SAFEPROMPTAI

Challenge: Auto Correct and Prompt Validation Before AI Execution

Empathy Meets Technology : Safe Conversations Powered by AI.

By Arulselvi Amirrthalingam (Selvi)
+ Selvamani Ramasamy

USER PROMPT SAFETY CHALLENGES

- Ensuring user prompt safety in AI interactions is crucial but complex, especially with ethical and regulatory concerns.
- Organizations face challenges in moderating harmful, biased, or sensitive prompts while maintaining AI responsiveness.
- As AI systems handle diverse user inputs, the risks of misuse, misinformation, and compliance violations increase.
- Manually filtering unsafe prompts is inefficient and inconsistent, requiring automated moderation for scalable AI safety

OUR SOLUTION:

SafePromptAI- An innovation to filter user prompt in many layers before sending to AI

SafePromptAI is designed to act as a protective shield between user inputs and AI responses, ensuring safe and context-aware interactions. Think of it as an AI safeguarding another AI.

Our system processes user inputs through multiple layers using Azure AI services, starting with PII (Personally Identifiable Information) detection. Users are prompted for confirmation before any sensitive data is shared, and modifications can be made if necessary.

Next, the system performs sentiment analysis, harmful content detection, and automatically replaces unsafe terms while preserving the original intent of the message.

SafePromptAI represents an innovative approach to maintaining safety and context awareness in AI interactions before any execution takes place.

SafePromptAI Multi-Layered AI Filtering Approach

✅ Layer 1 :PII Detection (Azure Text Analytics Service)

We begin by identifying Personally Identifiable Information (PII) such as credit card numbers, email addresses, and driver's license numbers but keeping in mind while certain applications (e.g., banking or DMV services) may require users to provide this information, hence SafePromtAI don't automatically remove it. Instead prompt the user for confirmation, giving them the option to revise or proceed.

✅ Layer 2: Sentiment Analysis (Azure Text Analytics Service)

Analyze the sentiment of the input—positive, neutral, or negative. This step helps us understand the user's emotional tone before interacting with the AI, ensuring more appropriate responses.

✅ Layer 3: Harmful Term Identification (Azure Content Moderator)

Our third layer focuses on filtering harmful content. If no harmful language is detected, the prompt proceeds. However, if harmful terms are found, we identify and extract harmful terms from user prompt and send to layer 4 find suitable replacements.

✅ Layer 4: Safe Term Replacement (Openai)

In this layer, harmful terms identified in Layer 4 are replaced with the safe alternatives, ensuring that the prompt is both safe and contextually intact before being sent to the AI for processing. If a user prompt contains harmful content, such as ***"I am going to kill someone,"*** Azure Content Moderator detects the term ***"kill"*** and replaces it with a safer alternative using OpenAI, ensuring the intent remains intact.Our system also enhances clarity and grammar, transforming the prompt into a safer version like ***"I am extremely frustrated with someone."*** This multi-layered filtering ensures responsible AI interactions without unnecessary censorship. This approach maintains the core intent of the original message while eliminating the harmful or inappropriate language.

✅ Layer 5: Audio Accessibility (Azure Speech Services)

Finally, for accessibility purposes, we integrate Azure Speech Services to provide an audio version of the AI's response. This allows users with visual impairments or other accessibility needs to listen to the output, ensuring inclusivity and broad user access.

SafePromptAI Core Functionalities

User Friendly front end to enter the prompt

Detects PII & asks users if they want to share it

Analyzes sentiment to understand user intent.

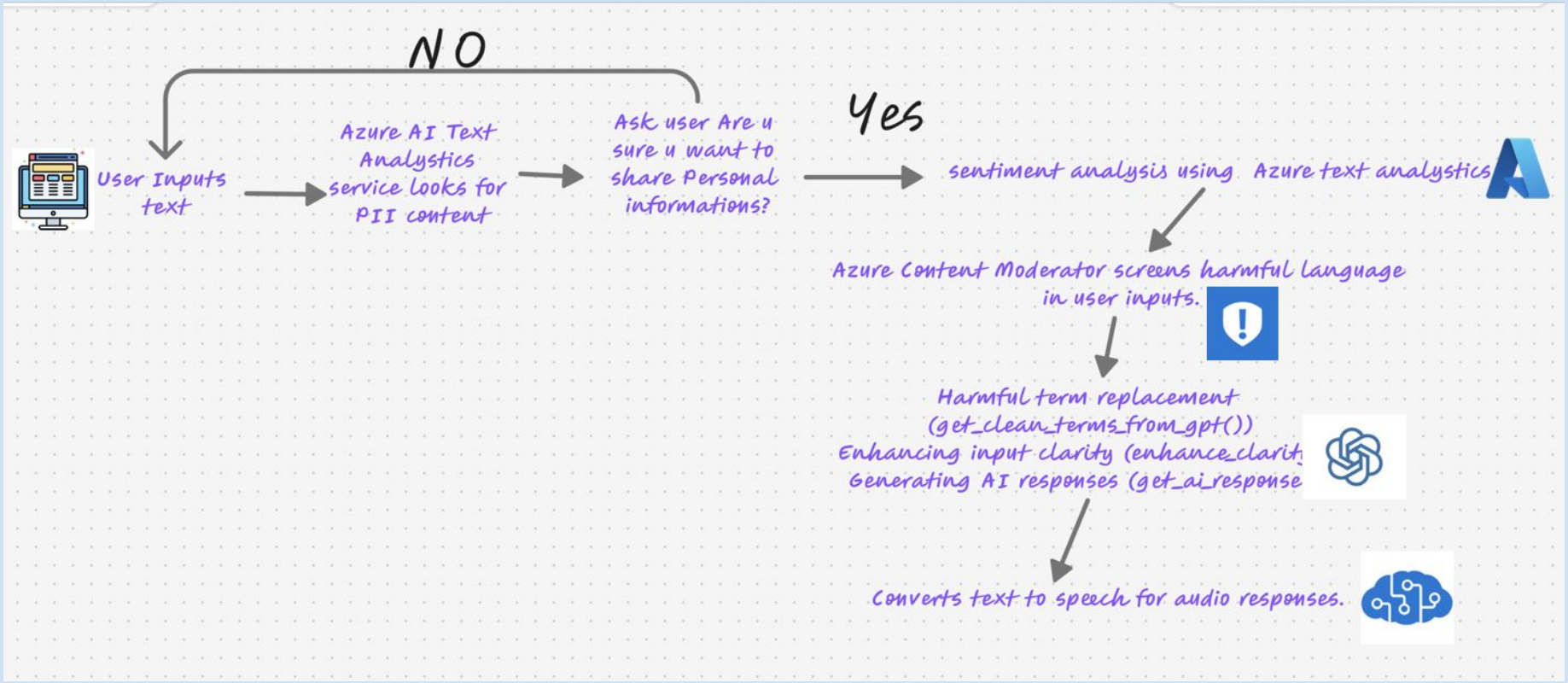
Checks for **harmful language** using Azure AI.

Replace harmful terms, enhance clarity & grammar correction

Send the Safe prompt to AI to fetch response

Azure Cognitive Speech Service ensuring accessibility for users who rely on spoken content

Project Architecture/ Workflow



TECH STACK

Frontend: React, Axios, CSS for UI styling

Backend: Flask, Azure AI Services, OpenAI GPT-4

APIs Used: Content Moderator, Text Analytics, Speech Service

Security: Environment variables (.env), data encryption

Responsible AI Features in SafePromptAI

✅ **Ethical AI Filtering**

SafePromptAI not just block harmful prompts—it refines and enhances user inputs while keeping AI responses ethical and meaningful.

✅ **Privacy & Compliance**

AI should be secure and privacy-compliant. SafePromptAI ensures that sensitive data is detected, flagged, and handled responsibly. User Consent for PII Sharing *– Ensures compliance with* ***GDPR, CCPA, and HIPAA****.*

✅ **Sentiment & Intent Analysis**

By analyzing sentiment, SafePromptAI can prevent harmful AI interactions while maintaining a natural user experience

✅ **Accessibility & Inclusion**

AI should be accessible for all users. Our built-in text-to-speech feature enhances usability and inclusivity."

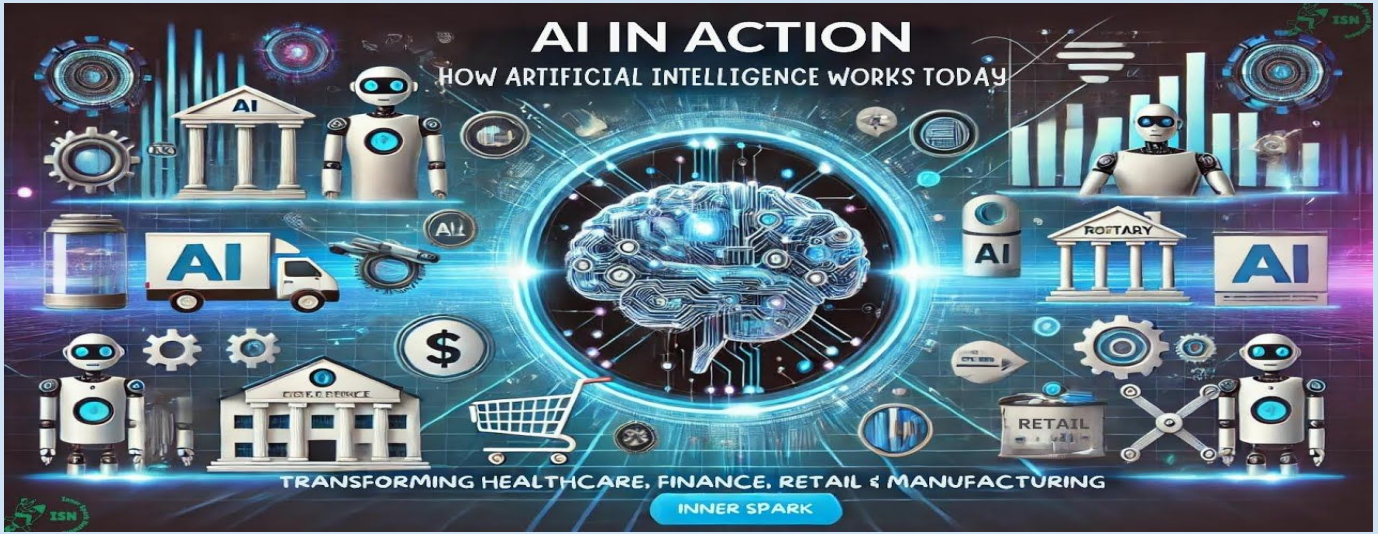
✅ **Bias Mitigation & Fairness**

Bias in AI can have unintended consequences. SafePromptAI acts as a safeguard, ensuring fairness and neutrality in AI responses.

🚀 Future Vision: Making AI Safer & Smarter

- 🔹 **Scalability & Real-Time Processing** – Deploying in cloud environments for **high-speed prompt moderation**.
- 🔹 **Multi-Language Support** – Expanding beyond English to **support diverse user interactions** globally.
- 🔹 **Integration with AI Chatbots & Virtual Assistants** – Seamless adoption across **AI-driven applications**.


AND NOW
FOR A LIVE DEMO


AI IN ACTION
HOW ARTIFICIAL INTELLIGENCE WORKS TODAY
AI
AI
AI
AI
AI
RETAIL
TRANSFORMING HEALTHCARE, FINANCE, RETAIL & MANUFACTURING
INNER SPARK