



BERKELEY LAB

Bringing Science Solutions to the World

Closed Domain QA System for LBL ScienceIT: Fine-Tuned and Retrieval Augmented Generation Models

Fengchen Liu, Jordan Jung

ScienceIT | Information Technology | Lawrence Berkeley National Laboratory

CS3-2024: March 13th, 2024



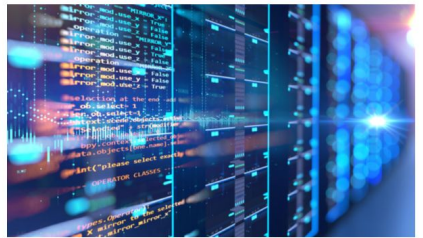
Information
Technology
BERKELEY LAB



You are here: Home / Services / SciencIT

SciencIT

Published on January 27, 2020 by OpsWeb3.



SciencIT is a comprehensive IT support program intended to facilitate the use of scientific computing and data to accomplish scientific research at the Lab. With a focus on the needs of scientists, the program not only provides traditional high-performance computing systems and services but is intent on providing a comprehensive portfolio of services including consulting, training, data management planning, cloud computing services and storage offerings. The goal is to help scientists move past computing barriers so that they can move forward in their research.

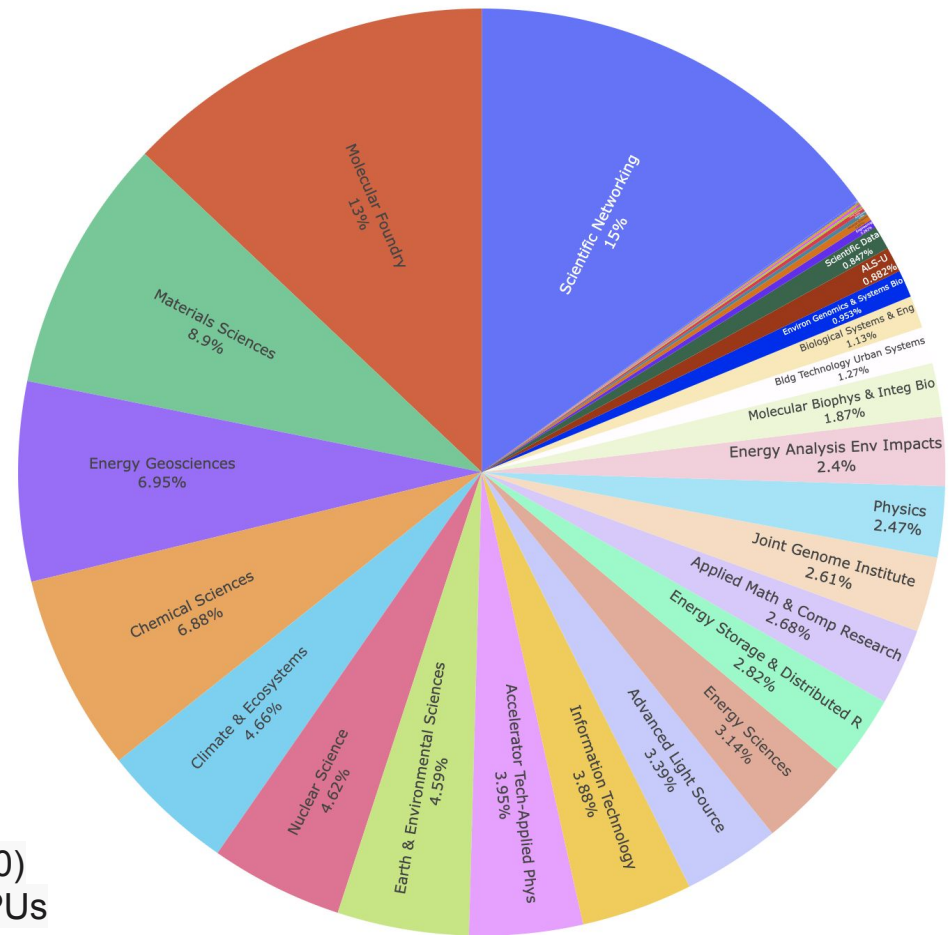
Working with SciencIT team is easy. Please peruse our services below. The place to start is to send an email to scien@lbl.gov.

Lawrencium HPC Cluster:

- Projects: 400+
- Users: 2000+
- Partitions: 38
- Nodes: 1000+
- CPU-Cores: 80000+
- Storage: 6 PB
- Memory (TB): 400+
- GPU: 200+ and 56 (NVIDIA-A40)
- Coming soon: NVIDIA H100 GPUs

Background of SciencIT at LBNL

- Lawrencium HPC Cluster, Storage, Scientific Virtual Machine(SVM), Cloud (AWS, GCP)



- Scientific Networking
- Molecular Foundry
- Materials Sciences
- Energy Geosciences
- Chemical Sciences
- Climate & Ecosystems
- Nuclear Science
- Earth & Environmental Sciences
- Accelerator Tech-Applied Phys
- Information Technology
- Advanced Light Source
- Energy Sciences
- Energy Storage & Distributed R
- Applied Math & Comp Research
- Joint Genome Institute
- Physics
- Energy Analysis Env Impacts
- Molecular Biophys & Integ Bio
- Bldg Technology Urban Systems
- Biological Systems & Eng
- Environ Genomics & Systems Bio
- ALS-U
- Scientific Data
- Engineering
- Physical Sciences
- Earth Sciences
- Energy Technologies
- Environ, Health, & Safety
- Life Sciences
- NERSC
- unknown



Closed Domain QA System for LBL ScienceIT

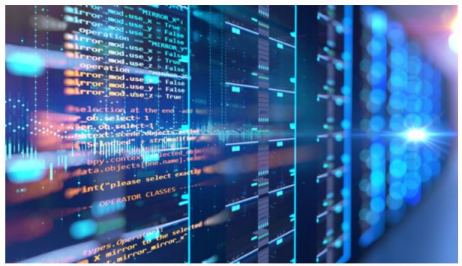
- **Generative AI:**
 - ScienceIT Chatbot with GCP [Vertex AI Search and Conversation](#)
 - ScienceIT Question-Answering (QA) System:
 - Fine-tuning foundation LLMs with Vertex AI [Generative AI Studio](#)
 - Retrieval-Augmented Generation (**RAG**) QA with:
 - LangChain
 - AWS **Bedrock Titan** and **Claude-v2**
 - GCP **PaLM-2 (text-bison)**
 - Meta **LLaMA-2** (self-hosting on **Lawrencium GPU node**)



You are here: [Home](#) / [Services](#) / [ScienceIT](#)

ScienceIT

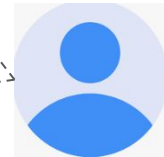
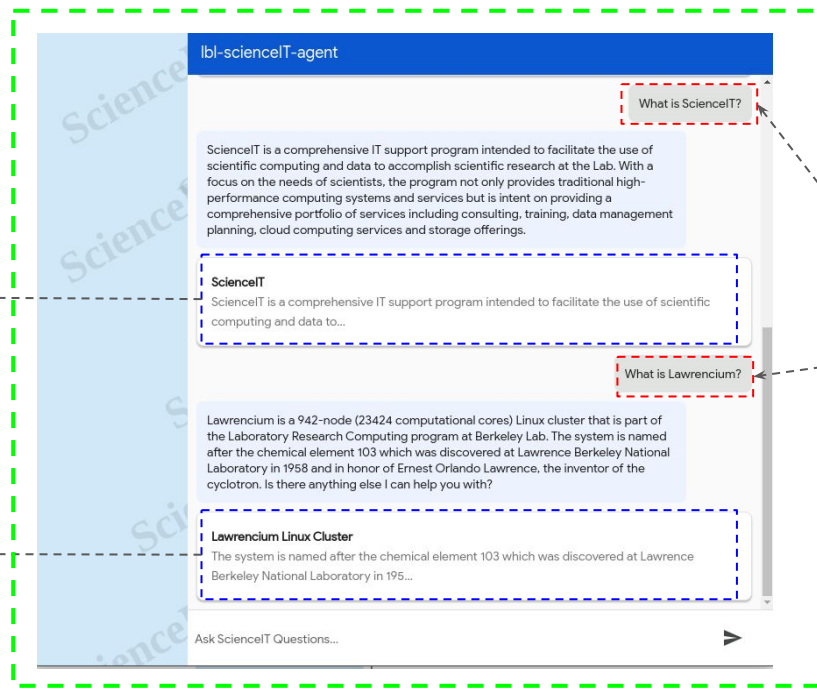
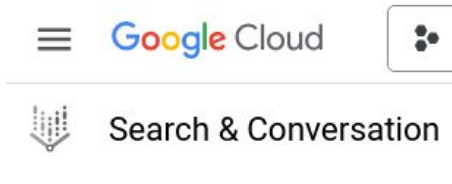
Published on January 27, 2020 by OpsWeb3.

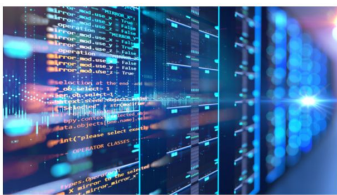


ScienceIT is a comprehensive IT support program intended to facilitate the use of scientific computing and data to accomplish scientific research at the Lab. With a focus on the needs of scientists, the program not only provides traditional high-performance computing systems and services but is intent on providing a comprehensive portfolio of services including consulting, training, data management planning, cloud computing services and storage offerings. The goal is to help scientists move past computing barriers so that they can move forward in their research.

Working with ScienceIT team is easy. Please peruse our services below. The place to start is to send an email to scienceit@lbl.gov or send us a ticket to hpcshelp@lbl.gov.

Office Hours: 10:30am - noon Wednesdays over zoom or in person 50B-3209.

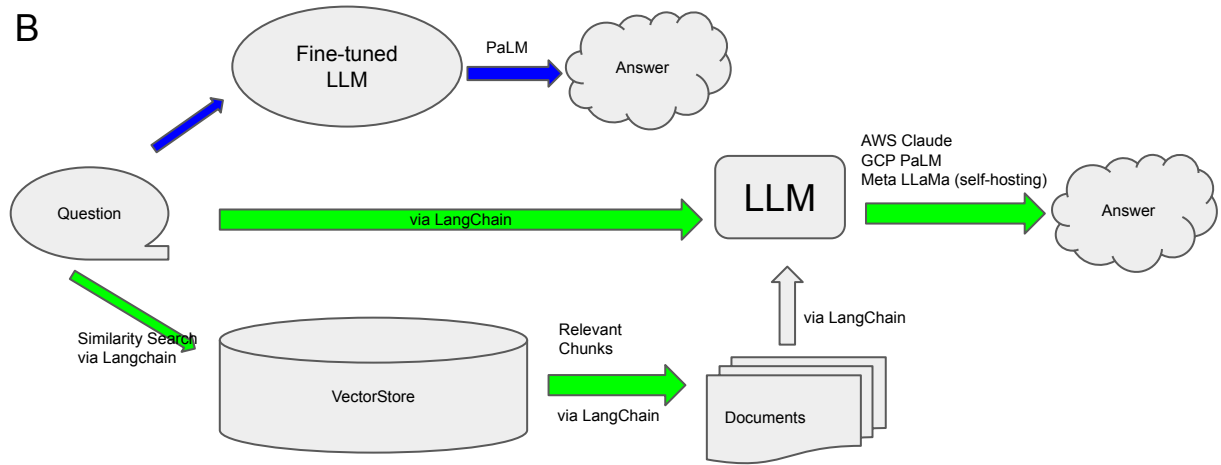
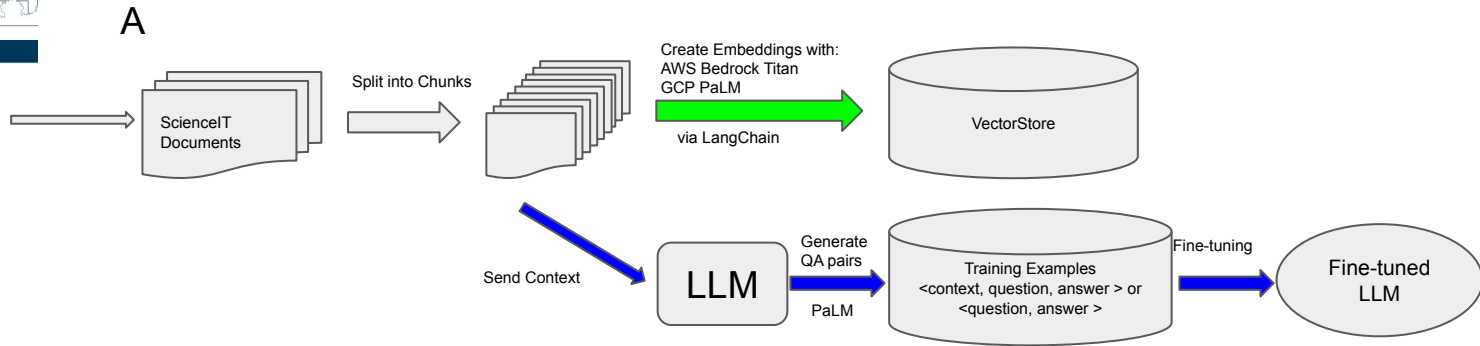
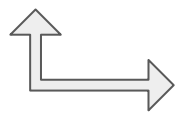




ScienceIT is a comprehensive IT support program intended to facilitate the use of scientific computing and data to accomplish scientific research at the Lab. With a focus on the needs of scientists, the program not only provides traditional high-performance computing systems and services but is intent on providing a comprehensive portfolio of services including consulting, training, data management planning, cloud computing services and storage offerings. The goal is to help scientists move past computing barriers so that they can move forward in their research.

Working with ScienceIT team is easy. Please peruse our services below. The place to start is to send an email to scienceit@lbl.gov or send us a ticket to hpcshelp@lbl.gov.

Office Hours: 10:30am – noon Wednesdays over zoom or in person 50B-3209.



You are here: [Home](#) / [Services](#) / [ScienceIT](#)

ScienceIT

Published on January 27, 2020 by OpsWeb3.



ScienceIT is a comprehensive IT support program intended to facilitate the use of scientific computing and data to accomplish scientific research at the Lab. With a focus on the needs of scientists, the program not only provides traditional high-performance computing systems and services but is intent on providing a comprehensive portfolio of services including consulting, training, data management planning, cloud computing services and storage offerings. The goal is to help scientists move past computing barriers so that they can move forward in their research.

Working with ScienceIT team is easy. Please peruse our services below. The place to start is to send an email to scienceit@lbl.gov or send us a ticket to hpcshelp@lbl.gov.

Office Hours: 10:30am - noon Wednesdays over [zoom](#) or in person 50B-3209.

[HPCS System Status](#)

Cloud Computing

Berkeley Lab researchers can make use of Amazon AWS and Google Cloud Platform (GCP) services through the Lab's Master Payer Agreement.

Search this site

QUICK LINKS

- [IT Help Desk](#)
- [IT Frequently Asked Questions](#)
- [IT Status](#)
- [Cyber Security Help](#)
- [Request a Computer](#)
- [Read the IT Newsletter](#)
- [Register for IT Training](#)
- [Search the IT Directory](#)

HPC ANNOUNCEMENTS

 **Video Replay: Data Science for Postdocs Workshop**
August 28, 2023

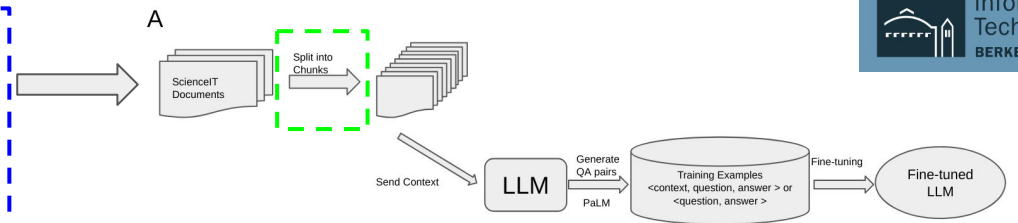
[More HPC Announcements »](#)

SCIENCEIT NEWS

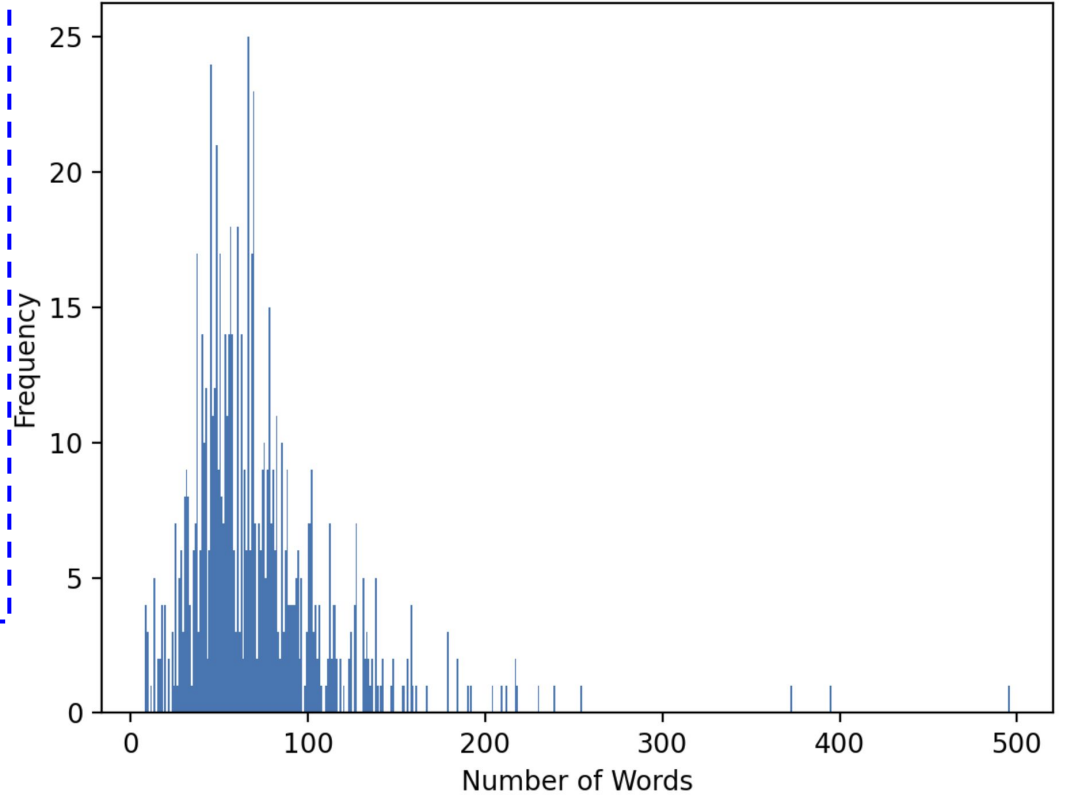
 **Gary Jung Recognized with 2023 Director's Lifetime Achievement Award**
October 27, 2023 By Arica Chhay

 **Expanding Storage Services for Research**

A



Histogram of Number of Words in Context



TOOLS

- Dashboard
- Model Garden
- Pipelines

NOTEBOOKS

- Colab Enterprise
- Workbench

GENERATIVE AI STUDIO

- Overview
- Language
- Vision
- Speech

DATA

- Feature Store
- Datasets
- Labeling tasks

Context (Optional) ?

Context: "SciencIT is a comprehensive IT support program intended to facilitate the use of scientific computing and data to accomplish scientific research at the Lab. With a focus on the needs of scientists, the program not only provides traditional high-performance computing systems and services but is intent on providing a comprehensive portfolio of services including consulting, training, data management planning, cloud computing services and storage offerings. The goal is to help scientists move past computing barriers so that they can move forward in their research."

Please summarize above context into question answer pairs

> Examples (Optional) ?

Responses

CLEAR CONVERSATION

USER Please summarize above context into 5 question answer pairs

- AI
- What is SciencIT?

Answer: SciencIT is a comprehensive IT support program intended to facilitate the use of scientific computing and data to accomplish scientific research at the Lab.
 - What is the focus of SciencIT?

Answer: The focus of SciencIT is on the needs of scientists.
 - What services does SciencIT provide?

Answer: SciencIT provides traditional high-performance computing systems and services, consulting, training, data management planning, cloud computing services and storage offerings.
 - What is the goal of SciencIT?

Answer: The goal of SciencIT is to help scientists move past computing barriers so that they can move forward in their research.
 - What is the benefit of SciencIT?

Answer: SciencIT helps scientists move past computing barriers so that they can move forward in their research.

We want your [feedback](#).

Model chat-bison (latest) ?

Temperature ?

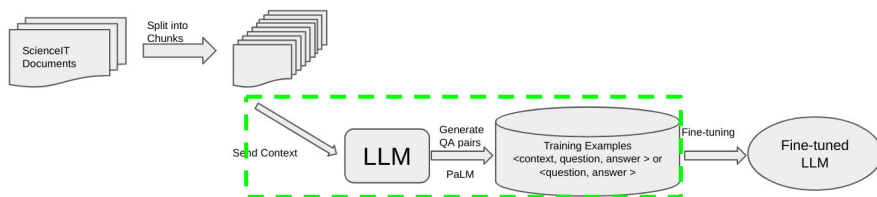
0 1

0.2

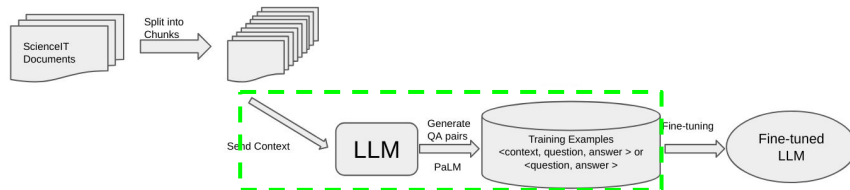
> Advanced

RESET PARAMETERS

A



A



```
In [1]: import vertexai
from vertexai.language_models import ChatModel, InputOutputTextPair
```

```
In [2]: vertexai.init(project="██████████", location="us-central1")
chat_model = ChatModel.from_pretrained("chat-bison")
parameters = {
    "candidate_count": 1,
    "max_output_tokens": 1024,
    "temperature": 0.2,
    "top_p": 0.8,
    "top_k": 40
}
```

```
In [3]: chat = chat_model.start_chat(
    context="""Context: \\"ScienceIT is a comprehensive IT support program intended to facilitate the use of scienc

Please summarize above context into question answer pairs""",
)
response = chat.send_message("""Please summarize above context into 5 question answer pairs""", **parameters)
print(f"Response from Model: {response.text}")
```

Response from Model: 1. What is ScienceIT?

Answer: ScienceIT is a comprehensive IT support program intended to facilitate the use of scientific computing and data to accomplish scientific research at the Lab.

2. What is the focus of ScienceIT?

Answer: The focus of ScienceIT is on the needs of scientists.

3. What services does ScienceIT provide?

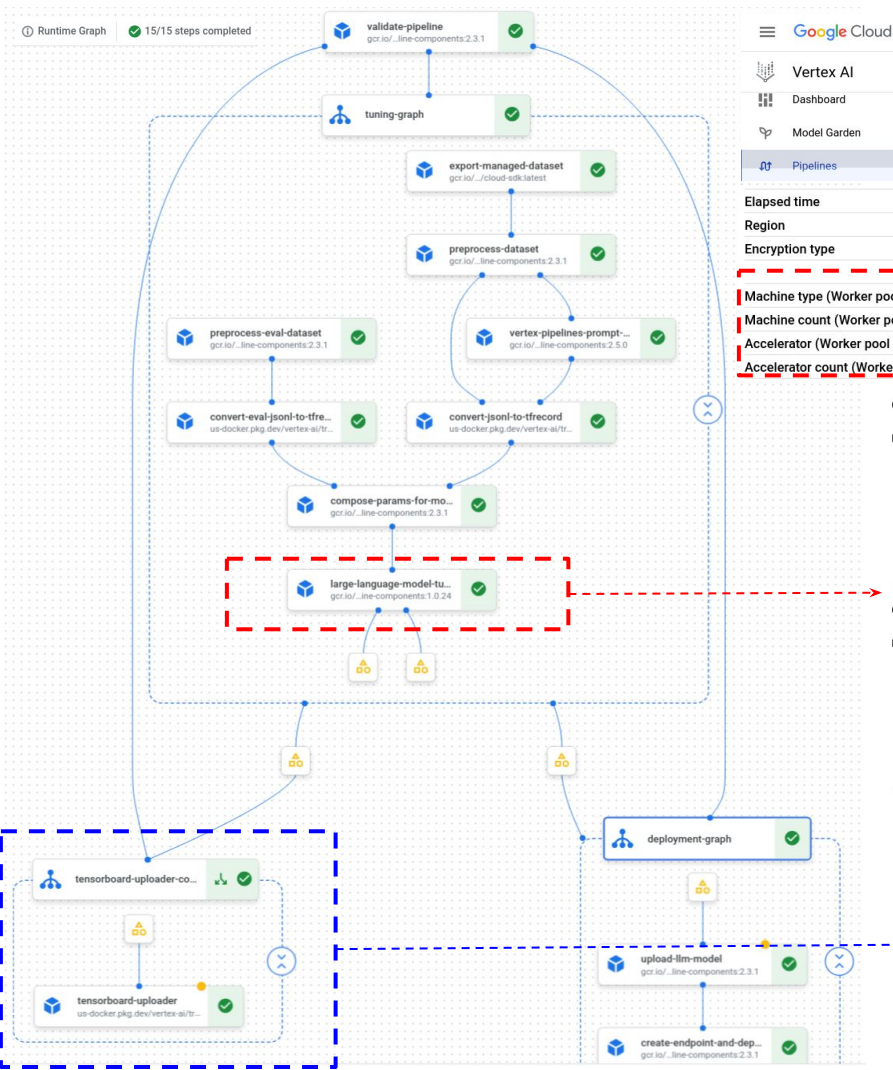
Answer: ScienceIT provides traditional high-performance computing systems and services, consulting, training, data management planning, cloud computing services, and storage offerings.

4. What is the goal of ScienceIT?

Answer: The goal of ScienceIT is to help scientists move past computing barriers so that they can move forward in their research.

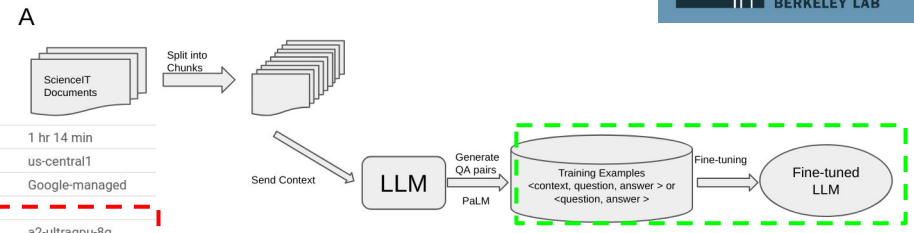
5. What is the benefit of ScienceIT?

Answer: The benefit of ScienceIT is that it provides scientists with the resources they need to conduct their research more efficiently and effectively.



Google Cloud
Vertex AI
Dashboard
Model Garden
Pipelines

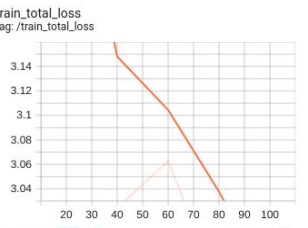
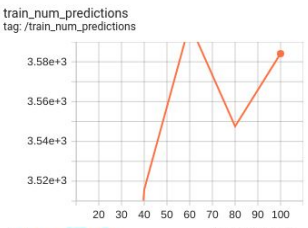
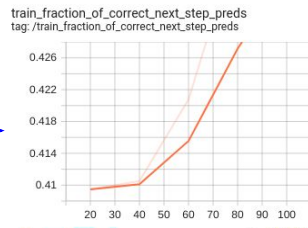
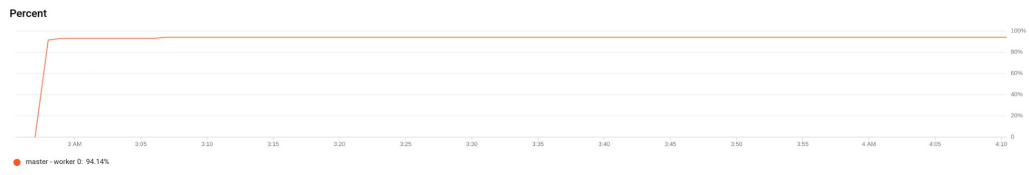
| | |
|---|------------------|
| Elapsed time | 1 hr 14 min |
| Region | us-central1 |
| Encryption type | Google-managed |
| Machine type (Worker pool 0 (chief)) | a2-ultragpu-8g |
| Machine count (Worker pool 0 (chief)) | 1 |
| Accelerator (Worker pool 0 (chief)) | NVIDIA_A100_80GB |
| Accelerator count (Worker pool 0 (chief)) | 8 |



GPU utilization



GPU memory utilization



Use your fine-tuned model to get predictions

```
import vertexai
vertexai.init(project=PROJECT_ID, location=REGION)

from vertexai.preview.language_models import TextGenerationModel
from google.cloud import aiplatform

deployed_model = 'projects/38[REDACTED]/locations/us-central1/models/878[REDACTED]04'
deployed_model = TextGenerationModel.get_tuned_model(deployed_model)
```

Now you can start send a prompt to the API. Feel free to update the following prompt.

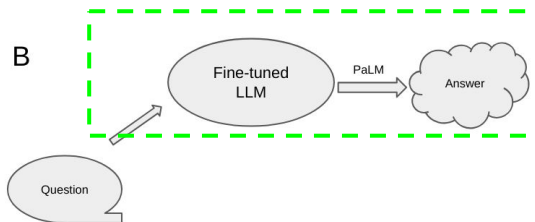
```
PROMPT = """
In context of ScienceIT
How do I load a software module?
"""
```

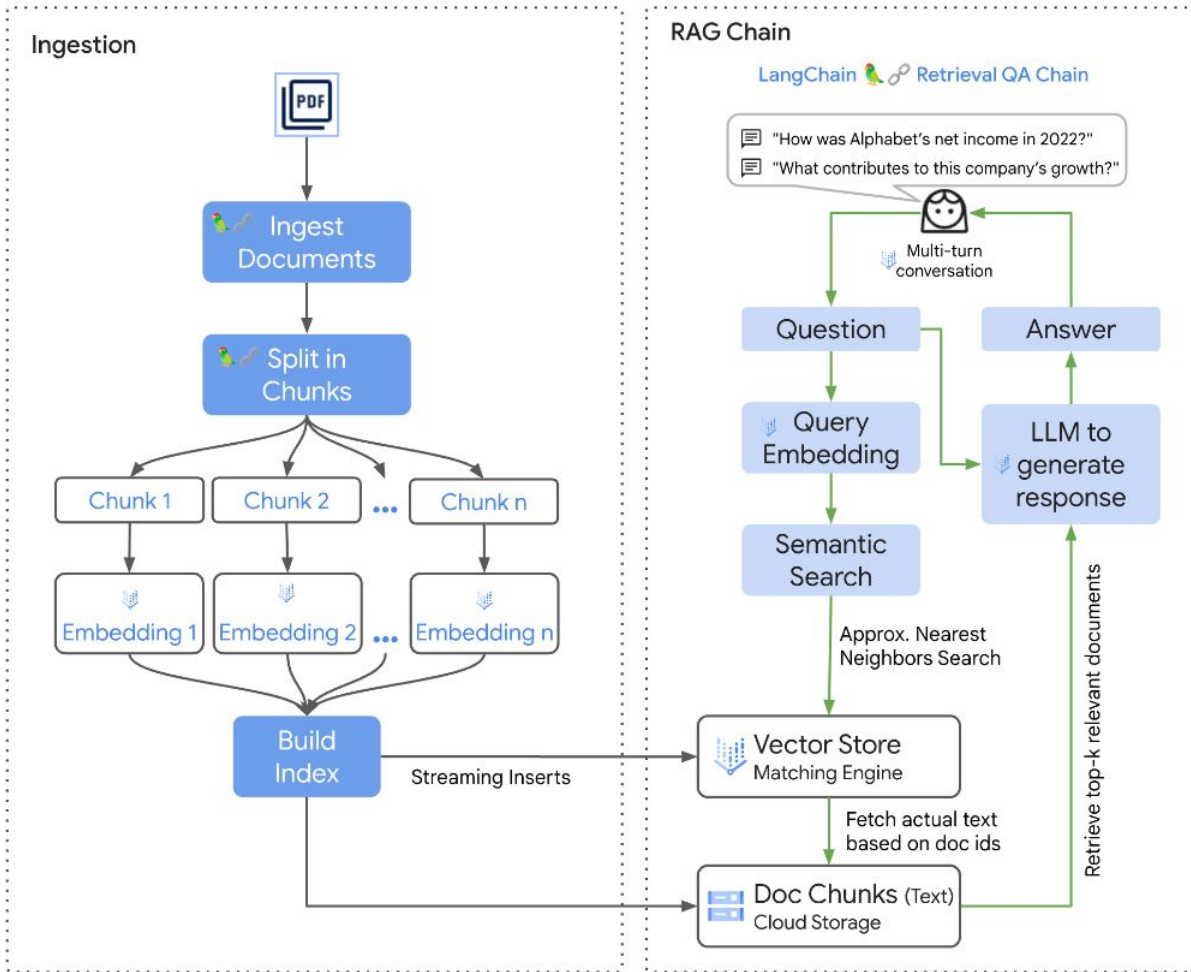
```
response = deployed_model.predict(PROMPT)
```

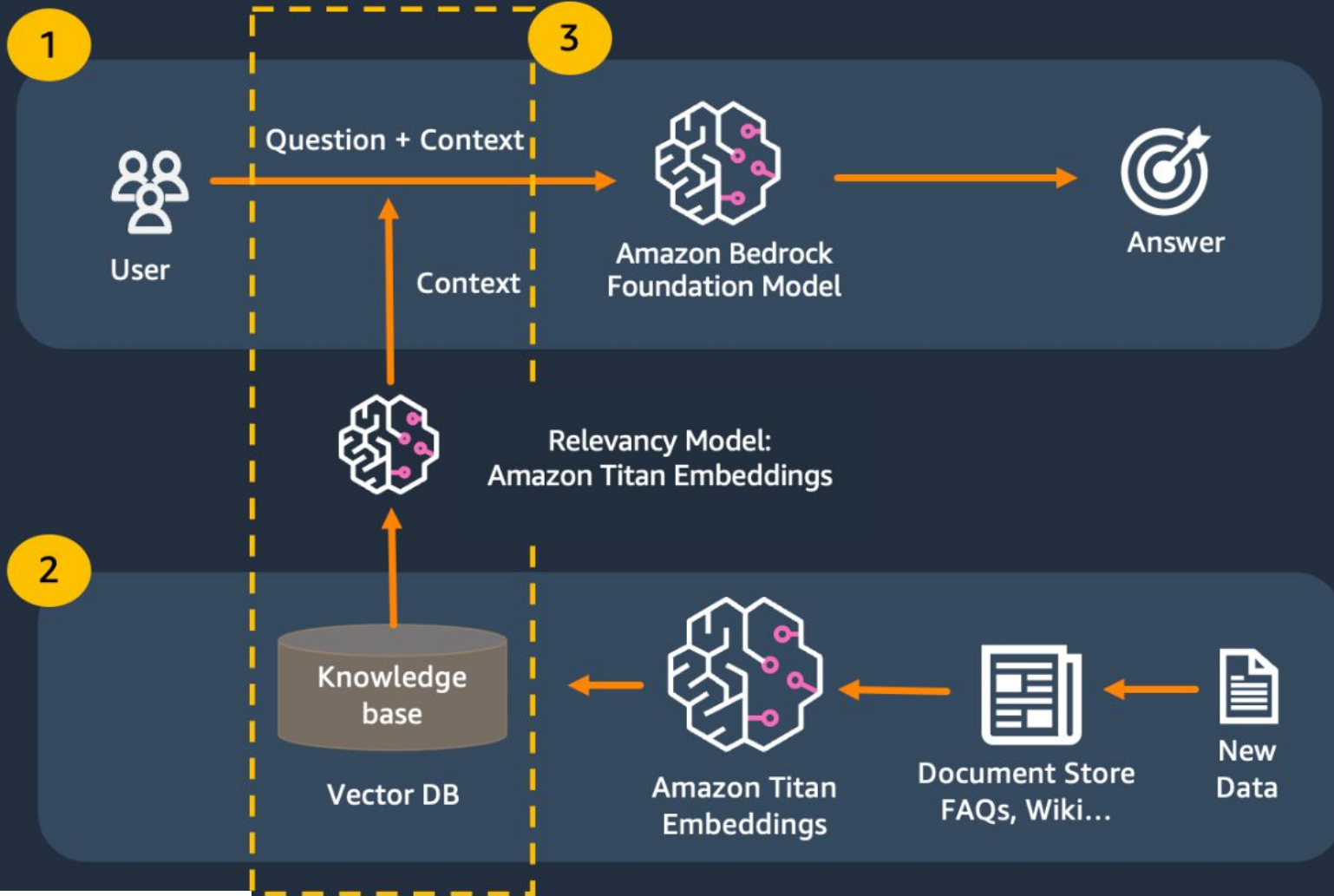
```
print(response.text)
```

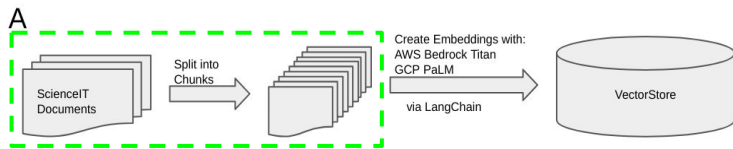
To load a software module, you can use the module command. For example, to load the module for the gnuplot software

```
module load gnuplot
```









Data Preparation

```

from langchain.text_splitter import CharacterTextSplitter, RecursiveCharacterTextSplitter
from langchain.document_loaders import PyPDFLoader, PyPDFDirectoryLoader, UnstructuredHTMLLoader

html_files = os.popen("ls ./scienceit_all_html/*.html").read().split('\n')[:-1]
html_documents = []
for html in html_files:
    loader = UnstructuredHTMLLoader(html)
    data = loader.load()
    html_documents.append(data)
html_documents_all = [h[0] for h in html_documents]
text_splitter = RecursiveCharacterTextSplitter(
    # Set a really small chunk size, just to show.
    chunk_size = 1000,
    chunk_overlap = 100,
)
docs = text_splitter.split_documents(html_documents_all)
  
```

Create VectorStore

```
from langchain.chains.question_answering import load_qa_chain
from langchain.vectorstores import FAISS
from langchain.indexes import VectorStoreIndexCreator
from langchain.indexes.vectorstore import VectorStoreIndexWrapper
```

Init Vertex-AI and Create GCP PaLM2 text-bison Embeddings & LLM

```
import vertexai
vertexai.init(project=PROJECT_ID, location="us-central1")
```

```
from langchain.embeddings import VertexAIEmbeddings
from langchain.llms import VertexAI
```

```
vertexai_embeddings = VertexAIEmbeddings()
vertexai_llm = VertexAI()

# VecotrStore with Vertex-AI embedding:
GCP_vectorstore_faiss = FAISS.from_documents(
    docs,
    vertexai_embeddings,
)
```

Init AWS and Create AWS Bedrock Titan Embeddings & Anthropic LLM

```
import boto3
boto3_bedrock = bedrock.get_bedrock_client(
    assumed_role=os.environ.get("BEDROCK_ASSUME_ROLE", None),
    region=os.environ.get("AWS_DEFAULT_REGION", None)
)
```

```
from langchain.embeddings import BedrockEmbeddings
from langchain.llms.bedrock import Bedrock
```

```
bedrock_llm = Bedrock(model_id="anthropic.claude-v2", client=boto3_bedrock, model_kwargs={'max_tokens_to_sample':200})
bedrock_embeddings = BedrockEmbeddings(model_id="amazon.titan-embed-text-v1", client=boto3_bedrock)
```

```
# VecotrStore with AWS Bedrock Titan embedding:
AWS_vectorstore_faiss = FAISS.from_documents(
    docs,
    bedrock_embeddings,
)
```

Create Meta LLaMA2 Embeddings & LLM

Self-hosting LLaMA2 on

Lawrencium ES1 GPU nodes

```
from langchain.embeddings import LlamaCppEmbeddings
from langchain.llms import LlamaCpp
from langchain.callbacks.manager import CallbackManager
from langchain.callbacks.streaming_stdout import StreamingStdOutCallbackHandler
```

```
# Make sure the model path is correct for your system!
llama_embeddings = LlamaCppEmbeddings(model_path="/global/scratch/users/fengchenliu/meta/llama/llama-2-7b.Q2_K.gguf")
```

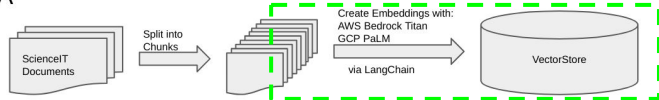
```
n_gpu_layers = 1 # Metal set to 1 is enough.
n_batch = 512 # Should be between 1 and n_ctx, consider the amount of RAM of your Apple Silicon Chip.
callback_manager = CallbackManager([StreamingStdOutCallbackHandler()])
```

```
# Make sure the model path is correct for your system!
```

```
llama_llm = LlamaCpp(
    model_path="/global/scratch/users/fengchenliu/tools/meta/llama-2-13b-chat.Q5_K_M.gguf",
    n_gpu_layers=n_gpu_layers,
    n_batch=n_batch,
    n_ctx=2048,
    f16_kv=True, # MUST set to True, otherwise you will run into problem after a couple of calls
    callback_manager=callback_manager,
    verbose=True,
)
```

```
# VecotrStore with Meta LLaMA2 embedding:
META_vectorstore_faiss = FAISS.from_documents(
    docs,
    llama_embeddings,
)
```

A



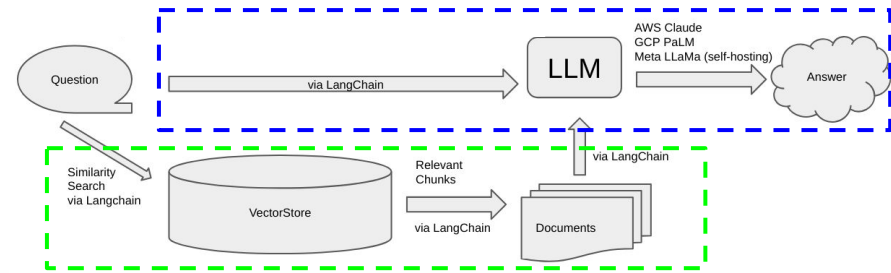
Question Answering w RAG

Replace `llm` with:

- `vertexai_llm`
- `bedrock_llm`
- `llama_llm`

Replace `VectorStore` `retriever` with

- `GCP_vectorstore_faiss`
- `AWS_vectorstore_faiss`
- `META_vectorstore_faiss`



```
from langchain.chains import RetrievalQA
from langchain.prompts import PromptTemplate

prompt_template = """
Human: Use the following pieces of context to provide a concise answer to the question at the end. If you don't know the answer, just say that you don't know, don't try to make up a
<context>
{context}
</context>
Question: {question}
Assistant: ""

PROMPT = PromptTemplate(
    template=prompt_template, input_variables=["context", "question"]
)

qa = RetrievalQA.from_chain_type(
    ## REPLACE LLM with: (1) vertexai_llm, (2) bedrock_llm, (3) llama_llm
    llm=vertexai_llm,
    chain_type="stuff",
    ## REPLACE LLM with: (1) AWS_vectorstore_faiss, (2) AWS_vectorstore_faiss, (3) META_vectorstore_faiss
    retriever=GCP_vectorstore_faiss.as_retriever(
        search_type="similarity", search_kwargs={"k": 1}
    ),
    return_source_documents=True,
    chain_type_kwargs={"prompt": PROMPT}
)

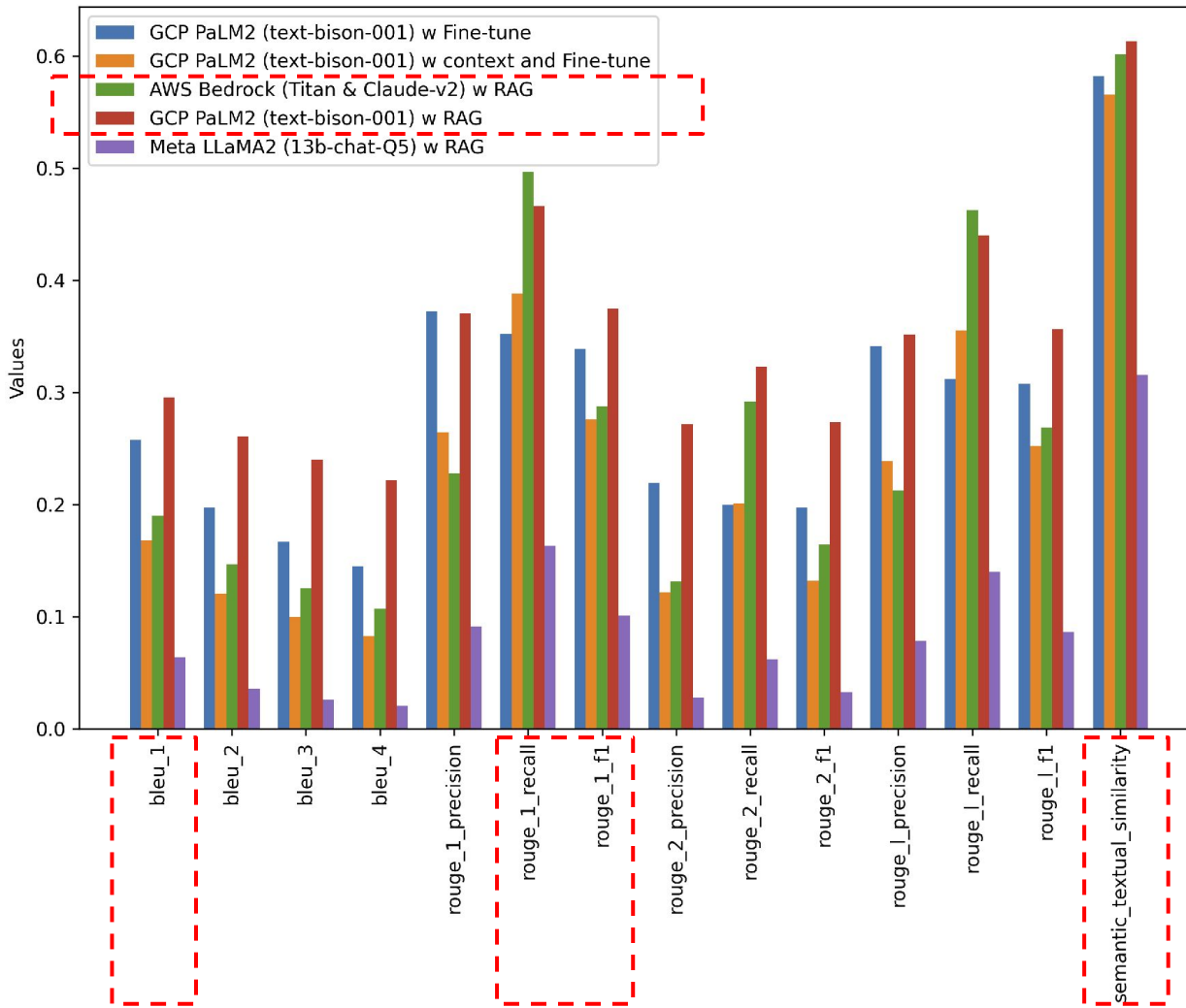
query = "What is the discount on Google Cloud Storage usage?"
result = qa({"query": query})

print_w(result['result'])
28% discount on Google Cloud Storage usage

result['source_documents']

[Document(page_content='Discounted Pricing \n \n \n \n \n 8-12% overall discount across all on AWS services \n Global Data Egress Waiver \xa0agreement is already in place for all users
on the LBL master payer program. \n \n \n \n \n Significant spending-based discounts for Cloud Storage and Cloud Engine, compared to list pricing in addition to any standard applicab
le sustained use discounts.\n \n 19% off Compute VM Engine and attached storage services \n 28% discount on Google Cloud Storage usage \n 41% discount on data egress fees when cumula
tive egress exceeds the waiver limit \n \n \n \n Data egress fee waiver for all users on the LBL master payer program is already in place. \n Direct agreement between DOE and Google cov
ers all GCP usage for all DOE National Labs through 2025. \n \n \n \n \n Data Egress \n \n \n \n \n Agreements in place that waive all egress charges up to an amount calculated as 15% o
f the total monthly bill.\xa0 Current egress amounts are 5-7% of the monthly bill so all egress charges have been waived entirely.', metadata={'source': './scienceit_all_html/resourc
e_cloud-computing_index.html'})]
```

Evaluation of ScienceIT QA



| Metrics | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---------------------|---------|---------|---------|---------|---------|
| BLEU-1 | 0.2576 | 0.168 | 0.1897 | 0.2956 | 0.0636 |
| BLEU-2 | 0.1972 | 0.1208 | 0.1467 | 0.2604 | 0.0355 |
| BLEU-3 | 0.1666 | 0.0997 | 0.1252 | 0.2397 | 0.0261 |
| BLEU-4 | 0.1445 | 0.083 | 0.107 | 0.2215 | 0.0202 |
| ROUGE-1 (Precision) | 0.3724 | 0.2641 | 0.2276 | 0.3709 | 0.0909 |
| ROUGE-1 (Recall) | 0.3519 | 0.3882 | 0.4967 | 0.4663 | 0.1633 |
| ROUGE-1 (F1) | 0.3391 | 0.2759 | 0.2873 | 0.3748 | 0.101 |
| ROUGE-2 (Precision) | 0.2194 | 0.122 | 0.1315 | 0.2714 | 0.0279 |
| ROUGE-2 (Recall) | 0.1997 | 0.2013 | 0.2919 | 0.3228 | 0.0622 |
| ROUGE-2 (F1) | 0.1975 | 0.132 | 0.1643 | 0.2734 | 0.0326 |
| ROUGE-L (Precision) | 0.3415 | 0.2387 | 0.2124 | 0.3516 | 0.0784 |
| ROUGE-L (Recall) | 0.3121 | 0.3552 | 0.4626 | 0.4405 | 0.1403 |
| ROUGE-L (F1) | 0.3077 | 0.2521 | 0.2688 | 0.3567 | 0.0865 |
| STS | 0.5824 | 0.5655 | 0.6019 | 0.6134 | 0.3159 |

Table 1: Performance metrics across different models. Model 1: GCP PaLM2 (text-bison-001) w Fine-tune. Model 2: GCP PaLM2 (text-bison-001) w context and Fine-tune. Model 3: AWS Bedrock (Titan & Claude-v2) w RAG. Model 4: GCP PaLM2 (text-bison-001) w RAG. Model 5: Meta LLaMA2 (13b-chat-Q5) w RAG

| | question | answer_groundtruth | answer_prediction1 | answer_prediction2 | answer_prediction3 | answer_prediction4 | answer_prediction5 |
|----|---|---|---|---|--|--|--|
| 0 | What is SCG? | SCG (Scientific Computing Group) is a group th... | SCG is a company that provides software and se... | SCG is the Science Computing Group. | SCG is not defined or explained, so I don't k... | SCG is Scientific Computing Group | I don't know. The context does not mention SCG. |
| 1 | What are the benefits of using Subscription Ag... | Subscription Agreements can provide a consiste... | Subscription Agreements allow you to manage yo... | Subscription Agreements are used to ensure tha... | some of the key benefits of using the Lab's c... | Benefits of using Subscription Agreements inc... | Greetings! I'm here to help answer your quest... |
| 2 | Does it need to be moved elsewhere for the com... | Yes, it needs to be moved elsewhere for the co... | No Answer | The answer is no.\n\nThe context is about Scie... | Based on the context, I do not have enough in... | Yes, the data needs to be moved elsewhere for... | Based on the information provided, it does no... |
| 3 | What is the billing method? | LBL pays the bill at the end of the month, and... | The billing method is a credit card. | The billing method is a flat fee. | the billing method is:\n\nCharges are applied... | Both options feature centralized billing. LBL... | I don't know. Based on the context, there is ... |
| 25 | What is Lawrencium? | Lawrencium is a x86 Intel processor general pu... | Lawrencium is a chemical element with the symb... | Lawrencium is a chemical element with the symb... | Lawrencium is a x86 Intel processor general p... | Lawrencium is a x86 Intel processor general p... | I don't know the answer to that question. |
| 26 | How long does it take to procure compute nodes... | Generally, it takes about three months from st... | 1-2 weeks | The answer is: 12-18 months | it takes about three months from start to fin... | It takes about three months from start to fin... | I don't know, but I can try to help you find ... |
| 57 | What is SVM? | SVM stands for Scientific Virtual Machine, whi... | SVM stands for Support Vector Machine. | SVM is a machine learning algorithm that can b... | SVM (Scientific Virtual Machine) is an integr... | SVM stands for Scientific Virtual Machine Ser... | I'm not sure what you're referring to with "S... |
| 58 | What information should be included in the email? | The existing Lawrencium project name and the P... | The email should include the following informa... | The email should include the following informa... | I do not have enough information to determine... | The email should include the name of the pers... | it seems that you would need to include the f... |
| 59 | What are the benefits of using AutoML and Clou... | Access to services like AutoML and Cloud ML En... | AutoML and Cloud ML Engine can help you build ... | AutoML and Cloud ML Engine are two machine lea... | some of the key benefits of using AutoML and ... | The benefits of using AutoML and Cloud ML Eng... | I don't know. |

Open-Source Repo & Demo



AWS Generative AI Bedrock Workshop:

- <https://github.com/aws-samples/amazon-bedrock-workshop>

GCP Generative AI Examples:

- <https://github.com/GoogleCloudPlatform/generative-ai>



ScienceIT Chatbot (Demo):

- https://hpcs-stats.lbl.gov/cluster_utilization/state_hpc