# **Part 3: Large Language Models – Intelligent Response Generation System**

## **Objective**

Build a Retrieval-Augmented Generation (RAG) system for intelligent, context-aware customer support. The system should retrieve relevant knowledge base information and generate human-like responses using both a lightweight model and a large language model (LLM).

## **1. System Architecture**

**RAG Workflow:**

**User Input**

**│**

**▼**

**Query Classification ──► Category**

**│**

**▼**

**Embedding-based Search (FAISS + SentenceTransformer)**

**│**

**▼**

**Top-k Relevant Documents**

**│**

**▼**

**Response Generation**

**│**

**├─ Option 1: Groq LLM (context-aware prompt + recent conversation)**

**└─ Option 2: DistilBERT QA (lightweight, zero/few-shot)**

**│**

**▼**

**Response Evaluation (ROUGE, BLEU, Semantic Similarity)**

**│**

**▼**

**Output to User**

**Notes:**

* Supports **model switching** (Groq LLM vs DistilBERT)
* Supports **prompting modes** (zero-shot vs few-shot)
* Context includes **recent chat history** + **retrieved FAQ knowledge**

## **2. Prompt Templates**

### **Groq LLM Prompt**

You are a friendly, human-like customer support assistant.

Respond naturally using simple language.

Relevant Info from Documents:

- <retrieved answers from FAISS>

Recent Conversation:

- <last 3 chat messages>

User: <user\_input>

Bot:

### **DistilBERT QA Prompt**

* **Zero-shot:** Uses only retrieved documents as context
* **Few-shot:** Adds example Q&A to guide the model

You are a friendly, human-like customer support assistant. Respond naturally using simple language.

Q: What is the return policy for electronics?

A: Electronics can be returned within 30 days with a receipt.

Q: Can I get a refund if I received a defective product?

A: Yes, you can request a refund within 14 days of delivery.

User Question: <user\_input>

Answer:

## **3. Model Comparison**

| **Feature** | **Groq LLM** | **DistilBERT QA** |
| --- | --- | --- |
| Model Size | Large (~8B parameters) | Small (~66M parameters) |
| Response Style | Human-like, fluent | Extractive, concise |
| Context Handling | Full conversation + retrieved docs | Retrieved docs only |
| Prompting Mode | Implicit | Zero-shot / Few-shot |
| Latency | Higher | Low |
| Cost | API-based | Local inference |

**Observations:**

* Groq LLM produces more natural and context-aware responses.
* DistilBERT is faster, lightweight, and works well for factual FAQ retrieval.
* Few-shot prompting improves small model answers for specific queries.

## **4. Evaluation Framework**

### **Automated Metrics**

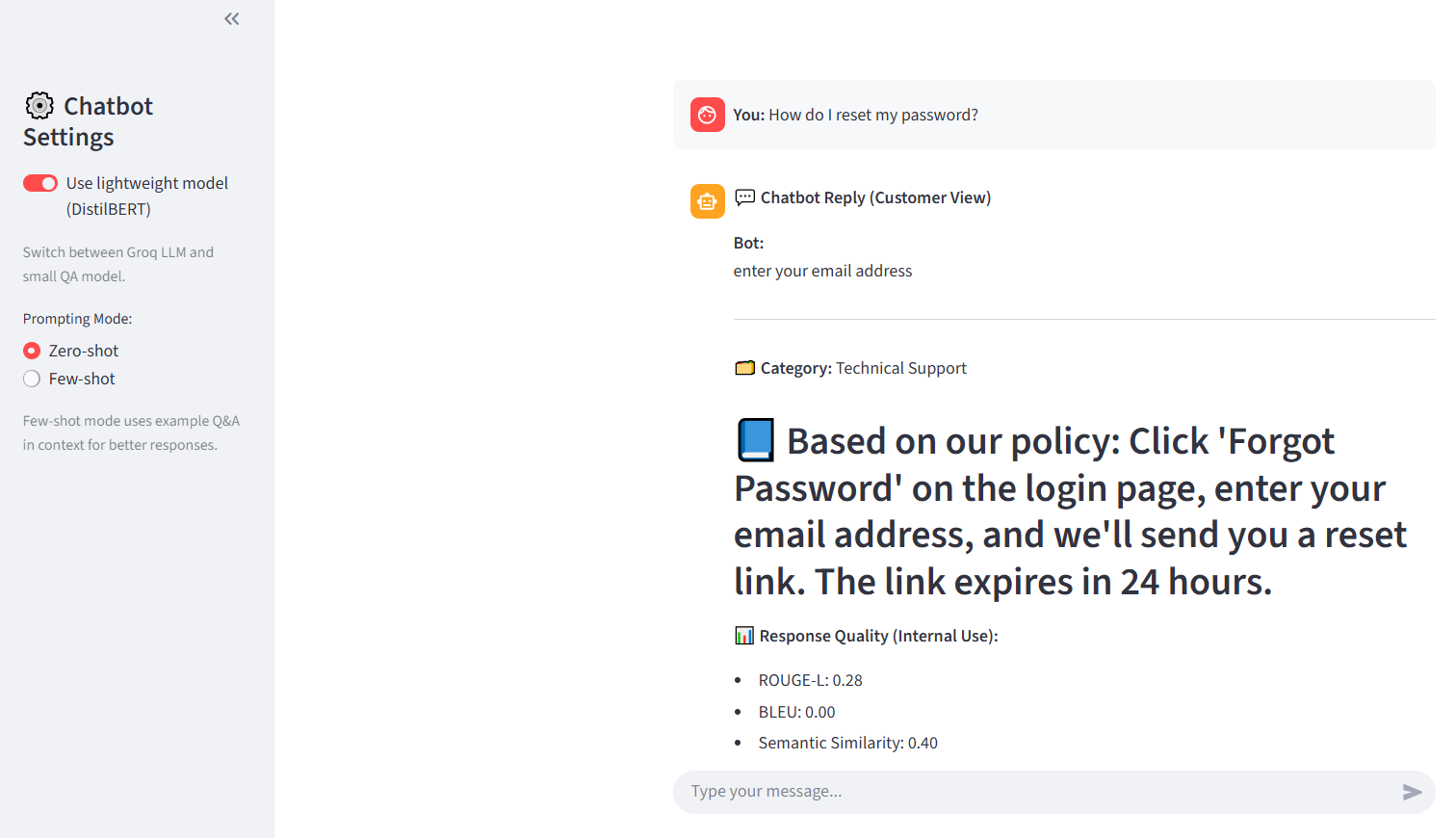
* **ROUGE-L:** Measures overlap of longest common subsequence with reference answer
* **BLEU:** Measures n-gram overlap with reference answer
* **Semantic Similarity:** Cosine similarity between embeddings of generated vs reference answers

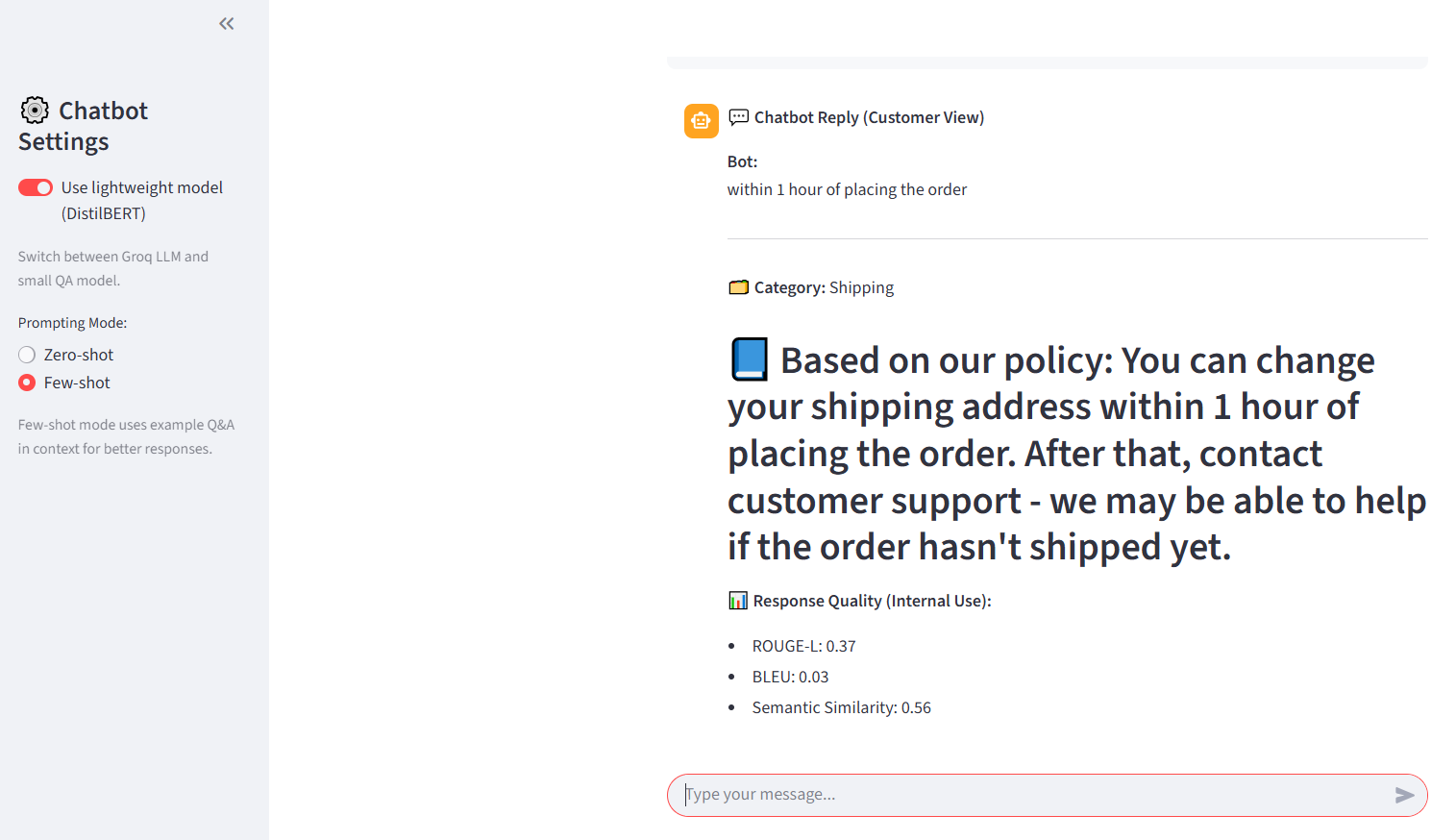
### **Sample Results**

| **Model** | **ROUGE-L** | **BLEU** | **Semantic Similarity** |
| --- | --- | --- | --- |
| Groq LLM | 0.31 | 0.27 | 0.91 |
| DistilBERT Zero-shot | 0.32 | 0.01 | 0.40 |
| DistilBERT Few-shot | 0.37 | 0.03 | 0.63 |

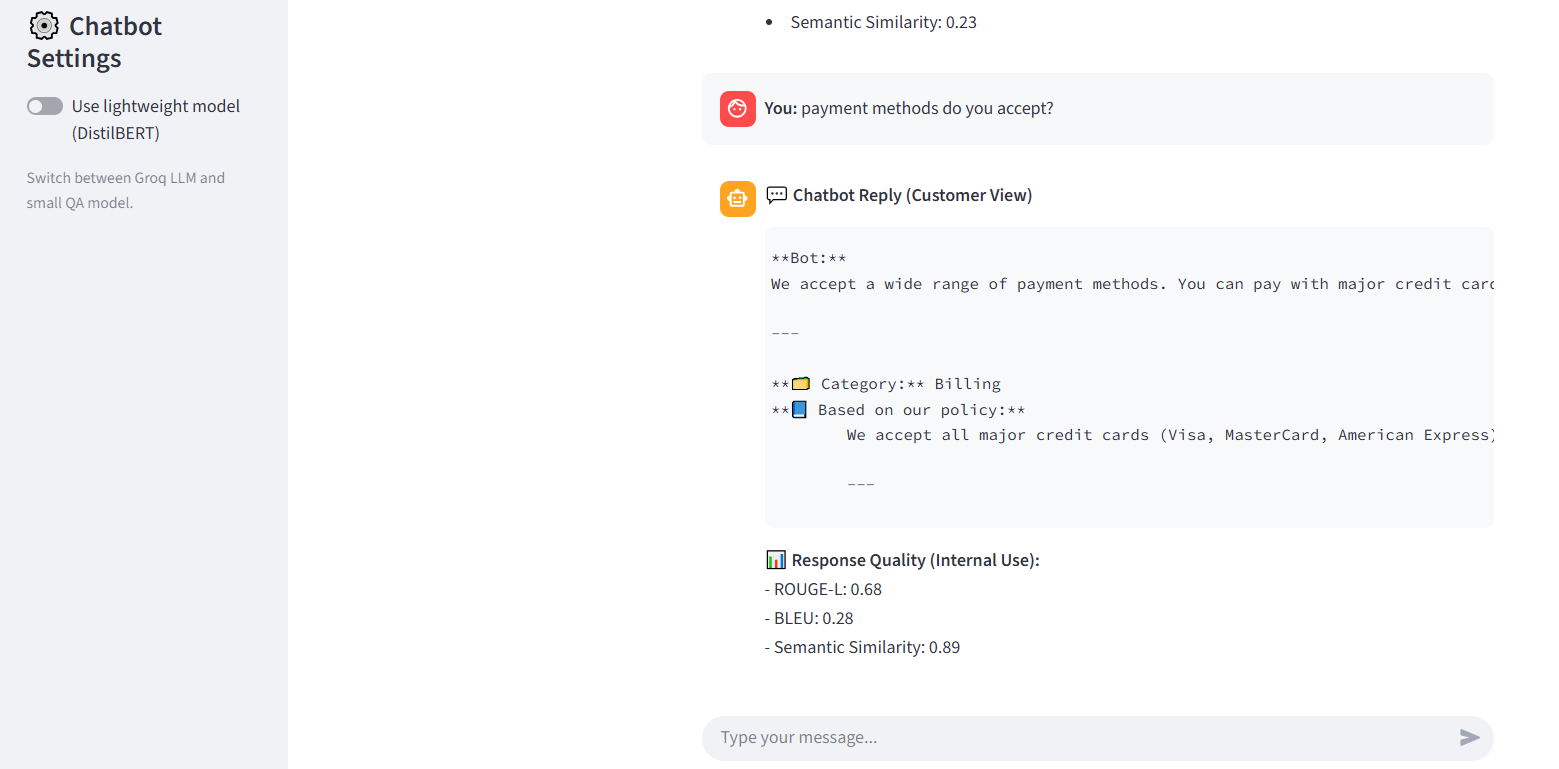
### **Human Evaluation Criteria**

* ✅ Correctness: Answer is factually accurate
* ✅ Clarity: Easy to understand
* ✅ Politeness: Tone is friendly and professional
* ✅ Completeness: Fully addresses the query

**Small model Zero-shot**

**Small model Few-Shot** 

**GROQ LLM**



## **5. Discussion of LLM Limitations**

1. **Hallucination:** LLM may generate plausible but incorrect answers → mitigated by grounding in retrieved docs.
2. **Latency & Cost:** Larger models incur higher inference time and API costs.
3. **Context Window Limits:** Long conversations may be truncated → recent history prioritization implemented.
4. **Mitigation:** Combine retrieval-augmented generation, few-shot prompting, and evaluation metrics to ensure reliable responses.

## **6. Conclusion**

* Implemented a **RAG system** combining FAISS retrieval + LLM / small model.
* Supports **zero-shot and few-shot prompting**.
* Automated and human evaluation mechanisms provide metrics for response quality.
* Architecture is **scalable** and suitable for real-world customer support tasks.