SummarizeSafe

Hallucination Detection for Conversational Summaries

Kavin Dev R (21z224)
Likith Sai G (21z216)
AswinKumar V (21z212)

PROBLEM STATEMENT

Inaccurate Summarization in Conversational AI

AI-generated summaries can sometimes include hallucinations, leading to misinformation and reducing reliability in automated systems like customer service or virtual assistants. Manually verifying these summaries is time-consuming, highlighting the need for an automated solution to detect hallucinations and ensure accuracy.

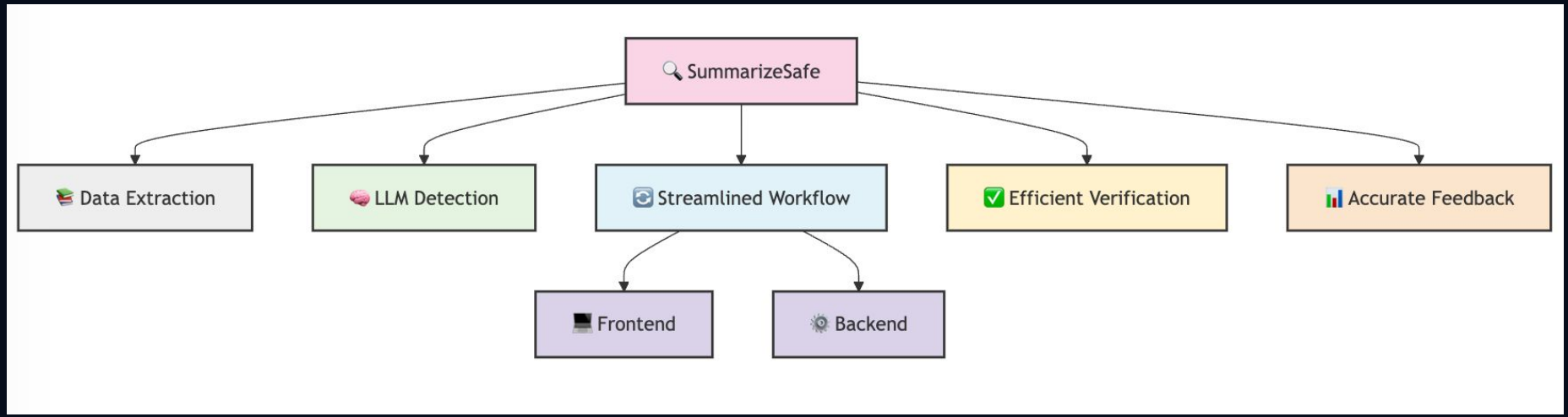
VISION

To develop a robust, scalable system that ensures the accuracy and reliability of automated summaries by detecting hallucinations in text summarization, ultimately enhancing trust in AI-driven conversational tools.

OUR SOLUTION- SummarizeSafe

Advanced AI System for Hallucination Detection in Summaries

Our Solution



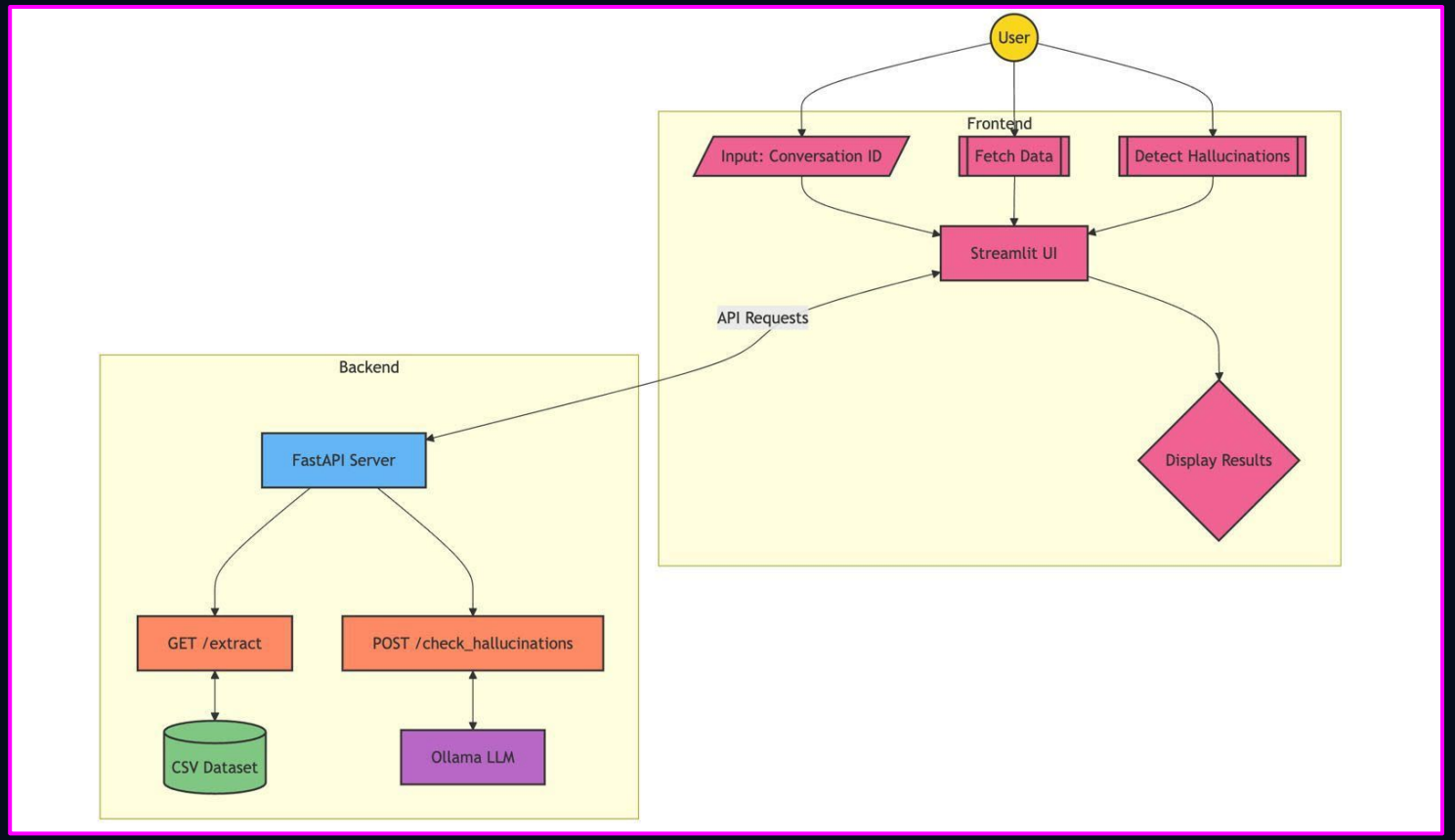
TechStack Used

FastAPI

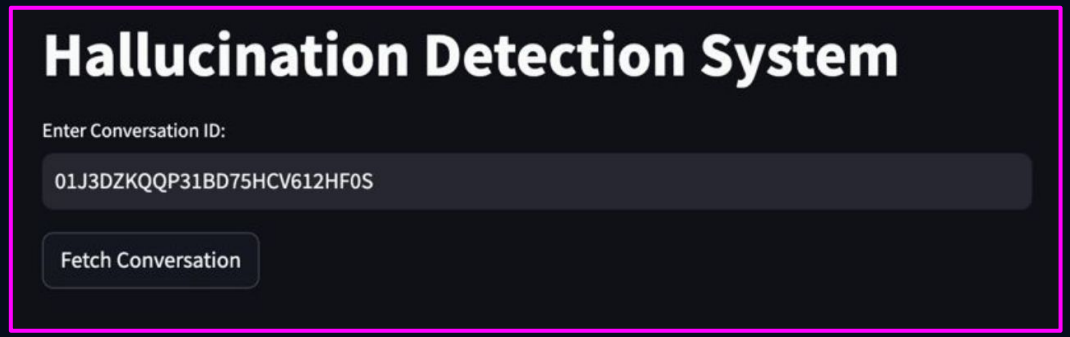
Streamlit

pandas

System Architecture



Implementation Screenshots



Extracted Text

siete hola bienvenido al departamento de lealtad de PSG mi nombre es juan con quién tengo el gusto hablar el día de hoy ocho o la buenas tardes habla con javier gusto señor javier dígame en qué le puedo servir estoy llamando por mi internet que no no no hay internet se cortó no se se cayó la línea okay más o menos cuánto es que usted está presentando este problema pues ya va en la segunda vez que presenta ese problema ya internet okay gracias por ello entonces bueno me puede dar por favor el número de teléfono asociado la cuenta o el número de cuenta para verificar qué es lo que está pasando y pues por le solucionar así es sí ocho tres dos tres siete dos cinco cinco nueve dos javier listo y me puede corroborar nombre y apellido el dueño de la cuenta y la clave de {voice.ssn.digitsTranscribed:} **números por favor muchas gracias la clave es el {voice.ssn.digitsTranscribed:}** ***le dicen que le confirmo rápidamente es el número {voice.ssn.digitsTranscribed: ******* * *** **** ******* * *** ********* * ***** ******* * } y el correo bolívar de ciento {voice.ssn.digitsTranscribed:*** }arroba gmail punto com un buen número y un buen

Summary

Customer called to report internet connectivity issues for the second time. Agent offered a free technician visit to check the wiring and replace the modem if necessary. An appointment was scheduled for Friday, and steps were taken to address a previous concern regarding a contract. Customer has an upcoming technician visit scheduled and will receive assistance regarding a previous contract issue.

Detect Hallucinations

Hallucination Detection Result

Hallucinations present

Hallucinated Parts: Conclusion: The summary appears to contain hallucinated content.

Hallucinated parts:

- There is no mention of "internet connectivity issues" or a scheduled technician visit in the original text. The summary seems to fabricate this information, which is not supported by the text.
- The summary also includes details about a contract issue being addressed, which is not mentioned in the original text.

Overall analysis: The original text appears to be a transcription of a customer's conversation with an agent regarding a rewards program. The summary seems to take creative liberties by adding fabricated information and details that are not present in the original text. This suggests that the summary contains hallucinated content.

Unique Selling Propositions

- Real-Time Hallucination Detection

  The project provides real-time hallucination detection in generated summaries, ensuring that users get instant feedback about the reliability of condensed text.

- Granular Hallucination Identification

  The system not only flags whether hallucinations exist but also identifies and highlights specific parts of the summary that are hallucinated.

- LLaMA-3 Based Accuracy

  By leveraging the power of a state-of-the-art LLM (LLaMA-3), your project offers high accuracy and sophisticated natural language understanding in detecting nuanced hallucinations.

Unique Selling Propositions

- Custom Prompt Optimization

  The project employs advanced prompt engineering techniques to ensure that the LLM provides accurate, context-sensitive outputs for hallucination detection.

- Domain-Agnostic Functionality

  The system can be applied across multiple domains (news, customer service, academia) where accurate summaries are crucial.

FUTURE SCOPE

Automated Summary Refinement: Auto-correct hallucinated parts to generate factually accurate summaries without human intervention.

Continuous Learning: Implement a feedback loop to improve LLM performance using human corrections over time.

Real-Time Integration: Enable real-time hallucination detection in live conversational AI systems for instant corrections.

Scalable Cloud Deployment: Deploy on cloud platforms (AWS, Azure, GCP) for scalable, enterprise-level analysis and high-volume applications.

GITHUB Link and Demo

https://github.com/kavinDEV15/SummarizeSafe

https://drive.google.com/file/d/11n3bfBPJwN3UvanXyrgIFvpgqaVo58bg/view?usp=sharing

Thank You!