Name: Sahal Shrestha

Date: 30/01/2026

Section A:

1) B
2) C
3) B
4) B
5) B
6) A
7) C
8) C
9) D
10) B
11) B
12) B
13) B
14) B
15) A
16) B
17) B
18) B
19) B
20) B

Section B

Link: [Click here]

1.

1. Chunking Strategy

To ensure high retrieval accuracy, we used a **Recursive Content-Aware** chunking strategy.

Implementation Details

- **Tool**: LangChain's `RecursiveCharacterTextSplitter`.
- **Chunk Size**: `500` characters. This provides a balance between sufficient context for the model and focused retrieval.
- **Chunk Overlap**: `50` characters. This ensures that information spanning across chunks maintains continuity.
- **Separators**: `["\n\n", "\n", " ", ""]`
  - **Priority 1 (`\n\n`)**: We first attempt to split by double newlines to keep individual product entries as atomic units.
  - **Priority 2 (`\n`)**: If a product entry is too long, we split by line breaks.
  - **Priority 3 (` `)**: As a final resort, we split at word boundaries.

2. Embedding Configuration

Embeddings convert text into high-dimensional vectors that the chatbot uses to find relevant information.

Implementation Details

- **Model**: `models/embedding-001` via **Google Generative AI (Gemini)**.
- **Provider**: `langchain-google-genai`.
- **Vector Store**: **ChromaDB**.
- **Persistence**: Data is persisted locally in the /home/labuser/assessment/chroma_db directory.

3. Workflow Integration

During a query:

1. The user's question is embedded using the same `embedding-001` model.
2. A **Similarity Search** is performed against the ChromaDB vector store.
3. The top `k=3` most relevant chunks are retrieved and injected into the LLM prompt as context.

2.

Walkthrough - Product Knowledge Base RAG

I have completed the end-to-end implementation of the Product Knowledge Base RAG system.

Changes Made

1. Knowledge Base Ingestion

- Developed ingest_data.py to process product details.
- Used `RecursiveCharacterTextSplitter` for chunking.
- Stored embeddings in a persistent ChromaDB local instance.

2. RAG Inference Chain

- Created rag_chain.py using LangChain Expression Language (LCEL).
- **Model**: Configured `gemini-2.5-flash` as requested.
- **Retriever**: Connected to the ChromaDB vector store.
- **Prompting**: Implemented a system prompt that enforces answering strictly based on the provided context.

3.

1. LangGraph Workflow Structure

Developed langgraph_workflow.py which implements the following nodes:

- **Node 1: Classifier**: Uses `gemini-2.5-flash` to categorize queries into `products`, `returns`, or `general`.
- **Node 2: RAG Responder**: Reuses our RAG logic to answer product-specific questions when the classifier detects a product query.
- **Node 3: Escalation**: Intercepts `returns` or `general` queries and provides a structured escalation response.

2. Conditional Routing

Implemented routing logic that directs the flow dynamically:

- `products` → **RAG Responder**
- `returns` or `general` → **Escalation**

4.

localhost:8000/docs#/

Product Knowledge Base Chatbot 0.1.0 OAS 3.1

/openapi.json

default

GET / Root

POST /query Query Chatbot

Schemas

HTTPValidationError › Expand all object

Code	Details
200	Response body

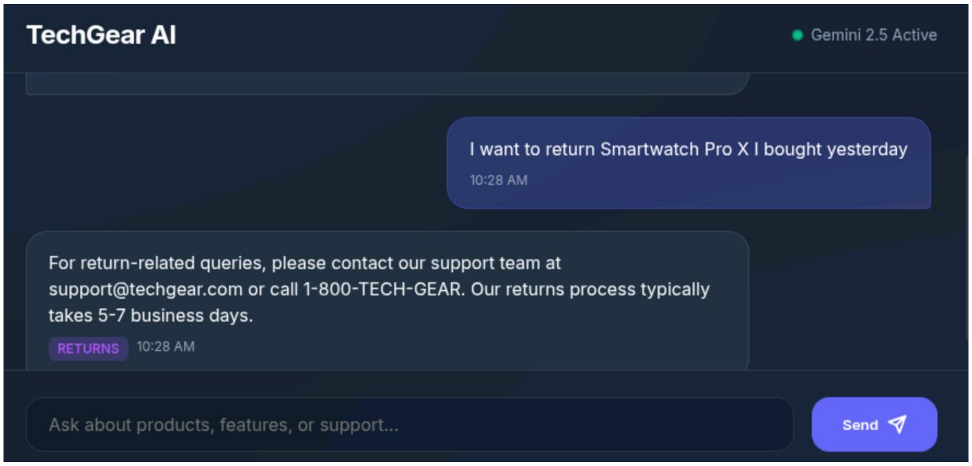

```
{
  "question": "what is the price of Smartwatch Pro X",
  "category": "products",
  "answer": "The price of SmartWatch Pro X is ₹15,999."
}
```

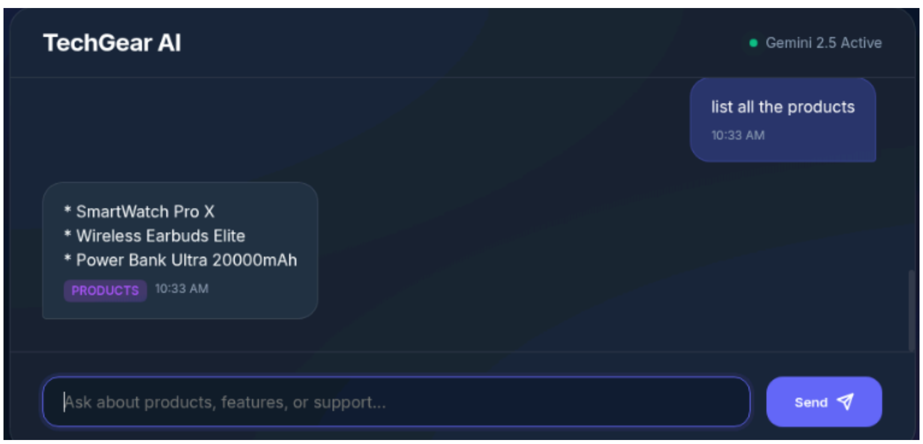

System Architecture

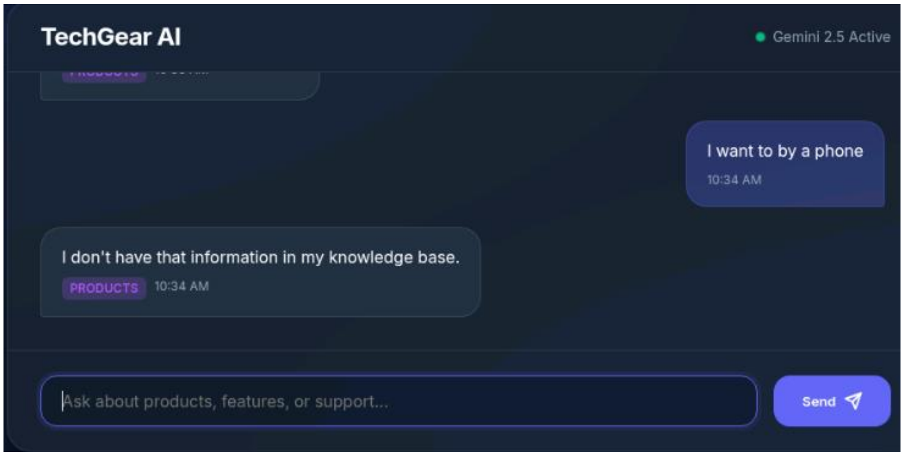
The system is built with a layered architecture for reliability and intelligence:

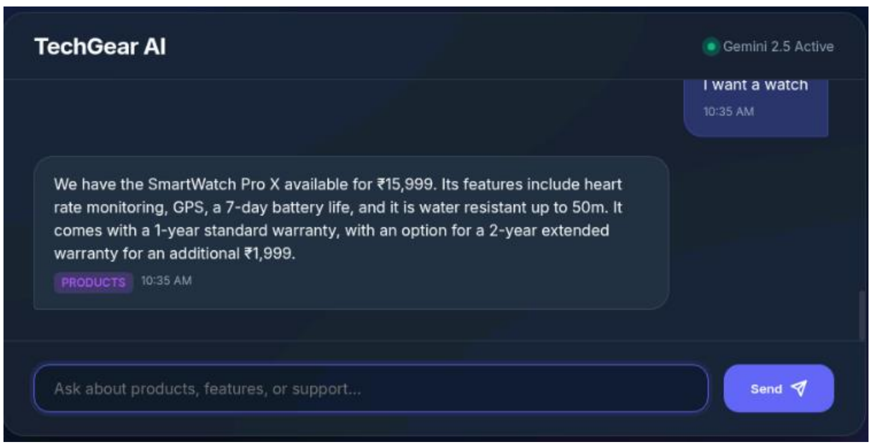
1. **Ingestion Layer**: Processes product_details.txt into high-quality vector embeddings stored in a local ChromaDB .
2. **Orchestration Layer**: A LangGraph workflow that classifies user intent and routes queries to either the RAG Responder or an Escalation handler.
3. **Service Layer**: A FastAPI application that provides a robust web interface for user interaction.

Outputs:


TechGear AI
Gemini 2.5 Active
I want to return Smartwatch Pro X I bought yesterday
10:28 AM
For return-related queries, please contact our support team at support@techgear.com or call 1-800-TECH-GEAR. Our returns process typically takes 5-7 business days.
RETURNS
10:28 AM
Ask about products, features, or support...
Send


TechGear AI
Gemini 2.5 Active
list all the products
10:33 AM
* SmartWatch Pro X
* Wireless Earbuds Elite
* Power Bank Ultra 20000mAh
PRODUCTS
10:33 AM
Ask about products, features, or support...
Send


TechGear AI
Gemini 2.5 Active
I want to by a phone
10:34 AM
I don't have that information in my knowledge base.
PRODUCTS
10:34 AM
Ask about products, features, or support...
Send


TechGear AI
Gemini 2.5 Active
I want a watch
10:35 AM
We have the SmartWatch Pro X available for ₹15,999. Its features include heart rate monitoring, GPS, a 7-day battery life, and it is water resistant up to 50m. It comes with a 1-year standard warranty, with an option for a 2-year extended warranty for an additional ₹1,999.
PRODUCTS
10:35 AM
Ask about products, features, or support...
Send