

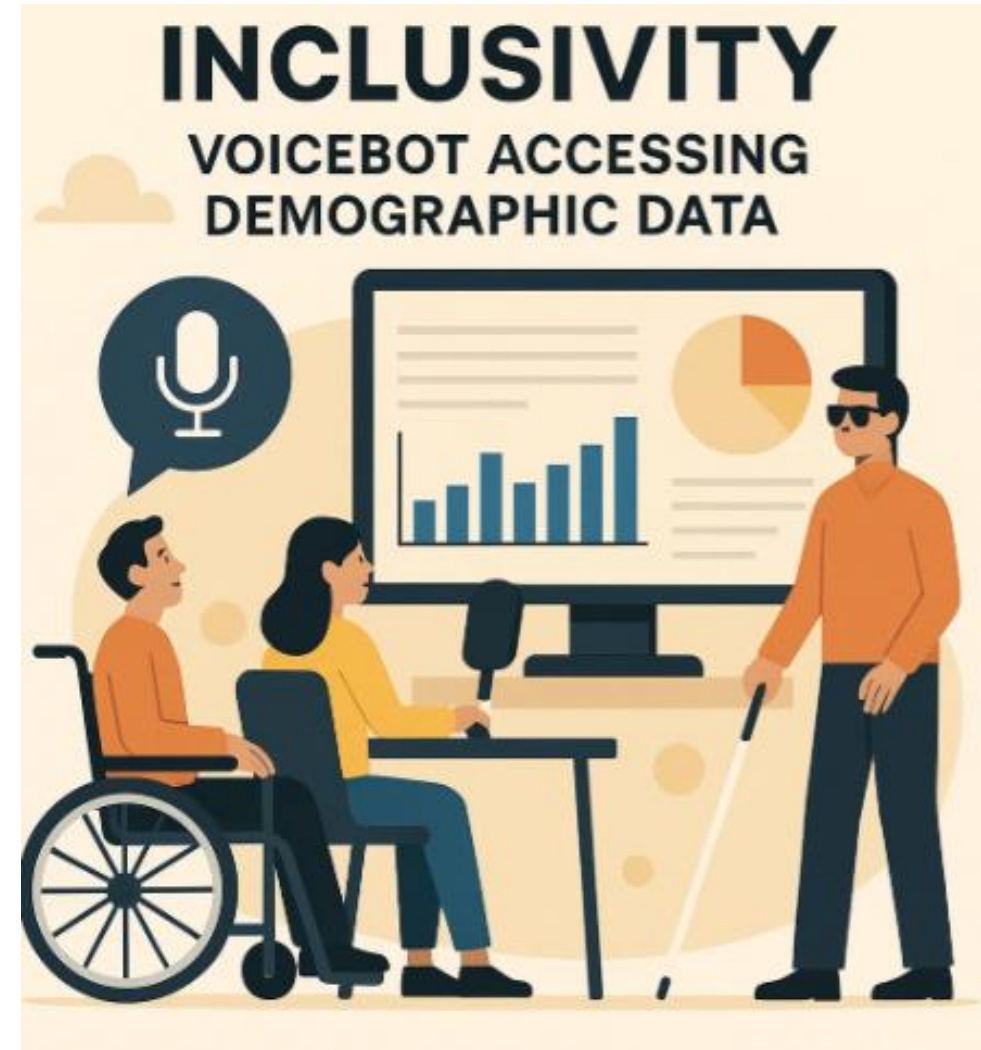
Talk to the data: a voice assistant to explore population statistics

Simona Cafieri

Istat

Outline

- ✓ Background
- ✓ Our project
- ✓ Data
- ✓ System Architecture
- ✓ Results
- ✓ Performance analysis
- ✓ Metodological contribution
- ✓ Conclusions



Background

- ✓ Generative AI refers to the application of AI capable of creating unique content using stochastic generators and transformer models.
- ✓ Thanks to recent developments in GenAI, combining dynamic external data with LLMs, Retrieval Augmented Generation (RAG) has emerged as a technique that improves responses by extracting pertinent information from additional sources. This approach combines retrieval-based and generative models to improve the effectiveness of knowledge-intensive tasks
- ✓ AI-powered chatbots are generated here: they may use LLMs and RAG innovations for intelligent, context-aware conversations, analyzing difficult questions, and producing human-like responses. Voice agent systems have emerged as a major form of interaction
- ✓ Voice-based assistants are AI applications that can interpret and respond to user requests through speech recognition and natural language processing .
- ✓ By combining dynamic external data with LLMs and using Retrieval Augmented Generation (RAG), thereby enhancing the overall user experience and satisfaction.



Our project



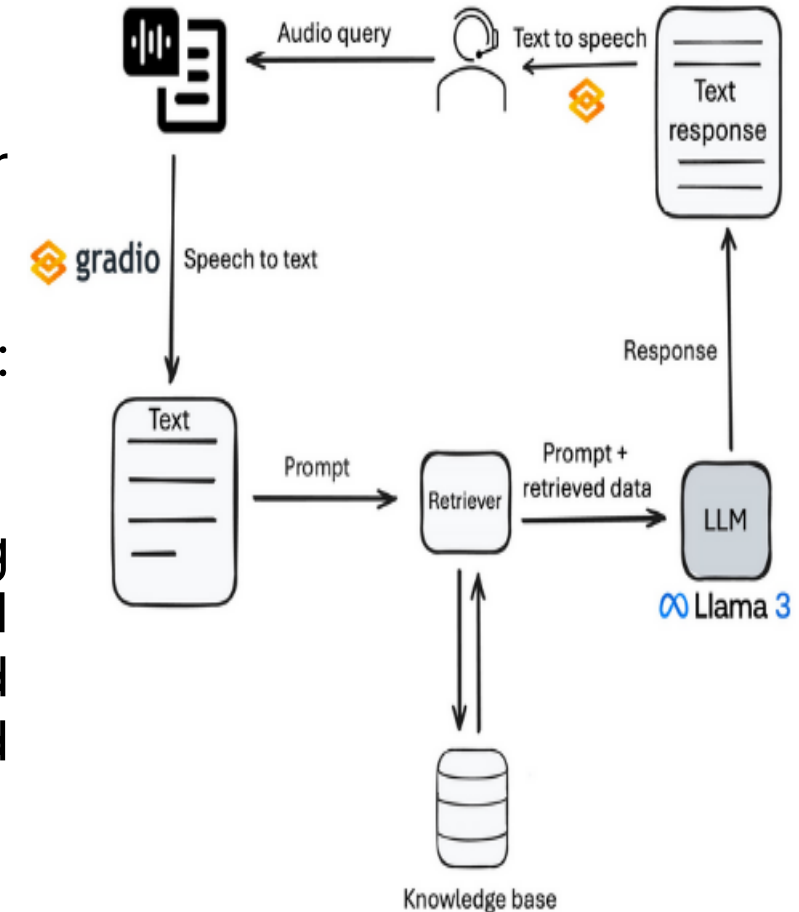
- ✓ Present voicebot technologies lack significant humanization
- ✓ As a result, some users,, are dissatisfied with their interactions. To address this, it is critical to create intelligent speech agent systems that generate user satisfaction
- ✓ Our project presents the creation of a voice-activated chatbot intended to retrieve ISTAT population data. Using cutting-edge artificial intelligence technology can help improve data accessibility and usability for users and academics. New speech recognition and LLM-based methods will be used by the proposed voicebot to understand and naturally and attractively answer user inquiries. The voicebot seeks to offer intelligent, context-aware conversations

Data

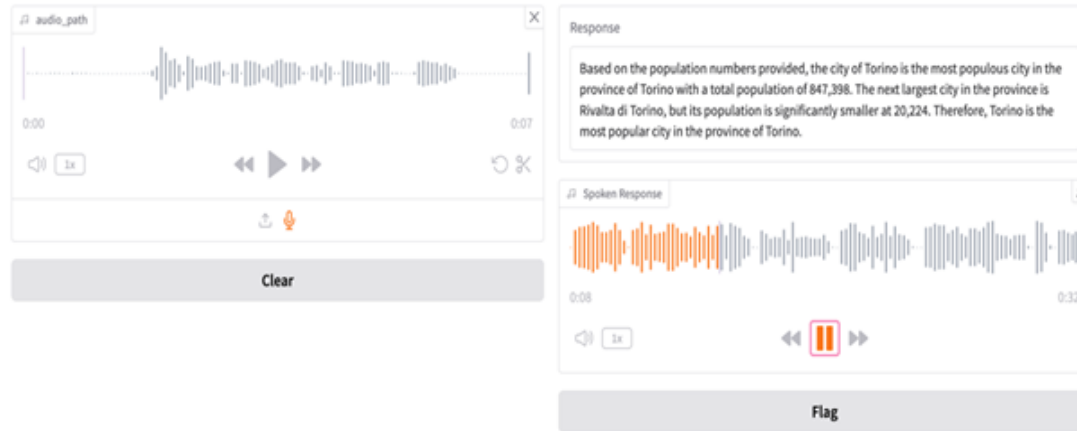
- ✓ The Istat database demo.istat.it provides extensive demographic data, accessible through web interfaces and downloadable datasets.
- ✓ The data covers key aspects such as gender distribution, total population counts, age groups, and migration trends, providing information for 7.904 Italian municipalities.
- ✓ This granularity makes the database a valuable resource for researchers, policymakers, and analysts seeking to understand and analyze population dynamics across different geographical scales in Italy, as it contains data also at the provincial and regional levels
- ✓ This work focuses on the 2023 population dataset

System architecture

- ✓ Retrieval Augmented Generation (RAG) strategy is used to develop a voicebot.
- ✓ LangChain orchestration framework is used for implementing Generative AI (GenAI) applications.
- ✓ The RAG process with LangChain involves two main steps: indexing and retrieval, followed by generation.
- ✓ The document is loaded and split into smaller chunks using RecursiveCharacterTextSplitter, then embedded and stored in a VectorStore. At query time, similar vectors are retrieved and Llama3 generates a response based on the matched texts.
- ✓ The system exhibits a form of implicit memory, allowing it to retain conversational context and generate coherent responses across multiple user turns.



Results



Demonstration of Voicebot during answer to a query as an example


The system correctly transformed the spoken inquiry to text, extracted the required information, searched the ISTAT database.

Additionally, can perform comparative analyses, such as evaluating demographic differences between cities, enabling a deeper understanding of statistical data.

It demonstrates how effectively the voicebot uses text-to-speech synthesis, natural language comprehension, database interaction, and speech recognition.

Results

Speak your question



0:00 0:06


1x

Clear Submit

Latest Answer

I don't know. The provided context only gives information about cities in the Nord-ovest region of Italy, specifically in the province of Torino and region of Piemonte, but it does not provide any information about Napoli. Therefore, it is not possible to make a comparison between Torino and Napoli based on the provided context.

Spoken Answer



0:00 0:05

1x

Conversation History

👤: Tell me about the population of Torino.

🤖: According to the provided context, the population of Torino (the city) is:

- * Male population: 406,718
- * Female population: 440,680
- * Total population: 847,398

Note that there is also a separate entry for Torino (the province), which has a different population:

- * Male population: 1,069,885
- * Female population: 1,134,747
- * Total population: 2,204,632

👤: Which one is more populous? Turina or Napoli?

🤖: According to the provided information, Napoli has a total population of 917510 persons, while Torino has a total population of 847398 persons. Therefore, Napoli is more populous than Torino.

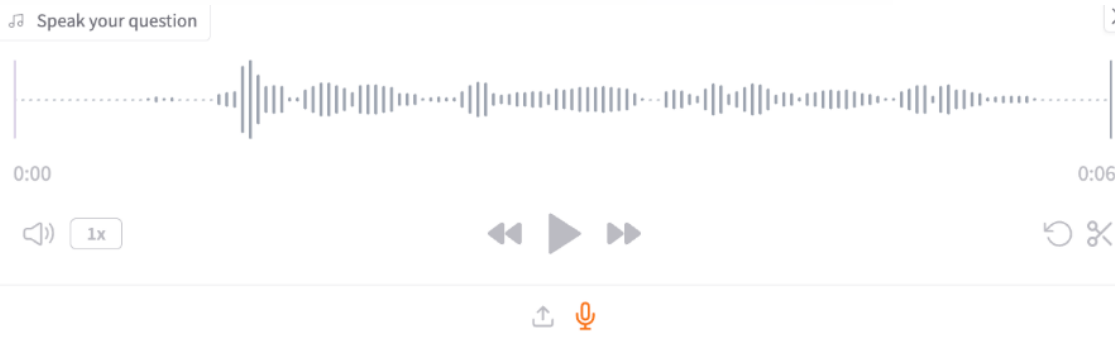
👤: which is a better place for a student.

🤖: I don't know. The provided context only gives information about cities in the Nord-ovest region of Italy, specifically in the province of Torino and region of Piemonte, but it does not provide any information about Napoli. Therefore, it is not possible to make a comparison between Torino and Napoli based on the provided context.

Flag

Results

Speak your question



0:00 0:06

1x

Clear Submit

Latest Answer

I don't know. The provided context only gives information about the population of certain cities in Italy, but it does not mention international students or provinces with the highest number of international students.

Spoken Answer



0:00 0:05

1x

Conversation History

🗣️: Please tell me about the population of Roma province.

🌐: According to the provided context, the population of Roma province is:

* Male population: 2033837

* Female population: 2193222

* Total population: 4227059

🗣️: How about Turino province? Which one is more populous?

🌐: According to the provided context, the population of Turin province is 2204632 persons.

Comparing the population of Roma province (4227059 persons) with Turin province (2204632 persons), Roma province has a larger population.

🗣️: Which province has more international students?

🌐: I don't know. The provided context only gives information about the population of certain cities in Italy, but it does not mention international students or provinces with the highest number of international students.

Results

Speak your question



0:00

0:26



Clear

Submit

Latest Answer

I apologize, but I don't understand what you're saying. It seems like there might be some language or cultural barrier. The context provided is about three cities in Italy, but I'm not able to decipher the meaning behind your statement. Could you please rephrase or provide more context so I can better understand and assist you?

Spoken Answer



0:00

0:05



Conversation History

👤: Nevit? Ah! Alaa, They've metkeen! Lyodon Isjana!

🤖: I apologize, but I don't understand what you're saying. It seems like there might be some language or cultural barrier. The context provided is about three cities in Italy, but I'm not able to decipher the meaning behind your statement. Could you please rephrase or provide more context so I can better understand and assist you?

Performance analysis

The experimental evaluation combining quantitative (statistics, execution times) and qualitative methods (feedback, bias detection) yielded the following performance metrics:

- ✓ Speech Recognition Accuracy: 92% transcription accuracy
- ✓ Query Precision: 95% response accuracy compared to ISTAT's official records (evaluated via precision-recall analysis)
- ✓ Processing Time: An average response time of 1.5 seconds per query
- ✓ User Satisfaction: 98% of test users found the system effective for statistical retrieval
- ✓ Bias Assessment: Analysis of discrepancies between generated responses and ISTAT ground-truth data confirmed the system's statistical integrity

Methodological contribution

- ✓ The integration of advanced artificial intelligence methods into statistical retrieval systems represents a significant step forward.
- ✓ Combining RAG with a speech interface increases the accuracy and efficiency of data retrieval while improving accessibility.
- ✓ The hybrid strategy uses generative models and vector-based retrieval to ensure accurate and contextual answers.
- ✓ Dataset segmentation and indexing with a vector store enables the extraction of relevant information, reducing errors and processing time.
- ✓ Gradio's use of speech technologies improves accessibility, making the system particularly useful for people with disabilities and contributing to a more inclusive digital environment.
- ✓ The conversion between speech and text preserves data accuracy, making statistical information more accessible.
- ✓ The modular architecture ensures scalability and adaptability, enabling applications in areas such as healthcare, business and education, driving innovation across multiple sectors.
- ✓ A key aspect is the assessment of bias and performance*,

The aim of this project is :

- ✓ to **bridge the gap** between **complex statistical data** and a wider, more inclusive audience,
- ✓ **fostering statistical literacy,**
- ✓ contributing to **a more informed and engaged society.**

Thank you!

cafieri@istat.it