

# Helpmate AI using Qdrant DB

## Problem statement

## Project goal

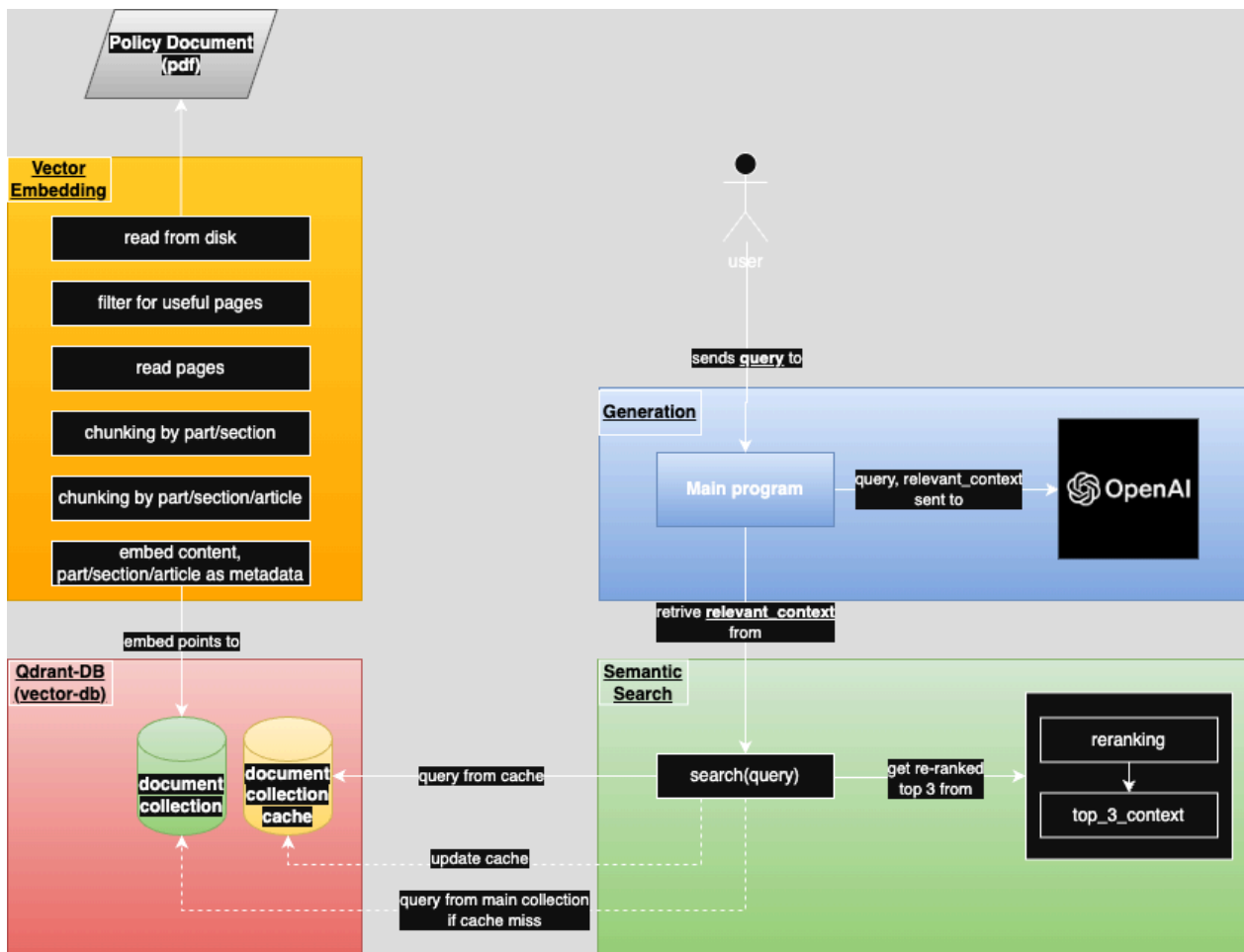
The life insurance policy documents are very elaborate and lengthy. It's a challenge to go through the entire document at once and understand it well. This project provides an easy way to understand the document by asking questions and queries to the RAG based application.

## Project setup

- Install poetry for dependency management
- Install pyenv for python version management
- Install docker and docker-compose to start qdrant db locally
- run `docker-compose up -d` to start qdrant db, then view the qdrant dashboard at <http://localhost:6333/dashboard>
- run `poetry install` to install the dependencies
- run `poetry shell` to activate the virtual environment
- run `jupyter notebook` to start the jupyter notebook server

# Solution Architecture

## Architecture diagram



# Architecture description & Design choices

## Vector Embedding Pipeline

### 1. Read from disk

Load the PDF policy document.

### 2. Filter useful pages

Retain only pages with meaningful content (those which are within a part and section).

### 3. Read pages

Extract text content using a PDF parser (using `pdfplumber`).

### 4. Chunk by part/section

Split document hierarchically to maintain structure.

### 5. Chunk by part/section/article

Finer-level segmentation for better retrieval precision.

### 6. Embed content

Use sentence embedding model ( `all-MiniLM-L6-v2` ).

### 7. Add metadata

Attach `part` , `section` , and `article` as payload.

---

## Qdrant Vector DB

### 8. Create `document collection`

Stores all document chunks with embeddings and metadata.

### 9. Create `document collection cache`

Stores previously seen query results for reuse.

### 10. Embed points

Vectorized chunks are inserted into the main collection.

---

## Semantic Search & Caching

### 11. `search(query)`

- Checks `cache` collection first.
- If found → return cached result.
- If not → search main `document collection` .

### 12. Update cache

Stores new query + result in cache for future reuse.

**13. Retrieve `relevant_context`**

Search result passed back to main app for response generation.

---

## Re-ranking

**14. Re-rank results**

Use a cross-encoder (e.g., `ms-marco-MiniLM` ) to refine ranking.

**15. Select top 3**

Final `top_3_context` chosen for LLM prompt.

---

## Generation Flow

**16. User sends query**

A natural language question is submitted.

**17. Main program dispatches search**

Sends query to semantic search module.

**18. Send to OpenAI**

Passes `query` + `relevant_context` to OpenAI for generation.

---

# Solution Implementation

## Imports

import all the required libraries and modules

## Embedding Layer

### Text Processing

- **is\_useful\_page**: Filters out pages that are not relevant.
- **get\_part**: Extracts the part from the text.
- **get\_section**: Extracts the section from the text.
- **process\_document**: Reads the PDF and organizes the text into a structured format.

### Chunking

- **is\_not\_same**: Compares two strings to check if they are not the same (ignoring case and spaces).
- **get\_chunks**: Processes the structured data to create chunks of text, each associated with a part, section, and article.

### Embedding Logic

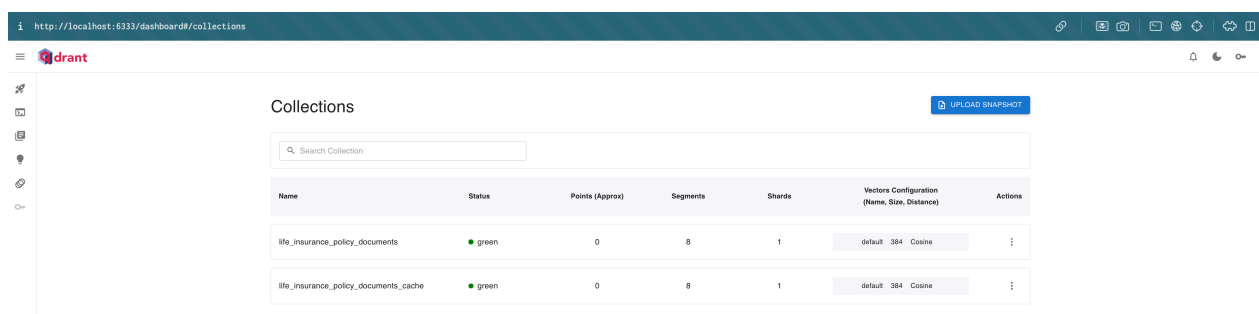
- **get\_embeddable\_points**: Converts the chunks into a format suitable for embedding, including vector representation and metadata.

### Prepare Vector DB

- **QdrantClient**: Connects to the Qdrant database.
- **delete\_collection**: Deletes any existing collections to start fresh.
- **create\_collection**: Creates a new collection for storing the document chunks.

### output

- **empty\_collections.png**: Shows the empty collections in Qdrant.

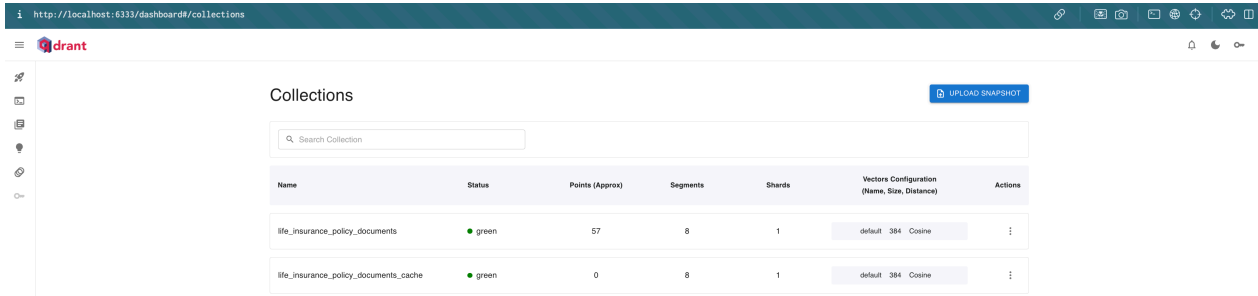


## Embedding the chunks

- **upsert:** Inserts the chunks into the Qdrant collection.

## output

- **filled\_collection.png:** Shows the filled collection in Qdrant.



The screenshot shows the Qdrant dashboard at <http://localhost:6333/dashboard/collections>. The dashboard displays a table of collections. The first collection, 'life\_insurance\_policy\_documents', has a status of 'green', approximately 57 points, 8 segments, and 1 shard. The second collection, 'life\_insurance\_policy\_documents\_cache', also has a status of 'green', 0 points, 8 segments, and 1 shard. Both collections use a 'default' vector configuration with a size of 384 and a cosine distance metric. An 'UPLOAD SNAPSHOT' button is visible in the top right corner of the dashboard.

Name	Status	Points (Approx)	Segments	Shards	Vectors Configuration (Name, Size, Distance)	Actions
life_insurance_policy_documents	green	57	8	1	default 384 Cosine	
life_insurance_policy_documents_cache	green	0	8	1	default 384 Cosine	

## Search Layer

### Search & Cache

- **hash\_query**: Generates a hash for the query string.
- **save\_to\_cache**: Saves the query and its results to the cache collection.
- **query\_cache\_collection**: Queries the cache collection for previously seen queries.
- **query\_collection**: Queries the main collection for the given query.
- **search**: Main function that checks the cache first, then queries the main collection if not found.

### search test

- **search:** Searches for the query in the Qdrant collection.
- Displays a **Cache miss!**

Cache miss!



```
[
{
  "id": 1,
  "version": 0,
  "score": 0.5939084,
  "part": "PART II - POLICY ADMINISTRATION",
  "section": "Section A - Contract",
  "article": "Article 2 - Policy Changes",
  "content": "Insurance under this Group Policy runs annually to the Policy Anniversary, unless so
oner terminated. No agent, employee, or person other than an officer of The Principal has authorit
y to change this Group Policy, and, to be effective, all such changes must be in Writing and Signe
d by an officer of The Principal. The Principal reserves the right to change this Group Policy as foll
ows: a. Any or all provisions of this Group Policy may be amended or changed at any time, includi
ng retroactive changes, to the extent necessary to meet the requirements of any law or any regula
tion issued by any governmental agency to which this Group Policy is subject. b. Any or all provisi
ons of this Group Policy may be amended or changed at any time when The Principal determines t
hat such amendment is required for consistent application of policy provisions. c. By Written agre
ement between The Principal and the Policyholder, this Group Policy may be amended or changed
at any time as to any of its provisions. Any change to this Group Policy, including, but not limited t
o, those in regard to coverage, benefits, and participation privileges, may be made without the co
nsent of any Member or Dependent. Payment of premium beyond the effective date of the change
constitutes the Policyholder's consent to the change. ",
  "text_length": 1308
},
{
  "id": 9,
  "version": 0,
  "score": 0.47650385,
  "part": "PART II - POLICY ADMINISTRATION",
  "section": "Section A - Contract",
  "article": "Article 10 - Policy Interpretation",
  "content": "T he Principal has complete discretion to construe or interpret the provisions of this
group insurance policy, to determine eligibility for benefits, and to determine the type and extent
of benefits, if any, to be provided. The decisions of The Principal in such matters shall be controlli
ng, binding and final as between The Principal and persons covered by this Group Policy, subject t
o the Claims Procedures in PART IV, Section D. ",
  "text_length": 434
},
{
  "id": 5,
  "version": 0,
  "score": 0.46555924,
  "part": "PART II - POLICY ADMINISTRATION",
  "section": "Section A - Contract",
  "article": "Article 6 - Information to be Furnished",
```

"content": "The Policyholder must, upon request, give The Principal all information needed to administer this Group Policy. If a clerical error is found in this information, The Principal may at any time adjust premium to reflect the facts. An error will not invalidate insurance that would otherwise be in force. Neither will an error continue insurance that would otherwise be terminated. The Principal may inspect, at any reasonable time, all Policyholder records, which relate to this Group Policy. ",

"text\_length": 491

}

]

### cache test

- **search:** Searches for the query in the Qdrant collection.
- Displays a **Cache hit!**

Cache hit!

```
[
{
  "id": 1,
  "version": 0,
  "score": 0.5939084,
  "part": "PART II - POLICY ADMINISTRATION",
  "section": "Section A - Contract",
  "article": "Article 2 - Policy Changes",
  "content": "Insurance under this Group Policy runs annually to the Policy Anniversary, unless so
oner terminated. No agent, employee, or person other than an officer of The Principal has authorit
y to change this Group Policy, and, to be effective, all such changes must be in Writing and Signe
d by an officer of The Principal. The Principal reserves the right to change this Group Policy as foll
ows: a. Any or all provisions of this Group Policy may be amended or changed at any time, includi
ng retroactive changes, to the extent necessary to meet the requirements of any law or any regula
tion issued by any governmental agency to which this Group Policy is subject. b. Any or all provisi
ons of this Group Policy may be amended or changed at any time when The Principal determines t
hat such amendment is required for consistent application of policy provisions. c. By Written agre
ement between The Principal and the Policyholder, this Group Policy may be amended or changed
at any time as to any of its provisions. Any change to this Group Policy, including, but not limited t
o, those in regard to coverage, benefits, and participation privileges, may be made without the co
nsent of any Member or Dependent. Payment of premium beyond the effective date of the change
constitutes the Policyholder's consent to the change. ",
  "text_length": 1308
},
{
  "id": 9,
  "version": 0,
  "score": 0.47650385,
  "part": "PART II - POLICY ADMINISTRATION",
  "section": "Section A - Contract",
  "article": "Article 10 - Policy Interpretation",
  "content": "T he Principal has complete discretion to construe or interpret the provisions of this
group insurance policy, to determine eligibility for benefits, and to determine the type and extent
of benefits, if any, to be provided. The decisions of The Principal in such matters shall be controlli
ng, binding and final as between The Principal and persons covered by this Group Policy, subject t
o the Claims Procedures in PART IV, Section D. ",
  "text_length": 434
},
{
  "id": 5,
  "version": 0,
  "score": 0.46555924,
  "part": "PART II - POLICY ADMINISTRATION",
  "section": "Section A - Contract",
  "article": "Article 6 - Information to be Furnished",
```

"content": "The Policyholder must, upon request, give The Principal all information needed to administer this Group Policy. If a clerical error is found in this information, The Principal may at any time adjust premium to reflect the facts. An error will not invalidate insurance that would otherwise be in force. Neither will an error continue insurance that would otherwise be terminated. The Principal may inspect, at any reasonable time, all Policyholder records, which relate to this Group Policy. ",

"text\_length": 491

}

]

output

- **cache\_test\_output.png**: Shows that the queries are cached

http://localhost:6333/dashboard/collections

drant

COLLECTIONS

Search Collection

Name	Status	Points (Approx)	Segments	Shards	Vectors Configuration (Name, Size, Distance)	Actions
life_insurance_policy_documents	green	57	8	1	default 384 Cosine	
life_insurance_policy_documents_cache	green	1	8	1	default 384 Cosine	

## Re-Ranking

- **get\_df\_from\_points:** Converts the list of points into a DataFrame for easier manipulation.
- **get\_re\_ranked\_results:** Uses a cross-encoder model to re-rank the results based on their relevance to the query.
- **search\_with\_re\_ranking:** Main function that performs the search and re-ranking.

### re-ranking comparison with distance score

- **search:** Searches for the query in the Qdrant collection.
- **get\_re\_ranked\_results:** Re-ranks the results based on their relevance to the query.
- **similaritysearch\_reranking\_comparision\_df:** DataFrame showing the comparison between the original scores and the re-ranking scores.

Cache miss!

	score	re_ranking_scores	id
0	0.548356	4.623034	1
2	0.427931	-5.191046	5
1	0.448610	-6.051162	9

### observation

- re-ranking produces a different result than regular semantic search
- based on regular semantic search, the chosen top 3 chunks are 1, 9 and 5
- based on re-ranking, the chosen top 3 chunks are 1, 5 and 9

## Generation Layer

### openai connection and initial system prompt

- **openai:** Connects to the OpenAI API.
- **get\_chat\_model\_completions:** Sends the messages to the OpenAI API and retrieves the response.
- **get\_insurance\_answers:** Prepares the system prompt and user question, and sends it to the OpenAI API for generation.
- **user\_query:** Main function that takes the user query, performs the search, and generates the response.

# Query Search

## First Query

Who is allowed to make changes to this group policy and how are those changes validated?

Cache miss!

### search

```
[
  {
    "part": "PART II - POLICY ADMINISTRATION",
    "section": "Section A - Contract",
    "article": "Article 2 - Policy Changes"
  },
  {
    "part": "PART II - POLICY ADMINISTRATION",
    "section": "Section A - Contract",
    "article": "Article 10 - Policy Interpretation"
  },
  {
    "part": "PART II - POLICY ADMINISTRATION",
    "section": "Section A - Contract",
    "article": "Article 6 - Information to be Furnished"
  }
]
```

### generation

The Principal has complete discretion to consent to any changes to the group policy. Changes are validated through the Principal's authority as outlined in the policy.

Part	Section	Article	Content Summary
PART II - POLICY ADMINISTRATION	Section A - Contract	Article 2 - Policy Changes	The Principal has complete discretion to consent to any changes to the policy.



## Second Query

When does my dependent life insurance coverage begin and are there any conditions?  
Cache miss!

### search

```
[
{
  "part": "PART III - INDIVIDUAL REQUIREMENTS AND RIGHTS",
  "section": "Section A - Eligibility",
  "article": "Article 3 - Dependent Life Insurance"
},
{
  "part": "PART III - INDIVIDUAL REQUIREMENTS AND RIGHTS",
  "section": "Section B - Effective Dates",
  "article": "Article 3 - Dependent Life Insurance"
},
{
  "part": "PART III - INDIVIDUAL REQUIREMENTS AND RIGHTS",
  "section": "Section C - Individual Terminations",
  "article": "Article 3 - Dependent Life Insurance"
}
]
```

### generation

Your dependent life insurance coverage begins when you are eligible for it, which is available only with your own life insurance coverage. Additionally, there are conditions related to the effective dates of the insurance.

Here is the relevant information:

Part	Section	Article	Content Summary
PART III - INDIVIDUAL REQUIREMENTS AND RIGHTS	Section A - Eligibility	Article 3 - Dependent Life Insurance	A person will be eligible for Dependent Life Insurance only if they have their own life insurance coverage.
PART III - INDIVIDUAL REQUIREMENTS AND RIGHTS	Section B - Effective Dates	Article 3 - Dependent Life Insurance	Dependent Life Insurance is available only with your own life insurance coverage.

In summary, your dependent life insurance coverage begins when you have your own life insurance, and it is subject to the conditions outlined in the policy.

# Third Query

## query

What happens to my life insurance benefit if I become terminally ill?  
Cache miss!

## search

```
[
  {
    "part": "PART IV - BENEFITS",
    "section": "Section A - Member Life Insurance",
    "article": "Article 7 - Accelerated Benefits"
  },
  {
    "part": "PART IV - BENEFITS",
    "section": "Section C - Dependent Life Insurance",
    "article": "Article 2 - Death Benefits Payable"
  },
  {
    "part": "PART IV - BENEFITS",
    "section": "Section A - Member Life Insurance",
    "article": "Article 2 - Death Benefits Payable"
  }
]
```

## generation

If you become terminally ill, you may be eligible for accelerated benefits under your life insurance policy. This means that you can receive a portion of your life insurance benefit while you are still alive, which can help cover medical expenses or other costs associated with your illness.

Here is the relevant information:

Part	Section	Article	Content Summary
PART IV - BENEFITS	Section A - Member Life Insurance	Article 7 - Accelerated Benefits	You may qualify for accelerated benefits if you become terminally ill, allowing you to access part of your life insurance benefit early.

This information indicates that you can access your life insurance benefit if you are diagnosed with a terminal illness.

# Conclusions

## Challenges faced

- Identifying the right chunking strategy for the document.
- Understanding the api for Qdrant and how to use it.
- Caching the points changes the data structure from ScoredPoints to dict.