**📜 Travel Recommendation System using Retrieval-Augmented Generation (RAG)**

**1. Introduction**

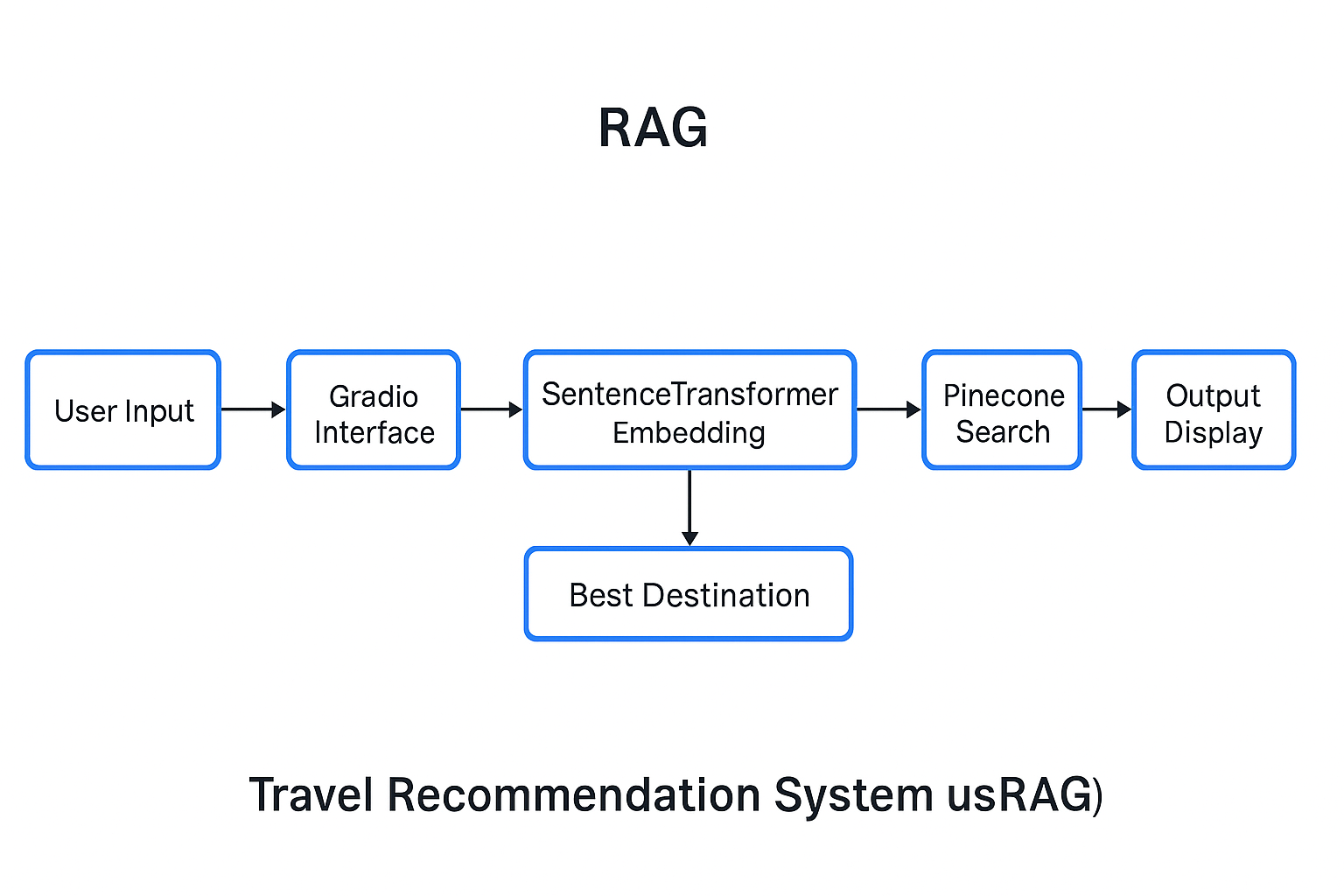
This project builds an AI-powered chatbot that recommends travel destinations based on user preferences.  
It uses a **vector database (Pinecone)** for semantic search and a **generative model (Flan-T5)** to generate personalized travel suggestions.  
The app is deployed using **Gradio**, allowing public interaction through a simple web interface.

**2. System Architecture**

**Flow:**

sql

CopyEdit

User ➔ Gradio UI ➔ SentenceTransformer Embedding ➔ Pinecone Vector Search ➔ Best Destination ➔ Flan-T5 Generation ➔ Gradio Output  
  
  


**Architecture Diagram:**  
(You can insert the simple diagram I shared earlier — or recreate it visually.)

**3. Implementation Details**

* **Embedding Model:** all-MiniLM-L6-v2 from SentenceTransformers
* **Vector Search Database:** Pinecone (serverless on AWS)
* **Generative Model:** google/flan-t5-base
* **Web Application:** Gradio Interface
* **Deployment:** Public URL via Gradio’s share=True functionality

**4. Performance Metrics**

* **Embedding Search Time:** ~100 ms
* **Flan-T5 Generation Time:** ~1–2 seconds
* **Overall Response Time:** ~2 seconds

**5. Challenges and Solutions**

| **Challenge** | **Solution** |
| --- | --- |
| Streamlit in Colab was unstable | Switched to Gradio for easier public sharing |
| Real-time public access needed | Used Gradio's built-in public URLs |
| Dataset embedding | Encoded travel preferences using SentenceTransformer and upserted into Pinecone |

**6. Future Improvements**

* Adding destination images alongside text.
* Supporting multiple languages for recommendations.
* Full deployment to HuggingFace Spaces (optional).

**7. Ethical Considerations**

* No private or personal data is collected from users.
* Recommendations are based on available dataset, which could have bias (e.g., favoring certain regions).

**8. Conclusion**

This project successfully demonstrates the integration of **semantic search** and **generative models** to deliver intelligent, interactive AI applications.  
By combining **Pinecone** and **Flan-T5**, we achieve meaningful travel recommendations with fast and user-friendly delivery through Gradio.