

---

---

# Understanding Retrieval Augmented Generation (RAG)

— By: Tahreem Rasul —

---

---

# About Me

- **Senior ML Engineer** at *Red Buffer*, a services-based ML and AI company
- Previously worked as a medical physicist in cancer radiotherapy research at Siemens Healthineers, Erlangen, Germany
- [LinkedIn](#)

## Educational Background:

- Masters in Medical AI from Erlangen, Germany
- Bachelors in Electrical Engineering from Islamabad, Pakistan

# Introduction

# Introduction to RAG

## What is RAG?

- Retrieval Augmented Generation (RAG) is a very popular paradigm.
- Combines retrieval mechanisms with generative models.
- Enables language models to access and use external data.

## Why RAG?

- LLMs (Large Language Models) may not have seen all relevant data.
- RAG allows incorporating private, latest, or specialized information.

# Overview of RAG Workflow

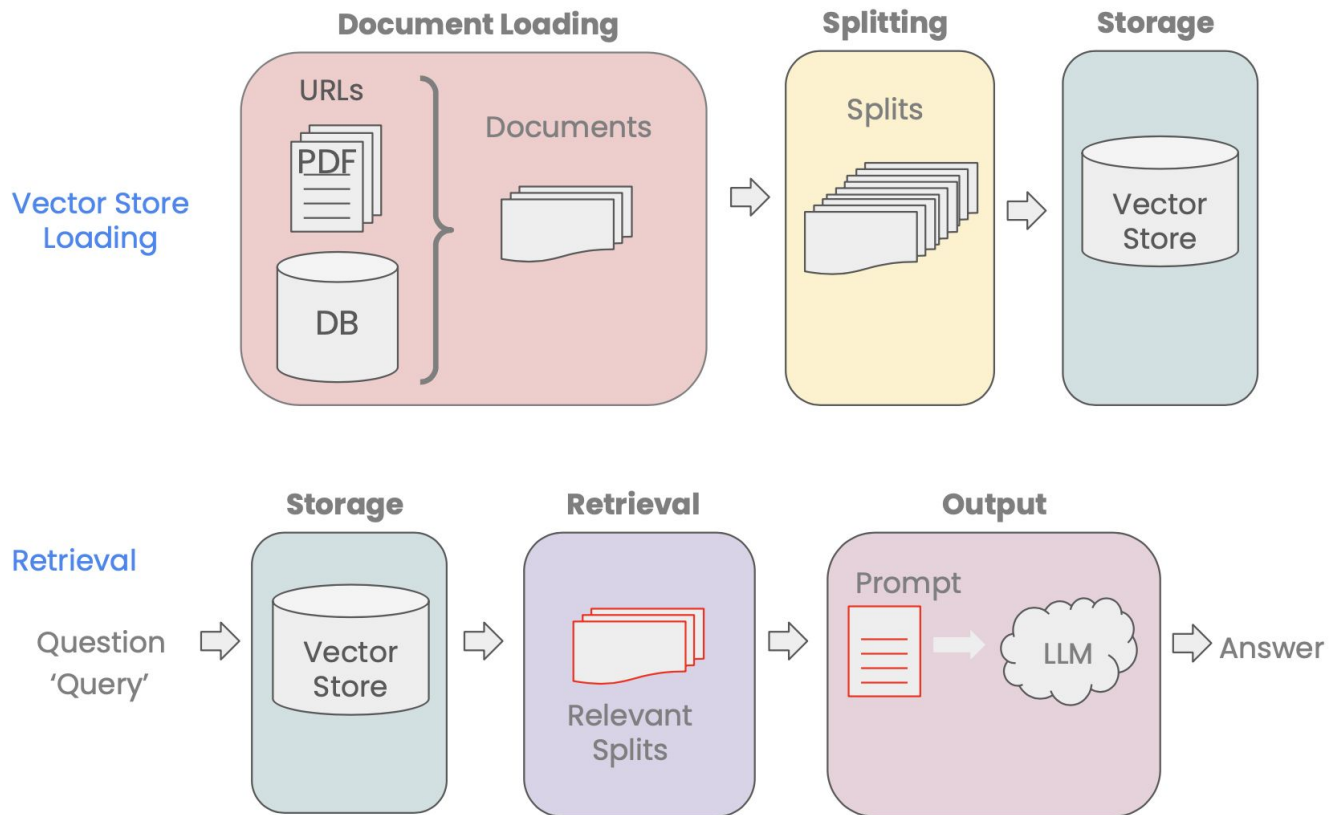
# RAG Pipeline

Retrieves relevant documents and loads into working memory/context window

Three Main Components:

1. Indexing/Storage
2. Retrieval
3. Generation

# RAG Pipeline



# Indexing



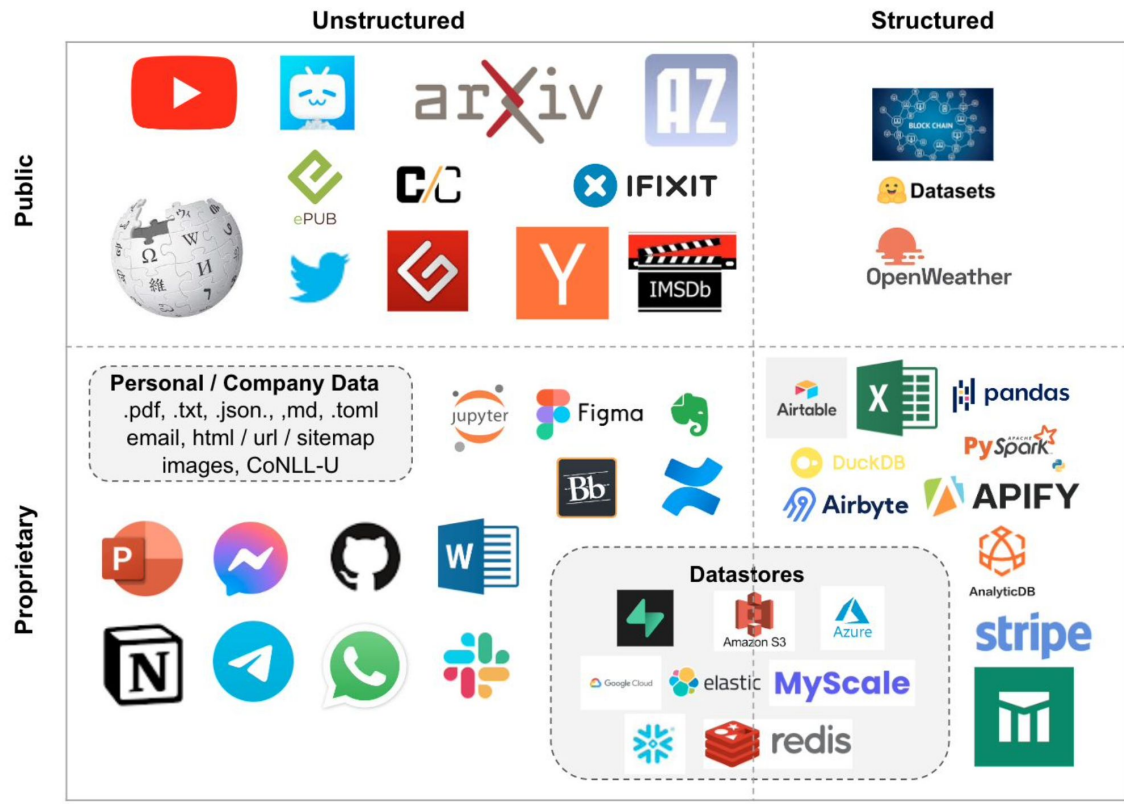
# LangChain Loaders

Loaders deal with the specifics of accessing and converting data

- Accessing
  - Web Sites
  - Data Bases
  - Youtube
  - arXiv
  - ...
- Data Types
  - PDF
  - HTML
  - JSON
  - Word, PowerPoint, ...

Returns a list of `Document` objects

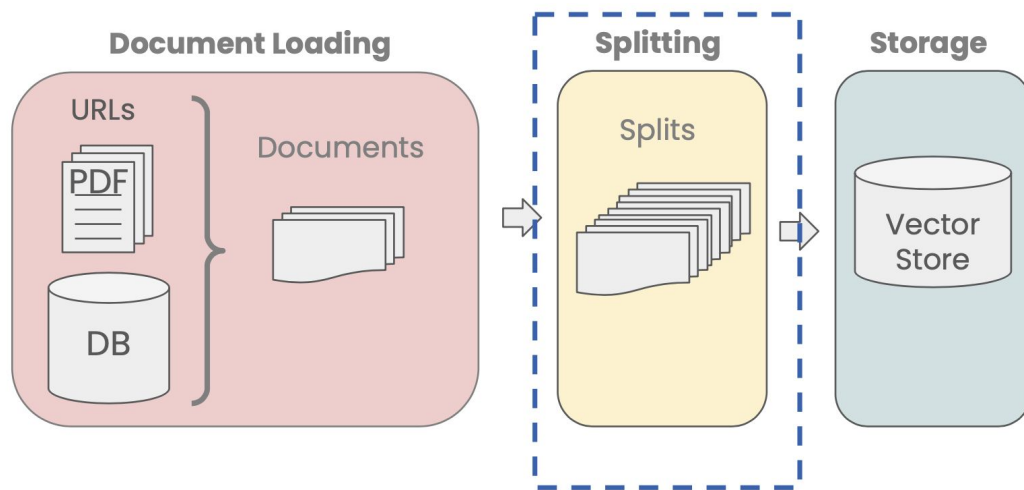
# Document Loaders



# Document Splitting

Splitting documents into smaller chunks

- Important to retain meaningful relationships!



...  
on this model. The Toyota Camry has a head-snapping  
80 HP and an eight-speed automatic transmission that will  
...

[Chunk 1:](#) on this model. The Toyota Camry has a head-snapping

[Chunk 2:](#) 80 HP and an eight-speed automatic transmission that will

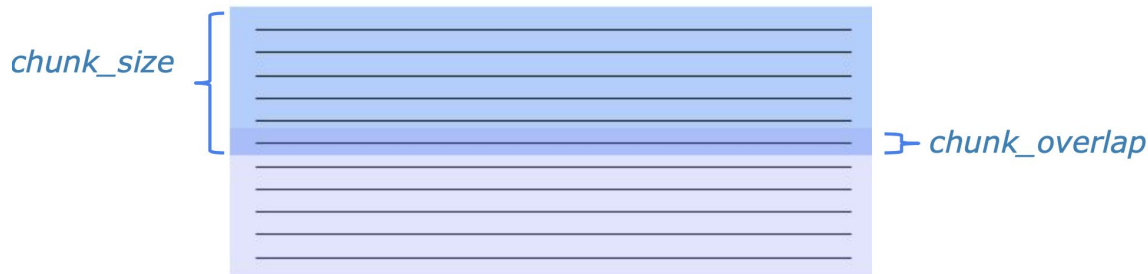
[Question:](#) What are the specifications on the Camry?

# Example Splitter

```
langchain.text_splitter.CharacterTextSplitter(separator: str="\n\n",  
                                              chunk_size=4000,  
                                              chunk_overlap=200,  
                                              length_function=<builtin function len>)
```

methods:

- **create\_documents():** create documents from a list of texts
- **split\_documents():** split documents



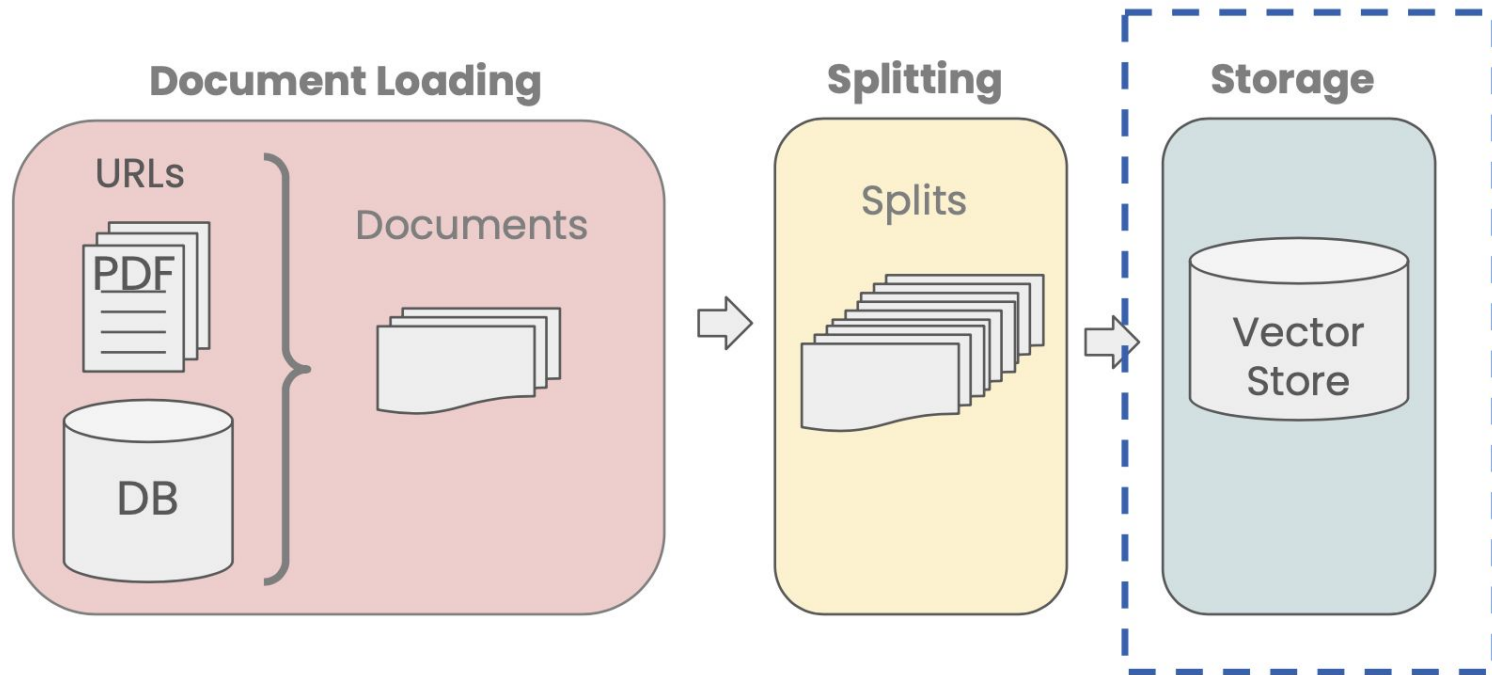
# Types of splitters

langchain.text\_splitter:

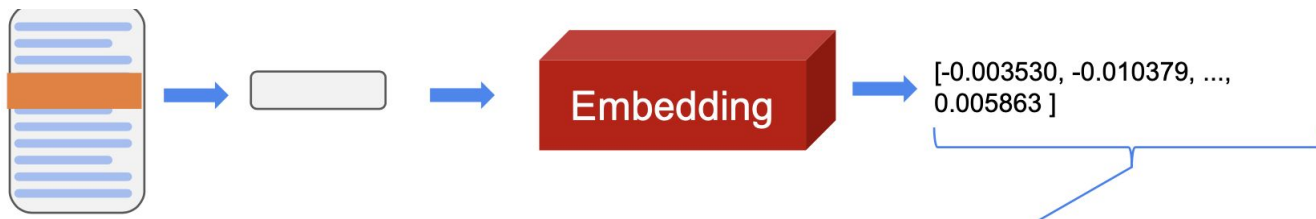
- **CharacterTextSplitter()**- Implementation of splitting text that looks at characters.
- **MarkdownHeaderTextSplitter()** - Implementation of splitting markdown files based on specified headers.
- **TokenTextSplitter()** - Implementation of splitting text that looks at tokens.
- **SentenceTransformersTokenTextSplitter()** - Implementation of splitting text that looks at tokens.
- ***RecursiveCharacterTextSplitter()*** - Implementation of splitting text that looks at characters. Recursively tries to split by different characters to find one that works.
- **Language()** – for CPP, Python, Ruby, Markdown etc
- **NLTKTextSplitter()** - Implementation of splitting text that looks at sentences using NLTK (Natural Language Tool Kit)

# Vector Stores and Embeddings

# Vector Stores

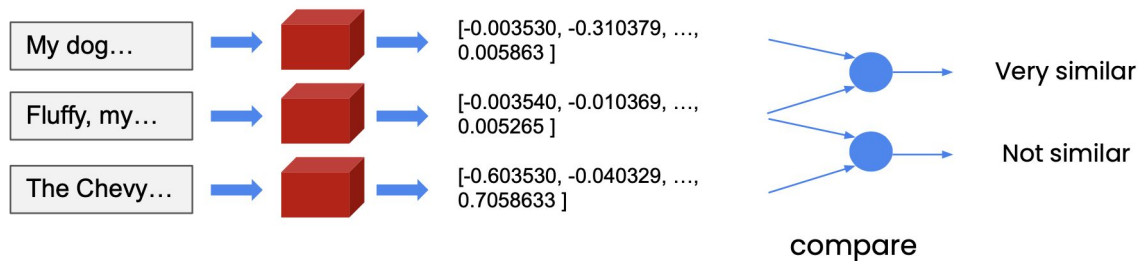


# Embeddings



- Embedding vector captures content/meaning
- Text with similar content will have similar vectors

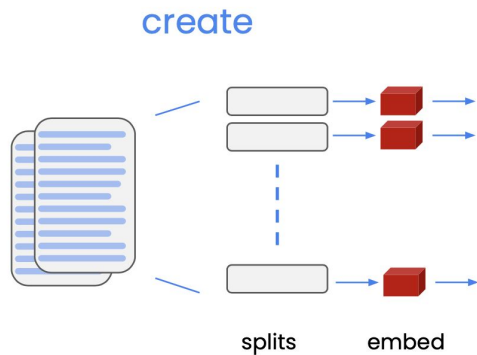
- 1) My dog Rover likes to chase squirrels.
- 2) Fluffy, my cat, refuses to eat from a can.
- 3) The Chevy Bolt accelerates to 60 mph in 6.7 seconds.





# Vector Store

create

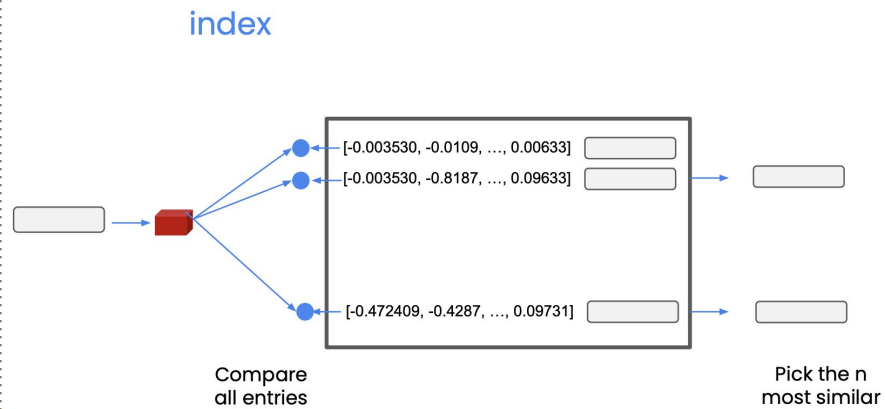


Vector Store

embedding  
vector

original  
spits

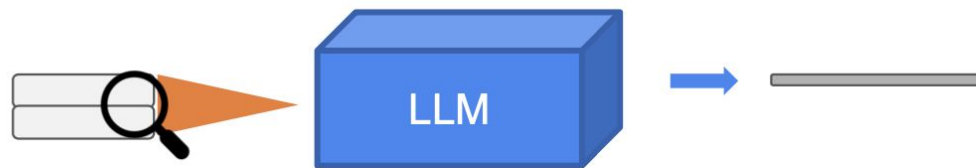
index



# Vector Store/Database

Process with LLM

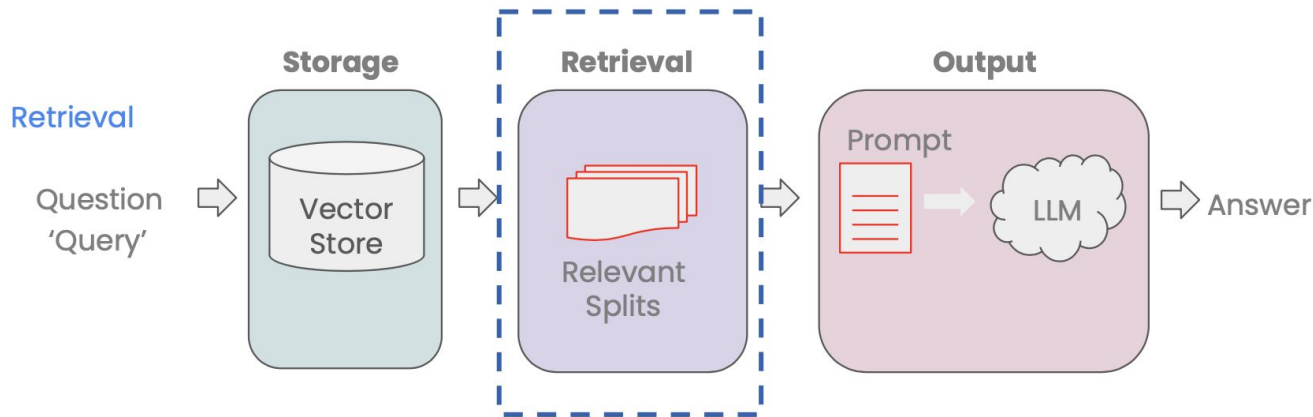
Process with llm



The returned values can now fit in the LLM context

# Retrieval

# Retrieval Pipeline

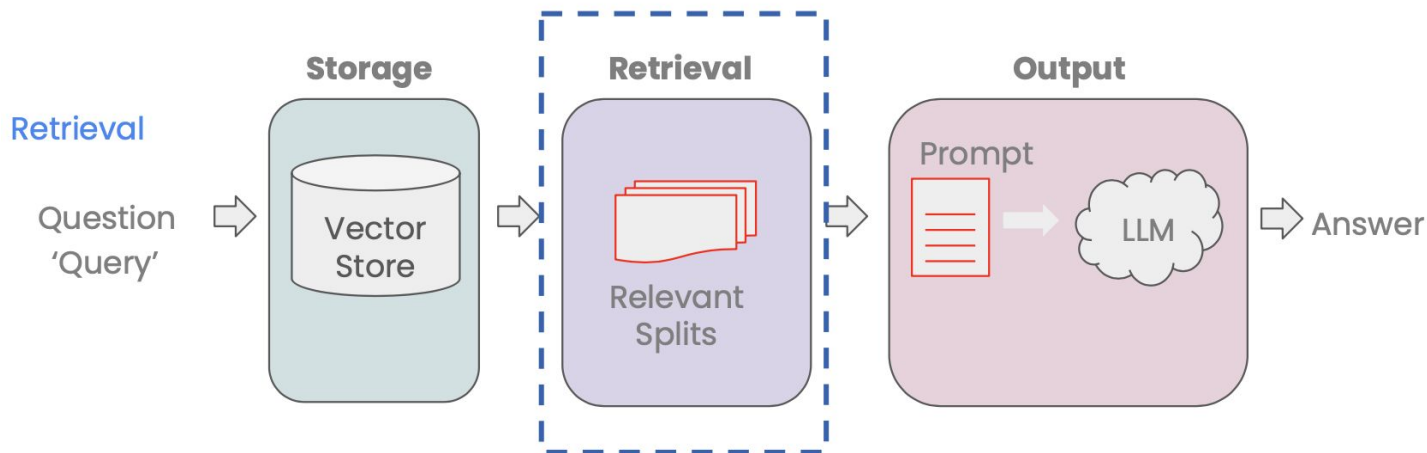


- Accessing/indexing the data in the vector store
  - Basic semantic similarity
  - Maximum marginal relevance
  - Including Metadata
- LLM Aided Retrieval

# Generation

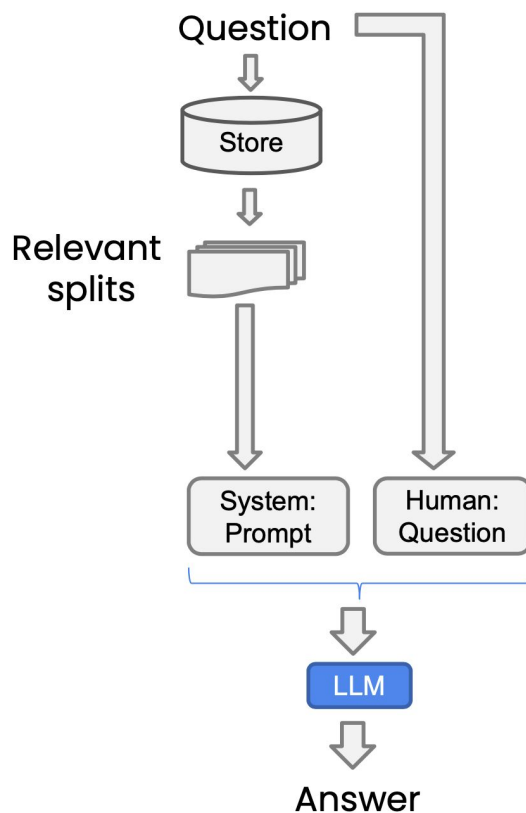
# Question Answering

- Multiple relevant documents have been retrieved from the vector store
- Potentially compress the relevant splits to fit into the LLM context
- Send the information along with our question to an LLM to select and format an answer



# RetrievalQA Chain

```
RetrievalQA.from_chain_type(  
    chain_type="stuff", ..  
)
```



Question is applied to the Vector Store as a query

Vector store provides k relevant documents

Docs and original question are sent to an LLM

# Discussion



# Google Cloud Project & Credits Setup

# Try Google Cloud for Free!

Visit [trygcp.dev/e/build-ai-IS03](https://trygcp.dev/e/build-ai-IS03) to sign up with your Google Account and get access to the Free Tier.

Learn about Free Tier products at [cloud.google.com/free](https://cloud.google.com/free)

## Free Tier products

There is no charge to use these products up to their specified [free usage limit](#). The free usage limit does not expire, but is subject to change. Available for eligible customers.

### Compute Engine

Scalable, high-performance virtual machines.

1 e2-micro instance per month



### Cloud Storage

Best-in-class performance, reliability, and pricing for all your storage needs.

5 GB-months Standard Storage



### BigQuery

Fully managed, petabyte scale, analytics data warehouse.

1 TB queries per month



### Google Kubernetes Engine

One-click container orchestration via Kubernetes clusters, managed by Google.

One Autopilot or Zonal cluster per month



### App Engine

Platform for building scalable web applications and mobile back ends.

28 instance hours per day



### Cloud Run

A fully managed environment to run stateless containers.

2 million requests per month



and more!

# trygcp.dev/e/build-ai-IS03



Sign in

Hi, welcome Cloud Community

(gcloudcommunity@gmail.com)



## GenAI workshop Arizona

Your credit will allow you to use Google Cloud [Free Tier products](#).

It has an amount of \$5.

Once redeemed, it will be valid for **180 days** or until the balance is depleted if you use non-free services.

CLICK HERE TO ACCESS YOUR CREDITS



## GCP credit application

Fill in the following information below to apply GCP credits to your account listed below.

First name \*

Sara

Last name \*

Presentation

Account email

sararob.presentation@gmail.com

Credits will be applied to this account. If you'd like to apply credits to a different account, specify your preference [here](#).

Coupon code

3BJV-DK5N-QR15-6FLR

### Terms and conditions

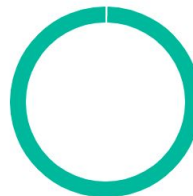
The following [Terms of Service](#) apply to the credit you received for Google Cloud products.

ACCEPT AND CONTINUE

\* Indicates required

1. Click here

## Credits ?



\$10.00

Remaining credits

Out of \$10.00

### Remaining credits

DevRel - Instrumentless Credits for IO workshop (sararob) 203082262	\$10.00
---	---------

→ Credit details

Credit successfully applied



Last step to apply credit!

Set your project's billing account to **Google Cloud Platform Trial Billing Account**



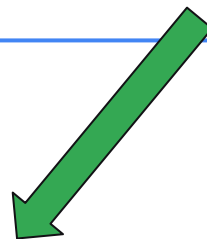
### Set the billing account for project "My First Project"

This project pays for both Google Cloud Platform and Maps Platform. Select a billing account that supports both Google Cloud Platform and Maps Platform. [Learn more](#)

**Billing account \***  
Google Cloud Platform Trial Billing Account

Any charges for this project will be billed to the account you select here.

CANCEL SET ACCOUNT



My First Project

stable-device-229520

Google Cloud Platform Trial Billing Account

0112C1-8A6B69-4258F2



# Issues When attempting to redeem credits

It may be unclear where the credits land. If your users are confused, have them navigate to the Credits tab on the page on of the

The screenshot shows the Google Cloud Billing Account Overview page. The left sidebar has a red arrow pointing to the 'Credits' tab. The main content area displays the following information:

- Your total cost (April 1 – 18, 2024):** Cost \$0.00, Credits used \$0.00, Total cost \$0.00.
- FinOps hub:** Save up to \$0.00, 0% of last month's total cost.
- Create a budget alert:** Set a monthly budget for your billing account. You'll get email notifications as your costs approach and exceed the budget amount.

That is where they will see the credits

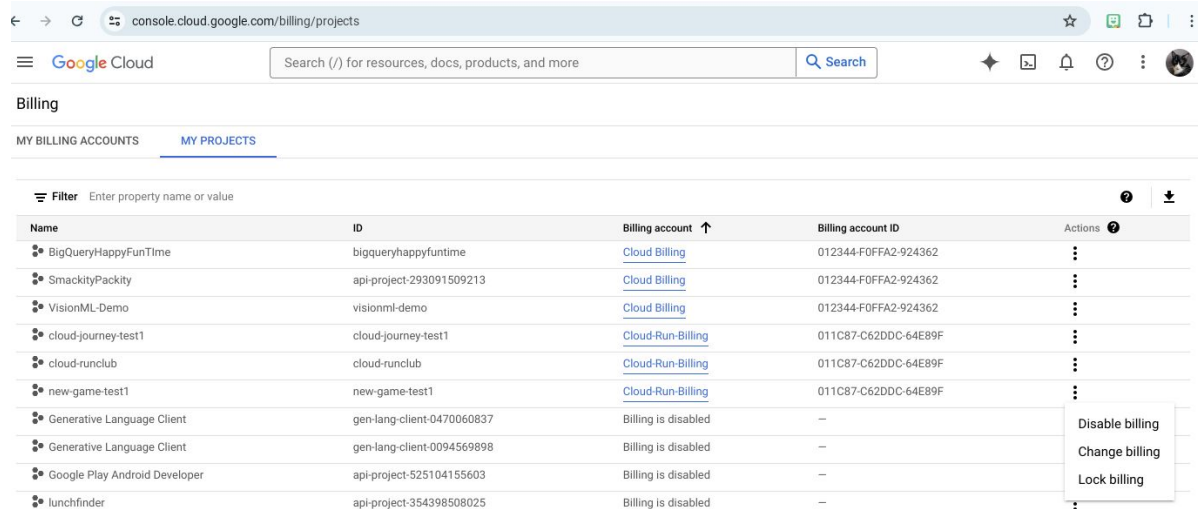
The screenshot shows the Google Cloud Credits page. The left sidebar has the 'Credits' tab selected. The main content area displays the following information:

- View and download credit details here.** Active committed use discounts are not included here and can be viewed on the [Commitments page](#).
- Filter:** Filter credits
- Table:**

Credit name	Status	Percent remaining	Remaining value	Original value	Type	Credit ID	Scope
Frictionless access to Google Cloud	Available	100%	\$5.00	\$5.00	One-time	N0LED275...	Any service on this billing account.

# Issues if the user ALREADY has a existing billing account make sure they use the correct one

Click on the 3 Dot menu on the project, and select “Change Billing,”  
from the drop down select that Google Cloud Platform Trial Billing

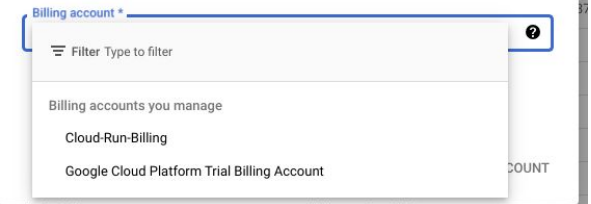


The screenshot shows the Google Cloud Billing console. The URL is `console.cloud.google.com/billing/projects`. The page title is "Billing". Under "MY BILLING ACCOUNTS", the "MY PROJECTS" tab is selected. A table lists projects with columns: Name, ID, Billing account, Billing account ID, and Actions. The project "new-game-test1" is highlighted. The Actions menu is open, showing options: "Disable billing", "Change billing", and "Lock billing".

Name	ID	Billing account	Billing account ID	Actions
BigQueryHappyFunTime	bigqueryhappyfuntime	Cloud Billing	012344-F0FFA2-924362	⋮
SmackityPackity	api-project-293091509213	Cloud Billing	012344-F0FFA2-924362	⋮
VisionML-Demo	visionml-demo	Cloud Billing	012344-F0FFA2-924362	⋮
cloud-journey-test1	cloud-journey-test1	Cloud-Run-Billing	011C87-C62DDC-64E89F	⋮
cloud-runclub	cloud-runclub	Cloud-Run-Billing	011C87-C62DDC-64E89F	⋮
new-game-test1	new-game-test1	Cloud-Run-Billing	011C87-C62DDC-64E89F	⋮
Generative Language Client	gen-lang-client-0470060837	Billing is disabled	—	⋮
Generative Language Client	gen-lang-client-0094569898	Billing is disabled	—	⋮
Google Play Android Developer	api-project-525104155603	Billing is disabled	—	⋮
lunchfinder	api-project-354398508025	Billing is disabled	—	⋮

## Set the billing account for project “new-game-test1”

This project pays for both Google Cloud Platform and Maps Platform. Select a billing account that supports both Google Cloud Platform and Maps Platform. [Learn more](#)



The screenshot shows a "Billing account" selection dialog. It has a search bar with the text "Filter Type to filter". Below the search bar, there are two options: "Cloud-Run-Billing" and "Google Cloud Platform Trial Billing Account". The "Google Cloud Platform Trial Billing Account" is selected.

# Workshop

