Final Project Python

Team Members

- **Lý Vĩnh Thuận - 22280092**
- **Nguyễn Nhựt Trường - 22280099**
- **Nguyễn Phạm Anh Văn - 22280104**

Quest 1

LLM integration

Purpose: To translate input text from one language to another.

Config

Including Single Text and Multiple Texts Translation

```
input_language = "english"
output_language = "vietnamese"

#if the input is a file containing text
# input_path = "......."

#split_string to split long text from file
# split_string = "\n\n"

input_text = "Hello, I am Ben"
input_list_1 = ["Hello", "I am Ben", "tôi là người Việt Nam", "Tôi không thể win được vì tôi thiếu lucky!"]
```

Generate API URL with API key to call Gemini model

```
url = "https://generativelanguage.googleapis.com/v1beta/models/gemini-1.5-flash:generateContent?key=GEMINI_API_KEY"
headers = {
    "Content-Type": "application/json"
}
api_key = ""
url = url.replace("GEMINI_API_KEY", api_key)
```

Prompt

system_prompt: "You are a translator that translates text from {input_language} to {output_language}. You receive a string written in {input_language} and solely return the same string in {output_language} without losing any of the original formatting. Your translations should be accurate, aiming not to deviate from the original structure, content, writing style and tone. Consider all input text, if the input text contains any word other than {output_language} then translate to {output_language}, the remaining words in the input text remain unchanged. Input text is: "

```
system_prompt = f"You are a translator that translates text from
{input_language} to {output_language}. You receive a string written in
{input_language} and solely return the same string in
{output_language} without losing any of the original formatting. Your
translations should be accurate, aiming not to deviate from the
original structure, content, writing style and tone. Consider all
input text, if the input text contains any word other than
{output_language} then translate to {output_language}, the remaining
words in the input text remain unchanged. Input text is: "

system_prompt
```

```
'You are a translator that translates text from english to vietnamese.
You receive a string written in english and solely return the same
string in vietnamese without losing any of the original formatting.
Your translations should be accurate, aiming not to deviate from the
original structure, content, writing style and tone. Consider all
input text, if the input text contains any word other than vietnamese
then translate to vietnamese, the remaining words in the input text
remain unchanged. Input text is: '
```

Translation Request

```
import time
import requests

def translate_texts(input_list, system_prompt, url, headers,
max_attempts, retry_gap):

    # init results list and counter
    results = []
    counter = 0

    # iterate through each text in the input list
    for text in input_list:
        payload = {
            "contents": [{
                "parts": [{"text": f"{system_prompt} {text}"}]
            }]
        }
        #try sending the request multiple times if get an error
        for attempt in range(max_attempts):
            try:

                response = requests.post(url, headers = headers, json
= payload)

                if response.status_code == 200:      # check the
response from server, code 200 means the request was successful
                    completion = response.json()
                    print(f"Full response: {completion}")

                    #extract translated text
                    translated_text = (
                        completion.get("candidates", [{}])[0]
                        .get("content", {})
                        .get("parts", [{}])[0]
                        .get("text", "No translation output")
                        .strip()
                    )

                    #add translated text
                    results.append(translated_text)
                    counter += 1
                    print(f"Translated: {counter}/{len(input_list)}")
                    break
                else:
                    # if the response is not code 200 (error)
                    print(f"Error: {response.status_code},
{response.text}")
                    results.append("API Error")
                    break

            # if other errors occur
            except Exception as e:
                print(f"Attempt {attempt + 1} failed with error: {e}")
                if attempt < max_attempts - 1:
                    #time before retry
                    time.sleep(retry_gap)
                    retry_gap *= 1.5
                else:
                    #can not translate
                    print("Max attempts reached. Skipping this text.")
                    results.append("No translation output")

    print("\nTranslation Results:")
    for i, (original, translated) in enumerate(zip(input_list,
results)):
        print(f"{i + 1}. Original: {original}\n   Translated:
```

```
{translated}\n")

    return results

# configuration
max_attempts = 2  # number of retries if an error occurs
retry_gap = 3.0  # waiting time between retries

translate_texts(input_list_1, system_prompt, url, headers,
max_attempts, retry_gap)
```

```
Full response: {'candidates': [{'content': {'parts': [{'text': 'Xin
chào\n'}], 'role': 'model'}, 'finishReason': 'STOP', 'avgLogprobs': -
2.813120469606171e-05}], 'usageMetadata': {'promptTokenCount': 97,
'candidatesTokenCount': 3, 'totalTokenCount': 100}, 'modelVersion':
'gemini-1.5-flash'}
Translated: 1/4
Full response: {'candidates': [{'content': {'parts': [{'text': 'Tôi là
Ben\n'}], 'role': 'model'}, 'finishReason': 'STOP', 'avgLogprobs': -
2.4733806640142575e-05}], 'usageMetadata': {'promptTokenCount': 99,
'candidatesTokenCount': 4, 'totalTokenCount': 103}, 'modelVersion':
'gemini-1.5-flash'}
Translated: 2/4
Full response: {'candidates': [{'content': {'parts': [{'text': 'tôi là
người Việt Nam\n'}], 'role': 'model'}, 'finishReason': 'STOP',
'avgLogprobs': -0.018310656150182087}], 'usageMetadata':
{'promptTokenCount': 101, 'candidatesTokenCount': 6,
'totalTokenCount': 107}, 'modelVersion': 'gemini-1.5-flash'}
Translated: 3/4
Full response: {'candidates': [{'content': {'parts': [{'text': 'Tôi
không thể thắng được vì tôi thiếu may mắn!\n'}], 'role': 'model'},
'finishReason': 'STOP', 'avgLogprobs': -7.948116641879702e-07}],
'usageMetadata': {'promptTokenCount': 106, 'candidatesTokenCount': 12,
'totalTokenCount': 118}, 'modelVersion': 'gemini-1.5-flash'}
Translated: 4/4

Translation Results:
1. Original: Hello
   Translated: Xin chào

2. Original: I am Ben
   Translated: Tôi là Ben

3. Original: tôi là người Việt Nam
   Translated: tôi là người Việt Nam

4. Original: Tôi không thể win được vì tôi thiếu lucky!
   Translated: Tôi không thể thắng được vì tôi thiếu may mắn!
```

```
['Xin chào',
 'Tôi là Ben',
 'tôi là người Việt Nam',
 'Tôi không thể thăńg được vì tôi thiêú may măń!']
```

Question 2: Chatbot Development

Access the Product

To use the product, please visit the following link:

Product on Render

We have an error reporting page available. If you encounter any issues or bugs, feel free to report them, and we will get back to you with a solution. (Database: MongoDB)

You can provide feedback, and we will respond to your queries via your provided email.

Your feedback is valuable to us, and we strive to improve the product continuously.

Deployment Details

When deploying the product, we have configured environment variables on the Render server, which are read through `OS` in the code. For detailed code and configuration, please visit the link below. GitHub

Performance Considerations

Due to the free tier on Render, there can be a slowdown when deploying. To mitigate this, we have implemented a solution using **Better Stack Uptime**. This service sends a request every 3 minutes to check the web service status and ensure it remains active, preventing any idle shutdown.

Tools and Technologies Used

- **SparseVec model**: BM25 (Best Matching 25)
- **Dense model**: BGE M3 (using Cosine similarity)
- **LLM**: Llama3 70B
- **VectorDB**: Qdrant
- **ErrorReportingDB**: MongoDB

2.1 Data Access and Indexing

Crawl Data

```
import requests
from bs4 import BeautifulSoup
import spacy
import os
import json

class PresightScraper:
    def __init__(self, save_dir="data"):
        self.session = requests.Session()
        self.base_url = "https://www.presight.io/privacy-policy.html"
        self.headers = {
            'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36'
        }
        self.save_dir = save_dir
        os.makedirs(self.save_dir, exist_ok=True)

        # Load English language model
        self.nlp = spacy.load("en_core_web_sm")

    def get_page_content(self, url):
        try:
            response = self.session.get(url, headers=self.headers)
            response.raise_for_status()
            soup = BeautifulSoup(response.text, 'html.parser')


            for unwanted_h2 in soup.find_all('h2', class_="chakra-heading css-18j379d"):
                if "Last updated" in unwanted_h2.get_text(strip=True):
                    unwanted_h2.decompose()

            return soup
        except requests.RequestException as e:
            print(f"Error fetching {url}: {e}")
            return None

    def extract_ordered_content(self, soup):
        ordered_content = []
        current_section = None

        # Get the main content area - adjust selector based on the website structure
        main_content = soup.find('body')

        if main_content:
            for element in main_content.descendants:
                if element.name in ['h1', 'h2', 'h3', 'h4'] and element.string:
                    # Start a new section for each header
                    header = element.get_text(strip=True)
                    current_section = {
                        'header': header,
                        'content': []
                    }
                    ordered_content.append(current_section)
                elif element.name in ['p', 'li'] and current_section:
                    # Get all text, including text from child elements
like <span>
                    text = element.get_text(strip=True)
                    if text:
                        # Process text with spaCy
                        doc = self.nlp(text)
                        for sent in doc.sents:
                            clean_sentence = sent.text.strip()
                            if clean_sentence:

current_section['content'].append(clean_sentence)

        return ordered_content

    def scrape_page(self):
        print(f"Scraping: {self.base_url}")
        soup = self.get_page_content(self.base_url)
        if soup:
            return self.extract_ordered_content(soup)
        return []

    def save_to_json(self, content, filename):
        file_path = os.path.join(self.save_dir, f"{filename}.json")
        with open(file_path, 'w', encoding='utf-8') as f:
            json.dump(content, f, ensure_ascii=False, indent=4)
        print(f"Data saved to JSON: {file_path}")

def main():
    scraper = PresightScraper(save_dir="output")
    content = scraper.scrape_page()
    scraper.save_to_json(content, 'privacy_policy_final')

if __name__ == "__main__":
    main()
```

Insert chunks of text

Here, the team has decided to have two different versions of indexing into the database to evaluate performance.

Create collection, add Docs function

BM25 (Best Matching 25)

Key Concepts:

1. **Term Frequency (TF)**:
   - Measures how often a term appears in a document. The more times a term appears, the more relevant the document is considered for that term.
2. **Inverse Document Frequency (IDF)**:
   - Measures how important a term is across the entire corpus. Rare terms (those that appear in fewer documents) are given more weight than common terms.
3. **Document Length Normalization**:
   - BM25 accounts for the length of the document by introducing a normalization factor to prevent longer documents from being unfairly favored.

BM25 Formula:

The BM25 score for a document d with respect to a query q is calculated as:

$$\text{BM25}(d,q)=\sum_{t\in q}\text{IDF}(t)\cdot\frac{\text{TF}(t,d)\cdot(k_1+1)}{\text{TF}(t,d)+k_1\cdot\left(1-b+b\cdot\frac{|d|}{\text{avg\_doc\_length}}\right)}$$

Where:

- t is a term in the query q.
- $\text{TF}(t,d)$ is the term frequency of t in document d.
- $\text{IDF}(t)$ is the inverse document frequency of term t.
- $|d|$ is the length of document d.
- avg_doc_length is the average document length in the corpus.
- k_1 and b are tuning parameters, usually set to $k_1=1.5$ and $b=0.75$.

Hybrid Search

Hybrid Search combines both **dense retrieval** and **sparse retrieval** to search for documents, optimizing the search results by leveraging the strengths of both methods.

How it works:

1. **Dense Retrieval**:
   - The query is converted into **dense embeddings** (semantic vectors), which helps search for documents with similar semantic meanings to the query.
2. **Sparse Retrieval**:
   - The query is converted into **sparse embeddings** (keyword-based vectors), which helps search for documents with keywords that match the query.

3. **Combining Dense and Sparse**:

   - Two queries are performed in parallel: one for **dense embeddings** and one for **sparse embeddings**.
   - The results from both methods are **combined** and **re-ranked** using an algorithm like **Reciprocal Rank Fusion (RRF)** or a similar approach.

4. **Reciprocal Rank Fusion (RRF)**:

   - RRF calculates a score for each document based on its rank in the result list from each method.
   - The **Reciprocal Rank Score** for a document at rank position `rank` in the result list is calculated as:

$$\text{score} = \frac{1}{\text{rank} + k}$$

   Where:

   - `rank` is the position of the document in the result list.
   - `k` is a constant, typically chosen as **60**.

   Documents with higher **score** are prioritized in the final results.

5. **Benefits of Hybrid Search**:

   - **High accuracy**: Combines both keyword-based and semantic search.
   - **Optimization**: Leverages the strengths of both dense and sparse retrieval methods.

Maximal Marginal Relevance (MMR)

Maximal Marginal Relevance (MMR) is a technique used in information retrieval and document ranking to improve the diversity of search results while maintaining relevance to the query. MMR aims to select documents that are both relevant to the query and diverse from each other, reducing redundancy in the returned results.

How it works:

1. **Relevance and Diversity**:

   - MMR attempts to balance two key factors:
     - **Relevance**: The degree to which a document is related to the query.
     - **Diversity**: The degree to which a document differs from the ones already selected.

2. **MMR Formula**:

   - For a given query and set of candidate documents, MMR selects documents based on both relevance and diversity. The formula for MMR is as follows:

$$\text{MMR}(d) = \lambda \cdot \text{Relevance}(d, q) - (1 - \lambda) \cdot \max_{d' \in S} \text{Similarity}(d, d')$$

Where:

- (d) is the candidate document.
- (q) is the query.
- (S) is the set of documents already selected.
- **Relevance(d, q)** is the relevance score between document (d) and the query (q).
- **Similarity(d, d')** is the similarity score between document (d) and the previously selected document (d').
- (λ) is a parameter that controls the trade-off between relevance and diversity. It is typically a value between 0 and 1.

3. **Selection Process**:

   - MMR selects documents by considering their relevance to the query and their similarity to the documents already selected.
   - At each step, MMR picks the document with the highest score from the remaining candidates, which is a balance between relevance and diversity.

4. **Benefits of MMR**:

   - **Reduced Redundancy**: MMR helps in minimizing the overlap between the returned documents, ensuring that the results cover a broader range of information.
   - **Improved Diversity**: By considering diversity, MMR can offer a more comprehensive set of results, especially when the query has multiple facets or is vague.
   - **Effective for Ranking**: MMR is especially useful in scenarios like search engines, document summarization, and recommendation systems, where both relevance and diversity matter.

5. **Applications**:

   - **Search Engines**: MMR can be applied to rank search results by ensuring diverse and relevant documents are returned.
   - **Document Summarization**: In extractive summarization, MMR helps select diverse sentences or paragraphs that cover different aspects of the document.
   - **Recommendation Systems**: MMR helps recommend diverse items based on user preferences.

By balancing relevance and diversity, MMR ensures that the final set of documents returned provides a comprehensive and varied response to the query.

```
from langchain.embeddings import HuggingFaceInferenceAPIEmbeddings
from langchain_qdrant import QdrantVectorStore,FastEmbedSparse
from qdrant_client.http.models import Distance, VectorParams
from qdrant_client import QdrantClient, models
from langchain_community.embeddings.fastembed import
FastEmbedEmbeddings
from langchain.document_loaders import TextLoader

def Create_collection(collectionName):
    # Qdrant client
    QDRANT_URL="https://9ba55ee0-09ef-4c78-8d04-72c6392c0425.us-east4-
0.gcp.cloud.qdrant.io"
    QDRANT_API_KEY=""
    client = QdrantClient(
          url=QDRANT_URL,
          api_key=QDRANT_API_KEY,
          prefer_grpc=False
      )

    vector_name = "sparse_vector"
    client.create_collection(
    collection_name=collectionName,
    vectors_config=VectorParams(size=1024, distance=Distance.COSINE),
    sparse_vectors_config={
        vector_name: models.SparseVectorParams(
            index=models.SparseIndexParams(
                on_disk=False,
            )
        )
    }
    )
def add_Documents(collectionName,docs):
    QDRANT_URL="https://9ba55ee0-09ef-4c78-8d04-72c6392c0425.us-east4-
0.gcp.cloud.qdrant.io"
    QDRANT_API_KEY=""
    embeddings = HuggingFaceInferenceAPIEmbeddings(
        model_name="BAAI/bge-m3",
        api_key = "",
        model_kwargs = {'device': 'auto'}
    )
    sparse_embeddings = FastEmbedSparse(model_name="Qdrant/bm25",
                      model_kwargs = {'device':
'auto'})
    qdrant = QdrantVectorStore.from_documents(
        docs,
        embeddings,
        sparse_embedding=sparse_embeddings,
        sparse_vector_name="Qdrant/bm25",
        url=QDRANT_URL,
        prefer_grpc=False,
        collection_name=collectionName,
        api_key=QDRANT_API_KEY,
        timeout=300
    )
```

Indexing Based on Chunksize

The popular chunking method is commonly used in applications related to RAG (Retrieval-Augmented Generation).

```
from langchain.document_loaders import TextLoader
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.vectorstores import Qdrant
from langchain.schema import Document

loader_sotay = TextLoader(r"C:\Users\Vinh Thuan\Downloads\
privacy_policy.txt", encoding='utf-8')
documents = loader_sotay.load()
text_splitter = RecursiveCharacterTextSplitter(chunk_size=300,
chunk_overlap=100)
texts = text_splitter.split_documents(documents)

texts
```

```
[Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='By Role\nBy Team\n\nPRIVACY
POLICY\n\n\nLast updated 15 Sep 2023\n\nAt Presight, we are committed
to protecting the privacy of our customers and visitors to our
website.\nThis Privacy Policy explains how we collect, use, and
disclose information about our customers and visitors.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='Information Collection and Use\n\
nWe collect several different types of information for various
purposes to provide and improve our Service to you.\n\nTypes of Data
Collected'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='While using our Service, we may
ask you to provide us with certain personally identifiable information
that can be used to contact or identify you ("Personal Data").\
nPersonally identifiable information may include, but is not limited
to:\n  • Email address\n  • First name and last name'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='• Email address\n  • First name
and last name\n  • Phone number\n  • Address, State, Province,
ZIP/Postal code, City\n  • Cookies and Usage Data'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='• Phone number\n  • Address,
State, Province, ZIP/Postal code, City\n  • Cookies and Usage Data\nWe
may also collect information that your browser sends whenever you
visit our Service or when you access the Service by or through a
mobile device ("Usage Data").'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content="This Usage Data may include
information such as your computer's Internet Protocol address (e.g. IP
address), browser type, browser version, the pages of our Service that
```

```
you visit, the time and date of your visit, the time spent on those
pages, unique device identifiers, and other diagnostic data."),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='Use of Data'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='Presight uses the collected data
for various purposes:\n  • To provide and maintain our Service\n  • To
notify you about changes to our Service\n  • To allow you to
participate in interactive features of our Service when you choose to
do so\n  • To provide customer support'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='• To provide customer support\n  •
To gather analysis or valuable information so that we can improve our
Service\n  • To monitor the usage of our Service\n  • To detect,
prevent and address technical issues'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='Consent\n\nAs personal information
is collected, you will be asked to confirm that your information is
correct prior to submitting it to Presight.\n\nAccess to Personal
Information\n\n\nAccessing Your Personal Information'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='Accessing Your Personal
Information\n\nYou have the right to access all of your personal
information that we hold.\nThrough the application, you can correct,
amend, or append your personal information by logging into the
application and navigating to your settings and profile.\n\nAutomated
Edit Checks'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='Presight employs automated edit
checks to ensure that data entry fields are completed properly when
collecting personal information.\nThese edit checks help maintain data
integrity and accuracy.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='These edit checks help maintain
data integrity and accuracy.\nYou are encouraged to provide complete
and valid information to ensure the smooth processing of their
personal data.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='Disclosure of Information'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='We may disclose your application
data to third-party service providers who help us provide our services
such as Datadog, AWS, Google Cloud and Google Workspace.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='We may also disclose your
information in response to a legal request, such as a subpoena or
court order, or to protect our rights or the rights of others.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='Sharing of Personal Data\n\nYour
```

```
personal data will not be subject to sharing, transfer, rental or
exchange for the benefit of third parties, including AI models.\n\
nGoogle User Data and Google Workspace APIs'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='In all cases when users
authenticate the platform to Google Workspace, the following applies:\
n  • We do not retain or use Google User Data to develop, improve, or
train generalized/non-personalized AI and/or ML models.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='• We do not use Google Workspace
APIs to develop, improve, or train generalized/non-personalized AI
and/or ML models.\n  • We do not transfer Google User Data to third-
party AI tools for the purpose of developing, improving, or training
generalized or non-personalized AI and/or ML models.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='Data Security'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='• All data is encrypted both in
transit and at rest, using industry-standard encryption methods.\n  •
We regularly perform security audits and vulnerability assessments to
ensure the safety of our platform and the data stored within it.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='• Our employees are trained on
best practices for data security, and access to customer data is
restricted on a need-to-know basis.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='Data Retention & Disposal\n\
nCustomer data is retained for as long as the account is in active
status.\nData enters an â\x80\x9cexpiredâ\x80\x9d state when the
account is voluntarily closed.\nExpired account data will be retained
for 60 days.\nAfter this period, the account and related data will be
removed.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content="Quality, Including Data Subjects'
Responsibilities for Quality"),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='• We are committed to maintaining
the quality and accuracy of the personal information we collect and
process.\n  • We rely on data subjects to provide accurate and up-to-
date information.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='• We rely on data subjects to
provide accurate and up-to-date information.\n  • Data subjects have
the responsibility to inform us of any changes or inaccuracies in
their personal data.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='• If you believe that any
information we hold about you is inaccurate, incomplete, or outdated,
please contact us promptly to rectify the information.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='Monitoring and Enforcement'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='• We regularly monitor its data
processing activities to ensure compliance with this privacy policy
and applicable data protection laws.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='• In the event of a data breach or
any unauthorized access to your personal information, we will notify
you and the appropriate authorities as required by law.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='• We committed to cooperating with
data protection authorities and complying with their advice and
decisions regarding data protection and privacy matters.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='Cookies\n\nWe use cookies to
enhance your experience on our website.\nYou can control the use of
cookies through your web browser settings.\n\nThird-Party Websites\n\
nOur website may contain links to third-party websites.\nWe are not
responsible for the privacy practices or content of those websites.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content='Changes to Privacy Policy\n\nWe
may update this Privacy Policy from time to time.\nThe updated Privacy
Policy will be posted on our website.\n\nContact Us\n\n\nPurposeful
Use Only'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
privacy_policy.txt'}, page_content="Contact Us\n\n\nPurposeful Use
Only\n\nWe commit to only use personal information for the purposes
identified in the entity's privacy policy.\nPresight.io 2022 All
Rights Reserved\nHo Chi Minh City, Vietnam\nSingapore\nSeattle, WA,
USA")]
```

```python
Create_collection("Python_simple_rag")

add_Documents("Python_simple_rag",texts)
```

```
Fetching 1 files: 100%|██████████| 1/1 [00:00<?, ?it/s]
```

topic-based indexing

The chunking method based on topics.

```python
import json
from langchain_community.document_loaders import JSONLoader

with open(r"C:\Users\Vinh Thuan\Downloads\privacy_policy_final.json",
'r', encoding='utf-8') as file:
    json_data = json.load(file)

processed_data = []
```

```python
previous_header = None

for entry in json_data:
    if not entry['content']:
        previous_header = entry['header']
    else:
        if previous_header:
            entry['content'].insert(0, previous_header)
            previous_header = None
        entry['content'].insert(0, entry['header'])
        entry['content'] = ' '.join(entry['content'])
        processed_data.append(entry)

processed_data

processed_file_path = r"C:\Users\Vinh Thuan\Downloads\
processed_privacy_policy.json"

with open(processed_file_path, 'w', encoding='utf-8') as file:
    json.dump(processed_data, file, ensure_ascii=False, indent=4)

print(f"Processed data has been saved at {processed_file_path}")
```

```
Processed data đã được lưu tại C:\Users\Vinh Thuan\Downloads\
processed_privacy_policy.json
```

```python
def metadata_func(record: dict, metadata: dict) -> dict:
    metadata["header"] = record.get("header")
    return metadata

loader = JSONLoader(
    file_path=processed_file_path,
    jq_schema='.[]',
    content_key="content",
    metadata_func=metadata_func,
    json_lines=False
)

documents = loader.load()

documents
```

```
[Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 1, 'header': 'PRIVACY
POLICY'}, page_content='PRIVACY POLICY At Presight, we are committed
to protecting the privacy of our customers and visitors to our
website. This Privacy Policy explains how we collect, use, and
disclose information about our customers and visitors.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 2, 'header': 'Information
Collection and Use'}, page_content='Information Collection and Use We
```

```
collect several different types of information for various purposes to
provide and improve our Service to you.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 3, 'header': 'Types of Data
Collected'}, page_content='Types of Data Collected While using our
Service, we may ask you to provide us with certain personally
identifiable information that can be used to contact or identify you
("Personal Data"). Personally identifiable information may include,
but is not limited to: Email address First name and last name Phone
number Address, State, Province, ZIP/Postal code, City Cookies and
Usage Data We may also collect information that your browser sends
whenever you visit our Service or when you access the Service by or
through a mobile device ("Usage Data"). This Usage Data may include
information such as your computer\'s Internet Protocol address (e.g.
IP address), browser type, browser version, the pages of our Service
that you visit, the time and date of your visit, the time spent on
those pages, unique device identifiers, and other diagnostic data.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 4, 'header': 'Use of
Data'}, page_content='Use of Data Presight uses the collected data for
various purposes: To provide and maintain our Service To notify you
about changes to our Service To allow you to participate in
interactive features of our Service when you choose to do so To
provide customer support To gather analysis or valuable information so
that we can improve our Service To monitor the usage of our Service To
detect, prevent and address technical issues'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 5, 'header': 'Consent'},
page_content='Consent As personal information is collected, you will
be asked to confirm that your information is correct prior to
submitting it to Presight.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 6, 'header': 'Accessing
Your Personal Information'}, page_content='Accessing Your Personal
Information Access to Personal Information You have the right to
access all of your personal information that we hold. Through the
application, you can correct, amend, or append your personal
information by logging into the application and navigating to your
settings and profile.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 7, 'header': 'Automated
Edit Checks'}, page_content='Automated Edit Checks Presight employs
automated edit checks to ensure that data entry fields are completed
properly when collecting personal information. These edit checks help
maintain data integrity and accuracy. You are encouraged to provide
complete and valid information to ensure the smooth processing of
their personal data.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 8, 'header': 'Disclosure of
Information'}, page_content='Disclosure of Information We may disclose
your application data to third-party service providers who help us
provide our services such as Datadog, AWS, Google Cloud and Google
Workspace. We may also disclose your information in response to a
legal request, such as a subpoena or court order, or to protect our
rights or the rights of others.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 9, 'header': 'Sharing of
Personal Data'}, page_content='Sharing of Personal Data Your personal
data will not be subject to sharing, transfer, rental or exchange for
the benefit of third parties, including AI models.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 10, 'header': 'Google User
Data and Google Workspace APIs'}, page_content='Google User Data and
Google Workspace APIs In all cases when users authenticate the
platform to Google Workspace, the following applies: We do not retain
or use Google User Data to develop, improve, or train generalized/non-
personalized AI and/or ML models. We do not use Google Workspace APIs
to develop, improve, or train generalized/non-personalized AI and/or
ML models. We do not transfer Google User Data to third-party AI tools
for the purpose of developing, improving, or training generalized or
non-personalized AI and/or ML models.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 11, 'header': 'Data
Security'}, page_content='Data Security All data is encrypted both in
transit and at rest, using industry-standard encryption methods. We
regularly perform security audits and vulnerability assessments to
ensure the safety of our platform and the data stored within it. Our
employees are trained on best practices for data security, and access
to customer data is restricted on a need-to-know basis.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 12, 'header': 'Data
Retention & Disposal'}, page_content='Data Retention & Disposal
Customer data is retained for as long as the account is in active
status. Data enters an â\x80\x9cexpiredâ\x80\x9d state when the
account is voluntarily closed. Expired account data will be retained
for 60 days. After this period, the account and related data will be
removed.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 13, 'header': "Quality,
Including Data Subjects' Responsibilities for Quality"},
page_content="Quality, Including Data Subjects' Responsibilities for
Quality We are committed to maintaining the quality and accuracy of
the personal information we collect and process. We rely on data
subjects to provide accurate and up-to-date information. Data subjects
have the responsibility to inform us of any changes or inaccuracies in
their personal data. If you believe that any information we hold about
you is inaccurate, incomplete, or outdated, please contact us promptly
to rectify the information."),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 14, 'header': 'Monitoring
and Enforcement'}, page_content='Monitoring and Enforcement We
regularly monitor its data processing activities to ensure compliance
with this privacy policy and applicable data protection laws. In the
event of a data breach or any unauthorized access to your personal
information, we will notify you and the appropriate authorities as
required by law. We committed to cooperating with data protection
authorities and complying with their advice and decisions regarding
data protection and privacy matters.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 15, 'header': 'Cookies'},
page_content='Cookies We use cookies to enhance your experience on our
website. You can control the use of cookies through your web browser
settings.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 16, 'header': 'Third-Party
Websites'}, page_content='Third-Party Websites Our website may contain
links to third-party websites. We are not responsible for the privacy
practices or content of those websites.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 17, 'header': 'Changes to
Privacy Policy'}, page_content='Changes to Privacy Policy We may
update this Privacy Policy from time to time. The updated Privacy
Policy will be posted on our website.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 18, 'header': 'Contact
Us'}, page_content='Contact Us If you have any questions about this
Privacy Policy, please contact us through the customer portal or by
email atpresight@presight.io.'),
 Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 19, 'header': 'Purposeful
Use Only'}, page_content="Purposeful Use Only We commit to only use
personal information for the purposes identified in the entity's
privacy policy. Presight.io 2022 All Rights Reserved Ho Chi Minh City,
Vietnam Singapore Seattle, WA, USA")]
```

```
Create_collection("Python_simple_rag2")
```

```
add_Documents("Python_simple_rag2",documents)
```

```
Fetching 1 files: 100%|██████████| 1/1 [00:00<?, ?it/s]
```

2.2 Chatbot Development

Chatbot

```
from langchain.chains import create_retrieval_chain
from langchain.prompts import PromptTemplate
from langchain_groq import ChatGroq
from qdrant_client import QdrantClient
from langchain_qdrant import QdrantVectorStore
from langchain_qdrant import FastEmbedSparse, RetrievalMode
from langchain.chains.combine_documents import
create_stuff_documents_chain

class RAGPipelineSetup:
    def __init__(self, qdrant_url, qdrant_api_key,
huggingface_api_key, embeddings_model_name, groq_api_key):
        self.QDRANT_URL = qdrant_url
        self.QDRANT_API_KEY = qdrant_api_key
        self.HUGGINGFACE_API_KEY = huggingface_api_key
        self.EMBEDDINGS_MODEL_NAME = embeddings_model_name
        self.GROQ_API_KEY = groq_api_key
        self.embeddings = self.load_embeddings()
        self.pipe = self.load_model_pipeline()
        self.prompt = self.load_prompt_template()
        self.current_source = None

    def load_embeddings(self):
        # Load HuggingFace embeddings
        bge_embeddings = HuggingFaceInferenceAPIEmbeddings(
            model_name=self.EMBEDDINGS_MODEL_NAME,
            api_key=self.HUGGINGFACE_API_KEY,
        )
        return bge_embeddings

    def load_retriever(self, retriever_name):
        client = QdrantClient(
            url=self.QDRANT_URL,
            api_key=self.QDRANT_API_KEY,
            prefer_grpc=False
        )

        # Load sparse embedding
        sparse_embeddings = FastEmbedSparse(model_name="Qdrant/bm25")

        db = QdrantVectorStore(
            client=client,
            embedding=self.embeddings,
            sparse_embedding=sparse_embeddings,
            sparse_vector_name="sparse_vector",
            collection_name=retriever_name,
            retrieval_mode=RetrievalMode.HYBRID
        )

        retriever = db.as_retriever(search_kwargs={"k":
3},search_type="mmr")
        return retriever
    def load_model_pipeline(self, max_new_tokens=1024):
        llm = ChatGroq(
            temperature=0,
            groq_api_key=self.GROQ_API_KEY,
            model_name="llama3-70b-8192"
        )
        return llm

    def load_prompt_template(self, source=None):
        query_template = '''
        ### Context:
        {context}

        ### User Question:
        {input}

        ### Instructions for the Assistant:
        1. Carefully read the user's question and analyze the intent.
        2. Search the context provided above for the most accurate and
relevant information.
        3. Formulate a clear and concise response to address the
user's question.
        4. If the answer cannot be derived directly from the context,
politely inform the user and suggest an alternative.

        '''

        prompt = PromptTemplate(template=query_template,
input_variables=["context", "input"])
        return prompt

    def load_rag_pipeline(self, llm, retriever, prompt):
        # Retrieval Augmented Generation
        rag_chain = create_retrieval_chain(
            retriever=retriever,
            combine_docs_chain=create_stuff_documents_chain(llm,
prompt)
        )

        return rag_chain

    def rag(self, source):
        if source == self.current_source:
            return self.rag_pipeline
        else:
            self.retriever =
self.load_retriever(retriever_name=source)
            self.pipe = self.load_model_pipeline()
            self.prompt = self.load_prompt_template(source)
            self.rag_pipeline = self.load_rag_pipeline(llm=self.pipe, 
retriever=self.retriever, prompt=self.prompt)
            self.current_source = source
            return self.rag_pipeline
```

Example

```
rag_pipeline_setup = RAGPipelineSetup(
    qdrant_url="https://9ba55ee0-09ef-4c78-8d04-72c6392c0425.us-east4-
0.gcp.cloud.qdrant.io",
    qdrant_api_key="",
    huggingface_api_key="",
    embeddings_model_name="BAAI/bge-m3",
    groq_api_key=""
)

question = "What is the use of data in privacy policy?"

qdrant_collection_name="Python_simple_rag2"
rag_pipeline = rag_pipeline_setup.rag(source=qdrant_collection_name)

inputs = {
    "input": question
}

response = rag_pipeline.invoke(inputs)

response['answer']
```

```
Fetching 1 files: 100%|██████████| 1/1 [00:00<00:00, 813.01it/s]
```

```
'According to the privacy policy, the use of data is for the following 
purposes:\n\n1. To provide and maintain the Service\n2. To notify 
users about changes to the Service\n3. To allow users to participate 
in interactive features of the Service\n4. To provide customer 
support\n5. To gather analysis or valuable information to improve the 
Service\n6. To monitor the usage of the Service\n7. To detect, 
prevent, and address technical issues.\n\nThese are the specific uses 
of data as outlined in the privacy policy.'
```

```
response
```

```
{'input': 'What is the use of data in privacy policy?',
 'context': [Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\
Downloads\\processed_privacy_policy.json', 'seq_num': 4, 'header': 
'Use of Data', '_id': '27152d3b-a152-43a2-891b-2f9397f377d8', 
'_collection_name': 'Python_simple_rag2'}, page_content='Use of Data 
Presight uses the collected data for various purposes: To provide and 
maintain our Service To notify you about changes to our Service To 
allow you to participate in interactive features of our Service when
you choose to do so To provide customer support To gather analysis or
valuable information so that we can improve our Service To monitor the
usage of our Service To detect, prevent and address technical
issues'),
  Document(metadata={'source': 'C:\\Users\\Vinh Thuan\\Downloads\\
processed_privacy_policy.json', 'seq_num': 3, 'header': 'Types of Data
Collected', '_id': 'f33f6697-1944-47eb-9fdf-32da45edec70',
'_collection_name': 'Python_simple_rag2'}, page_content='Types of Data
Collected While using our Service, we may ask you to provide us with
certain personally identifiable information that can be used to
contact or identify you ("Personal Data"). Personally identifiable
information may include, but is not limited to: Email address First
name and last name Phone number Address, State, Province, ZIP/Postal
code, City Cookies and Usage Data We may also collect information that
your browser sends whenever you visit our Service or when you access
the Service by or through a mobile device ("Usage Data"). This Usage
Data may include information such as your computer\'s Internet
Protocol address (e.g. IP address), browser type, browser version, the
pages of our Service that you visit, the time and date of your visit,
the time spent on those pages, unique device identifiers, and other
diagnostic data.')],
 'answer': 'According to the provided privacy policy, the use of data
is for the following purposes:\n\n1. To provide and maintain the
Service\n2. To notify users about changes to the Service\n3. To allow
users to participate in interactive features of the Service\n4. To
provide customer support\n5. To gather analysis or valuable
information to improve the Service\n6. To monitor the usage of the
Service\n7. To detect, prevent, and address technical issues.\n\nThese
are the primary uses of the collected data as stated in the privacy
policy.'}
```

generate Question - Ground truth for Evaluation

```python
import csv
from langchain_core.prompts import ChatPromptTemplate
from langchain_groq import ChatGroq

chat = ChatGroq(temperature=0, model_name="mixtral-8x7b-32768",
groq_api_key="")


qa_pairs = []

# Iterate over each document and generate questions and answers
for doc in documents:
    header = doc.metadata['header']
    content = doc.page_content

    # Define the system and human messages
    system = "You are an AI designed to generate questions and answers
```

```
from text."
    human = (
        f"Given the following text:\n\n{content}\n\n"
        "1. Formulate a clear and concise question that a reader might
ask\n"
        "2. Provide a detailed answer to the question, ensuring the
response is informative and directly related to the content."
    )

    prompt = ChatPromptTemplate.from_messages([("system", system),
("human", human)])

    # Combine the prompt and chat model
    chain = prompt | chat

    # Invoke the chain with the document content
    response = chain.invoke({})

    # Process the response content
    response_text = response.content.strip()

    if 'Question:' in response_text and 'Answer:' in response_text:
        # Extract the question and answer
        question_part, answer_part = response_text.split('Answer:', 1)
        question = question_part.replace('Question:', '').strip()
        answer = answer_part.strip()
    else:
        # Handle unexpected formats
        question = "N/A"
        answer = response_text.strip()


    qa_pairs.append((question, answer))

    print(f"Header: {header}")
    print(response_text)
    print("-" * 80)

# Save the questions and answers to a CSV file
with open('questions_answers.csv', 'w', newline='', encoding='utf-8')
as csvfile:
    csvwriter = csv.writer(csvfile)
    csvwriter.writerow(['Question', 'Answer'])
    csvwriter.writerows(qa_pairs)
```

```
Header: PRIVACY POLICY
Question: What is Presight's commitment to privacy as stated in the
text?

Answer: Presight is committed to protecting the privacy of its
```

```
customers and visitors to its website. This commitment is outlined in
their Privacy Policy, which explains how they collect, use, and
disclose information about their customers and visitors.
----------------------------------------------------------------------
----------
Header: Information Collection and Use
Question: What is the purpose of information collection in the
provided text?

Answer: The purpose of information collection in the provided text is
to improve and provide the Service. The text mentions that various
types of information are collected for different purposes. This
implies that the collected data is used to enhance the user experience
and optimize the functionality of the Service. However, the text does
not specify the exact types of information collected or how they are
used.
----------------------------------------------------------------------
----------
Header: Types of Data Collected
Question: What types of personal data and usage data are collected
while using the service?

Answer: When using the service, the following personally identifiable
information, also known as Personal Data, may be collected: email
address, first name and last name, phone number, address,
state/province, ZIP/postal code, city, and cookies. Personal Data can
be used to contact or identify you.

Additionally, Usage Data is collected, which includes information sent
by your browser or mobile device when accessing the service. Usage
Data may consist of your computer's Internet Protocol address (e.g. IP
address), browser type, browser version, the pages of the service you
visit, the time and date of your visit, the time spent on those pages,
unique device identifiers, and other diagnostic data.
----------------------------------------------------------------------
----------
Header: Use of Data
Question: What are the purposes for which Data Presight utilizes the
collected data?

Answer: Data Presight uses the collected data for several key
purposes. First, they use it to provide and maintain their Service,
ensuring it remains functional and up-to-date. They also use the data
to notify users about any changes to the Service.

Furthermore, Data Presight allows users to participate in interactive
features of their Service, and they use the collected data to
facilitate this interaction. They provide customer support using the
data, which helps them address user issues more effectively.
```

In addition, Data Presight gathers analysis or valuable information from the data to improve their Service. They monitor the usage of their Service through the data, which helps them understand how users interact with their platform.

Lastly, Data Presight uses the collected data to detect, prevent, and address technical issues, ensuring a smooth and uninterrupted service for their users.

Header: Consent
Question: Why will I be asked to confirm that my personal information is correct before submitting it to Presight?

Answer: The text indicates that you will be asked to confirm the correctness of your personal information prior to submitting it to Presight because they want to ensure the accuracy of the data they collect. This step is crucial for maintaining the quality of their database and reducing errors that might arise from incorrect or outdated information. It also shows respect for your privacy and data ownership, as you have the opportunity to review and verify your information before it is submitted.

Header: Accessing Your Personal Information
Question: How can I access my personal information held by the application?

Answer: The text states that you have the right to access all of your personal information that the application holds. To do this, you need to log into the application. Once you are logged in, navigate to your settings and profile. From there, you will be able to access your personal information and make any corrections, amendments, or additions as necessary.

Header: Automated Edit Checks
Question: What is the purpose of the automated edit checks used by Presight when collecting personal information?

Answer: The automated edit checks employed by Presight serve to ensure that data entry fields are completed correctly and accurately when personal information is being gathered. This process helps maintain data integrity, which is crucial for the smooth and efficient processing of individuals' personal data. By providing complete and valid information, users can help ensure that these checks are carried out effectively and that their data is managed correctly.

Header: Disclosure of Information

```
Question: To whom does the company disclose application data?

Answer: The company discloses application data to third-party service
providers who assist them in providing their services. Specifically,
they mention using Datadog, AWS (Amazon Web Services), Google Cloud,
and Google Workspace. These service providers help the company in
various ways, such as managing infrastructure, analyzing data, and
facilitating communication and collaboration.

Additionally, the company may disclose information in response to a
legal request, such as a subpoena or court order. This disclosure is
intended to protect their rights or the rights of others, suggesting
that they may share information when required by law or when necessary
to ensure the safety and security of their users or operations.
----------------------------------------------------------------------
----------
Header: Sharing of Personal Data
Question: Will my personal data be shared or transferred to third
parties, including AI models?

Answer: No, your personal data will not be subject to sharing,
transfer, rental, or exchange for the benefit of third parties,
including AI models. This statement provides assurance that the
protection of your personal data is a priority and it will not be
disclosed to external entities without your consent.
----------------------------------------------------------------------
----------
Header: Google User Data and Google Workspace APIs
Question: What is the policy regarding the use of Google User Data and
Google Workspace APIs for developing, improving, or training
generalized/non-personalized AI and/or ML models?

Answer: The policy clearly states that the system does not retain or
use Google User Data to develop, improve, or train generalized/non-
personalized AI and/or ML models. This means that any data provided by
users when they authenticate the platform to Google Workspace is not
stored or used for the purpose of enhancing AI or machine learning
models that are not specific to the individual user.

Furthermore, the policy also specifies that Google Workspace APIs are
not used to develop, improve, or train generalized/non-personalized AI
and/or ML models. This implies that the tools and interfaces provided
by Google for interacting with Workspace are not utilized for the
aforementioned AI and machine learning development purposes.

Lastly, the policy clearly states that Google User Data is not
transferred to third-party AI tools for the purpose of developing,
improving, or training generalized or non-personalized AI and/or ML
models. This means that even if third-party tools are used, they do
not receive any user data from Google for AI or machine learning model
```

```
development purposes.
----------------------------------------------------------------------
----------
Header: Data Security
Question: How is data secured in this system, both when it is being
transferred and when it is stored?

Answer: In this system, data security is taken very seriously. To
protect data during transit, or when it is being transferred, the
system uses industry-standard encryption methods. This means that even
if the data is intercepted during transmission, it would be unreadable
and secure.

Similarly, when data is at rest, or stored in the system, it is also
encrypted using industry-standard encryption methods. This ensures
that the data remains secure even when it is not being actively used
or transferred.

In addition to these measures, the system regularly undergoes security
audits and vulnerability assessments. These processes help to identify
and address any potential security weaknesses, further ensuring the
safety of the platform and the data stored within it.

The system's employees are also trained on best practices for data
security. This includes understanding how to handle sensitive data,
the importance of strong passwords, and the risks of phishing and
other social engineering attacks.

Finally, access to customer data is restricted on a need-to-know
basis. This means that employees only have access to the data they
need to perform their job functions. This limits the potential for
data breaches and ensures that customer data is only accessed when
absolutely necessary.
----------------------------------------------------------------------
----------
Header: Data Retention & Disposal
Question: How long is customer data retained after an account is
closed?

Answer: After a customer voluntarily closes their account, the data
enters an "expired" state and is retained for 60 days. Once this 60-
day period has elapsed, the account and all related data are removed.
----------------------------------------------------------------------
----------
Header: Quality, Including Data Subjects' Responsibilities for Quality
Question: What is the role of data subjects in maintaining the quality
and accuracy of their personal information?

Answer: Data subjects play a significant role in ensuring the quality
and accuracy of their personal information. They are expected to
```

```
provide accurate and up-to-date information when it is collected.
Furthermore, data subjects have the responsibility to inform the
organization of any changes or inaccuracies in their personal data. If
a data subject believes that any information the organization holds
about them is inaccurate, incomplete, or outdated, they should contact
the organization promptly to rectify the information. This proactive
approach helps maintain the accuracy and quality of personal data,
which is crucial for both the data subject and the organization.
----------------------------------------------------------------------
----------
Header: Monitoring and Enforcement
Question: What measures does this entity take to ensure the security
and privacy of my personal information?

Answer: This entity is committed to protecting your personal
information by regularly monitoring its data processing activities to
ensure compliance with this privacy policy and applicable data
protection laws. In the unfortunate event of a data breach or any
unauthorized access to your personal information, they will promptly
notify you and the appropriate authorities as required by law.
Furthermore, they are dedicated to cooperating with data protection
authorities and complying with their advice and decisions regarding
data protection and privacy matters. This demonstrates their strong
commitment to maintaining the security and privacy of your personal
data.
----------------------------------------------------------------------
----------
Header: Cookies
Question: What are cookies used for on this website and how can I
control them?

Answer: Cookies on this website are used to enhance your experience,
which may include features like remembering your preferences or
improving the way the site presents information to you. To control the
use of cookies, you can adjust your web browser settings. Most
browsers allow you to manage your cookie preferences, including
blocking or deleting cookies. However, please note that disabling
cookies might affect your browsing experience and limit some
functionalities of the website.
----------------------------------------------------------------------
----------
Header: Third-Party Websites
Question: What is the responsibility of the website regarding the
privacy practices and content of third-party websites?

Answer: The website is not responsible for the privacy practices or
content of third-party websites. Even if our website contains links to
these external sites, we do not have control over their data
protection policies or the material they publish. Users should review
the individual privacy policies of any third-party websites they visit
```

```
to understand how their personal information may be handled.
----------------------------------------------------------------------
----------
Header: Changes to Privacy Policy
Question: When can the Privacy Policy be updated on the website?

Answer: The Privacy Policy can be updated on the website from time to
time. The company has the discretion to modify the policy as needed
and will post the updated version on the website when it is changed.
It is recommended that users periodically check the Privacy Policy on
the website to stay informed about any modifications.
----------------------------------------------------------------------
----------
Header: Contact Us
Question: How can I get in touch with Presight if I have questions
about the Privacy Policy?

Answer: You can contact Presight in two ways if you have questions
about their Privacy Policy. First, you can reach out to them through
the customer portal, which is a dedicated platform for interacting
with their support team. Alternatively, you can send an email to
presight@presight.io. This email address is specifically designated
for privacy policy inquiries, so you can expect a detailed and
relevant response.
----------------------------------------------------------------------
----------
Header: Purposeful Use Only
Question: What is the commitment of Presight.io regarding the use of
personal information?

Answer: Presight.io is committed to using personal information solely
for the purposes outlined in their privacy policy. This means they
will not use personal data for any other reasons than those explicitly
stated in their policy. This demonstrates a responsible and
transparent approach to handling personal information, which is
essential in today's digital world. The company's commitment to this
practice ensures that users can trust them with their personal data,
fostering a stronger relationship between the company and its users.
----------------------------------------------------------------------
----------
```

```python
import pandas as pd
QA=pd.read_csv('questions_answers.csv')

QA
```

```
                                             Question  \
0   What is Presight's commitment to privacy as st...
1   What is the purpose of information collection ...
2   What types of personal data and usage data are...
```

```
3    What are the purposes for which Data Presight ...
4    Why will I be asked to confirm that my persona...
5    How can I access my personal information held ...
6    What is the purpose of the automated edit chec...
7    To whom does the company disclose application ...
8    Will my personal data be shared or transferred...
9    What is the policy regarding the use of Google...
10   How is data secured in this system, both when ...
11   How long is customer data retained after an ac...
12   What is the role of data subjects in maintaini...
13   What measures does this entity take to ensure ...
14   What are cookies used for on this website and ...
15   What is the responsibility of the website rega...
16   When can the Privacy Policy be updated on the ...
17   How can I get in touch with Presight if I have...
18   What is the commitment of Presight.io regardin...

                                                Answer
0    Presight is committed to protecting the privac...
1    The purpose of information collection in the p...
2    When using the service, the following personal...
3    Data Presight uses the collected data for seve...
4    The text indicates that you will be asked to c...
5    The text states that you have the right to acc...
6    The automated edit checks employed by Presight...
7    The company discloses application data to thir...
8    No, your personal data will not be subject to ...
9    The policy clearly states that the system does...
10   In this system, data security is taken very se...
11   After a customer voluntarily closes their acco...
12   Data subjects play a significant role in ensur...
13   This entity is committed to protecting your pe...
14   Cookies on this website are used to enhance yo...
15   The website is not responsible for the privacy...
16   The Privacy Policy can be updated on the websi...
17   You can contact Presight in two ways if you ha...
18   Presight.io is committed to using personal inf...
```

Collect answer from Chatbot

Indexing by topic

```python
QA = pd.read_csv('questions_answers.csv')

answers = []

# Iterate through the questions
for _, row in QA.iterrows():
    question = row['Question']
    inputs = {
        "input": question
    }

    response = rag_pipeline.invoke(inputs)

    # Extract the answer from the response
    answer = response.get('answer', 'No answer found')

    # Append the answer to the list
    answers.append(answer)

QA['LLM_Answer'] = answers

QA.to_csv('questions_answers_with_llm_dtbtopic.csv', index=False)

QA
```

```
                                             Question  \
0   What is Presight's commitment to privacy as st...
1   What is the purpose of information collection ...
2   What types of personal data and usage data are...
3   What are the purposes for which Data Presight ...
4   Why will I be asked to confirm that my persona...
5   How can I access my personal information held ...
6   What is the purpose of the automated edit chec...
7   To whom does the company disclose application ...
8   Will my personal data be shared or transferred...
9   What is the policy regarding the use of Google...
10  How is data secured in this system, both when ...
11  How long is customer data retained after an ac...
12  What is the role of data subjects in maintaini...
13  What measures does this entity take to ensure ...
14  What are cookies used for on this website and ...
15  What is the responsibility of the website rega...
16  When can the Privacy Policy be updated on the ...
17  How can I get in touch with Presight if I have...
18  What is the commitment of Presight.io regardin...

                                               Answer  \
0   Presight is committed to protecting the privac...
1   The purpose of information collection in the p...
2   When using the service, the following personal...
3   Data Presight uses the collected data for seve...
4   The text indicates that you will be asked to c...
5   The text states that you have the right to acc...
6   The automated edit checks employed by Presight...
7   The company discloses application data to thir...
8   No, your personal data will not be subject to ...
9   The policy clearly states that the system does...
```

```
10  In this system, data security is taken very se...
11  After a customer voluntarily closes their acco...
12  Data subjects play a significant role in ensur...
13  This entity is committed to protecting your pe...
14  Cookies on this website are used to enhance yo...
15  The website is not responsible for the privacy...
16  The Privacy Policy can be updated on the websi...
17  You can contact Presight in two ways if you ha...
18  Presight.io is committed to using personal inf...

                                           LLM_Answer
0   According to the provided context, Presight's ...
1   The purpose of information collection is to pr...
2   According to the context, the types of persona...
3   According to the provided context, Data Presig...
4   You will be asked to confirm that your persona...
5   You can access your personal information held ...
6   The purpose of the automated edit checks used ...
7   According to the provided context, the company...
8   According to the provided context, your person...
9   According to the policy, Presight.io does not ...
10  According to our data security practices, all ...
11  According to the context, customer data is ret...
12  According to the context, data subjects have t...
13  According to the provided context, this entity...
14  According to the website's policy, cookies are...
15  According to the context, the website is not r...
16  According to the provided context, the Privacy...
17  You can get in touch with Presight if you have...
18  According to Presight.io's policy, the commitm...
```

Indexing by chunksize

```python
qdrant_collection_name="Python_simple_rag"
rag_pipeline = rag_pipeline_setup.rag(source=qdrant_collection_name)
```

```
Fetching 1 files: 100%|██████████| 1/1 [00:00<00:00, 1060.77it/s]
```

```python
answers = []

# Iterate through the questions
for _, row in QA.iterrows():
    question = row['Question']

    inputs = {
        "input": question
    }

    response = rag_pipeline.invoke(inputs)
    # Extract the answer from the response
    answer = response.get('answer', 'No answer found')

    # Append the answer to the list
    answers.append(answer)

QA['LLM_Answer'] = answers

# Optionally, save the updated DataFrame to a new CSV file
QA.to_csv('questions_answers_with_llm_dtbchunksize.csv', index=False)

# Print the DataFrame with the answers
QA
```

```
                                             Question  \
0   What is Presight's commitment to privacy as st...
1   What is the purpose of information collection ...
2   What types of personal data and usage data are...
3   What are the purposes for which Data Presight ...
4   Why will I be asked to confirm that my persona...
5   How can I access my personal information held ...
6   What is the purpose of the automated edit chec...
7   To whom does the company disclose application ...
8   Will my personal data be shared or transferred...
9   What is the policy regarding the use of Google...
10  How is data secured in this system, both when ...
11  How long is customer data retained after an ac...
12  What is the role of data subjects in maintaini...
13  What measures does this entity take to ensure ...
14  What are cookies used for on this website and ...
15  What is the responsibility of the website rega...
16  When can the Privacy Policy be updated on the ...
17  How can I get in touch with Presight if I have...
18  What is the commitment of Presight.io regardin...

                                               Answer  \
0   Presight is committed to protecting the privac...
1   The purpose of information collection in the p...
2   When using the service, the following personal...
3   Data Presight uses the collected data for seve...
4   The text indicates that you will be asked to c...
5   The text states that you have the right to acc...
6   The automated edit checks employed by Presight...
7   The company discloses application data to thir...
8   No, your personal data will not be subject to ...
9   The policy clearly states that the system does...
10  In this system, data security is taken very se...
11  After a customer voluntarily closes their acco...
12  Data subjects play a significant role in ensur...
13  This entity is committed to protecting your pe...
```

```
14  Cookies on this website are used to enhance yo...
15  The website is not responsible for the privacy...
16  The Privacy Policy can be updated on the websi...
17  You can contact Presight in two ways if you ha...
18  Presight.io is committed to using personal inf...

                                           LLM_Answer
0   According to the provided context, Presight's ...
1   The purpose of information collection is to pr...
2   According to the provided context, the types o...
3   According to the context, Presight utilizes th...
4   You will be asked to confirm that your persona...
5   You can access your personal information held ...
6   The purpose of the automated edit checks used ...
7   According to the provided context, the company...
8   According to our Privacy Policy, your personal...
9   According to our policy, we do not use Google ...
10  According to our security measures, all data i...
11  According to our data retention policy, custom...
12  According to our data management practices, da...
13  According to our privacy policy, we regularly ...
14  According to our website's policy, cookies are...
15  According to the provided context, the website...
16  According to the provided context, the Privacy...
17  According to the provided Privacy Policy, if y...
18  According to the provided Privacy Policy, Pres...
```

2.3 Evaluation - Bert Score

BERTScore

BERTScore is a method for evaluating the similarity between texts based on the BERT language model. It uses embeddings generated by BERT or similar language models to compare sentences or text passages, allowing for an evaluation of text quality based on semantic meaning rather than just keyword matching.

How it works:

1. **Embedding Calculation**:
   - BERTScore uses the BERT model to generate embeddings for words in the sentence, and then calculates the similarity between corresponding words in the reference sentence and the candidate sentence.
2. **Word-to-Word Comparison**:
   - To evaluate the similarity between the reference sentence and the candidate sentence, BERTScore uses **cosine similarity** between the embeddings of words in each sentence.
3. **BERTScore Metrics**:
   - **Precision**: Calculates the similarity between words in the candidate sentence and the reference sentence.

   - **Recall**: Calculates the similarity between words in the reference sentence and the candidate sentence.
   - **F1 Score**: Derived from precision and recall, it combines both metrics to provide an overall evaluation of the candidate sentence's quality.
4. **Benefits**:
   - **Semantic Evaluation**: BERTScore outperforms traditional methods like BLEU and ROUGE because it considers the meaning of the text rather than just keyword matching.
   - **More Accurate Evaluation**: BERTScore can detect synonymous sentences, where different structures convey the same meaning.
   - **Practicality**: BERTScore is easily applicable to various NLP tasks such as evaluating sentence quality, machine translation, and content generation.

BERTScore Formulas:

- **Precision**: Calculated using cosine similarity between embeddings of words in the candidate sentence and the reference sentence.

$$\text{Precision} = \frac{1}{|C|} \sum_{w_i \in C} \max_{w_j \in R} \text{cosine\_similarity}\left(w_i, w_j\right)$$

- **Recall**: Calculates the similarity between words in the reference sentence and the candidate sentence.

$$\text{Recall} = \frac{1}{|R|} \sum_{w_i \in R} \max_{w_j \in C} \text{cosine\_similarity}\left(w_i, w_j\right)$$

- **F1 Score**: The harmonic mean of Precision and Recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Advantages of BERTScore:

- **Vocabulary Independence**: BERTScore doesn't rely on a fixed vocabulary, reducing issues related to vocabulary shifts between sentences.
- **Semantic Handling**: Since BERTScore is based on BERT, it better captures the meaning of the sentence, rather than just focusing on specific words.

Applications:

- BERTScore is widely used in **evaluating machine translation quality**, **automatic text generation**, **question-answering systems**, and **text summarization**.

Eval Chunksize Verse

```
QA=pd.read_csv('questions_answers_with_llm_dtbchunksize.csv')

!pip install bert-score
```

```
Collecting bert-score
  Downloading bert_score-0.3.13-py3-none-any.whl.metadata (15 kB)
Requirement already satisfied: torch>=1.0.0 in
/usr/local/lib/python3.10/dist-packages (from bert-score)
(2.5.1+cu121)
Requirement already satisfied: pandas>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from bert-score) (2.2.2)
Requirement already satisfied: transformers>=3.0.0 in
/usr/local/lib/python3.10/dist-packages (from bert-score) (4.47.1)
Requirement already satisfied: numpy in
/usr/local/lib/python3.10/dist-packages (from bert-score) (1.26.4)
Requirement already satisfied: requests in
/usr/local/lib/python3.10/dist-packages (from bert-score) (2.32.3)
Requirement already satisfied: tqdm>=4.31.1 in
/usr/local/lib/python3.10/dist-packages (from bert-score) (4.67.1)
Requirement already satisfied: matplotlib in
/usr/local/lib/python3.10/dist-packages (from bert-score) (3.10.0)
Requirement already satisfied: packaging>=20.9 in
/usr/local/lib/python3.10/dist-packages (from bert-score) (24.2)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.10/dist-packages (from pandas>=1.0.1->bert-
score) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.10/dist-packages (from pandas>=1.0.1->bert-
score) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in
/usr/local/lib/python3.10/dist-packages (from pandas>=1.0.1->bert-
score) (2024.2)
Requirement already satisfied: filelock in
/usr/local/lib/python3.10/dist-packages (from torch>=1.0.0->bert-
score) (3.16.1)
Requirement already satisfied: typing-extensions>=4.8.0 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.0.0->bert-
score) (4.12.2)
Requirement already satisfied: networkx in
/usr/local/lib/python3.10/dist-packages (from torch>=1.0.0->bert-
score) (3.4.2)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.0.0->bert-
score) (3.1.5)
Requirement already satisfied: fsspec in
/usr/local/lib/python3.10/dist-packages (from torch>=1.0.0->bert-
score) (2024.10.0)
Requirement already satisfied: sympy==1.13.1 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.0.0->bert-
score) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.10/dist-packages (from sympy==1.13.1-
>torch>=1.0.0->bert-score) (1.3.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.24.0 in
```

```
/usr/local/lib/python3.10/dist-packages (from transformers>=3.0.0-
>bert-score) (0.27.1)
Requirement already satisfied: pyyaml>=5.1 in
/usr/local/lib/python3.10/dist-packages (from transformers>=3.0.0-
>bert-score) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.10/dist-packages (from transformers>=3.0.0-
>bert-score) (2024.11.6)
Requirement already satisfied: tokenizers<0.22,>=0.21 in
/usr/local/lib/python3.10/dist-packages (from transformers>=3.0.0-
>bert-score) (0.21.0)
Requirement already satisfied: safetensors>=0.4.1 in
/usr/local/lib/python3.10/dist-packages (from transformers>=3.0.0-
>bert-score) (0.5.0)
Requirement already satisfied: contourpy>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->bert-score)
(1.3.1)
Requirement already satisfied: cycler>=0.10 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->bert-score)
(0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->bert-score)
(4.55.3)
Requirement already satisfied: kiwisolver>=1.3.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->bert-score)
(1.4.8)
Requirement already satisfied: pillow>=8 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->bert-score)
(11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->bert-score)
(3.2.1)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests->bert-score)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.10/dist-packages (from requests->bert-score)
(3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests->bert-score)
(2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests->bert-score)
(2024.12.14)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2-
>pandas>=1.0.1->bert-score) (1.17.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from jinja2->torch>=1.0.0-
>bert-score) (3.0.2)
```

```
Downloading bert_score-0.3.13-py3-none-any.whl (61 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 61.1/61.1 kB 3.3 MB/s eta
0:00:00
```

```python
from bert_score import score

def compute_bertscore(row):
    P, R, F1 = score([row['LLM_Answer']], [row['Answer']], lang="en",
verbose=True)
    return pd.Series([P.mean().item(), R.mean().item(),
F1.mean().item()], index=["Precision", "Recall", "F1"])


QA[['Precision', 'Recall', 'F1']] = QA.apply(compute_bertscore,
axis=1)


print(QA[['Question', 'Precision', 'Recall', 'F1']])
```

```
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.

{"model_id":"d43c83fd2dc64f9495ecae55960eab5d","version_major":2,"vers
ion_minor":0}

computing greedy matching.

{"model_id":"6bf39b2a01f349e580664e8d9ea5c384","version_major":2,"vers
ion_minor":0}

done in 1.42 seconds, 0.71 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.

{"model_id":"5358bddcfb9b4b3aa32e518e06522612","version_major":2,"vers
ion_minor":0}

computing greedy matching.
```

```
{"model_id":"d8a4e5b18c59400eb48cf5295bb5ecc7","version_major":2,"version_minor":0}

done in 2.06 seconds, 0.49 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.

{"model_id":"751b380ad8304c5684ab38c05f8e5a92","version_major":2,"version_minor":0}

computing greedy matching.

{"model_id":"6b5d2f7dc3ae4716913dc9ecadf7b536","version_major":2,"version_minor":0}

done in 2.79 seconds, 0.36 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.

{"model_id":"d734a0b7d6524877a7cb2c7f9089b34d","version_major":2,"version_minor":0}

computing greedy matching.

{"model_id":"4c84876693e24669a03b71fd83a8e1d8","version_major":2,"version_minor":0}

done in 3.79 seconds, 0.26 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.
```

```
{"model_id":"c48266fe59074a1b857a5ec38d801adb","version_major":2,"version_minor":0}

computing greedy matching.

{"model_id":"411c67cb9fd34d62aa4ccec284959467","version_major":2,"version_minor":0}

done in 2.57 seconds, 0.39 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.

{"model_id":"fd0967f4cffe4f62974982000afc4188","version_major":2,"version_minor":0}

computing greedy matching.

{"model_id":"c793530615d64f8f96922e291cae03c2","version_major":2,"version_minor":0}

done in 1.38 seconds, 0.73 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.

{"model_id":"45b773bbedd04b2791eadb4cc37b00cc","version_major":2,"version_minor":0}

computing greedy matching.

{"model_id":"67d93d93b6d040a6af61a50bbd326f3f","version_major":2,"version_minor":0}

done in 1.45 seconds, 0.69 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
```

```
calculating scores...
computing bert embedding.

{"model_id":"7103d4beaabb4629a8c7521a23b5d73d","version_major":2,"vers
ion_minor":0}

computing greedy matching.

{"model_id":"d107673f65434b1691c979f71bb163c6","version_major":2,"vers
ion_minor":0}

done in 2.24 seconds, 0.45 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.

{"model_id":"f6a2b10fbecb49db98740d6d1dd2f130","version_major":2,"vers
ion_minor":0}

computing greedy matching.

{"model_id":"097950bbc5384122b667a76ff461a884","version_major":2,"vers
ion_minor":0}

done in 1.17 seconds, 0.86 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.

{"model_id":"7af81268b7bf4e739cc911154a81e5dc","version_major":2,"vers
ion_minor":0}

computing greedy matching.

{"model_id":"ba24f0f1792a417f8404b94694552dde","version_major":2,"vers
ion_minor":0}

done in 4.91 seconds, 0.20 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
```

```
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.

{"model_id":"94946fe129794665b80a1acb24f18676","version_major":2,"vers
ion_minor":0}

computing greedy matching.

{"model_id":"108e38a8b35f42b49767d2aea0c57fe6","version_major":2,"vers
ion_minor":0}

done in 3.74 seconds, 0.27 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.

{"model_id":"9a3f98fcae084b51bd3f5f5c5be25319","version_major":2,"vers
ion_minor":0}

computing greedy matching.

{"model_id":"7291720f43644f0b918660dbafafd706","version_major":2,"vers
ion_minor":0}

done in 0.91 seconds, 1.10 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.

{"model_id":"ef4fcf0596b2483e8ddf0acb9ce2d519","version_major":2,"vers
ion_minor":0}

computing greedy matching.

{"model_id":"cf29fc2175494284aa9c63c96c3e02be","version_major":2,"vers
ion_minor":0}
```

```
done in 1.91 seconds, 0.52 sentences/sec
```

```
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
```

```
calculating scores...
computing bert embedding.
```

{"model_id":"5a68391419d84a298f080a0c5b7a9f96","version_major":2,"version_minor":0}

```
computing greedy matching.
```

{"model_id":"a2cbf3b8f84a4eebadb5f897052f6469","version_major":2,"version_minor":0}

```
done in 2.30 seconds, 0.44 sentences/sec
```

```
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
```

```
calculating scores...
computing bert embedding.
```

{"model_id":"0c2dd540ae894a4e82f19a53aac2e925","version_major":2,"version_minor":0}

```
computing greedy matching.
```

{"model_id":"ce9b9f39e4a44c15b46634fc243c0a73","version_major":2,"version_minor":0}

```
done in 2.11 seconds, 0.47 sentences/sec
```

```
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
```

```
calculating scores...
computing bert embedding.
```

{"model_id":"ddb33f4f4433412284fabaa66a335495","version_major":2,"version_minor":0}

```
computing greedy matching.
```

```
{"model_id":"df73f8432c0d4f39aae5b675736f18f0","version_major":2,"vers
ion_minor":0}

done in 2.66 seconds, 0.38 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.

{"model_id":"c708ac690b7446868b856245b5d8919b","version_major":2,"vers
ion_minor":0}

computing greedy matching.

{"model_id":"f70d6433e8374dfdbf2c932f06f70776","version_major":2,"vers
ion_minor":0}

done in 2.24 seconds, 0.45 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.

{"model_id":"81d2bb2e185e4731b2a425568d02b43d","version_major":2,"vers
ion_minor":0}

computing greedy matching.

{"model_id":"f303fd856fe846a6aef13707917fd1e4","version_major":2,"vers
ion_minor":0}

done in 4.82 seconds, 0.21 sentences/sec

Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

calculating scores...
computing bert embedding.
```

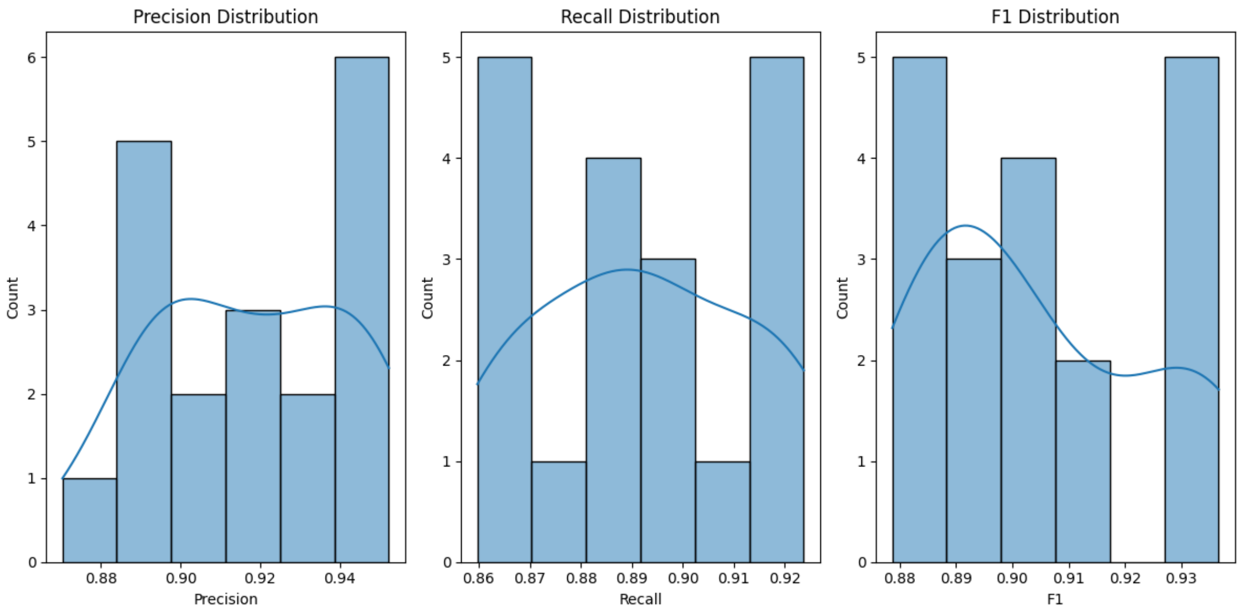
```
{"model_id":"3843d21dab214128a764184403ecf9dd","version_major":2,"vers
ion_minor":0}

computing greedy matching.

{"model_id":"8bee05f76d0848eabbc87860c41d4d8e","version_major":2,"vers
ion_minor":0}

done in 3.09 seconds, 0.32 sentences/sec
                                             Question  Precision
Recall  \
0   What is Presight's commitment to privacy as st...   0.903213
0.899906
1   What is the purpose of information collection ...   0.925107
0.863681
2   What types of personal data and usage data are...   0.887342
0.870203
3   What are the purposes for which Data Presight ...   0.896942
0.883106
4   Why will I be asked to confirm that my persona...   0.920337
0.906216
5   How can I access my personal information held ...   0.948165
0.922915
6   What is the purpose of the automated edit chec...   0.952372
0.915265
7   To whom does the company disclose application ...   0.932282
0.868267
8   Will my personal data be shared or transferred...   0.949958
0.923714
9   What is the policy regarding the use of Google...   0.941722
0.890742
10  How is data secured in this system, both when ...   0.918203
0.859742
11  How long is customer data retained after an ac...   0.940812
0.921975
12  What is the role of data subjects in maintaini...   0.943113
0.916499
13  What measures does this entity take to ensure ...   0.895074
0.901182
14  What are cookies used for on this website and ...   0.904275
0.862858
15  What is the responsibility of the website rega...   0.915620
0.893295
16  When can the Privacy Policy be updated on the ...   0.891050
0.885671
17  How can I get in touch with Presight if I have...   0.870295
0.888311
18  What is the commitment of Presight.io regardin...   0.892899
0.880587
```

```
          F1
0   0.901557
1   0.893339
2   0.878689
3   0.889971
4   0.913222
5   0.935369
6   0.933450
7   0.899136
8   0.936652
9   0.915523
10  0.888012
11  0.931298
12  0.929615
13  0.898117
14  0.883081
15  0.904320
16  0.888353
17  0.879211
18  0.886700
```

QA

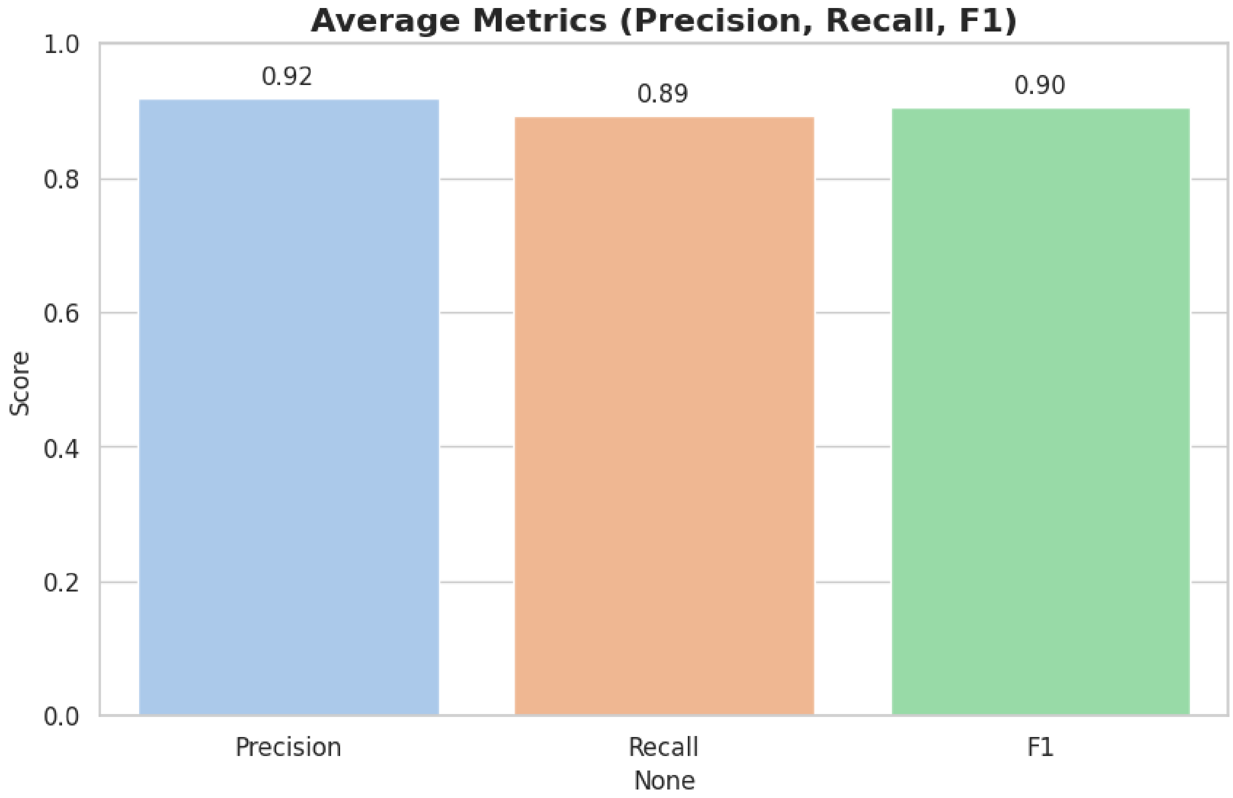
```
{"summary":"{\n  \"name\": \"QA\",\n  \"rows\": 19,\n  \"fields\": [\n
{\n      \"column\": \"Question\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 19,\n
\"samples\": [\n          \"What is Presight's commitment to privacy
as stated in the text?\",\n          \"How can I access my personal
information held by the application?\",\n          \"How long is
customer data retained after an account is closed?\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Answer\",\n      \"properties\":
{\n        \"dtype\": \"string\",\n        \"num_unique_values\": 19,\
n        \"samples\": [\n          \"Presight is committed to
protecting the privacy of its customers and visitors to its website.
This commitment is outlined in their Privacy Policy, which explains
how they collect, use, and disclose information about their customers
and visitors.\",\n          \"The text states that you have the right
to access all of your personal information that the application holds.
To do this, you need to log into the application. Once you are logged
in, navigate to your settings and profile. From there, you will be
able to access your personal information and make any corrections,
amendments, or additions as necessary.\",\n          \"After a
customer voluntarily closes their account, the data enters an
\\\"expired\\\" state and is retained for 60 days. Once this 60-day
period has elapsed, the account and all related data are removed.\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"LLM_Answer\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 19,\n        \"samples\": [\n
\"According to the provided context, Presight's commitment to privacy
is to \\\"protect the privacy of our customers and visitors to our
website.\\\"\",\n          \"You can access your personal information
held by the application by logging into the application and navigating
to your settings and profile. From there, you can correct, amend, or
append your personal information as needed.\",\n          \"According
to our data retention policy, customer data is retained for 60 days
after an account is closed. After this period, the account and related
data will be removed.\"\n        ],\n        \"semantic_type\": \"\",\
n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Precision\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 0.024755323845372972,\n
\"min\": 0.8702953457832336,\n        \"max\": 0.9523715972900391,\n
\"num_unique_values\": 19,\n        \"samples\": [\n
0.9032131433486938,\n          0.9481645822525024,\n
0.9408115148544312\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Recall\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0.021531344765895383,\n        \"min\": 0.8597424030303955,\n
\"max\": 0.9237141013145447,\n        \"num_unique_values\": 19,\n
\"samples\": [\n          0.8999060392379761,\n
0.9229145050048828,\n          0.9219750761985779\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"F1\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 0.02027458981431754,\n
\"min\": 0.8786889910697937,\n        \"max\": 0.9366522431373596,\n
\"num_unique_values\": 19,\n        \"samples\": [\n
0.9015565514564514,\n          0.9353691935539246,\n
0.9312980771064758\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"QA"}
```

```
import matplotlib.pyplot as plt
import seaborn as sns
from bert_score import score

metrics = ['Precision', 'Recall', 'F1']

plt.figure(figsize=(12, 6))

for i, metric in enumerate(metrics):
    plt.subplot(1, 3, i+1)
    sns.histplot(QA[metric], kde=True)
    plt.title(f'{metric} Distribution')

plt.tight_layout()
plt.show()
```



```
sns.set_theme(style="whitegrid")

average_metrics = QA[['Precision', 'Recall', 'F1']].mean()

plt.figure(figsize=(10, 6))
sns.barplot(x=average_metrics.index, y=average_metrics.values,
palette='pastel')

plt.title('Average Metrics (Precision, Recall, F1)', fontsize=16,
fontweight='bold')
plt.ylabel('Score', fontsize=12)
plt.ylim(0, 1)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)

for i, value in enumerate(average_metrics.values):
    plt.text(i, value + 0.02, f'{value:.2f}', ha='center',
fontsize=12)

# Hiển thị biểu đồ
plt.show()
```

```
<ipython-input-29-4cc8fe3366a0>:12: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  sns.barplot(x=average_metrics.index, y=average_metrics.values,
palette='pastel')
```



```
average_metrics = QA[['Precision', 'Recall', 'F1']].mean()
print("Average Precision:", average_metrics['Precision'])
print("Average Recall:", average_metrics['Recall'])
print("Average F1 Score:", average_metrics['F1'])
```

```
Average Precision: 0.9173041487994947
Average Recall: 0.892322910459418
Average F1 Score: 0.904506012013084
```

Eval Topic Verse

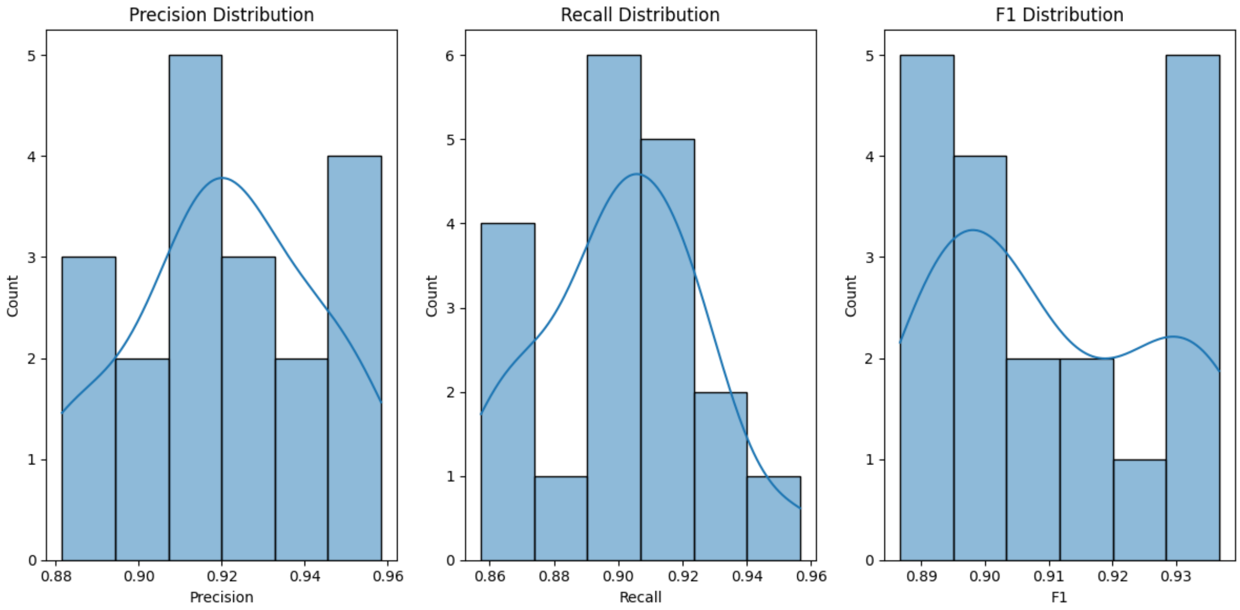
```
QA1=pd.read_csv('/content/questions_answers_with_llm_dtbtopic.csv')

from bert_score import score

def compute_bertscore(row):
    P, R, F1 = score([row['LLM_Answer']], [row['Answer']], lang="en",
verbose=False)
    return pd.Series([P.mean().item(), R.mean().item(),
F1.mean().item()], index=["Precision", "Recall", "F1"])

QA1[['Precision', 'Recall', 'F1']] = QA1.apply(compute_bertscore,
axis=1)
print(QA1[['Question', 'Precision', 'Recall', 'F1']])
```

```
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
```

```
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model
checkpoint at roberta-large and are newly initialized:
['roberta.pooler.dense.bias', 'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

                                             Question  Precision
Recall  \
0   What is Presight's commitment to privacy as st...   0.881522
0.919768
```

```
1   What is the purpose of information collection ...   0.925107
0.863681
2   What types of personal data and usage data are...   0.883264
0.904956
3   What are the purposes for which Data Presight ...   0.903239
0.895055
4   Why will I be asked to confirm that my persona...   0.926444
0.907010
5   How can I access my personal information held ...   0.948165
0.922915
6   What is the purpose of the automated edit chec...   0.951104
0.909682
7   To whom does the company disclose application ...   0.935201
0.897993
8   Will my personal data be shared or transferred...   0.883455
0.927413
9   What is the policy regarding the use of Google...   0.911787
0.867348
10  How is data secured in this system, both when ...   0.918165
0.857337
11  How long is customer data retained after an ac...   0.937892
0.924377
12  What is the role of data subjects in maintaini...   0.948584
0.918683
13  What measures does this entity take to ensure ...   0.917700
0.956552
14  What are cookies used for on this website and ...   0.915412
0.869793
15  What is the responsibility of the website rega...   0.921033
0.893140
16  When can the Privacy Policy be updated on the ...   0.902070
0.899059
17  How can I get in touch with Presight if I have...   0.958394
0.898630
18  What is the commitment of Presight.io regardin...   0.918409
0.881561

          F1
0   0.900239
1   0.893339
2   0.893978
3   0.899128
4   0.916624
5   0.935369
6   0.929932
7   0.916219
8   0.904901
9   0.889013
10  0.886709
```

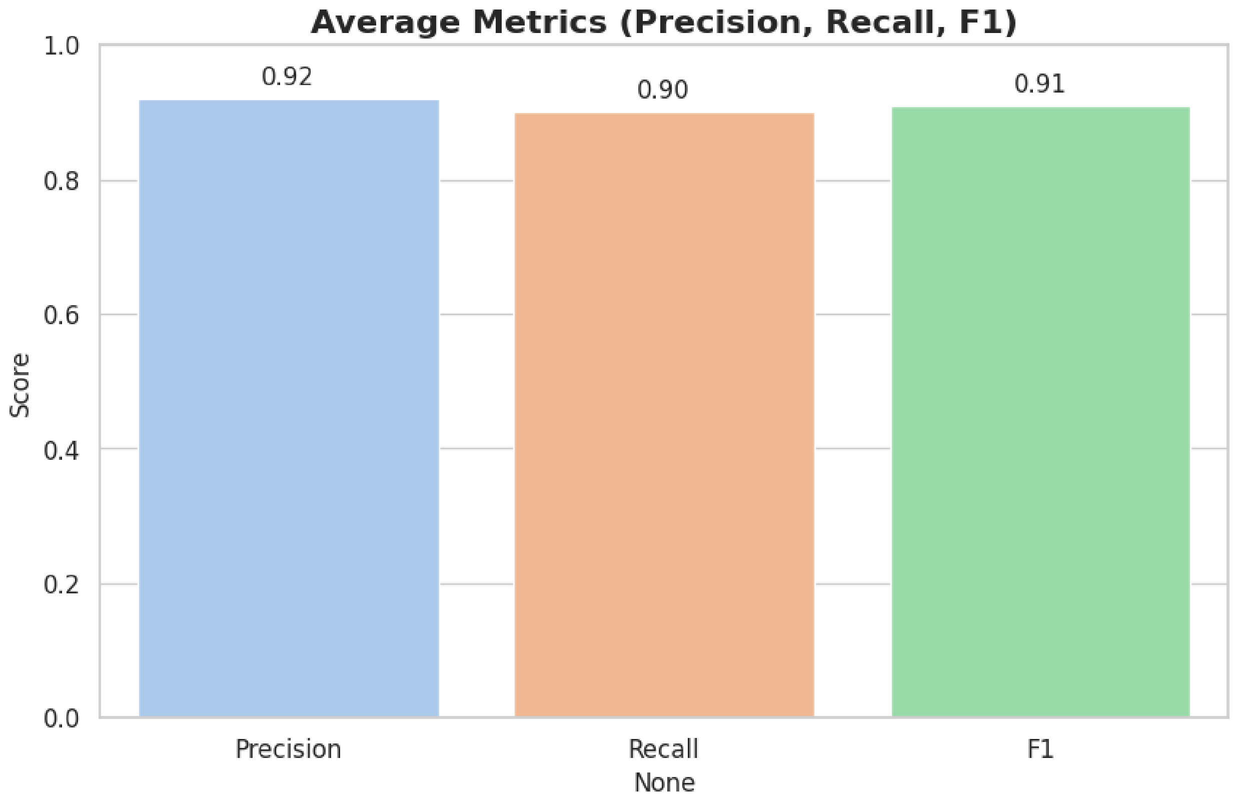
```
11   0.931085
12   0.933394
13   0.936724
14   0.892019
15   0.906872
16   0.900562
17   0.927550
18   0.899608

metrics = ['Precision', 'Recall', 'F1']

plt.figure(figsize=(12, 6))

for i, metric in enumerate(metrics):
    plt.subplot(1, 3, i+1)
    sns.histplot(QA1[metric], kde=True)
    plt.title(f'{metric} Distribution')

plt.tight_layout()
plt.show()
```



```
sns.set_theme(style="whitegrid")

average_metrics = QA1[['Precision', 'Recall', 'F1']].mean()

plt.figure(figsize=(10, 6))
sns.barplot(x=average_metrics.index, y=average_metrics.values,
palette='pastel')

plt.title('Average Metrics (Precision, Recall, F1)', fontsize=16,
fontweight='bold')
```

```
plt.ylabel('Score', fontsize=12)
plt.ylim(0, 1)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)

for i, value in enumerate(average_metrics.values):
    plt.text(i, value + 0.02, f'{value:.2f}', ha='center',
fontsize=12)

plt.show()
```

```
<ipython-input-28-c6e0f1f68935>:12: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  sns.barplot(x=average_metrics.index, y=average_metrics.values,
palette='pastel')
```



```
average_metrics = QA1[['Precision', 'Recall', 'F1']].mean()
print("Average Precision:", average_metrics['Precision'])
print("Average Recall:", average_metrics['Recall'])
print("Average F1 Score:", average_metrics['F1'])
```

```
Average Precision: 0.9203656347174394
Average Recall: 0.9007868515817743
Average F1 Score: 0.9101718726911043
```

Conclusion

Comments on Indexing Results

The results indicate that **topic-based indexing** performs better than **chunksize-based indexing** in this case.

Reasons:

1. **Data Consistency**:
   - When indexing by **topic**, documents that are semantically related are grouped together, making information retrieval more accurate and easier.
   - Grouping documents by topic ensures that queries find documents with the same context, which improves the relevance of the search results.
2. **Scalability and Flexibility**:
   - Topic-based indexing allows for easy expansion with new topics, while chunksize-based indexing may struggle to handle long documents or documents with high complexity.
   - Using topics as indexes also helps in managing documents more efficiently when needing to classify and organize data into semantic groups.
3. **Query Efficiency**:
   - When using **chunksize** as an index, queries may return irrelevant documents or lack semantic depth, lowering the quality of search results.
   - On the other hand, topic-based indexing improves the relevance of search results, particularly when users want to search for detailed information on a specific topic.

Conclusion:

With this dataset, **topic-based indexing** provides clear benefits in improving accuracy and efficiency of querying, as well as enhancing document organization and management. Therefore, topic-based indexing is the better choice compared to chunksize-based indexing in this specific case.

Drawbacks of Topic-Based Indexing When Scaling Up

When scaling up and applying **topic-based indexing**, there are several important drawbacks to consider, especially when working with large data chunks or uneven topics. One significant issue is:

1. **Uneven Distribution Across Chunks**:

- When documents are classified by **topic**, not every chunk contains evenly distributed information about a specific topic. This can result in **inconsistent context** in each chunk, making it more difficult to handle the data when performing a search.
- **Uneven information depth and length**: Some topics may have a large volume of data, while others may have less, causing an imbalance across chunks. Chunks with insufficient or incomplete information about a topic may reduce the effectiveness of the query, especially when inputting data into models like LLMs.

2. **Difficulties in Managing Context for LLMs**:

- When context is uneven across chunks, providing the full and accurate information to a **Large Language Model (LLM)** becomes more challenging. Some chunks may lack essential parts of the context needed for the LLM to generate a precise response, while others might contain excessive, irrelevant information.
- **Context overload**: Feeding too much or too little information into the LLM can reduce the model's effectiveness, particularly when there isn't a clear method for adjusting and selecting the appropriate context.

3. **Challenges in Selecting the Right Context**:

- With chunks containing different topics, choosing the correct context for the LLM becomes more complex, as the model may not fully understand the broader context of the topic. The lack of a filtering or selection strategy could result in important information being lost, affecting the quality of the LLM's output.

4. **Difficulty in Optimizing Search**:

- Searching for relevant documents within a set of uneven chunks can lead to inaccurate results. Some chunks may contain rich information about a specific topic, while others may have little to no relevance. This creates inconsistency in search results and reduces the overall effectiveness of the system.

Conclusion:

When using **topic-based indexing**, especially as the scale increases, issues like **uneven distribution across chunks** can affect the overall performance of the search system and the ability to provide context for LLMs. To mitigate these drawbacks, methods for efficient chunk segmentation and accurate context selection should be implemented, ensuring that each chunk contains sufficient and relevant information to support effective model operation.

Indexing Based on Chunksize

Chunksize-based indexing is a method where documents are split into smaller segments (chunks) based on a defined size, often with a specific number of characters or words. This approach is particularly useful in situations where the documents are too large or complex to process as a whole.

How It Works:

1. **Document Segmentation**:

- Documents are split into smaller, manageable chunks. Each chunk is indexed separately, making it easier to search through large documents without needing to process them all at once.
- This segmentation is typically done by character count or word count, with chunks designed to maintain a certain balance between size and relevance.

2. **Efficient Retrieval**:
   - By breaking documents into smaller pieces, the system can quickly retrieve the relevant chunk of text that matches the user's query. This makes the search process faster, especially for very large datasets.
   - Each chunk is indexed separately, which allows for targeted search within specific parts of the document.
3. **Flexibility**:
   - Chunksize indexing can adapt to documents of varying lengths. Since each document is divided into chunks, the system can scale with the size of the data and retrieve smaller, more focused pieces of information.
   - This flexibility is particularly useful when working with documents that have different formats or content structures.

Drawbacks:

1. **Loss of Context**:
   - One of the main challenges with chunksize-based indexing is the potential loss of context. A chunk may contain part of a concept or idea, but without the surrounding context, it may not make sense in relation to the full document.
   - Queries may return fragmented information that doesn't provide enough context, which could reduce the quality of search results.
2. **Reduced Semantic Relevance**:
   - Since chunks are created based on size rather than semantic meaning, the search results may lack the depth of understanding of the topic. This may lead to queries returning chunks that are technically relevant but not fully aligned with the user's intent.
3. **Difficulties in Managing Large Chunks**:
   - When dealing with large or complex documents, choosing the appropriate chunksize can be challenging. Too small a chunk may result in too many fragments, while too large a chunk may still miss context or detail.
   - Managing and organizing these chunks effectively requires more processing and careful consideration of chunk sizes to balance efficiency and relevance.

Conclusion:

While **chunksize-based indexing** offers flexibility and efficiency, particularly for large datasets, it may struggle with maintaining contextual relevance and semantic understanding. This approach is best suited for documents where context is less crucial, but care must be taken to ensure that the chunk sizes are optimized for the search system’s needs.