



HALLUCINATION CORRECTION ASSISTANT

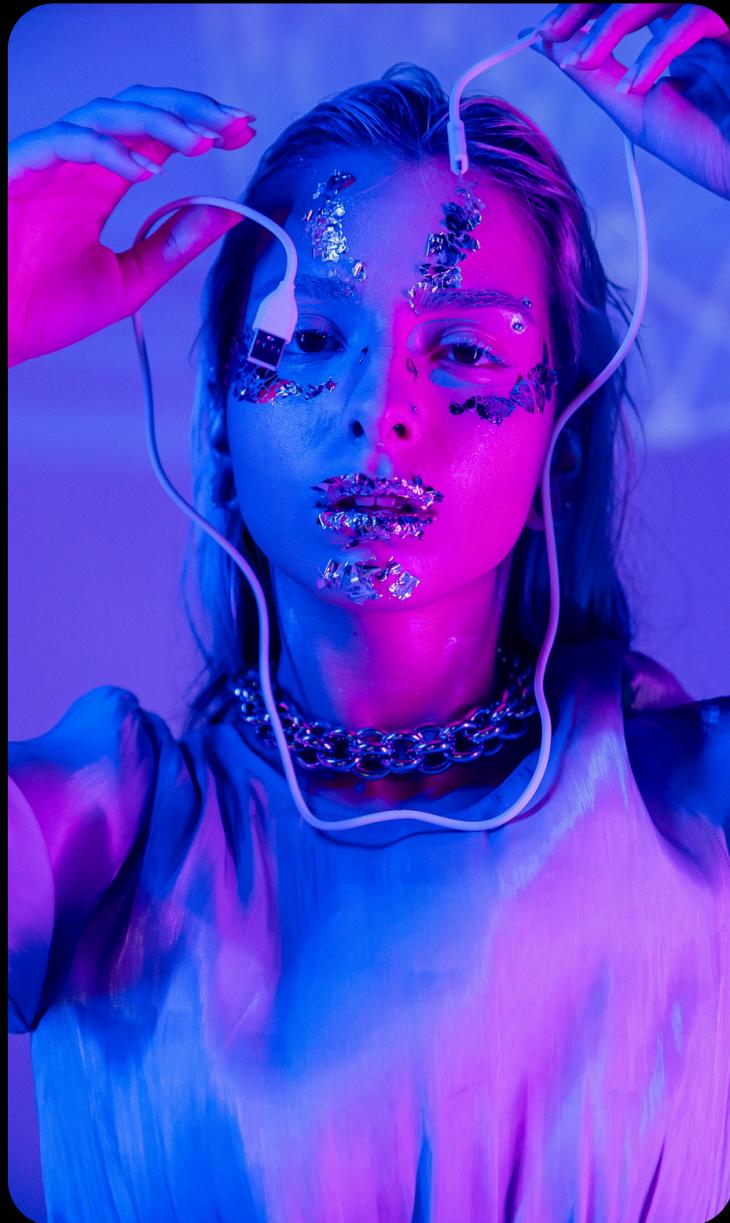
AI-GENERATED HALLUCINATIONS USING A FINE-TUNED GPT-3.5 TURBO MODEL

Presented By: Darius Brown

Bellevue
University

www.reallygreatsite.com





Why Hallucinations Matter

- Large Language Models (LLMs) often generate false but fluent-sounding outputs
- These “hallucinations” pose risks in education, healthcare, and enterprise AI
- Need for tools to verify and correct hallucinated content



Project Goals

- **Detect and correct hallucinations using a fine-tuned GPT model**
- **Offer clear and accurate outputs**
- **Measure and visualize model confidence**
- **Compare model responses with external LLMs for validation**

Bellevue University

www.reallygreatsite.com



@



Model Architecture

- **Fine-Tuned Model Name:** ft:gpt-3.5-turbo-0125:dares-apis:hallucination-logical-v1:BvwmRqwc
- **Tools Used:**
 1. OpenAI fine-tuning
 2. Streamlit for UI
 3. Fuzzy string matching for response similarity
- **Diagram Flow:** Simple flow: User Input → Model → Correction + Confidence Score

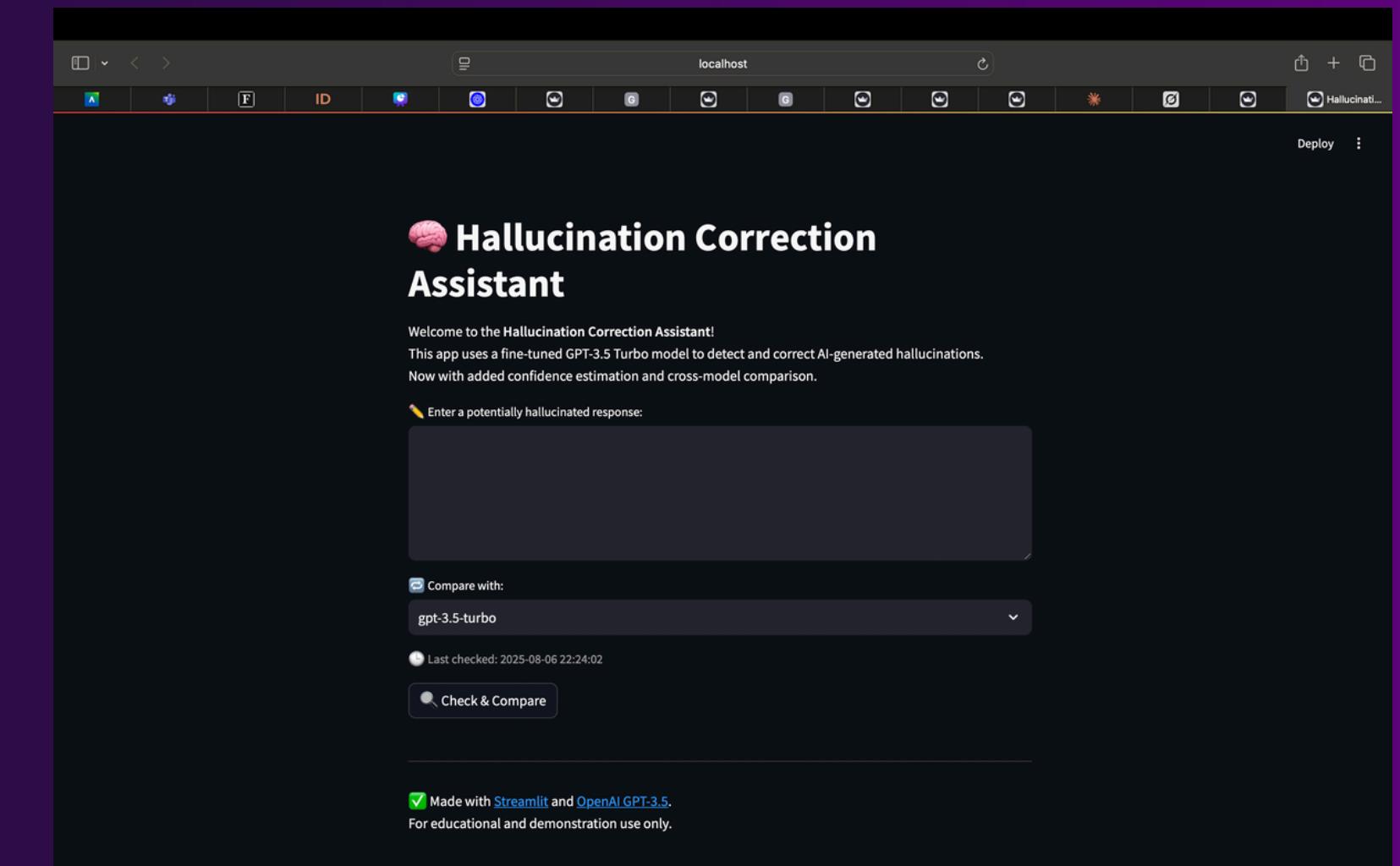


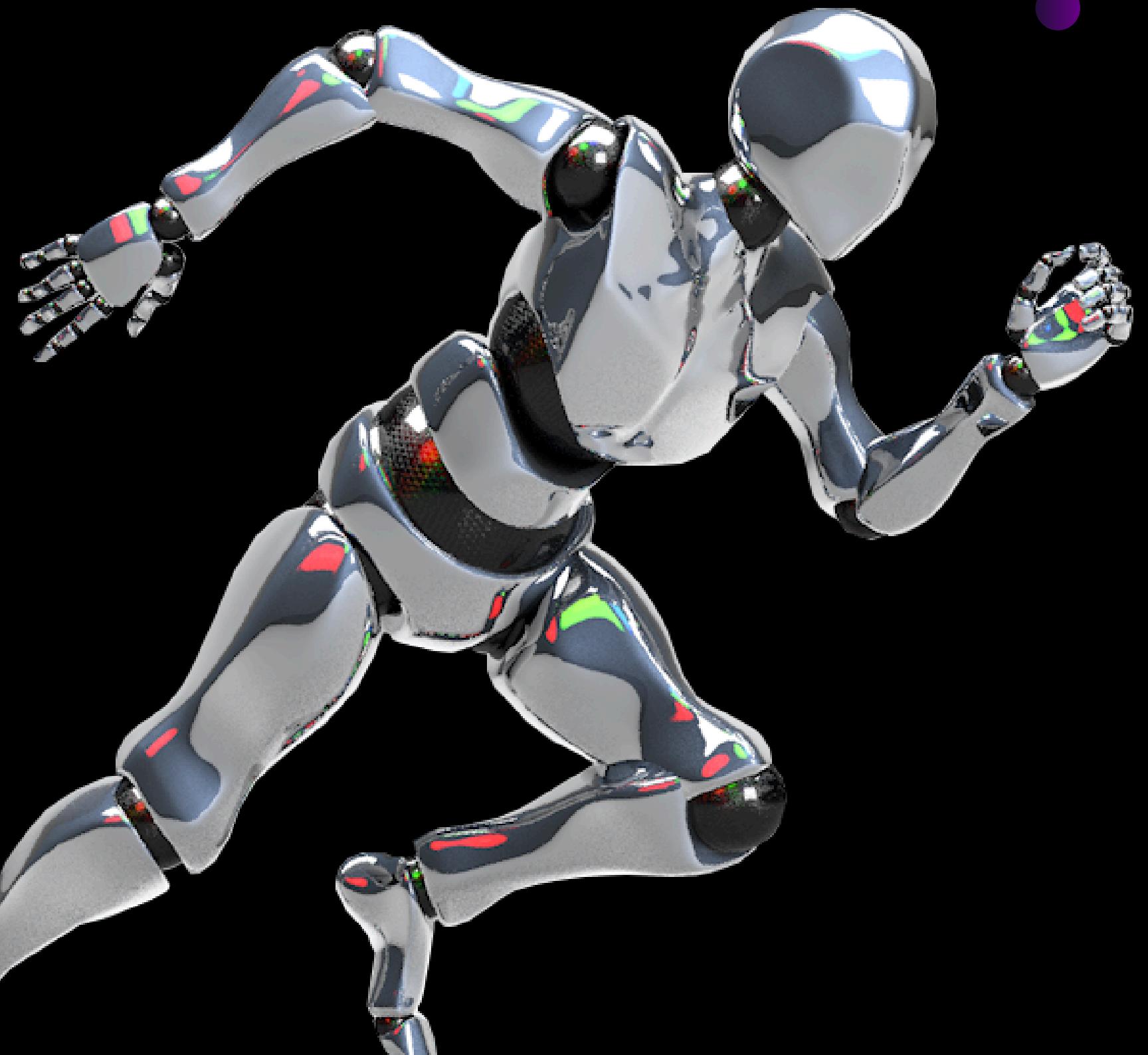
User Interface (Streamlit)



Labels for:

- Input field
- Corrected output
- Confidence percentages
- Compare button
- Mention: Real-time interaction, no coding required





• Enhancements Summary

- Confidence score based on fuzzy string matching
- External model response comparison
- Timestamped analysis for auditability
- Error handling for clean UX
- OpenAI best practices (separating roles, safety layers)





EXAMPLE INPUT

The screenshot shows a Streamlit application running on localhost. The title bar says "localhost". The main heading is "🧠 Hallucination Correction Assistant". Below it, a sub-headline reads: "Welcome to the Hallucination Correction Assistant! This app uses a fine-tuned GPT-3.5 Turbo model to detect and correct AI-generated hallucinations. Now with added confidence estimation and cross-model comparison." A text input field contains the question "what is was the reason that caused the war between Russia and Ukraine?". Below the input field is a note: "Press ⌘+Enter to apply". A dropdown menu labeled "Compare with:" shows "gpt-3.5-turbo" selected. At the bottom, there's a checkbox checked with the text "Made with Streamlit and OpenAI GPT-3.5. For educational and demonstration use only."

www.reallygreatsite.com





EXAMPLE OUTPUT

localhost

Deploy :

Compare with:

gpt-3.5-turbo

Last checked: 2025-08-06 22:59:03

Check & Compare

Corrected Output (Fine-Tuned Model)

The war between Russia and Ukraine was caused by a complex interplay of factors, including historical, political, economic, and social issues.

The war between Russia and Ukraine was caused by a complex interplay of factors, i

Output from gpt-3.5-turbo

The war between Russia and Ukraine began in 2014 following Russia's annexation of Crimea, a region that was part of Ukraine. The conflict escalated due to political unrest in Ukraine, disagreements over the country's direction, and Russia's support for separatist movements in eastern Ukraine.

Confidence Score (Similarity to gpt-3.5-turbo)

57%

Made with Streamlit and OpenAI GPT-3.5.
For educational and demonstration use only.

www.reallygreatsite.com



DARIOUS BROWN

The screenshot shows a Streamlit application titled "Hallucination Correction Assistant". The interface includes a header bar with various icons and a "Deploy" button. Below the header, there's a section titled "Hallucination Correction Assistant" with a brain icon. A message box displays the question "what is was the reason that caused the war between Russia and Ukraine?". A dropdown menu labeled "Compare with:" is set to "gpt-3.5-turbo". A timestamp "Last checked: 2025-08-06 22:59:35" and a "Check & Compare" button are also visible. A green checkmark indicates "Corrected Output (Fine-Tuned Model)". The output text states: "The war between Russia and Ukraine was caused by a complex interplay of factors, including historical, political, economic, and social issues."

As questions were repeated the model showed a direct positive relationship through an increased confidence percentage

Bellevue University

www.reallygreatsite.com

Model Growth Through RAG

This screenshot shows the same Streamlit application after multiple iterations. The "Check & Compare" button is now active. The "Corrected Output (Fine-Tuned Model)" section remains the same. In the "Output from gpt-3.5-turbo" section, the response is identical to the previous one. A "Confidence Score (Similarity to gpt-3.5-turbo)" is displayed as "80%". The footer notes "Made with Streamlit and OpenAI GPT-3.5. For educational and demonstration use only."



Conclusion

- **AI hallucination correction is critical in LLM deployment**
- **Our app demonstrates a practical, user-friendly solution**
- **Future work:**
 - Support multi-sentence documents
 - Cross-model benchmarking
 - Extend to medical/financial domains





Get in Touch

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed at ipsum vitae lacus lobortis lacinia. Donec tristique arcu massa, at.

Phone Number

267-309-5099

Website

www.reallygreatsite.com

Email

darius.brown3@icloud.com

GitHub Repository & App Link

[@reallygreatsite](#)

www.reallygreatsite.com

