

```
embedding_fn.py >  embedding_fn
                                   def embedding_fn():
> __pycache__
> .pytest_cache

∨ chroma

 ∨ f727b2d2-b3ed-4b70-ba73...
  ≡ data_level0.bin

    Iength.bin

    Iink_lists.bin

✓ data

                                       #use netstat cmd onterminal
 Second Amended Credit Ag...
> venv
                              25
                                       # try pulling it first"}
gitignore
embedding_fn.py
q_n_a.py
                                       embeddings = OllamaEmbeddings(model="nomic-embed-text")

 README.md

                                       return embeddings
test_rag_response.py
vectorize_docs.py
```

∨ RAG-LANGCHAIN-LOCALLLM-TEST

```
∨ RAG-LANGCHAIN-LOCALLLM-TEST
                                 vectorize_docs.py > \( \phi \) calculate_chunk_ids
                                       import argparse
  > __pycache__
                                       import os
 > .pytest_cache
                                       import shutil
 ∨ chroma
                                       from langchain.document_loaders.pdf import PyPDFDirectoryLoader

√ f727b2d2-b3ed-4b70-ba73...

                                       from langchain_text_splitters import RecursiveCharacterTextSplitter

    data_level0.bin

                                       from langchain.schema.document import Document
                                       from embedding_fn import embedding_fn
    from langchain.vectorstores.chroma import Chroma

    Iength.bin

    CHROMA_PATH = "chroma"
   DATA_PATH = "data"

✓ data

                                       def main():
  Second Amended Credit Ag...
 > venv
 gitignore
                                           parser = argparse.ArgumentParser()
 embedding_fn.py
                                           parser.add_argument("--reset", action="store_true", help="Reset the database.")
                                           args = parser.parse_args()
 q_n_a.py
                                           if args.reset:
 ① README.md
                                               print("** Clearing Database")
 ≡ requirements.txt
                                               clear_database()
 test_rag_response.py
 vectorize_docs.py
                                           documents = load_documents_from_dir()
                                           print(documents[84])
                                           # Wild Oats Markets Inc. credit agreement with wells fargo and other banks - 2nd amended restated
                                           chunks = split_documents(documents)
                                           print(f"** Printing First Chunk for Reference: {chunks[0]}")
                                           add_to_chroma(chunks)
                                       def load_documents_from_dir():
                                           # https://python.langchain.com/docs/modules/data_connection/document_loaders/
                                           document_loader = PyPDFDirectoryLoader(DATA_PATH)
                                           return document_loader.load()
> OUTLINE
> TIMELINE
                                       def split documents(documents: list[Document]):
```

```
∨ RAG-LANGCHAIN-LOCALLLM-TEST
                                 vectorize_docs.py > \( \frac{1}{2} \) calculate_chunk_ids
                                        def split_documents(documents: list[Document]):
  > __pycache__
                                                chunk_overlap=80, # 100 from 80
  > .pytest_cache
                                                length_function=len,

∨ chroma

                                                is_separator_regex=False,
  f727b2d2-b3ed-4b70-ba73...
                                            return text_splitter.split_documents(documents)

≡ data_level0.bin

    ■ length.bin
                                        def add_to_chroma(chunks: list[Document]):

    Iink_lists.bin

                                            # Load the existing database.
  db = Chroma(
                                                persist_directory=CHROMA_PATH, embedding_function=embedding_fn()

∨ data

  Second Amended Credit Ag...
  > venv
                                            # Calculate Page IDs.
 .gitignore
                                            chunks_with_ids = calculate_chunk_ids(chunks)
 embedding_fn.py
                                            # Add or Update the documents.
 q_n_a.py
                                            existing_items = db.get(include=[]) # IDs are always included by default

 README.md

                                            existing_ids = set(existing_items["ids"])
 ≡ requirements.txt
                                            print(f"Number of existing documents in DB: {len(existing_ids)}")
 test_rag_response.py
                                            # Only add documents that don't exist in the DB.
 vectorize_docs.py
                                            new_chunks = []
                                            for chunk in chunks_with_ids:
                                                if chunk.metadata["id"] not in existing_ids:
                                                    new_chunks.append(chunk)
                                            if len(new_chunks):
                                                print(f"* Adding new documents: {len(new_chunks)}")
                                                new_chunk_ids = [chunk.metadata["id"] for chunk in new_chunks]
                                                db.add_documents(new_chunks, ids=new_chunk_ids)
                                                db.persist()
                                            else:
                                                print("* No new documents to add")
                                        def calculate_chunk_ids(chunks):
> OUTLINE
                                            # This will create IDs like "data/<pdf_name>:6:2"
> TIMELINE
                                            # Page Source : Page Number : Chunk Index
```

```
∨ RAG-LANGCHAIN-LOCALLLM-TEST

                                  🕏 vectorize_docs.py > 😭 calculate_chunk_ids
  > __pycache__
  > .pytest_cache
                                        def calculate_chunk_ids(chunks):

∨ chroma

√ f727b2d2-b3ed-4b70-ba73...

                                            # This will create IDs like "data/<pdf_name>:6:2"
                                            # Page Source : Page Number : Chunk Index
    ≡ data_level0.bin
    last_page_id = None

    Iength.bin

                                            current_chunk_index = 0

    Iink_lists.bin

   for chunk in chunks:
                                                source = chunk.metadata.get("source")

✓ data

                                                page = chunk.metadata.get("page")
  Second Amended Credit Ag...
                                                current_page_id = f"{source}:{page}"
  > venv
 .gitignore
                                                # If the page ID is the same as the last one, increment the index.
 embedding_fn.py
                                                if current_page_id == last_page_id:
                                                    current_chunk_index += 1
 q_n_a.py
                                                else:
 ① README.md
                                                    current_chunk_index = 0
 ≡ requirements.txt
                                  98
 test_rag_response.py
                                                # Calculate the chunk ID.
                                                chunk_id = f"{current_page_id}:{current_chunk_index}"
                                 100
 vectorize_docs.py
                                                last_page_id = current_page_id
                                                # Add it to the page meta-data.
                                                chunk.metadata["id"] = chunk_id
                                 104
                                            return chunks
                                        def clear_database():
                                 110
                                            if os.path.exists(CHROMA_PATH):
                                                shutil.rmtree(CHROMA_PATH)
                                 112
                                 113
                                        if __name__ == "__main__":
                                 115
                                            main()
                                 116
> OUTLINE
> TIMELINE
```

```
∨ RAG-LANGCHAIN-LOCALLLM-TEST

                                 q_n_a.py > ...
                                       import argparse
 > __pycache__
                                       from langchain.vectorstores.chroma import Chroma
 > .pytest_cache
                                       from langchain.prompts import ChatPromptTemplate

∨ chroma

                                   4
                                       from langchain_community.llms.ollama import Ollama
  ∨ f727b2d2-b3ed-4b70-ba73...
                                       from embedding_fn import embedding_fn

    data level0.bin

                                       CHROMA_PATH = "chroma"

    Iength.bin

                                       PROMPT_TEMPLATE = """

    Iink_lists.bin

                                       Given the following context:

∨ data

  Second Amended Credit Ag...
 > venv
 gitignore
                                       Answer the question based on only the above context: {question}
 embedding_fn.py
 q_n_a.py

    README.md

                                       def main():
 ≡ requirements.txt
                                            # Create CLI.
 test_rag_response.py
                                            parser = argparse.ArgumentParser()
                                            parser.add_argument("query_text", type=str, help="The query text.")
 vectorize_docs.py
                                            args = parser.parse_args()
                                            query_text = args.query_text
                                            query_rag(query_text)
                                       def query_rag(query_text: str):
                                            # Prepare the DB.
                                            embedding_function = embedding_fn()
                                            db = Chroma(persist_directory=CHROMA_PATH, embedding_function=embedding_function)
                                            # Search the DB.
                                            results = db.similarity_search_with_score(query_text, k=5) # k nearest neighbours
                                            context_text = "\n\n---\n\n".join([doc.page_content for doc, _score in results])
                                            prompt_template = ChatPromptTemplate.from_template(PROMPT_TEMPLATE)
                                            prompt = prompt_template.format(context=context_text, question=query_text)
> OUTLINE
                                            print(f" **Prompt is: {prompt}")
```

> TIMELINE

```
∨ RAG-LANGCHAIN-LOCALLLM-TEST
                             🕏 q_n_a.py > ...
                                  def query_rag(query_text: str):
 > __pycache__
 > .pytest_cache

∨ chroma

                                      model = Ollama(model="llama2") # select the model
  ∨ f727b2d2-b3ed-4b70-ba73...
                                      response_text = model.invoke(prompt) # run the model with prompt

≡ data_level0.bin

    header.bin

                                      sources = [doc.metadata.get("id", None) for doc, _score in results] # k nearest neighbours source chunks

    Iength.bin

   ≣ link_lists.bin
                                      formatted_response = f"** Response: {response_text}\n\n** Sources: {sources}"
  print(formatted_response)
                                      print(" ")

✓ data

                                      Second Amended Credit Ag...
 > venv
                                      return response_text
 gitignore
 embedding_fn.py
                                  if __name__ == "__main__":
 🕏 q_n_a.py
                                      main()
 README.md
 ≡ requirements.txt
 test_rag_response.py
 vectorize_docs.py
```

```
# It only mentions that the Administrative Agent acts on behalf of the Lenders and that Participants, who are different from the Lenders,
# have certain rights related to payments and setoffs.
# Llama2 Response: Based on the context provided, the names of the Lenders in the WILD OATS MARKETS credit agreement are:
# 1. Wells Fargo Bank, National Association
# 2. U.S. Bank National Association
# Expected Answer: Wells Fargo Bank, National Association AND U.S. Bank National Association AND Vectra Bank Colorado N.A. from SCHEDULE I
# sample query = What are the names of the borrowers in the WILD OATS MARKETS credit agreement? Give the names only.
# Mistral Response: The name of the borrower in the WILD OATS MARKETS credit agreement is Wild Oats Markets, Inc.
# Llama2 Response: * Wild Oats Markets, Inc.
# Expected Answer: Wild Oats Markets, Inc.
# sample query = From 'SCHEDULE I', List the names of the Lenders in the WILD OATS MARKETS credit agreement? Give the names only.
# Mistral Response: The context provided does not directly contain a list of the Lenders' names from 'SCHEDULE I'.
# However, it mentions that each Lender has an office designated as their Eurodollar Lending Office in Schedule I.
# Llama2 Response: Based on the context provided, the names of the Lenders in the WILD OATS MARKETS credit agreement are:
# 1. Wells Fargo Bank, National Association
# Expected Answer: Wells Fargo Bank, National Association AND U.S. Bank National Association AND Vectra Bank Colorado N.A. from SCHEDULE I
# sample query = What is the arbitration amount beyond which 3 arbitrators are required? Give the amount only.
# Mistral Response: The arbitration proceeding requires three arbitrators when the amount in controversy exceeds $5,000,000.00.
# Llama2 Response: Based on the provided context, the answer to the question is: $10,000,000. --- but unit test RUN gave correct answer
# Expected Answer: $5,000,000.00
# sample query = Who is the Chief Executive Officer of Wild Oats? Give the name only.
  as well as some details regarding loan agreements and repurchase rights.
# Llama2 Response: Perry D. Odak
# Sources: ['data/Second Amended Credit Agreement WildOats Wells.pdf:86:1', 'data/Second Amended Credit Agreement WildOats Wells.pdf:18:1',
  'data/Second Amended Credit Agreement WildOats Wells.pdf:70:3', 'data/Second Amended Credit Agreement WildOats Wells.pdf:64:4',
  'data/Second Amended Credit Agreement WildOats Wells.pdf:52:0'] --- correct answer in page 64 4th chunk
# Expected Answer: Perry Odak
```

sample query = What are the names of the Lenders in the WILD OATS MARKETS credit agreement? Give the names only.
Mistral Response: The text does not provide the names of the specific Lenders in the WILD OATS MARKETS credit agreement.

```
∨ RAG-LANGCHAIN-LOCALLLM-TEST
                                 test_rag_response.py >  test_arbitration_amount
                                        from q_n_a import query_rag
  > __pycache__
                                        from langchain_community.llms.ollama import Ollama
  > .pytest_cache

∨ chroma

                                        EVAL_PROMPT = """

√ f727b2d2-b3ed-4b70-ba73...

                                        Expected Response: {expected_response}
                                        Actual Response: {actual_response}

    data_level0.bin

    (Answer with 'true' or 'false') Does the actual response match the expected response?

    Iength.bin

    ≣ link_lists.bin

≡ chroma.sqlite3

√ data

                                        def test_ceo_name():
                                            assert query_and_validate(
  Second Amended Credit Ag...
                                                question="Who is the Chief Executive Officer of Wild Oats? Give the name only.",
 > venv
                                                expected_response="Perry Odak",
 gitignore
 embedding_fn.py
 q_n_a.py
                                        def test_arbitration_amount():
 ① README.md
                                            assert query_and_validate(

    □ requirements.txt

                                                question="What is the arbitration amount beyond which 3 arbitrators are required? Give the amount only.",
 test_rag_response.py
                                                expected_response="$5,000,000.00",
 vectorize_docs.py
                                        def query_and_validate(question: str, expected_response: str):
                                            response_text = query_rag(question)
                                            prompt = EVAL_PROMPT.format(
                                                expected_response=expected_response, actual_response=response_text
                                            #model = Ollama(model="mistral")
                                            model = Ollama(model="llama2")
                                            evaluation_results_str = model.invoke(prompt)
                                            evaluation_results_str_cleaned = evaluation_results_str.strip().lower()
                                            print("pytest -s")
                                            print(prompt)
> OUTLINE
> TIMELINE
                                            if "true" in evaluation results str cleaned:
```

```
∨ RAG-LANGCHAIN-LOCALLLM-TEST
                                    test_rag_response.py > \(\operatorname{\textit{?}}\) test_arbitration_amount
                                          def query_and_validate(question: str, expected_response: str):
  > __pycache__
  > .pytest_cache
                                               if "true" in evaluation_results_str_cleaned:

∨ chroma

   ∨ f727b2d2-b3ed-4b70-ba73...
                                                   print("\033[92m" + f"Response: {evaluation_results_str_cleaned}" + "\033[0m")

≡ data_level0.bin

                                                   return True
                                               elif "false" in evaluation_results_str_cleaned:

    header.bin

    Iength.bin

                                                   print("\033[91m" + f"Response: {evaluation_results_str_cleaned}" + "\033[0m")

    Iink_lists.bin

                                                   return False
   ≡ chroma.sqlite3
                                               else:
                                                   raise ValueError(

✓ data

                                                       f"Invalid evaluation result. Cannot determine if 'true' or 'false'."
  Second Amended Credit Ag...
  > venv
 gitignore
 embedding_fn.py
 🕏 q_n_a.py
 ① README.md

≡ requirements.txt

 test_rag_response.py
 vectorize_docs.py
```

Exity, State:
ABA Number:
Account Number:
Acco

** Sources: ['data/Second Amended Credit Agreement WildOats Wells.pdf:86:1', 'data/Second Amended Credit Agreement WildOats Wells.pdf:6:2', 'data/Second Amended Credit Agreement WildOats Wells.pdf:41:0']

• (venv) (base) dechamma@192 rag-langchain-localllm-test % python3 q_n_a.py "What are the names of the borrowers in the WILD DATS MARKETS credit agreement? Give the names only."

**Prompt is: Human: Given the following context:

== test session starts === platform darwin —— Python 3.12.3, pytest—8.1.1, pluggy—1.5.0 rootdir: /Users/dechamma/Documents/rag-langchain-localllm-test plugins: anyio-4.3.0 collected 2 items test_rag_response.py **Prompt is: Human: Given the following context: Denver, Colorado 102003154 6966001586 City, State: ABA Number: Account Number: Ref: Wild Oats Markets, Inc. that all payments received at a time not otherwise permitted shall be turned over to the Administrative Agent on behalf of the Lenders and applied to that all payments received at a time not otherwise permitted such that the turned over to the Administrative Agent on behalf of the Lenders and applied to reduce the outstanding Obligations) and (iii) customary bankruptcy provisions.

"Person " shall mean and include an individual, a partnership, a corporation (including a business trust), a joint stock company, an unincorporated association, a limited liability company, a joint venture, a trust or other entity or a Governmental Authority.

"Pledge Agreement " shall mean the Amended and Restated Pledge Agreement to be initially entered into among the Administrative Agent and the Borrower and Wild Oats Financial, Inc., substantially in the form of Exhibit H, as the same may be from time to time be supplemented, modified, of this Agreement, the outstanding principal amount of the Executive Note (plus any interest accrued thereon), so long as (A) the consideration given by the Borrower for such shares consists only of a non-cash char ge related to the for giveness of all or any portion of the indebtedness arising under the Executive Note or (B) such repurchase is otherwise permitted by clause (v) below; and

(v) In addition to the foregoing, the Borrower may purchase shares of its capital stock from any shareholder (including Perry D. Odak) with cash and Permitted Subordinated Indebtedness in an amount not to exceed \$5,000,000 during any Fiscal Year, provided that after giving ef fect to above or clause (x) below, provided that (A) the principal amount of any such refinancing does not exceed the principal amount of the Indebtedness being refinanced and (B) the material terms and provisions of any such refinancing (including maturity, redemption, prepayment, default and subordination provisions) are no less favorable to the Lenders than the Indebtedness being refinanced;

(vii) Indebtedness of the Loan Parties with respect to workers' compensation claims, self-insurance obligations, Permitted Insurance Premium Indebtedness, surety, appeal, indemnity, performance, or other similar bonds in the ordinary course of business (including surety or similar bonds issued in connection with the stay of a proceeding of the type described in Section 6.01(h)); Answer the question based on only the above context: Who is the Chief Executive Officer of Wild Oats? Give the name only. $\star\!\star$ Response: Based solely on the provided context, the answer to the question is: Perry D. Odak ** Sources: ['data/Second Amended Credit Agreement WildOats Wells.pdf:86:1', 'data/Second Amended Credit Agreement WildOats Wells.pdf:70:3', 'data/Second Amended Credit Agreement WildOats Wells.pdf:64:4']

(venv) (base) dechamma@192 rag-langchain-localllm-test % pytest -s

(Answer with 'true' or 'false') Does the actual response match the expected response?

	**P	rompt	is:	Huma	an:	
Gi	iven	the	follo	wina	context:	

(d) Arbitrator Qualifications and Powers. Any arbitration proceeding in which the amount in controversy is \$5,000,000.00 or less will be decided by a single arbitrator selected accordin g to the Rules, and who shall not render an awar er an award of greater than \$5,000,000.00. Any dispute in which the amount in controversy exceeds \$5,000,000.00 shall be decided by majority vote of a panel of three arbitrat ors; provided however, that all hearings and deliberations. The arbitrator will be a neutral attorney licensed in the State of Colorado or a neutral retired judge of the state or federal judiciary of

which case the arbitration shall be conducted in accordance with the AAA's optional procedures for large, complex commercial disputes (the commercial dispute resolution procedures or the o ptional procedures for large, complex commercial disputes to be referred to, as applicable, as the "Rules"). If there is any inconsistency between the terms hereof and the Rules, the terms and procedures set for th herein shall control. Any party who fails or refuses to submit to arbitration following a demand by any other party shall bear all costs and expenses incurred by such other party in compelling arbitration of any di spute. Nothing contained herein shall be a waiver by any party that is a bank of the protections afforded to it under 12 U.S.C. 91 or any similar applicable state law.

subject to final determination by the arbitrator upon a showing that the request for discovery is essential for the party's presentation and that no alternative means for obtaining informat ion is available.

ion is available.

(f) Class Proceedings and Consolidations. The resolution of the r Class Proceedings and Consolidations. The resolution of any dispute arising pursuant to the terms of this Agreement shall be determined by a separate arbitration proceedi

Miscellaneous. To the maximum extent practicable, the AAA, the arbitrators and the parties shall take all action required to conclude any arbitration proceeding within 180 days of the filing of the dispute
with the AAA. No arbitrator or other party to an arbitration proceeding may disclose the existence, content or results thereof, except for disclosures of information by a party required in the ordinary course of its
business or by applicable law or regulation. If more than one agreement for arbitration by or between the parties potentially applies to a dispute, the arbitration provision most directly related to the Credit Documents

entitled to receive an amount in excess of all amounts previously paid by the Borrower pursuant to this Section 2.12, to such Person (less the aggregate amount of all prior payments by such Person to the Borrower under this Section 2.12(d)) (but any excess amount described in this proviso shall reduce pro tanto any amount that the Borrower is subsequently obligated to pay to such Person pursuant to this Section 2.12), and provided further that any amount otherwise due to the Borrower pursuant to this sentence shall not be paid to the extent of any payment or indemnity then due and owing from the Borrower to such Person (and any amount not so paid shall become payable at such time and to such extent as the Borrower shall

Answer the question based on only the above context: What is the arbitration amount beyond which 3 arbitrators are required? Give the amount only.

Expected Response: \$5,000,000.00
Actual Response: Based on the provided context, the arbitration amount beyond which 3 arbitrators are required is \$5,000,000.00.

(Answer with 'true' or 'false') Does the actual response match the expected response?

Response: true. the actual response of "based on the provided context, the arbitration amount beyond which 3 arbitrators are required is \$5,000,000.00" matches the expected response of "\$5,000,000.00".

Output

Output

Description

**Des

** Sources: ['data/Second Amended Credit Agreement WildOats Wells.pdf:82:1', 'data/Second Amended Credit Agreement WildOats Wells.pdf:81:12', 'data/Second Amended Credit Agreement WildOats Wells.pdf:82:5', 'data/Second Amended Credit Agreement WildOats Wells.pdf:82:6', 'data/Second Amended Credit Agreement WildOats Wells.pdf:48:0']

** Response: Based on the provided context, the arbitration amount beyond which 3 arbitrators are required is \$5,000,000.00.