# Tesla RAG System

## Retrieval-Augmented Generation for Policy & Product Knowledge Assistant

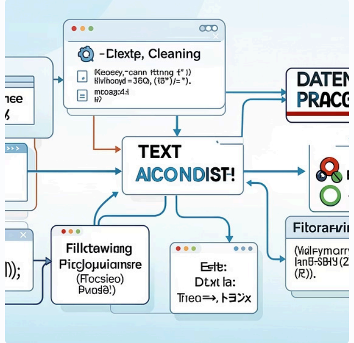
Built with: FAISS, Sentence-Transformers, Groq LLM, Streamlit

Author: Tamanna Yadav

# System Architecture Overview

### Ingestion Layer

PDF extraction using pdfplumber for precise text capture from Tesla documentation

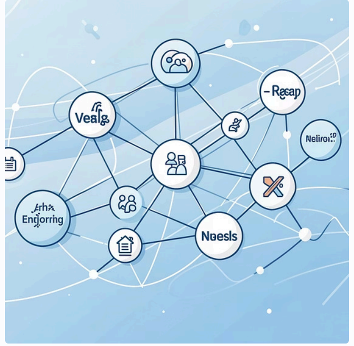


### Preprocessing

Text cleaning and normalisation ensure consistent, high-quality input for embedding generation

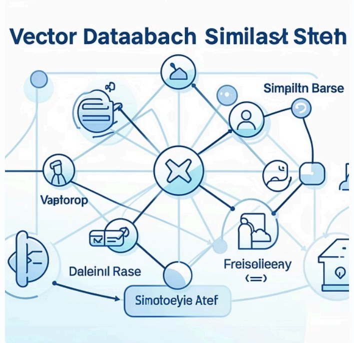
### Chunking Strategy

Recursive splitter with 512 characters per chunk and 50-character overlap maintains context continuity



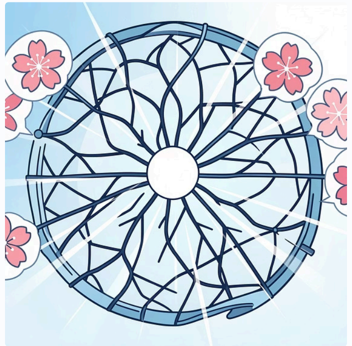
### Embeddings

all-MiniLM-L6-v2 generates 384-dimensional vectors for semantic similarity search



### Vector Database

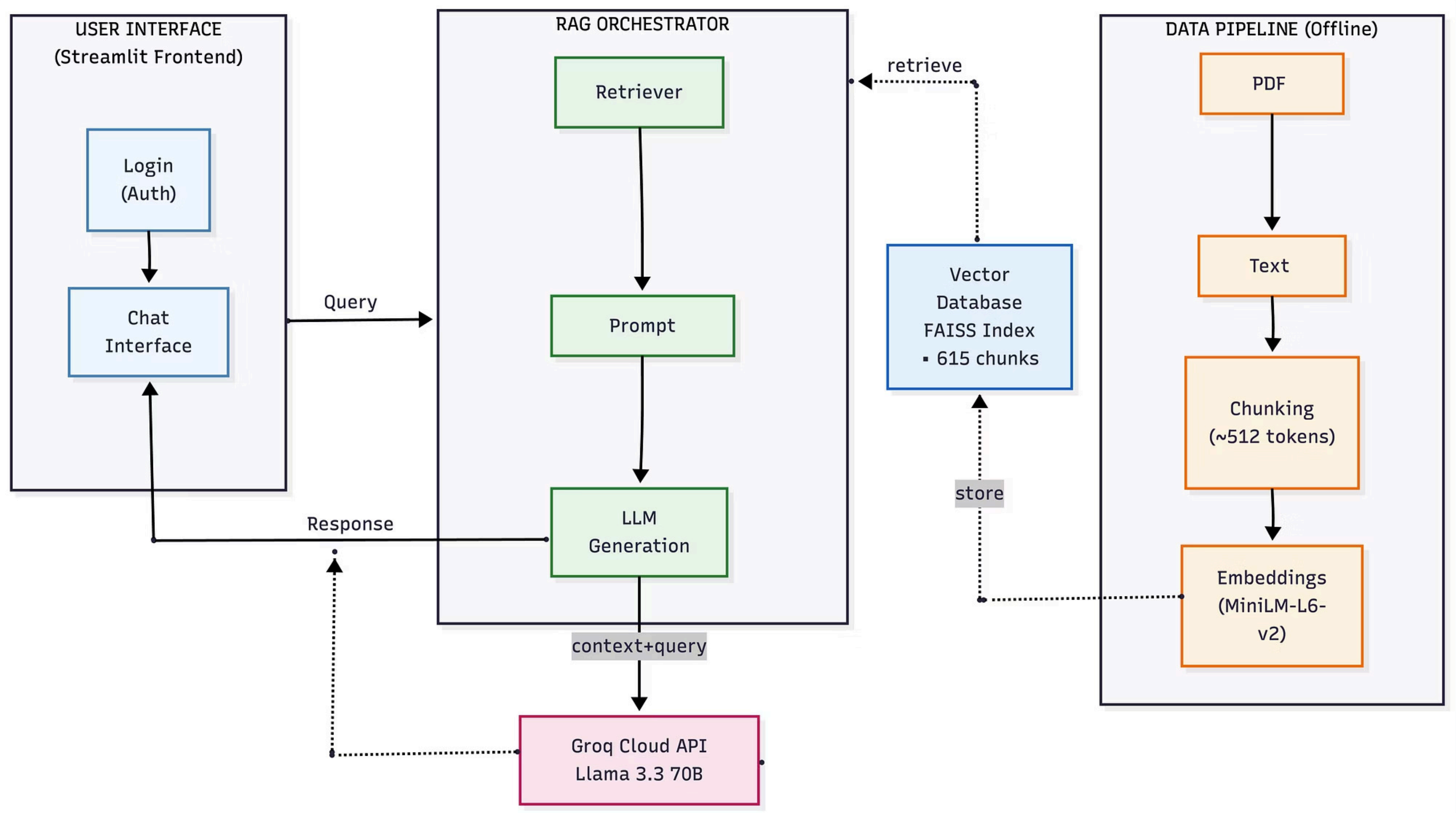
FAISS with cosine similarity enables lightning-fast retrieval from 615 indexed chunks



### Generation Model

Groq Llama 3.3 70B Versatile produces high-quality, contextually grounded responses

# Architecture Diagram



# Hybrid Architecture: SLM + LLM

My system leverages a strategic hybrid approach that combines the speed of small language models with the reasoning power of large language models, optimising for both performance and quality.

## Why Groq Llama 3.3 70B?

01

### Fastest Inference

Hardware-optimised LPU architecture delivers industry-leading speed

02

### Cost-Effective

Free tier enables development without budget constraints

03

### 70B Parameters

Excellent reasoning and instruction-following capabilities

04

### Extended Context

Large context window handles multiple retrieved chunks effectively

| Component | Choice | Reasoning |
|---|---|---|
| Embedding Model | all-MiniLM-L6-v2 (SLM) | Fast, 384 dimensions, runs locally, excellent for semantic search |
| Generation Model | Llama 3.3 70B (LLM) | High-quality responses, handles complex queries with superior reasoning |

# LLM vs SLM: Strategic Trade-offs

Understanding the strengths and limitations of small versus large language models informed my architectural decisions. Each model type excels in different dimensions, making the hybrid approach optimal.

### Inference Speed

**SLM:** Lightning-fast local execution with minimal latency

**LLM:** Slower inference due to model size and complexity

### Cost Structure

**SLM:** Free or minimal cost for local deployment

**LLM:** API costs scale with usage volume

### Response Quality

**SLM:** Limited reasoning and contextual understanding

**LLM:** Superior quality with advanced reasoning capabilities

### Deployment

**SLM:** Local or edge deployment with minimal infrastructure

**LLM:** Cloud-based deployment requires robust infrastructure

### SLM for Embeddings

MiniLM runs locally with no API calls, enabling rapid retrieval operations

### LLM for Generation

Groq's Llama 3.3 70B delivers high-quality, contextually grounded responses

### Optimal Result

Fast retrieval combined with premium generation quality

# RAG vs Base LLM: Evaluation Results

Our comprehensive evaluation demonstrates that retrieval-augmented generation significantly outperforms base LLM approaches across critical metrics. The RAG system's ability to ground responses in actual Tesla documentation translates to measurably better performance.

### Reduced Hallucination

RAG reduces hallucination by 37% (0.201 vs 0.320) by grounding answers in actual Tesla documents rather than relying on potentially outdated training data

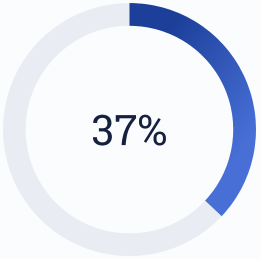
### Traceable Answers

Faithfulness score of 0.396 ensures every claim can be traced directly to source documents, whilst base LLM shows 0.0 grounding

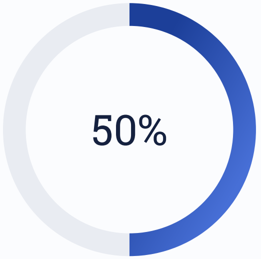
### Strong Retrieval

MRR of 0.500 indicates relevant documents consistently appear in top 2 positions, with ROUGE-L of 0.215 showing good textual overlap

| Metric | RAG System | Base LLM | Winner | Improvement |
|---|---|---|---|---|
| Answer Relevance | 0.980 | 1.000 | Tie | — |
| Faithfulness | 0.396 | 0.000 | **RAG** | ∞ |
| Hallucination Risk | 0.201 | 0.320 | **RAG** | 37% lower |
| ROUGE-L | 0.215 | 0.000 | **RAG** | ∞ |
| MRR | 0.500 | 0.000 | **RAG** | ∞ |



Hallucination Reduction



MRR Score

Top-2 retrieval

# Why RAG Outperforms Base Models

## Base LLM Limitations

Large language models operating without retrieval mechanisms face fundamental constraints that limit their effectiveness for enterprise knowledge systems.

### Knowledge Cutoff

Relies solely on training data with fixed knowledge cutoff dates

### Hallucination Risk

May fabricate specific Tesla policies or technical specifications

### Generic Knowledge

Lacks Tesla-specific details and proprietary information

### No Grounding

Cannot verify claims against actual source documents

## RAG System Advantages

Our retrieval-augmented approach fundamentally transforms how the system accesses and utilises information, delivering superior reliability.

01

### Active Retrieval

Retrieves relevant chunks from current Tesla PDFs in real-time

02

### Document Grounding

Answers grounded in actual, verifiable Tesla documentation

03

### Source Attribution

Cites specific sources with page numbers for transparency

04

### Reduced Hallucination

37% lower hallucination risk through factual grounding

# Technical Specifications

## Core System Configuration

Our carefully tuned parameters balance retrieval quality, generation accuracy, and system performance across the entire RAG pipeline.

| Parameter | Value |
|---|---|
| Chunk Size | 512 characters |
| Chunk Overlap | 50 characters |
| Embedding Model | all-MiniLM-L6-v2 |
| Embedding Dimension | 384 |
| Vector Store | FAISS (Flat Index) |
| Similarity Metric | Cosine Similarity |
| Top-K Retrieval | 5 chunks |
| LLM | Llama 3.3 70B Versatile |
| Temperature | 0.1 (low for factual) |
| Top-P | 0.9 |

## Dataset Overview

**7**

### PDF Documents

Tesla source files

**615**

### Indexed Chunks

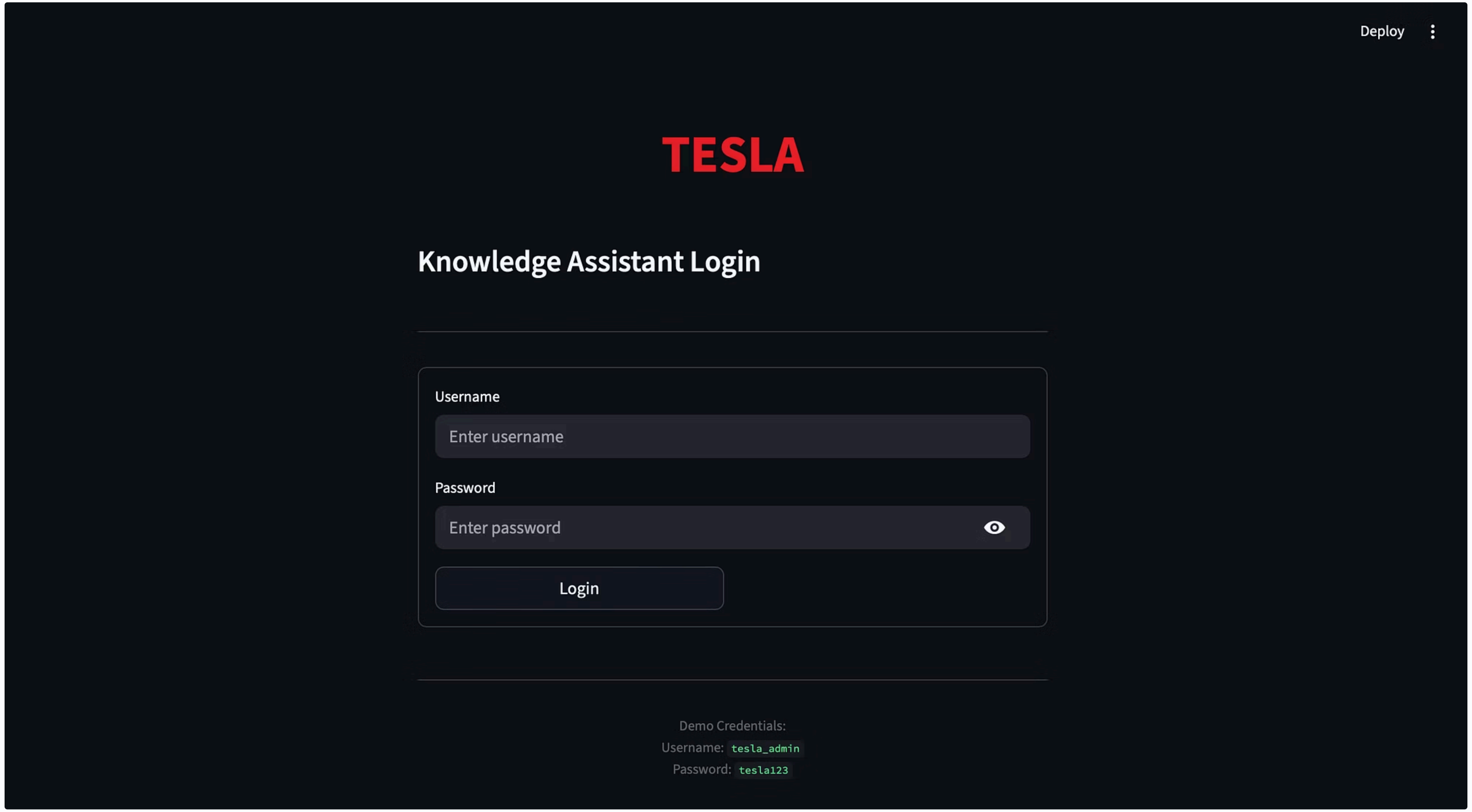
Searchable segments

Our comprehensive dataset spans multiple document types to provide broad coverage of Tesla knowledge:

- Privacy Policy documentation
- Terms of Service agreements
- Owner's Manual technical specs
- Impact Report sustainability data

# Streamlit Application

## 1. Login Page



Knowledge Assistant Login

Username

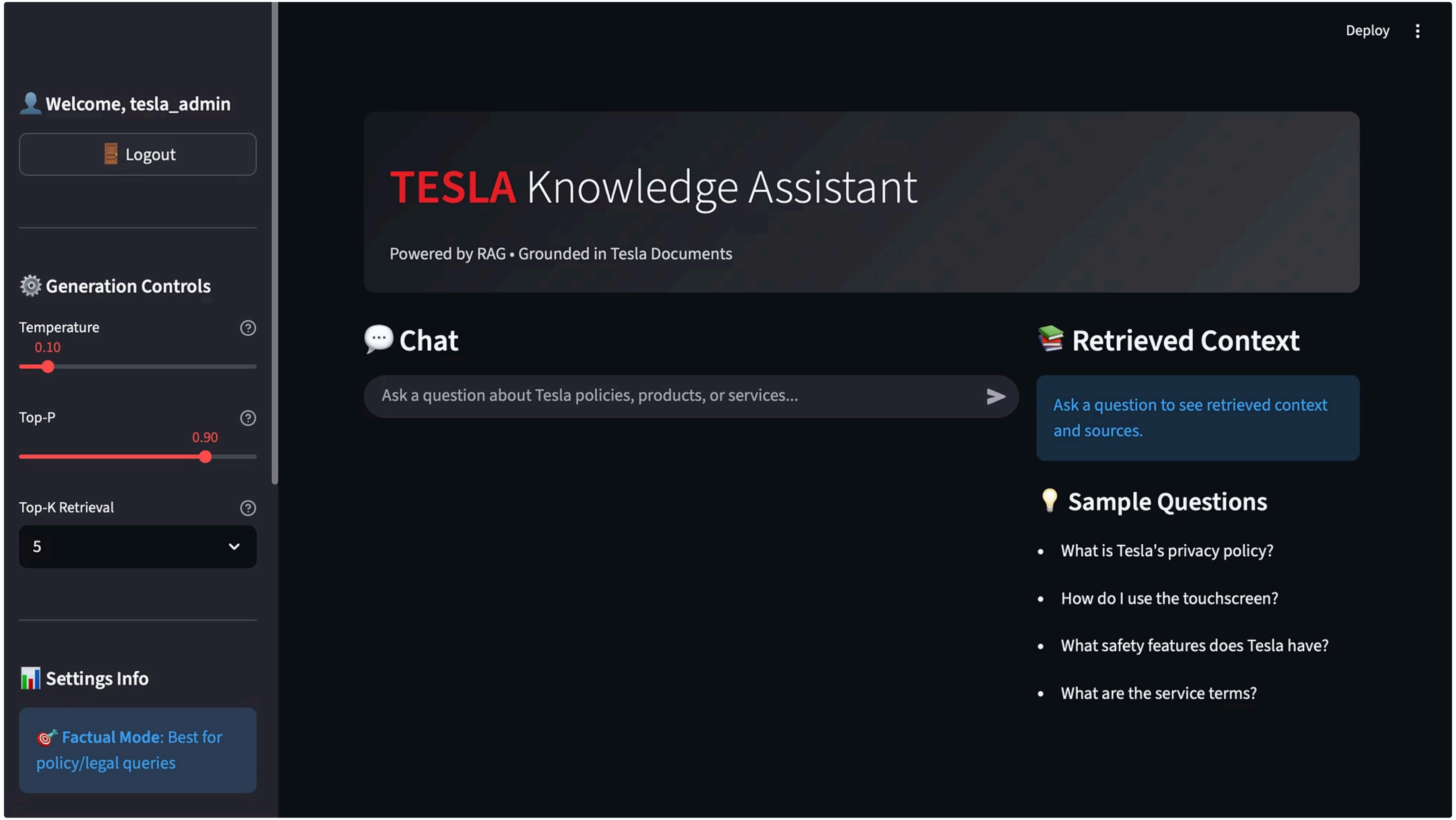
Enter username

Password

Enter password

Demo Credentials:
Username: `tesla_admin`
Password: `tesla123`

## 2. Home Page



👤 Welcome, tesla_admin

⚙️ Generation Controls

Temperature ⓘ
0.10

Top-P ⓘ
0.90

Top-K Retrieval ⓘ
5

📊 Settings Info

🎯 **Factual Mode**: Best for policy/legal queries

# TESLA Knowledge Assistant

Powered by RAG • Grounded in Tesla Documents

💬 Chat

Ask a question about Tesla policies, products, or services...

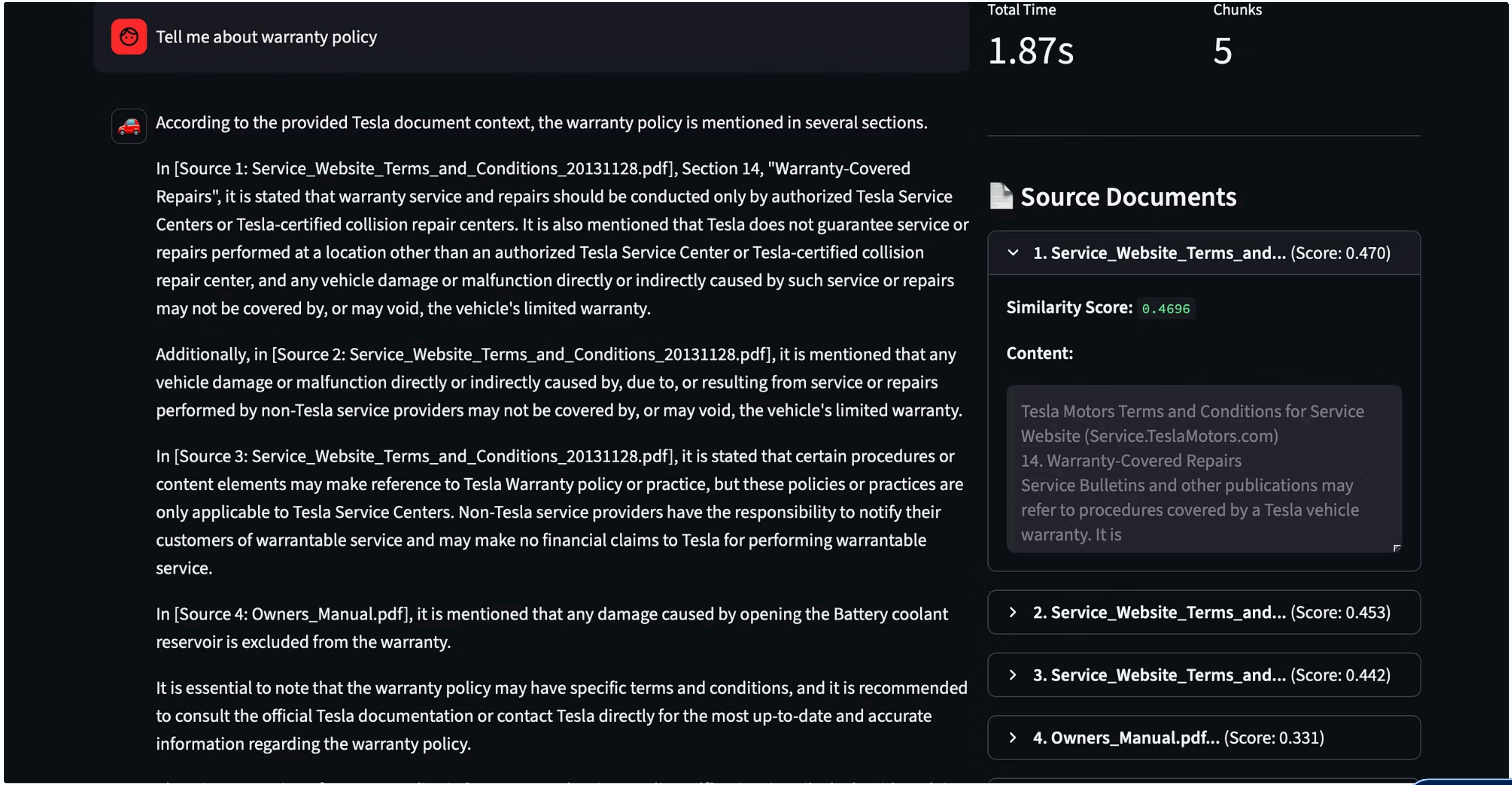
📚 Retrieved Context

Ask a question to see retrieved context and sources.

💡 Sample Questions

- What is Tesla's privacy policy?
- How do I use the touchscreen?
- What safety features does Tesla have?
- What are the service terms?

# 3.Query -response along with source documents



Tell me about warranty policy

According to the provided Tesla document context, the warranty policy is mentioned in several sections.

In [Source 1: Service_Website_Terms_and_Conditions_20131128.pdf], Section 14, "Warranty-Covered Repairs", it is stated that warranty service and repairs should be conducted only by authorized Tesla Service Centers or Tesla-certified collision repair centers. It is also mentioned that Tesla does not guarantee service or repairs performed at a location other than an authorized Tesla Service Center or Tesla-certified collision repair center, and any vehicle damage or malfunction directly or indirectly caused by such service or repairs may not be covered by, or may void, the vehicle's limited warranty.

Additionally, in [Source 2: Service_Website_Terms_and_Conditions_20131128.pdf], it is mentioned that any vehicle damage or malfunction directly or indirectly caused by, due to, or resulting from service or repairs performed by non-Tesla service providers may not be covered by, or may void, the vehicle's limited warranty.

In [Source 3: Service_Website_Terms_and_Conditions_20131128.pdf], it is stated that certain procedures or content elements may make reference to Tesla Warranty policy or practice, but these policies or practices are only applicable to Tesla Service Centers. Non-Tesla service providers have the responsibility to notify their customers of warrantable service and may make no financial claims to Tesla for performing warrantable service.

In [Source 4: Owners_Manual.pdf], it is mentioned that any damage caused by opening the Battery coolant reservoir is excluded from the warranty.

It is essential to note that the warranty policy may have specific terms and conditions, and it is recommended to consult the official Tesla documentation or contact Tesla directly for the most up-to-date and accurate information regarding the warranty policy.

Total Time

1.87s

Chunks

5

## Source Documents

1. Service_Website_Terms_and... (Score: 0.470)

**Similarity Score:** `0.4696`

**Content:**

Tesla Motors Terms and Conditions for Service Website (Service.TeslaMotors.com)
14. Warranty-Covered Repairs
Service Bulletins and other publications may refer to procedures covered by a Tesla vehicle warranty. It is

2. Service_Website_Terms_and... (Score: 0.453)

3. Service_Website_Terms_and... (Score: 0.442)

4. Owners_Manual.pdf... (Score: 0.331)

# Conclusion

## Summary

✅ Built a complete RAG system for Tesla knowledge base

✅ Hybrid SLM + LLM architecture for optimal performance

✅ 37% reduction in hallucination risk vs Base LLM

✅ Interactive Streamlit frontend with authentication

✅ Comprehensive evaluation framework

## Key Takeaway:

> *"Grounding matters. RAG ensures answers are traceable to actual Tesla documents, making it safer and more reliable for enterprise use."*