5. Wang, J.; Jiang, H.; Liu, Y.; Ma, C.; Zhang, X.; Pan, Y.; Liu, M.; Gu, P.; Xia, S.; Li, W.; et al. A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks. *arXiv* **2024**, arXiv:2408.01319.
6. Huang, D.; Yan, C.; Li, Q.; Peng, X. From Large Language Models to Large Multimodal Models: A Literature Review. *Appl. Sci.* **2024**, *14*, 5068. [CrossRef]
7. Xu, M.; Yin, W.; Cai, D.; Yi, R.; Xu, D.; Wang, Q.; Wu, B.; Zhao, Y.; Yang, C.; Wang, S.; et al. A Survey of Resource-Efficient LLM and Multimodal Foundation Models. *arXiv* **2024**, arXiv:2401.08092.
8. Han, Y.; Zhang, C.; Chen, X.; Yang, X.; Wang, Z.; Yu, G.; Fu, B.; Zhang, H. A Multimodal LLM for Chart Understanding and Generation. *arXiv* **2023**, arXiv:2311.16483.
9. Zhan, J.; Dai, J.; Ye, J.; Zhou, Y.; Zhang, D.; Liu, Z.; Zhang, X.; Yuan, R.; Zhang, G.; Li, L.; et al. AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. *arXiv* **2024**, arXiv:2402.12226.
10. McKinzie, B.; Gan, Z.; Fauconnier, J.-P.; Dodge, S.; Zhang, B.; Dufter, P.; Shah, D.; Du, X.; Peng, F.; Weers, F.; et al. MM1: Methods, Analysis & Insights from Multimodal LLM Pre-Training. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024.
11. Hu, W.; Xu, Y.; Li, Y.; Li, W.; Chen, Z.; Tu, Z. BLIVA: A Simple Multimodal LLM for Better Handling of Text-Rich Visual Questions. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 3. [CrossRef]
12. Ge, Z.; Huang, H.; Zhou, M.; Li, J.; Wang, G.; Tang, S.; Zhuang, Y. WorldGPT: Empowering LLM as Multimodal World Model. *Assoc. Comput. Mach.* **2024**, *1*, 7346–7355.
13. Fu, C.; Lin, H.; Long, Z.; Shen, Y.; Zhao, M.; Zhang, Y.; Dong, S.; Wang, X.; Yin, D.; Ma, L.; et al. VITA: Towards Open-Source Interactive Omni Multimodal LLM. *arXiv* **2024**, arXiv:2408.05211.
14. Wu, S.; Fei, H.; Qu, L.; Ji, W.; Chua, T.-S. NEX-T-GPT: Any-to-Any Multimodal LLM. *arXiv* **2023**, arXiv:2309.05519.
15. Islam, R.; Moushi, O.M. GPT-4o: The Cutting-Edge Advancement in Multimodal LLM. *TechRxiv* **2024**. [CrossRef]
16. Byun, J.; Kim, B.; Cha, K.A.; Lee, E. Design and Implementation of an Interactive Question-Answering System with Retrieval-Augmented Generation for Personalized Databases. *Appl. Sci.* **2024**, *14*, 7995. [CrossRef]
17. Yao, C.; Fujita, S. Adaptive Control of Retrieval-Augmented Generation for Large Language Models Through Reflective Tags. *Electronics* **2024**, *13*, 4643. [CrossRef]
18. Han, B.; Susnjak, T.; Mathrani, A. Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview. *Appl. Sci.* **2024**, *14*, 9103. [CrossRef]
19. Xu, K.; Zhang, K.; Li, J.; Huang, W.; Wang, Y. CRP-RAG: A Retrieval-Augmented Generation Framework for Supporting Complex Logical Reasoning and Knowledge Planning. *Electronics* **2025**, *14*, 47. [CrossRef]
20. Miao, J.; Thongprayoon, C.; Suppadungsuk, S.; Garcia Valencia, O.A.; Cheungpasitporn, W. Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications. *Medicina* **2024**, *60*, 445. [CrossRef]
21. Nong, S.; Zhu, J.; Wu, R.; Jin, J.; Shan, S.; Huang, X.; Xu, W. MobileFlow: A Multimodal LLM For Mobile GUI Agent. *arXiv* **2024**, arXiv:2407.04346.
22. Caffagni, D.; Cocchi, F.; Barsellotti, L.; Moratelli, N.; Sarto, S.; Baraldi, L.; Cornia, M.; Cucchiara, R. The Revolution of Multimodal Large Language Models: A Survey. *arXiv* **2024**, arXiv:2402.12451.
23. Zhang, D.; Yu, Y.; Dong, J.; Li, C.; Su, D.; Chu, C.; Yu, D. MM-LLMs: Recent Advances in MultiModal Large Language Models. *arXiv* **2024**, arXiv:2401.13601.
24. 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. GitHub Repository. 2019. Available online: <https://github.com/deepset-ai/haystack> (accessed on 23 December 2024).
25. Kärkkäinen, K.; Joo, J. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv* **2019**, arXiv:1908.04913.
26. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. *Proc. Mach. Learn. Res.* **2021**, *139*, 8748–8763.
27. OpenAI Text-Embedding-Ada-002. 2003. Available online: <https://platform.openai.com/docs/guides/embeddings> (accessed on 23 December 2024).
28. OpenAI Simple Evals. 2025. Available online: <https://github.com/openai/simple-evals> (accessed on 23 December 2024).
29. Shahriar, S.; Lund, B.D.; Mannuru, N.R.; Arshad, M.A.; Hayawi, K.; Bevara, R.V.K.; Mannuru, A.; Batool, L. Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency. *Appl. Sci.* **2024**, *14*, 7782. [CrossRef]
30. Wu, Y.; Hu, X.; Fu, Z.; Zhou, S.; Li, J. GPT-4o: Visual Perception Performance of Multimodal Large Language Models in Piglet Activity Understanding. *arXiv* **2024**, arXiv:2406.09781.
31. Islam, R.; Moushi, O.M. EasyChair Preprint GPT-4o: The Cutting-Edge Advancement in Multimodal LLM GPT-4o: The Cutting-Edge Advancement in Multimodal LLM; 2024; 13757. Available online: <https://easychair.org/publications/preprint/z4TJ> (accessed on 23 December 2024).

32. OpenAI GPT-4o Mini: Advancing Cost-Efficient Intelligence. 2024. Available online: <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/> (accessed on 23 December 2024).
33. OpenAI GPT-4o. 2024. Available online: <https://openai.com/index/hello-gpt-4o> (accessed on 23 December 2024).
34. Phlischmid Flan-T5-Base-Samsum. 2024. Available online: <https://huggingface.co/phlischmid/flan-t5-base-samsum> (accessed on 23 December 2024).
35. 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*; Liu, Q., Schlangen, D., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 38–45.
36. Deepset Customizing Agent. 2024. Available online: https://haystack.deepset.ai/tutorials/25_customizing_agent (accessed on 23 December 2024).
37. European Commission Humanitarian Aid Survey. Available online: <https://europa.eu/eurobarometer/surveys/detail/2976> (accessed on 23 December 2024).
38. European Commission; Directorate-General for European Civil Protection and Humanitarian Aid Operations (ECHO). *Summary—Special Eurobarometer 542-EU Humanitarian Aid*; Publication Office of the European Union: Brussels, Belgium, 2024. [[CrossRef](#)]
39. European Commission; Directorate-General for European Civil Protection and Humanitarian Aid Operations (ECHO). *Report—Special Eurobarometer 542-EU Humanitarian Aid*; Publication Office of the European Union: Brussels, Belgium, 2024. [[CrossRef](#)]
40. Madiaga, T.; Chahri, S. *Artificial Intelligence Act; EU Legislation in Progress*; European Parliament: Brussels, Belgium, 2024.
41. European Parliament. EU AI Act: First Regulation on Artificial Intelligence: Topics: European Parliament. Available online: www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed on 1 July 2024).
42. Ma, Z.; Wu, W.; Zheng, Z.; Guo, Y.; Chen, Q.; Zhang, S.; Chen, X. Leveraging Speech PTM, Text LLM, and Emotional TTS for Speech Emotion Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Republic of Korea, 14 April 2024; pp. 11146–11150.

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.