|  |
| --- |
| Table of Contents  [2.1 Design the Chatbot 2](#_Toc201502746)  [2.2 Develop the Chatbot 2](#_Toc201502747)  [2.3 Develop a security assessment for the Chatbot 5](#_Toc201502748) |

# 2.1 Design the Chatbot

**Application Domain and Purpose**

We consider this chatbot to be used within an e-commerce-related return and refund support system. The intent is to automate all front-line customer support interactions, particularly in terms of return policy inquiries. This goes in favor of the general goal of Activity 2, which is demonstrating how generative AI can be put to use to automate a business operation. Thus, it is essentially a very first-line assistant trying to get through straightforward issues efficiently, thus having fewer issues needing human intervention.

**Main Components of the Chatbot**

* **User Interface:** Built using Gradio for a simple, browser-based interaction.
* **Language Model:** OpenAI GPT-3.5-turbo model, chosen for its balanced performance in language understanding and response generation.
* **RAG (Retrieval-Augmented Generation) Layer:** Uses FAISS for vector-based search and MiniLM for embedding return-policy content.
* **Moderation & Safety:** Implemented basic moderation to filter inappropriate queries.
* **Knowledge Base:** A short set of documents outlining the company’s return and refund policies.

**Technologies Used**

* pandas: To manage the knowledge base as structured data
* sentence\_transformers: To generate dense vector embeddings using MiniLM
* faiss: To perform efficient vector similarity search
* openai: To interact with the GPT-3.5 model
* gradio: To build an accessible and clean chatbot UI

**Block Diagram of Chatbot Architecture**



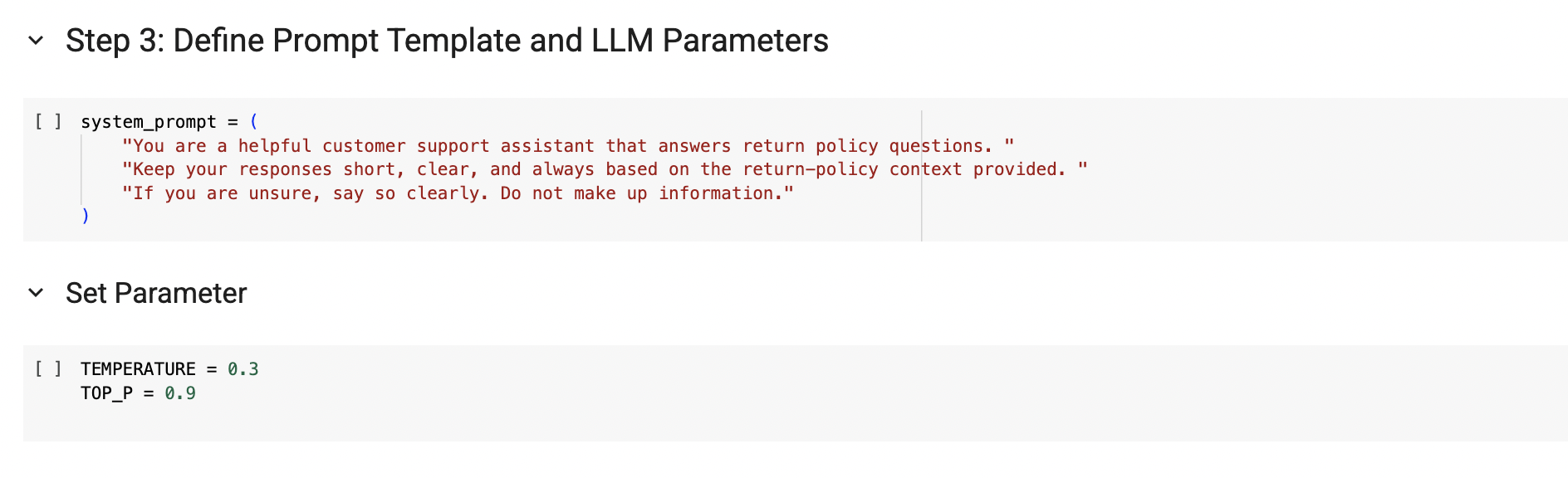
# 2.2 Develop the Chatbot

**Starting Prompt and LLM Parameters**  
A fixed system prompt was used to ensure consistency, safety, and helpfulness:

"You are a helpful customer support assistant that answers return policy questions. Keep your responses short, clear, and always based on the return-policy context provided. If you are unsure, say so clearly. Do not make up information."

**The GPT model was configured with the following parameters:**

* **Temperature**: 0.3 — to produce more deterministic and focused answers.
* **Top-p**: 0.9 — to retain diversity in responses while remaining relevant.
* **Max Tokens**: 512 — enough to return concise yet informative replies.



**Knowledge Base Creation and Embedding**  
A simple return-policy knowledge base was created using pandas. It included entries such as:

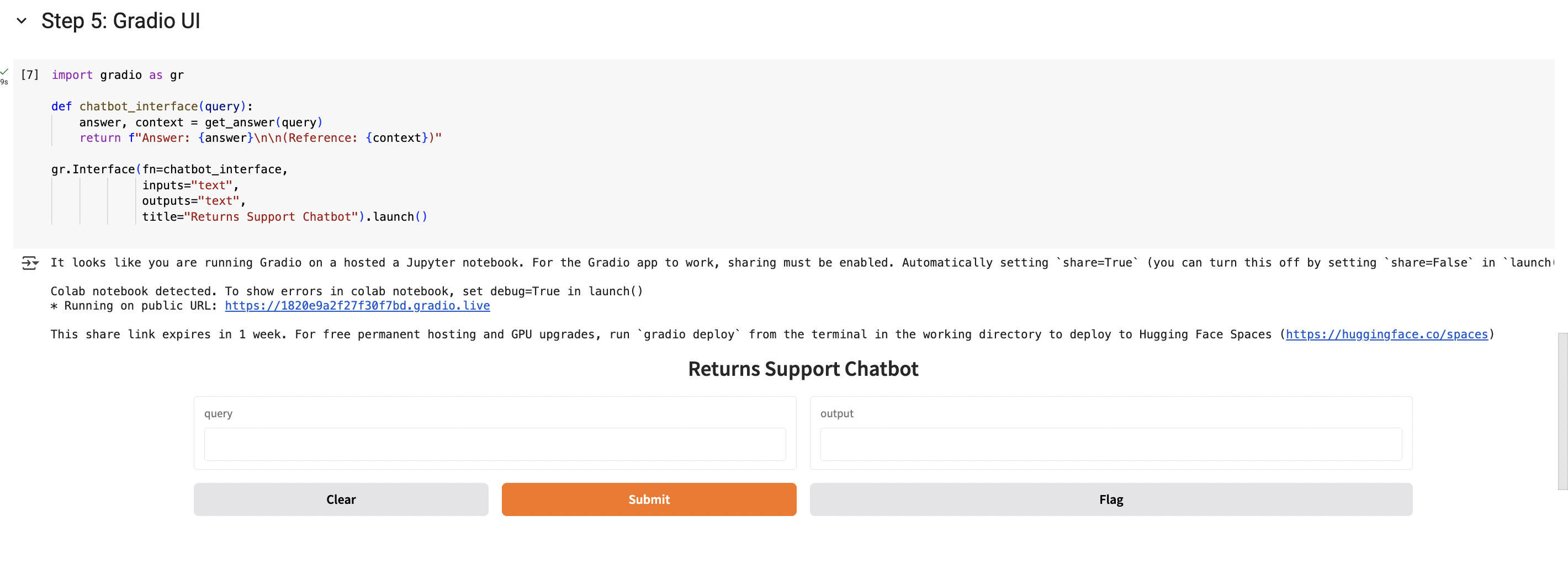
* Return timeframe
* Return conditions
* Refund processing
* Non-returnable items

These text entries were embedded using MiniLM and indexed using FAISS for efficient top-k document retrieval.

**Query Handling Process**  
A get\_answer() function was written to:

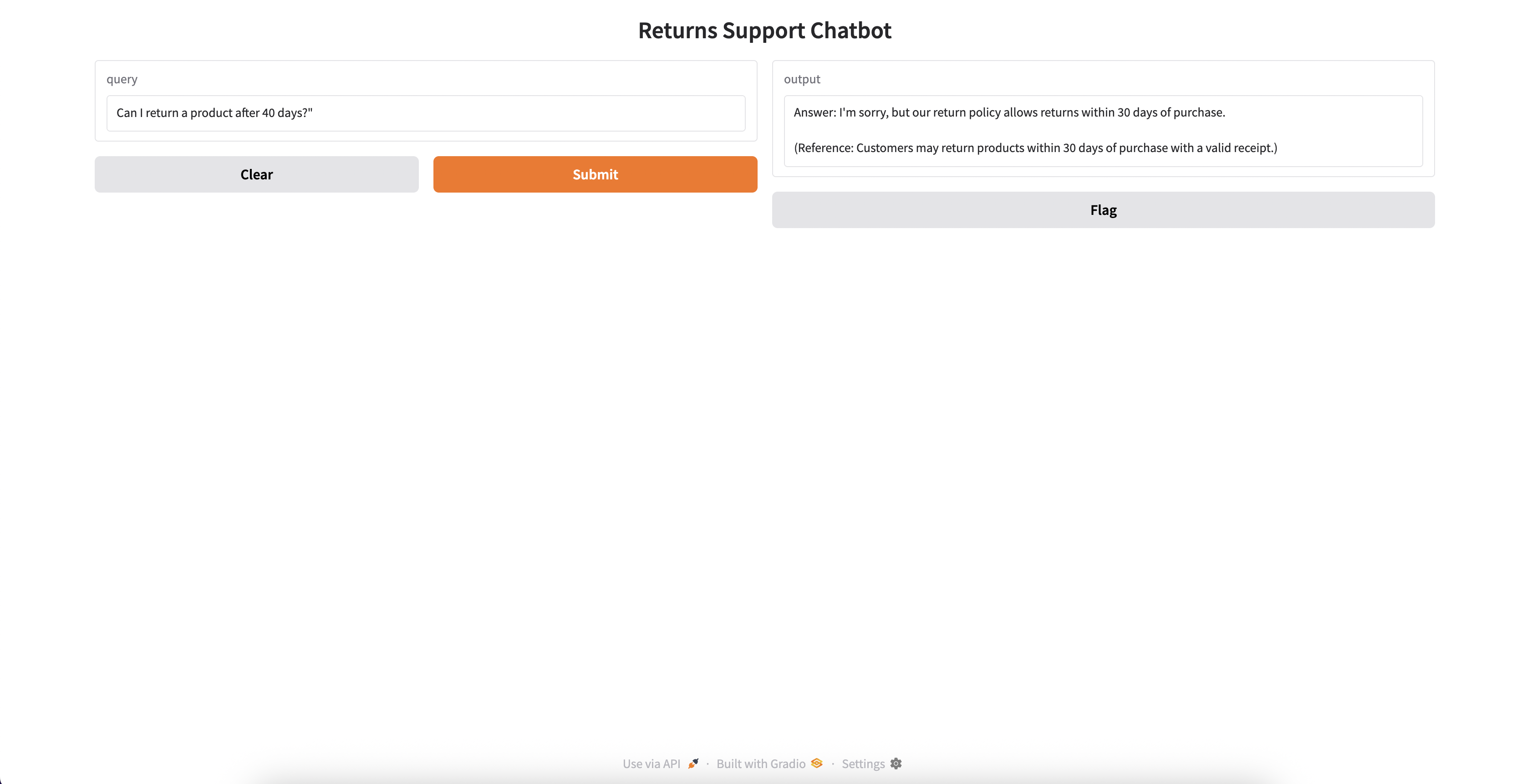
1. Encode the query using the same MiniLM model.
2. Search for the top-1 relevant snippet using FAISS.
3. Combine the query and context snippet with the system prompt.
4. Generate a natural-language answer using GPT-3.5.

**Live Interaction through Gradio**  
Gradio was used to build a real-time UI. The chatbot responded to user input and displayed both the generated answer and the supporting document for transparency.



**Demonstration and Examples**  
The chatbot was tested with the following sample queries:

**Query**: "Can I return a product after 40 days?"  
**Response**: "I'm sorry, but our return policy allows returns within 30 days of purchase with a valid receipt."  
**Reference**: "Customers may return products within 30 days of purchase with a valid receipt."



**Query**: "What’s your refund process for damaged items?"  
**Response**: "Refunds for damaged items are issued to the original payment method within 7 business days after return approval."

A screenshot of a chatbot

AI-generated content may be incorrect.

**Sample Test Queries and Outputs**

|  |  |  |
| --- | --- | --- |
| **User Query** | **Model Response** | **Referenced Policy** |
| Can I return a product after 40 days? | I'm sorry, but our return policy allows returns within 30 days of purchase with a valid receipt. | Customers may return products within 30 days of purchase with a valid receipt. |
| Do I need to keep the original receipt? | Yes, you need to keep the original receipt for returns. | Customers may return products within 30 days of purchase with a valid receipt. |
| What’s your refund process for damaged items? | Refunds are issued to the original payment method within 7 business days after return approval. | Refunds are issued to the original payment method within 7 business days after return approval. |
| How do I exchange something if the size is wrong? | You can exchange the item for a different size as long as it is unused and in its original packaging with all tags attached. | Items must be unused and in original packaging. Clothing must have all tags attached. |

These results confirmed the chatbot’s ability to retrieve the correct policy and generate grounded answers.

# 2.3 Develop a security assessment for the Chatbot

**Identified Security Challenges**

* Prompt Injection: Users might attempt to override the chatbot.
* Inappropriate Queries: Offensive questions, misinforming questions, ignoring the warnings for disclaimers, miss using for money disclaimers might impact the company's reputation.
* Context Misuse: False or absent context might lead to hallucinated answers.

**Vulnerability Analysis and Mitigation**

* Being a strict system prompt constrains the chatbot from deviating away from the intended role.
* The chatbot cannot return information from anything else other than approved policy snippets.
* Lacking collection or storage of any user data, the chatbot remains privacy-preserving by structure.

**Emerging Threats and Trends**  
Modern LLMs face evolving threats like zero-shot jailbreaks and semantic manipulation. This prototype minimizes such exposure by limiting the scope to factual recall from a static dataset. By using MiniLM + FAISS and not feeding in long contextual histories, the attack surface remains small.

**Relevance to the Case Study Goal**  
The main aim of this chatbot is for scalable and cost-effective customer service regarding returns. Security is critical for this end since any aggressor or misuse would erode the trust of the customer and the reliability of operations. The mitigations, embedded in the design and flow of interaction, keep the tool stable and compliant within its small but critical use case.

**Conclusion**

This chatbot is a working demonstration of how retrieval-augmented AI can enhance e-commerce customer support. Through the integration of MiniLM, FAISS, OpenAI GPT-3.5, and Gradio, it delivers quick and trustworthy responses grounded in real business policies. The development emphasized not just model performance, but also explainability, transparency, and operational security. As AI adoption in business expands, such focused tools can improve service quality and reduce human resource load while maintaining high levels of accountability.