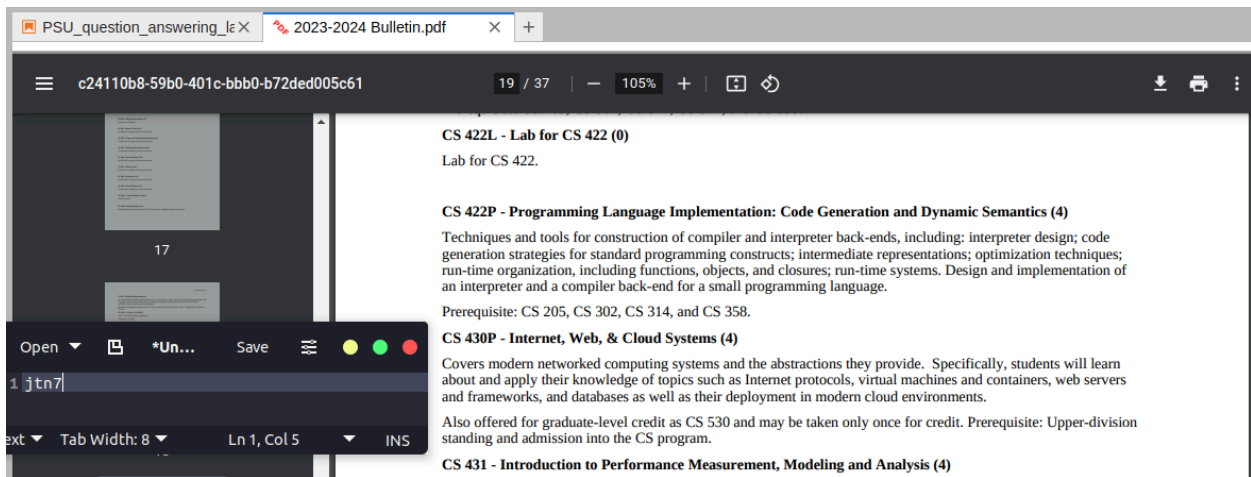
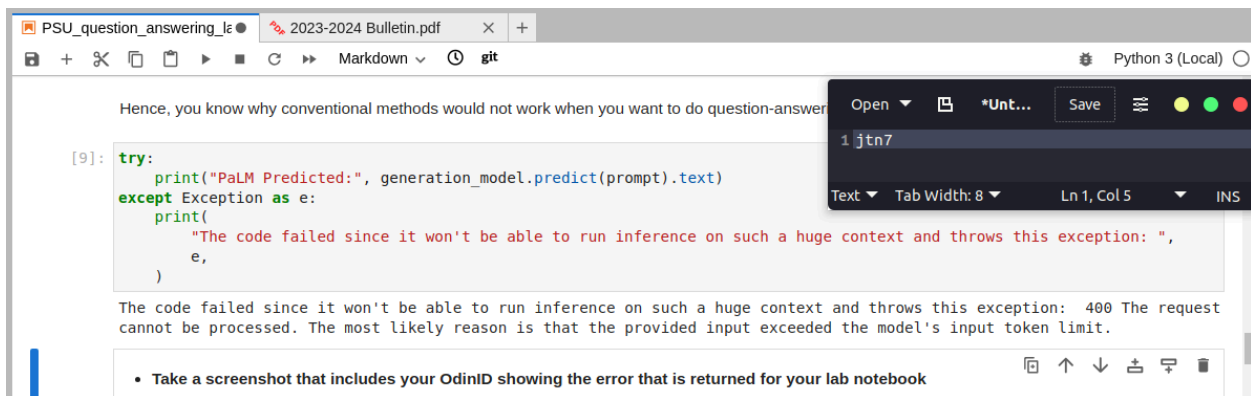


10.1g

- Take a screenshot that includes your OdinID showing the page number and the description of the class for your lab notebook



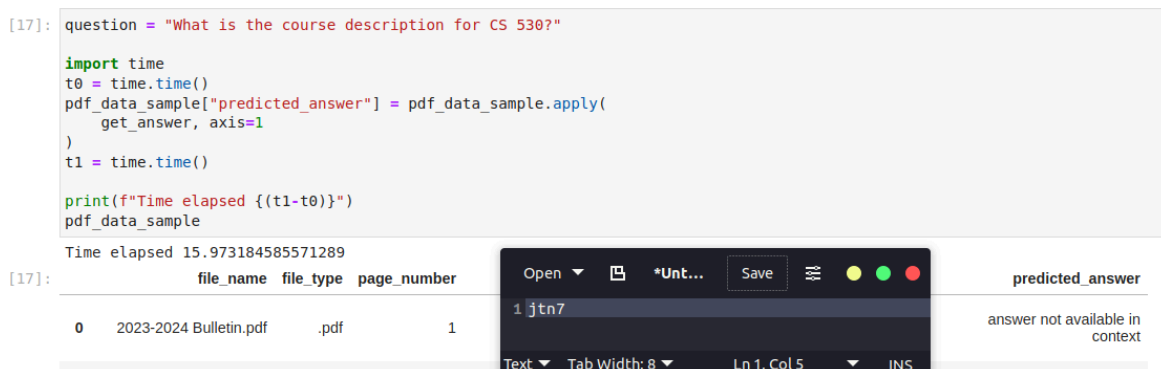
- Take a screenshot that includes your OdinID showing the error that is returned for your lab notebook



- Provide an explanation as to why the description is not returned for your lab notebook

The question passed in was asking about the course description for CS 530 but the first 5000 or so words in the document did not contain the answer to that question, thus it resulted in the program telling us that it did not find the answer.

- Take a screenshot including your OdinID that shows how long it took to perform the prediction across every chunk



- How many chunks returned predictions? 5
- Take a screenshot that includes your OdinID showing the result that is returned for your lab notebook

the prompt: Answer the question as precise as possible using the provided context. If the answer is not contained in the context, say "answer not available in context"

Context:

['Internet, Web, Cloud Systems', 'Internet, Web, Cloud Systems', 'Covers modern networked computing systems and the abstractions they provide Specifically, students will learn about and apply their knowledge of topics such as Internet protocols, virtual machines and containers, web servers and frameworks, and databases as well as their deployment in modern cloud environments', 'Covers modern networked computing systems and the abstractions they provide Specifically, students will learn about and apply their knowledge of topics such as Internet protocols, virtual machines and containers, web servers and frameworks, and databases as well as their deployment in modern cloud environments Also offered for graduate -level credit as CS 430P and may be taken only once for credit Prerequisite: Graduate - standing and admission into CS program', 'Advanced software design patterns using Java as the presentation language Course is suitable to software architects and developers who are already well -versed in this language In addition, it offers continuous opportunities for learning the most advanced features of the Java language and understanding some principles behind the design of its fundamental libraries Also offered as CS 653 and may be taken only once for credit Prerequisite: programming in Java and CS 520']?

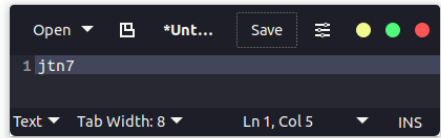
Question:

What is the course description for CS 530?

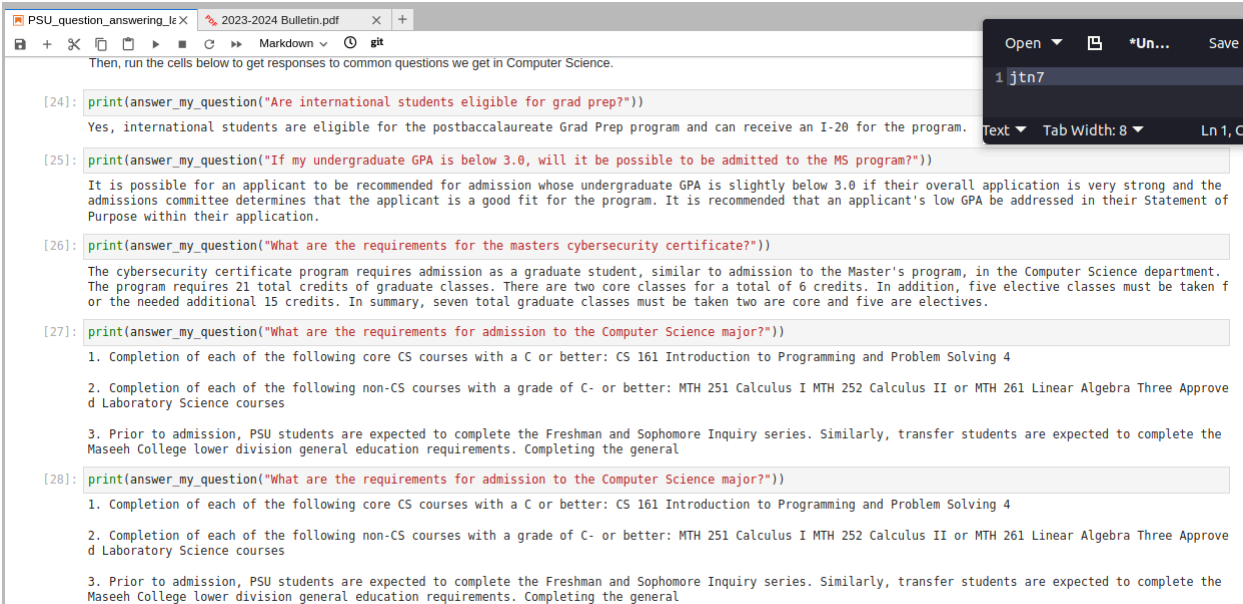
Answer:

the number of words in the prompt: 1623

PaLM Predicted: Covers modern networked computing systems and the abstractions they provide Specifically, students will learn about and apply their knowledge of topics such as Internet protocols, virtual machines and containers, web servers and frameworks, and databases as well as their deployment in modern cloud environments Also offered for graduate -level credit as CS 430P and may be taken only once for credit Prerequisite: Graduate - standing and admission into CS program



- Take a screenshot including your OdinID that shows the results of the queries



- Which of the approaches described would have issues with token limits on LLMs?

The first approach of stuffing where we did not set a limit to how many tokens to use and would pass all the data at once to query our question.

- Which of the approaches would result in the most queries for the LLM to handle? How many LLM requests are performed from a single user query in this approach?

The second approach of map reduce where we had multiple api calls by querying each individual chunk had a lot of queries for the LLM to handle and this could slow down the process drastically especially if the data to query is very large.

- Which of the approaches requires one to search a vector database for an appropriate context that is then sent to the LLM?

The map reduce with embeddings approach where we used an embedding model to create embedding from the chunks and then only passed in chunks that contained content similar to the question to the LLM when asking a question.

10.2g

- Take a screenshot of the output to include in your lab notebook. How many networks, subnetworks, and VM instances have been created?

```

Waiting for create [operation-1710471741096-613aa3c86c317-32080d3e-069a88d6]...done.

Create operation operation-1710471741096-613aa3c86c317-32080d3e-069a88d6 completed successfully.
NAME: asia-east1
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: asia1-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: e1-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: eu1-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: europe-west1
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: networking101
TYPE: compute.v1.network
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: us-east5
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: us-west-s1
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: us-west-s2
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: w1-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: w2-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

jtn7@cloudshell:~/networking101 (cloud-nguyen-jtn7)$

```

1 network and 5 subnetworks and 5 VM instances have been created.

- Visit the web console for VPC network and show the network and the subnetworks that have been created. Validate that it has created the infrastructure in the initial figure. Note the lack of firewall rules that have been created.

The screenshot shows the Google Cloud VPC network details page for a network named 'networking101'. The left sidebar lists various VPC network components, with 'VPC networks' selected. The main content area shows the 'SUBNETS' tab, displaying a table of subnets. The table has columns for Name, Region, Stack Type, Primary IPv4 range, and Secondary IPv4 ranges. Five subnets are listed: asia-east1, europe-west1, us-east5, us-west-s1, and us-west-s2.

Name	Region	Stack Type	Primary IPv4 range	Secondary IPv4 ranges
asia-east1	asia-east1	IPv4	10.40.0.0/16	
europe-west1	europe-west1	IPv4	10.30.0.0/16	
us-east5	us-east5	IPv4	10.20.0.0/16	
us-west-s1	us-west1	IPv4	10.10.0.0/16	
us-west-s2	us-west1	IPv4	10.11.0.0/16	

- Visit the web console for Compute Engine and show all VMs that have been created, their internal IP addresses and the subnetworks they have been instantiated on. Validate that it has created the infrastructure shown in the initial figure.

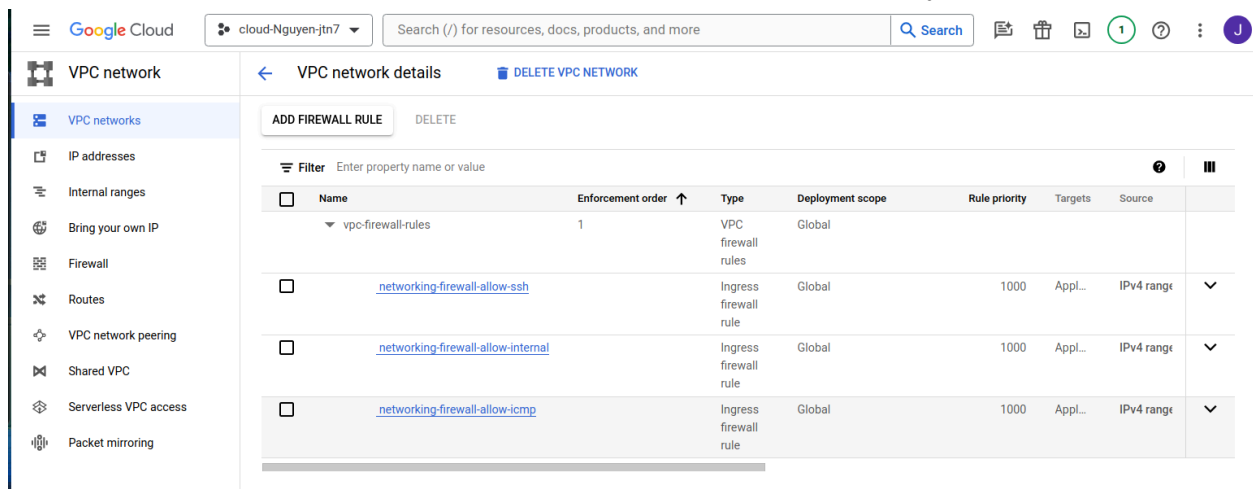
The screenshot shows the Google Cloud Compute Engine VM instances page. The left sidebar lists various Compute Engine components, with 'VM instances' selected. The main content area shows a table of VM instances. The table has columns for Status, Name, Zone, Recommendations, In use by, Internal IP, External IP, Network, and Connect. Five VM instances are listed: asia1-vm, e1-vm, eu1-vm, w1-vm, and w2-vm.

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Network	Connect
✓	asia1-vm	asia-east1-b			10.40.0.2 (nic0)	35.234.28.124 (nic0)	networking101	SSH
✓	e1-vm	us-east5-a			10.20.0.2 (nic0)	34.162.184.147 (nic0)	networking101	SSH
✓	eu1-vm	europe-west1-d			10.30.0.2 (nic0)	146.148.116.220 (nic0)	networking101	SSH
✓	w1-vm	us-west1-b			10.10.0.2 (nic0)	35.203.185.234 (nic0)	networking101	SSH
✓	w2-vm	us-west1-b			10.11.0.100 (nic0)	35.230.125.174 (nic0)	networking101	SSH

- Click on the ssh button for one of the VMs and attempt to connect. Did it succeed? No, the connection failed.

Visit the networking101 VPC network in the web UI

- Take a screenshot that indicates the new rules have been deployed



- Given this, fill in the table with the measured latencies between the 6 pairs and include it in your lab notebook. Use the shortest latency measured for each pair.

Location pair	Ideal latency	Measured latency
us-west1 us-east5	~45 ms	~53 ms
us-west1 europe-west1	~93 ms	~134 ms
us-west1 asia-east1	~114 ms	~119 ms
us-east5 europe-west1	~76 ms	~88 ms
us-east5 asia-east1	~141 ms	~169 ms
europe-west1 asia-east1	~110 ms	~251 ms

- Are the instances in the same availability zone or in different ones?

No, one is in us-east5-c while the other is in europe-west1-d zone.

- List all availability zones that your servers show up in for your lab notebook.

The us-east5 region only has one instance located in zone c while europe-west1 has 3 instances in zones b, c, d.

- Show a screenshot of the page that is returned. If you get an error, you may need to wait several minutes for the load balancer to finish deploying.

Google Cloud | cloud-nginx-jn7 | Search (/) for resources, docs, products, and more

Network services | Load balancer details | EDIT | DELETE | VIEW IN NETWORK TOPOLOGY

HTTP | 34.36.70.155:80 | Premium

Host and path rules

Hosts	Paths	Backend
All unmatched (default)	All unmatched (default)	webserver-backend-migs

Backend

Backend services

- webserver-backend-migs
 - Endpoint protocol: HTTP
 - Named port: http
 - Timeout: 30 seconds
 - Health check: [Instance health check](#)
 - Cloud CDN: Disabled
 - Logging: Disabled

[SHOW ADVANCED](#)

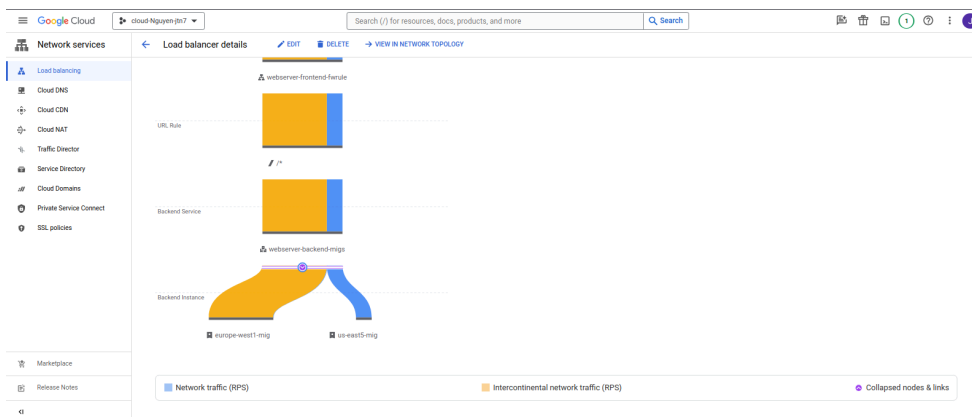
Backends

Name	Type	Scope	Healthy	Autoscaling	Balancing mode	Selected ports	Capacity
europe-west1-mig	Instance group	europe-west1	3 of 3	No configuration	Max backend utilization: 80%	80	100%
us-east5-mig	Instance group	us-east5	1 of 1	On: Target LB capacity fraction 80%	Max RPS: 50 (per instance)	80	100%

- Which availability zone does the server handling your request reside in?

us-east5-c

- Take a screenshot of the initial traffic distribution



Keep this window open for 5-10 minutes as the system adapts to the load and the UI updates.

- Take a screenshot of the UI as additional instances are brought up and show that the traffic distribution shifts



- Show a screenshot of the final traffic distribution.

