**Chatbot Comparison: Gemini AI, Ollama Phi, and Ollama Llama3**

**1. Introduction**

This report provides an in-depth comparison of three AI chatbot models: **Gemini AI**, **Ollama Phi**, and **Ollama Llama3**. The goal is to evaluate their performance across various metrics such as response quality, speed, usability, and scalability. By the end of the report, we will determine which model is most suitable for specific use cases.

**2. Project Overview**

In this project, we have developed a Streamlit application that integrates three different AI models: **Gemini AI**, **Ollama Phi**, and **Ollama Llama3**. The application allows users to interact with all three models simultaneously and compare their responses. The core objective of this project is to explore the capabilities of each model in terms of response generation, ease of integration, and suitability for different chatbot tasks.

**2.1 Key Features of the Application:**

* **Gemini AI**: Cloud-based, powerful language model for generating detailed and structured responses.
* **Ollama Phi**: Open-source model optimized for casual, informal conversations.
* **Ollama Llama3**: A versatile open-source model that strikes a balance between formal and informal responses.

The application displays the messages from all three chatbots side by side, allowing users to send the same input to all models and observe their responses in real-time.

**3. Methodology**

The methodology for this comparison involves running the chatbots in parallel on a Streamlit application. The user inputs a query, and each chatbot generates a response. The following steps outline the workflow:

1. **User Input**: The user enters a query in a text field.
2. **Model Response Generation**: Each chatbot (Gemini AI, Ollama Phi, and Ollama Llama3) generates a response based on the user's input.
3. **Display of Results**: The responses are displayed side-by-side for easy comparison.

Additionally, the user can send messages individually to each chatbot or use a shared input field to send messages to all three simultaneously.

**4. Technical Implementation**

The application leverages the following technologies:

* **Streamlit**: For creating the user interface, displaying chat histories, and allowing user input.
* **Gemini AI**: Integrated via the Google Generative AI API, this cloud-based model responds to user queries with high accuracy and coherence.
* **Ollama Phi and Llama3**: These are open-source models that generate conversational responses. They are integrated using the LangChain library.

The core functions in the code involve sending the user's input to each model and retrieving the response in real-time. The models are initialized with default messages, and the chat histories are maintained across multiple interactions.

**5. Screenshots and Visuals**

Here are some screenshots from the application:

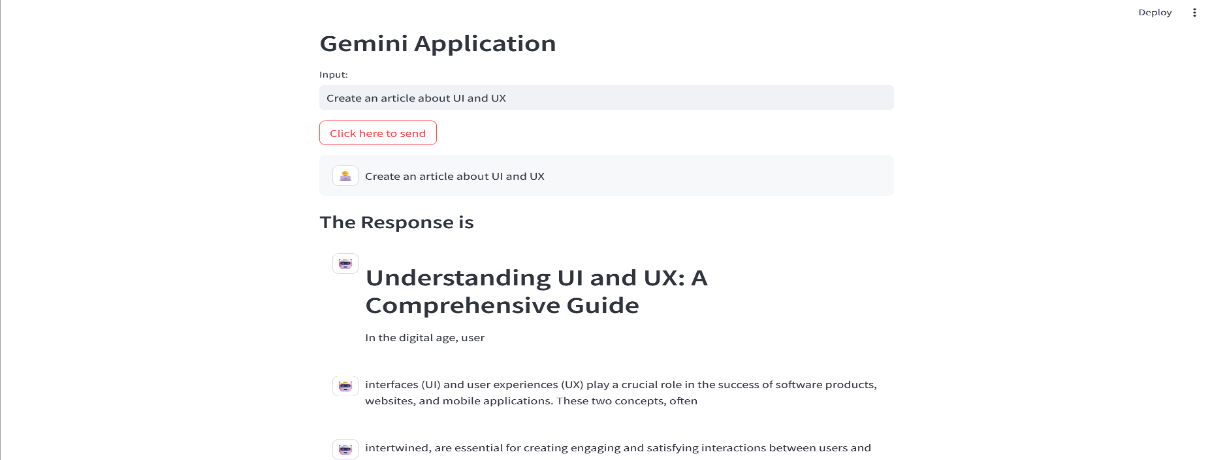


Fig: 1.1 Gemini AI chatbot

**5.1 Chatbot Interface**

* The interface shows three columns, each representing a different chatbot (Gemini, Phi, and Llama3).
* The user can type a message and see how each chatbot responds to the same input.

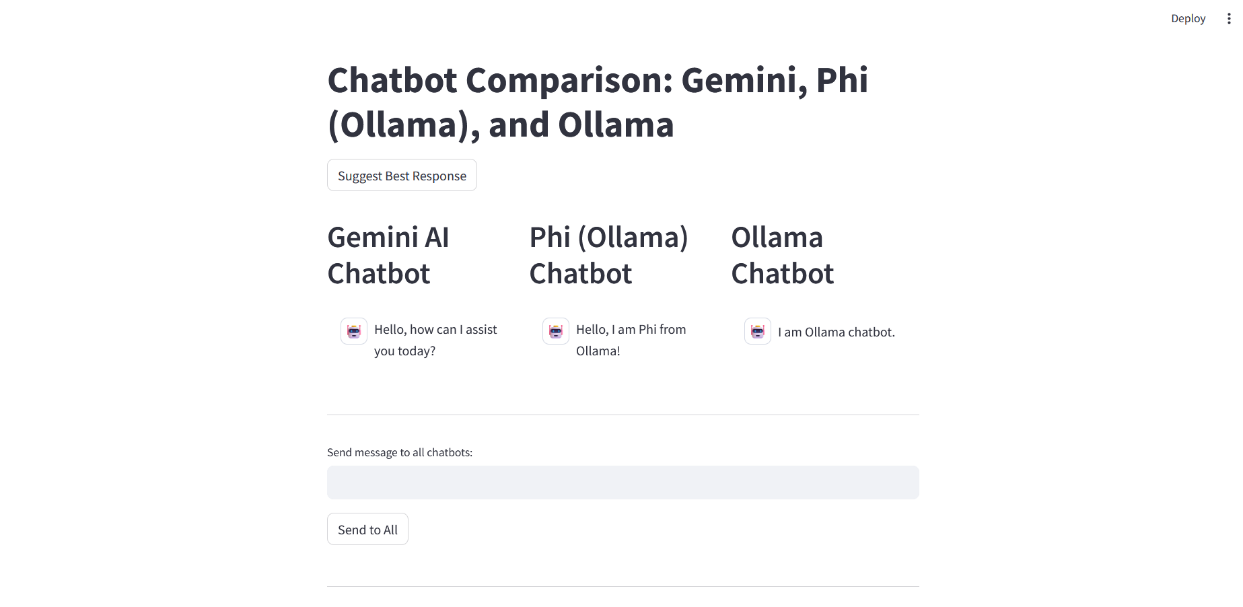


Fig: 1.2 All there chatbots in one Streamlit Interface

**5.2 Response Display**

* Once the user submits a query, the responses from all three chatbots are displayed below their respective sections.
* Make sure you **click the submit button twice**, one for loading the article information and one after the information loaded. So that the webpage can make sure that api key of Gemini, llama3 and ollama phi model works with streamlit.

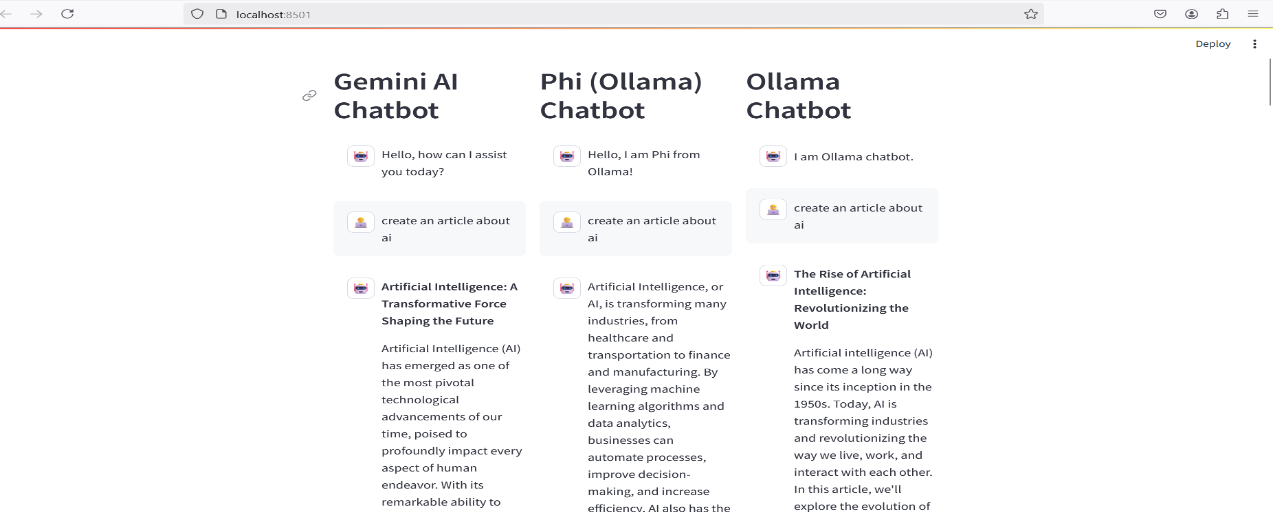


Fig: 1.3 Chatbot Response

**6. Response Comparison and Evaluation**

**Criteria for Comparison**

To evaluate and compare the three models, we consider the following key factors:

1. **Response Quality**: The relevance, coherence, and clarity of the generated response.
2. **Response Speed**: How quickly each model generates a response after receiving the user's input.
3. **Usability**: The ease with which the models can be integrated into applications and the flexibility they offer.
4. **Scalability**: How well the models handle different types of queries and the ability to scale for larger applications.
5. **Cost**: For cloud-based models like Gemini AI, pricing can influence the overall decision for integration.

**6.1 Gemini AI**

**Response Quality**:

* **Strengths**: Gemini AI produces highly coherent, accurate, and well-structured responses. It handles complex queries with in-depth context, making it ideal for applications that require detailed information.
* **Weaknesses**: It may not perform as well in generating informal or creative responses compared to models optimized for casual conversation.

**Response Speed**:

* **Moderate**: Gemini AI provides quality responses but may experience slight delays for complex requests.

**Usability**:

* **High**: Gemini AI is well-documented and supports a wide range of use cases. However, its integration requires proper API setup, and the cost of API calls can add up for large-scale use.

**Scalability**:

* **High**: Being cloud-based, Gemini AI can scale effectively to handle a large volume of requests.

**Cost**:

* **Moderate to High**: The cost depends on the usage, and cloud-based pricing may not be ideal for all projects.

**6.2 Ollama Phi**

**Response Quality**:

* **Strengths**: Ollama Phi is designed for casual, informal conversations, providing conversational responses that feel natural and engaging.
* **Weaknesses**: It struggles with highly detailed or specialized responses.

**Response Speed**:

* **Fast**: Ollama Phi generates responses quickly, making it suitable for real-time interactions.

**Usability**:

* **Moderate to High**: Ollama Phi is easy to integrate and supports streaming responses, which enhances the user experience for real-time conversations.

**Scalability**:

* **Moderate**: Ollama Phi is best suited for lightweight tasks and may not handle large-scale, complex queries as effectively as other models.

**Cost**:

* **Low to Moderate**: As an open-source tool, Ollama Phi is cost-effective for small to medium-sized projects.

**6.3 Ollama Llama3**

**Response Quality**:

* **Strengths**: Ollama Llama3 provides a balance between formal and informal responses, making it versatile for various applications.
* **Weaknesses**: While it can generate responses for a variety of queries, it may not always match Gemini AI’s level of detail or accuracy for specialized topics.

**Response Speed**:

* **Fast**: Similar to Ollama Phi, Llama3 provides quick responses, making it ideal for interactive applications.

**Usability**:

* **High**: Llama3 is easy to use and integrates seamlessly into applications.

**Scalability**:

* **Moderate to High**: Llama3 performs well for a wide range of queries but may not scale as well for large-scale, complex systems.

**Cost**:

* **Low to Moderate**: As an open-source model, Llama3 is cost-effective compared to cloud-based services.

**7. Overall Comparison: Which Model is the Best?**

| **Criteria** | **Gemini AI** | **Ollama Phi** | **Ollama Llama3** |
| --- | --- | --- | --- |
| **Response Quality** | High (complex queries, coherent) | Moderate (casual, informal) | Moderate to High (balanced) |
| **Response Speed** | Moderate | Fast | Fast |
| **Usability** | High (cloud, API-based) | Moderate to High (easy integration) | High (easy integration) |
| **Scalability** | High (cloud-based) | Moderate | Moderate to High (scalable) |
| **Cost** | Moderate to High | Low to Moderate | Low to Moderate |

**Winner: Gemini AI**

**Reasons:**

1. **Response Quality**: Gemini AI stands out with its ability to handle complex, detailed queries and provide structured, coherent responses.
2. **Scalability**: Being cloud-based, Gemini AI scales well to handle a high volume of requests and complex tasks.
3. **Versatility**: While Ollama Phi and Llama3 excel at informal conversations, Gemini AI is more versatile and can handle a wide range of use cases, including specialized knowledge tasks.

**Why Not Ollama Models?**

* While **Ollama Phi** and **Llama3** are excellent for casual and informal conversations, they do not match Gemini AI in terms of depth and the ability to provide highly accurate or specialized responses.

**8. Conclusion**

In conclusion, **Gemini AI** is the best model based on its superior response quality, scalability, and versatility. It is ideal for applications that require handling complex queries or need detailed information. However, for applications where cost is a primary concern or the chatbot is designed for casual conversations, **Ollama Phi** and **Llama3** are also great alternatives. Each model excels in different areas, and the choice depends on the specific requirements of the project.