Building a (RAG) AI Agent

Intelligent Document Q&A System

Components

Document Processing: Upload and extract text from PDFs, Word docs, etc.

Text Chunking: Break documents into smaller pieces (AI models have token limits)

Embeddings: Convert text to numerical vectors that capture meaning

Vector Storage: Store embeddings in a database for fast retrieval

Query Processing: Take user questions and find relevant document chunks

Answer Generation: Use an AI model to generate answers from relevant context

User Interface: A way for users to interact with the system

The Core Stack

- **LLM (Reasoning Layer)**
The brain of the system — understands the user’s question and composes answers strictly from retrieved context.

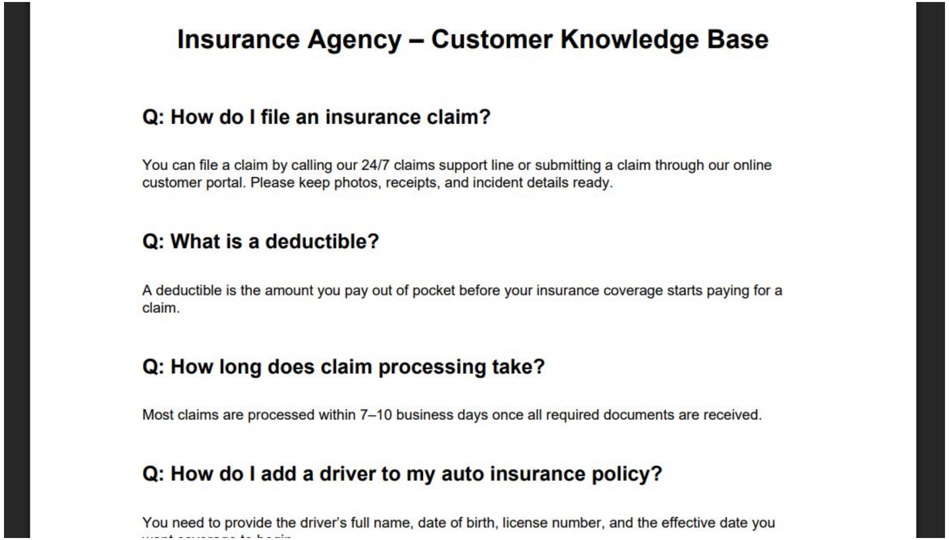
- **Chunking (Context Control)**
Breaks large documents into overlapping, token-aware pieces so meaning isn’t lost during retrieval.

- **Embeddings (Semantic Representation)**
Transforms text into numerical meaning vectors, enabling similarity search beyond keyword matching.

- **Vector Database (FAISS)**
Stores embeddings and retrieves the most relevant document chunks at high speed during queries.

Data Preparation

To bootstrap the system, I generated a **domain-style FAQ PDF** that mirrors how organizations actually document knowledge (clear questions, concise answers, minimal ambiguity).
The focus here is not content accuracy, but proving the end-to-end learning and retrieval flow works correctly.



Insurance Agency – Customer Knowledge Base

Q: How do I file an insurance claim?

You can file a claim by calling our 24/7 claims support line or submitting a claim through our online customer portal. Please keep photos, receipts, and incident details ready.

Q: What is a deductible?

A deductible is the amount you pay out of pocket before your insurance coverage starts paying for a claim.

Q: How long does claim processing take?

Most claims are processed within 7–10 business days once all required documents are received.

Q: How do I add a driver to my auto insurance policy?

You need to provide the driver's full name, date of birth, license number, and the effective date you

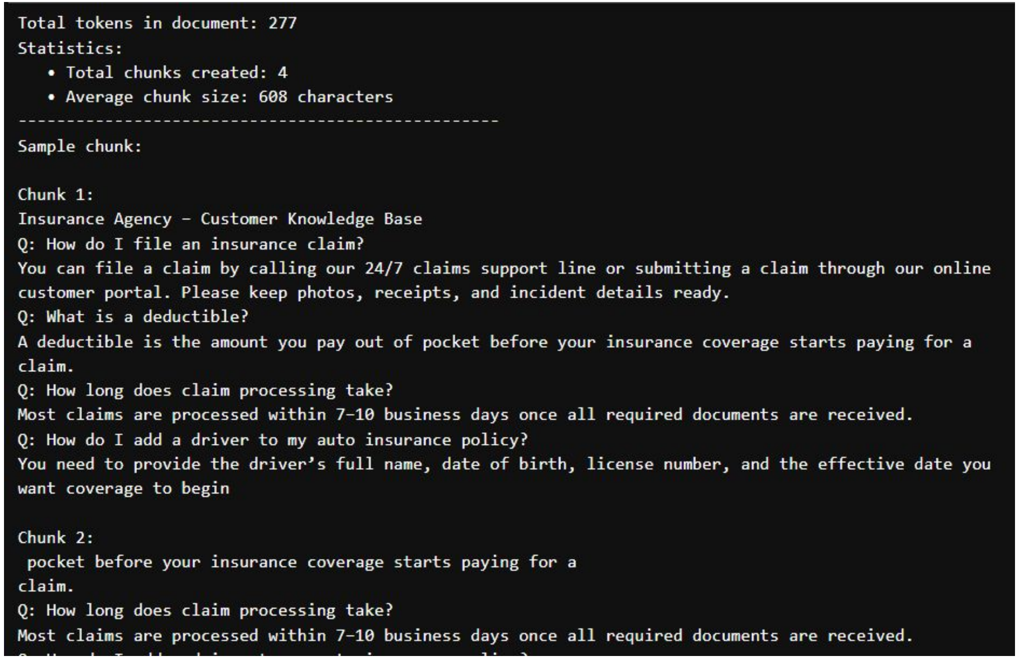
Chunking (The Most Critical RAG Step)

Large chunks blur unrelated topics and confuse similarity search. Tiny chunks lose context and break complete ideas.

To balance this, I used **token-based chunking with overlap**:

- Tokens ensure chunks align with how LLMs actually read text
- Overlap preserves sentence and idea continuity across boundaries

This prevents important information from being cut mid-thought while keeping retrieval precise and reliable.



```
Total tokens in document: 277
Statistics:
   • Total chunks created: 4
   • Average chunk size: 608 characters
--------------------------------------------------
Sample chunk:

Chunk 1:
Insurance Agency - Customer Knowledge Base
Q: How do I file an insurance claim?
You can file a claim by calling our 24/7 claims support line or submitting a claim through our online
customer portal. Please keep photos, receipts, and incident details ready.
Q: What is a deductible?
A deductible is the amount you pay out of pocket before your insurance coverage starts paying for a
claim.
Q: How long does claim processing take?
Most claims are processed within 7-10 business days once all required documents are received.
Q: How do I add a driver to my auto insurance policy?
You need to provide the driver’s full name, date of birth, license number, and the effective date you
want coverage to begin

Chunk 2:
 pocket before your insurance coverage starts paying for a
claim.
Q: How long does claim processing take?
Most claims are processed within 7-10 business days once all required documents are received.
```

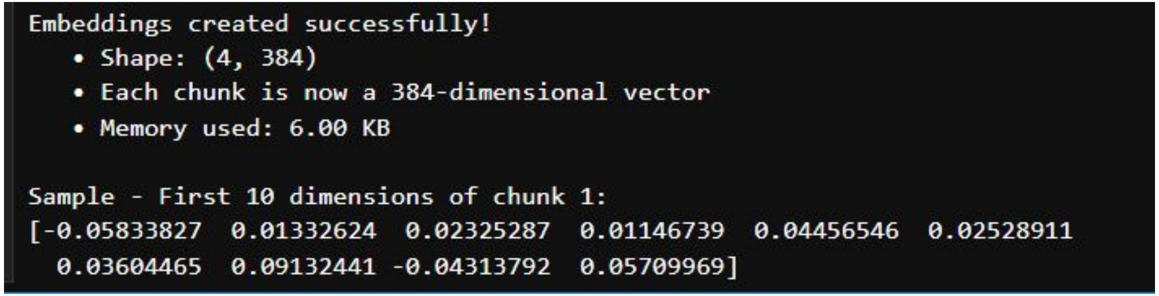
Embeddings (Turning Text into Searchable Numbers)

Embeddings are what make *semantic search* possible.

Each document chunk is converted into a numerical representation that captures **meaning, not keywords**.
These vectors are generated once, stored, and reused — keeping query-time retrieval fast and consistent.

This allows the system to:

- Match questions to relevant content even when wording differs
- Retrieve context based on intent, not exact phrases



```
Embeddings created successfully!
   • Shape: (4, 384)
   • Each chunk is now a 384-dimensional vector
   • Memory used: 6.00 KB

Sample - First 10 dimensions of chunk 1:
[-0.05833827  0.01332624  0.02325287  0.01146739  0.04456546  0.02528911
  0.03604465  0.09132441 -0.04313792  0.05709969]
```

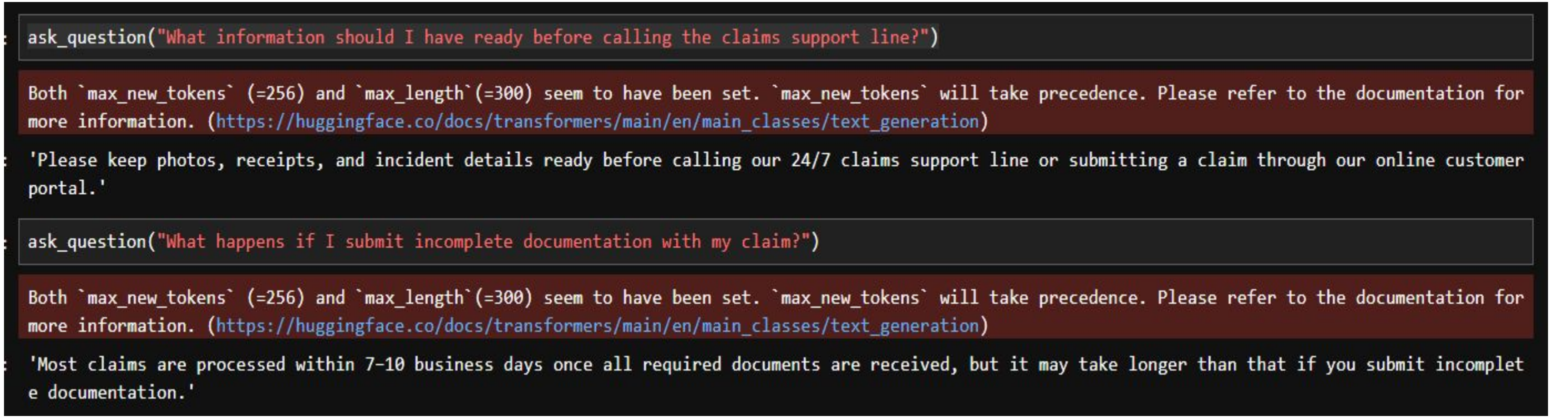
Retrieval (Finding best answers for user query)

At query time, the user’s question is embedded and matched against the vector database to retrieve the **most semantically relevant document chunks**.

These retrieved chunks are then injected as context for the LLM, ensuring responses are:

- Grounded in the source document
- Context-aware rather than speculative

The model isn’t asked to *invent* answers — it’s asked to **reason strictly over retrieved knowledge**. This is what turns a generic chatbot into a reliable document intelligence system.



```
ask_question("What information should I have ready before calling the claims support line?")
```

```
Both `max_new_tokens` (=256) and `max_length`(=300) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for
more information. (https://huggingface.co/docs/transformers/main/en/main_classes/text_generation)
'Please keep photos, receipts, and incident details ready before calling our 24/7 claims support line or submitting a claim through our online customer
portal.'
```

```
ask_question("What happens if I submit incomplete documentation with my claim?")
```

```
Both `max_new_tokens` (=256) and `max_length`(=300) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for
more information. (https://huggingface.co/docs/transformers/main/en/main_classes/text_generation)
'Most claims are processed within 7-10 business days once all required documents are received, but it may take longer than that if you submit incomplet
e documentation.'
```