

# HALLUCINATION DETECTION

## Team Members:

Arulpathi A	21Z209
Javagar M	21Z221
Roshini P	21Z245
Shiva Aravindha Samy A	21Z255
Vijayalakshmi P	21Z269

# TABLE OF CONTENTS

**01**

**PROBLEM STATEMENT**

**02**

**CONCEPTUAL DESIGN**

**03**

**PHASE-1**

**04**

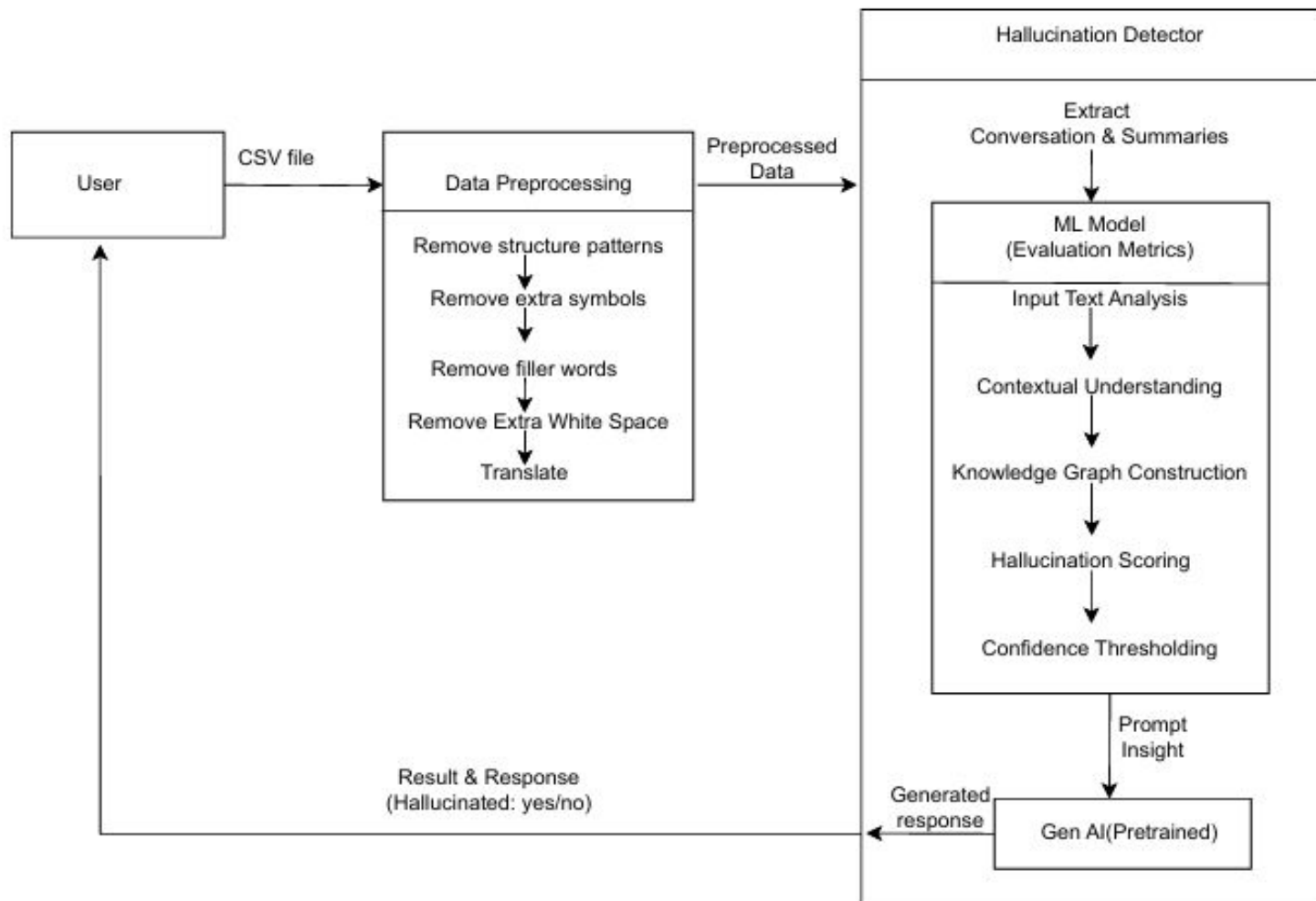
**PHASE-2**

# PROBLEM STATEMENT

In many natural language processing tasks, Large Language Models (LLMs) are employed to generate summaries from given text. However, these models sometimes introduce incorrect or fabricated information in the summaries, known as "hallucinations." This project aims to detect such hallucinations in generated summaries. The system is designed to take a text and its corresponding summary, and analyze whether hallucinated content exists in the summary. Additionally, the system can pinpoint the hallucinated sections and quantify hallucinations by comparing the summary with the original text.

We leverage the LLaMA model for this task, integrating advanced prompt engineering to accurately classify and highlight hallucinations. The final solution is packaged into an intuitive web application using Streamlit, offering users a simple interface to upload text and summaries, select conversation IDs, and receive results indicating whether hallucinations are present.

# CONCEPTUAL DESIGN



# PHASE 1

## Input Format:

- >Considering the columns : 'conversation\_id', 'speaker', 'text', 'summary\_a'
- >Choosing the particular conversation for conversation checking using 'conversation\_id'.
- >Extract Data: Retrieve the text and summary columns from your dataset, ensuring each conversation has corresponding data.

### Hallucination Checker

Upload conversation as CSV

 Drag and drop file here  
Limit 200MB per file • CSV

Browse files

Please upload a CSV file to proceed.

### Hallucination Checker

Upload conversation as CSV

 Drag and drop file here  
Limit 200MB per file • CSV

Browse files

 data.csv 1.6MB 

File uploaded successfully.

Process CSV

# PHASE 1

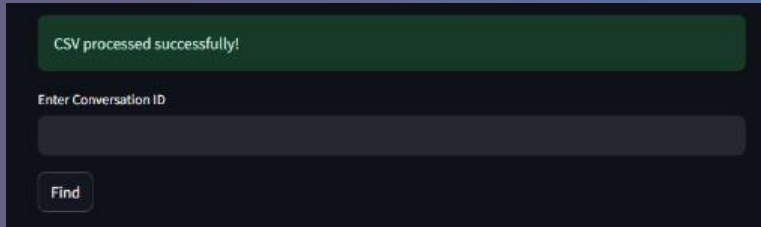
## IMPLEMENTATION:

>Prepare Input: For each conversation, take the original chat (text) and its generated summary.

>Process with LLaMA 3:

- Input the text and summary into the LLaMA 3 model.
- Let LLaMA 3 analyze the summary by comparing it to the original conversation.

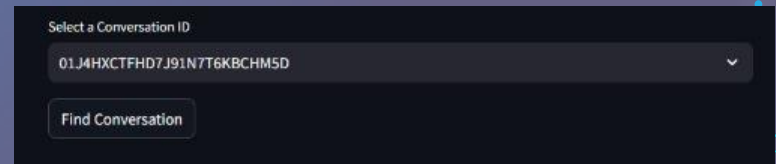
>Extract Data: Retrieve the text and summary columns from your dataset, ensuring each conversation has corresponding data.



CSV processed successfully!

Enter Conversation ID

Find



Select a Conversation ID

01J4HXCTFHD7J91N7T6KBCHM5D

Find Conversation

# PHASE 1

## IMPLEMENTATION:

>Hallucination Detection:

- LLaMA 3 will determine if there is any hallucinated (fabricated or misleading) information in the summary.
- It will provide a simple output: "Yes" (hallucination present) or "No" (no hallucination).

† and explanations for each conversation-summary pair

Hallucination result for conversation ID : 01J0NTYHP8SV09800XC4B3G68F

Hallucination : yes

Explanation : Summary A diverges significantly from the actual context. It mentions "Sunday" which has no relation to the conversation, and the tone of Summary A does not match the rest of the conversation. The actual summary should have been focused on the customer's interaction with the smart home manager app, but instead, it includes extraneous information like "Two nine nine four got." which is not present in the conversation.

# PHASE 1

## IMPLEMENTATION:

### >Generate Explanation:

- Along with the "Yes" or "No" output, LLaMA 3 will provide a detailed explanation or rationale, highlighting why it concluded that the summary is accurate or hallucinated.

### >Output Results:

- Save or display the results for further analysis, which includes the verdict and explanations for each conversation-summary pair

#### Hallucination : yes

Explanation: Summary A is a hallucination because it inaccurately summarizes several key points in the conversation. For instance, in response to the customer stating "because his passed away?", the agent should have provided more context or condolences rather than simply saying "Oh, I'm so sorry for your loss." This response does not fully capture the customer's request to cancel their cousin's phone service due to passing away.

Additionally, Summary A fails to convey the complexity of the situation, such as the need for a death certificate and a valid ID to authenticate and cancel the cellphone service. The agent's instructions are more detailed in the conversation, which is not reflected in Summary A.

Lastly, Summary A misrepresents the agent's suggestion that the customer should go to a corporate store with a valid ID and the death certificate, as if this was an entirely new request rather than part of the original instruction. The conversation clearly states that this is what the agent previously advised the customer to do.



# PHASE - 2

## Data Preprocessing

- **Emoji Conversion:** Convert emojis to text representation using `emoji.demojize` function.
- **Removal of Structured Patterns:** Identify and remove structured data patterns (e.g., `{voice.ssn.digitsTranscribed:*****}`) using regular expressions.
- **Punctuation Cleaning:** Remove unnecessary punctuation and repeated symbols (e.g., commas, ellipses, hashtags).
- **Filler Words Removal:** Eliminate common filler words (e.g., "um," "uh," "well," "yeah," "okay") from conversation text.
- **Correction of Misspellings:** Manually correct specific misspellings (e.g., 'four Oz four' → '404', 'They're tiff a kit' → 'the certificate').
- **Removal of Non-verbal Utterances:** Remove non-verbal or unclear words (e.g., "noise," "incomprehensible," "unclear").
- **Whitespace Normalization:** Remove extra spaces to ensure proper text spacing.
- **Translation to English:** Translate conversation text to English using Google Translate API (`googletrans` library), detecting source language automatically.

# PHASE - 2

## Implementation

### Objective:

- Detect hallucinations in text summaries by comparing with the original conversation using **cosine similarity** of sentence embeddings.

### Sentence Embeddings:

- Load **pre-trained SentenceTransformer** model (paraphrase-MiniLM-L6-v2).
- Encode both conversation and summary sentences into embeddings.

### Cosine Similarity Calculation:

- Use **cosine similarity** to measure the closeness between conversation and summary sentences.
- For each summary sentence, find the maximum similarity with any conversation sentence.

### Hallucination Detection:

- Sentences with similarity below a **threshold** (e.g., 0.7) are flagged as **hallucinations**.

# PHASE - 2

## Analysis

### Hallucination Detection Process

- **Comparison:** Compare the summary sentences to the conversation.
- **Detection:** Identify hallucinations where similarity is below the threshold.

### Metrics and Output:

- **Hallucinated Sentences:**
  - List of summary sentences that deviate from the conversation.
- **Hallucination Percentage:**
  - Calculate the percentage of hallucinated sentences in the entire summary:

$$\text{Hallucination Percentage} = \left( \frac{\text{Number of Hallucinated Sentences}}{\text{Total Number of Summary Sentences}} \right) \times 100$$

### Result Example:

- Hallucinations identified in the summary:
  - "The customer called to reschedule their cousin's appointment" (not grounded in the conversation).
- **Hallucination Percentage:** 50% (for example).

# PHASE - 2

## Result

Hallucinations identified in the summary:

- "The customer called to reschedule their cousin's appointment" (not grounded in the conversation).

**Hallucination Percentage:** 50% (for example).

```
PS C:\Users\USER\OneDrive\Desktop\asapp> c:: cd 'c:\Users\USER\OneDrive\Desktop\asapp'; & 'c:\Program Files\Python312\python.exe' 'c:\Users\USER\.vscode\extensions\ms-python.debugpy-2024.12.0-win32-x64\bundled\libs\debugpy\adapter/../../debugpy/launcher' '51399' '--' 'c:\Users\USER\OneDrive\Desktop\asapp\b1.py'
```

Hallucination: "The customer called to reschedule their cousin's appointment" does not closely match any sentence in the conversation.

Replacing with: "I need to cancel my cousin's appointment

Sure, I can help you with that" (Similarity: 0.61)

Hallucination: "The agent helped them with this" does not closely match any sentence in the conversation.

Replacing with: "What's the appointment date?

It's on the 23rd of this month

Got it" (Similarity: 0.14)

Hallucination Percentage: 100.00%

Final Corrected Summary:

I need to cancel my cousin's appointment

Sure, I can help you with that. What's the appointment date?

It's on the 23rd of this month

Got it.