JARVIS: AI Assistant Powered by Gemma 3n - Technical Writeup

Technical Writeup - JARVIS: AI Assistant Powered by Gemma 3n

Overview

JARVIS is a fully local AI assistant built using the open-source Gemma 3n model running in LM Studio, with a custom Python GUI built with Tkinter. It allows users to chat with an LLM without any internet dependency.

Problem Solved

Today, many AI assistants require cloud-based APIs and constant internet access. I wanted to build an assistant that is privacy-first, local, and doesn't rely on OpenAI/Google servers. Using Gemma 3n, I created an assistant that answers queries in real-time with a desktop interface.

Architecture

Frontend: Tkinter GUI with themed dark design

Backend:

- LM Studio serving the Gemma 3n model locally
- requests for communicating via HTTP
- pyttsx3 for offline TTS (text-to-speech)
- (Mic input using speech_recognition was built but disabled in final version)

Flow: User Input -> GUI -> HTTP POST -> Gemma 3n -> Response -> TTS + GUI Display

How Gemma 3n Was Used

- Hosted on LM Studio locally at http://127.0.0.1:1234

- Communication via HTTP POST to /v1/chat/completions

- Used parameters: temperature=0.7, max_tokens=512, and messages in OpenAI-style format

Features

- Offline Chat with Gemma

- Dark-mode GUI with background image

- Scrollable chat history

- Text-to-speech output

- Voice input code structure (optional)

Challenges Faced

- TTS engine not working consistently - resolved with pyttsx3

- Mic input causing lag - disabled in final version

- GUI resizing bugs - fixed with root.bind("<Configure>")

- Errors from Tkinter like bg="None" - replaced with valid color

Why These Tech Choices

Gemma 3n - Open-source, runs locally

LM Studio - Easiest way to serve Gemma models

Tkinter - Lightweight and fast for GUI

pyttsx3 - Works offline for voice

requests - Simple HTTP handling

Links

- YouTube Demo: [Insert Your Link]

- GitHub Repo: [Insert Your Link]

- (Optional) Kaggle Dataset or Public Link: [Insert Link]

Conclusion

JARVIS is a lightweight yet powerful example of what's possible using local AI and open models like Gemma 3n. No internet, no API keys - just Python, models, and creativity.