Team 2 - Spartans

Vasanth Kumaar S B
23z436

Santhosh T K
22z433

Praveen G
22z434

Problem Statement Chosen - 2

GCS - Google Cloud Services

We are asked to create a CLI (Command Line Interface) for Google Cloud Services (e.g., Google compute engine is a cloud service to perform computations). 🤔

Why don't we build a CLI that has some commands to interact with the GCS ?

"DON'T REINVENT THE WHEEL" 🤦

Google Cloud SDK : *It includes the Google Cloud Command Line Interface (gcloud CLI) and Cloud Client Libraries.*

Pain Points:

1. *Hard-to-Remember Commands & Syntax Complexity*
2. *Too Many Subcommands (e.g., compute, container)*

What we can bring to the table ?

Curious right !!

We have built a CLI tool, that reads ***Natural Language*** *and execute the corresponding gcloud command.*

- *Executes the generated gcloud commands automatically*
- *Ensures low latency and high correctness*
- *Works locally without relying on external APIs*

Attempt 1 - Rule-Based NLP (Regex + SpaCy)

Approach:

- *Used regex and rule-based parsing to extract keywords (e.g., "list TPU" → gcloud compute tpus list).*
- *Mapped input keywords to gcloud command templates.*

Limitations:

- ***Limited flexibility** – Cannot understand new sentence structures*
- ***High maintenance** – Adding new commands requires manual rule updates*
- ***Fails with variations** – Example: ("Show all TPU nodes" → Works, "Can you display all my TPUs?" → Fails)*

Why We Switched?

***Hard to scale** – Writing rules for every possible phrasing is not practical.*

***Not adaptable** – Cannot handle complex or unseen requests.*

Attempt 2 - Cloud-Based LLM (GPT-4 API, Google PaLM API)

Approach:

- *Used OpenAI GPT-4 and Google PaLM APIs to convert NL to gcloud CLI.*
- ***Example prompt:*** *Convert "List all TPU instances" into a valid gcloud command.*

Limitations:

- *Expensive – Each API request costs money per call.*
- *Latency Issues – Network-based APIs take 1-3 seconds per request.*

$$
\begin{aligned}
T_{\text{total}} &= T_{\text{conversion}} + T_{\text{execution}} \\
T_{\text{conversion}} &= \text{Time taken to convert natural language to gcloud command} \\
T_{\text{execution}} &= \text{Time taken to execute the generated gcloud command}
\end{aligned}
$$

Attempt 3 - Local LLM (Ollama + LLaMA 3)

Approach:

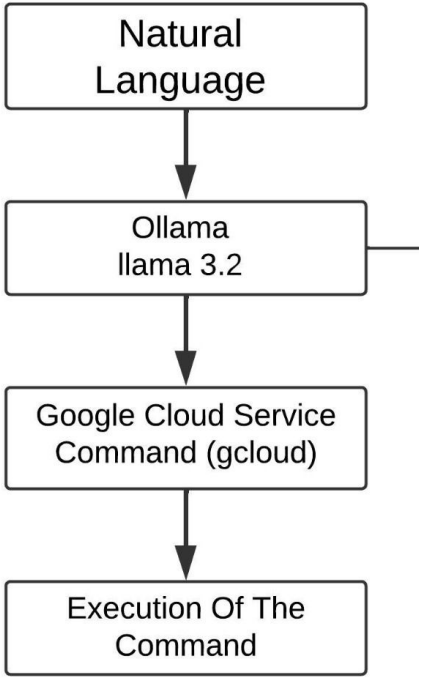
- *Used Ollama to run LLaMA 3 locally on the machine.*
- *Generates gcloud commands directly from natural language.*

Example conversion:

Input: "List all TPU nodes in zone us-central1"

Output: gcloud compute tpus list --zone=us-central1

High Level Architecture



1. *Latency Equation*

$$T_{\text{total}} = T_{\text{conversion}} + T_{\text{execution}}$$

Where:

- $T_{conversion}$ = *Time taken to convert natural language → gcloud command (Ollama processing time).*
- $T_{execution}$ = *Time taken to run the gcloud command and receive a response.*

2. Cost Equation

$$C_{\text{total}} = N_{\text{requests}} \times C_{\text{per\_request}}$$

Where:

$N_{requests}$ = *Number of API calls made.*

$C_{per\_request}$ = *Cost per API call (e.g., GPT-4 API costs $0.002 - $0.02 per request).*

For Ollama & Rule-Based NLP, cost is zero since they run locally:

$C_{total}=0$

3. CPU Usage

$$U_{\text{cpu}} = \frac{P_{\text{used}}}{P_{\text{available}}} \times 100$$

Where:

P_{used} = *CPU power consumed by the approach (LLM inference, regex, or API calls).*

$P_{available}$ = *Total available CPU power.*

Approximate CPU Usage by Approach:

$U_{cpu,Rule\text{-}Based} \approx 2\%$

$U_{cpu,Cloud\text{-}LLM} \approx 50\%$

$U_{cpu,Ollama\text{-}LLM} \approx 20\%$

4. Scalability Score

$$S = 1 - \frac{M_{\text{manual}}}{N_{\text{commands}}}$$

Where:

M_{manual} = *Number of manual rule updates needed for new commands.*

$N_{commands}$ = *Total number of commands supported.*

Scalability Estimates:

S_{Rule}*-Based≈0.3(Low, requires manual rule updates)*

S_{Cloud}*-LLM≈0.9(High, learns from cloud data)*

SOllama-LLM≈0.85(High, can be fine-tuned locally)

5. Accuracy Formula

$$A = \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100$$

Where:

$N_{correct}$ = *Number of correctly generated gcloud commands.*

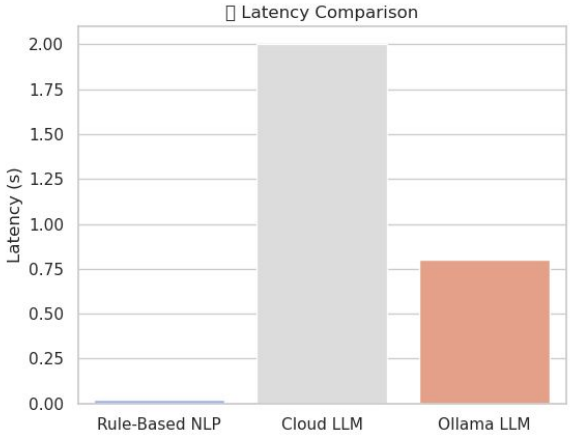
N_{total} = *Total test cases.*

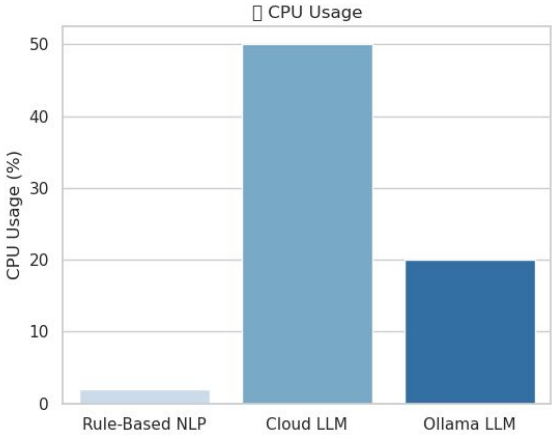
Estimated Accuracy by Approach:

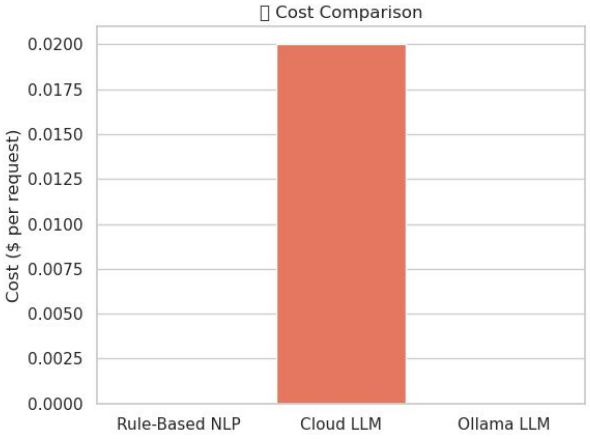
A_{Rule}*-Based≈50%(Fails on varied inputs)*

A_{Cloud}*-LLM≈95%(Very high, but costly)*

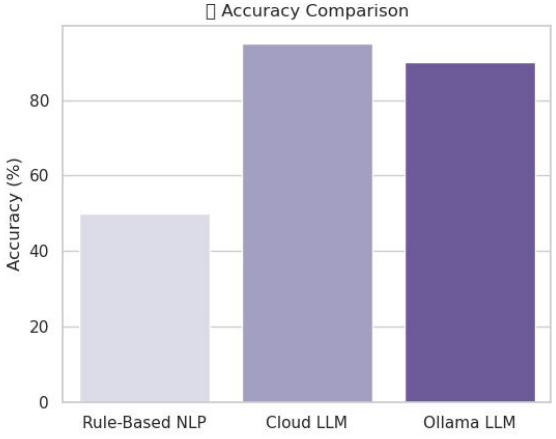
A_{Ollama}*-LLM≈90%(Nearly as good as Cloud, but offline)*


Latency Comparison
Latency (s)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Rule-Based NLP
Cloud LLM
Ollama LLM


CPU Usage
CPU Usage (%)
50
40
30
20
10
0
Rule-Based NLP
Cloud LLM
Ollama LLM


Cost Comparison
Cost ($ per request)
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
Rule-Based NLP
Cloud LLM
Ollama LLM


Scalability
Scalability Score (0-1)
0.8
0.6
0.4
0.2
0.0
Rule-Based NLP
Cloud LLM
Ollama LLM


Accuracy Comparison
Accuracy (%)
80
60
40
20
0
Rule-Based NLP
Cloud LLM
Ollama LLM

Windows PowerShell

```
PS C:\Users\spt\Desktop\Infinitum-Hackathon> python .\test.py
🔑 Authenticating with Google Cloud...
```

Sign in – Google accounts

https://accounts.google.com/o/oauth2/auth/oauthchooseaccount?response_type=code&client_id=764086051850-6qr4p6gpi6hn506pt8ejuq83di341hur.apps.googleusercontent.com&redirect_uri=http%3A%2F%

Sign in with Google

Choose an account

to continue to Google Auth Library

22Z262 - SRI DEV S
22z262@psgtech.ac.in
Signed out

22Z433 - SANTHOSH T K
22z433@psgtech.ac.in

Use another account

Before using this app, you can review Google Auth Library's privacy policy and Terms of Service.

English (United Kingdom)

Help Privacy Terms

Search

ENG IN

15:49
07-03-2025

Windows PowerShell

```
PS C:\Users\spt\Desktop\Infinitum-Hackathon> python .\test.py
🔑 Authenticating with Google Cloud...
✅ Authentication successful!


📝 Enter a command (or type 'exit' to quit): |
```

Windows PowerShell

```
PS C:\Users\spt\Desktop\Infinitum-Hackathon> python .\test.py
🔑 Authenticating with Google Cloud...
✅ Authentication successful!


📝 Enter a command (or type 'exit' to quit): start an already existing instance with name instance-20250307-092440 with zone us-central1-c

🔍 Raw LLaMA Output: 'gcloud compute instances start instance-20250307-092440 --zone us-central1-c'

🖥️ Generated Command: gcloud compute instances start instance-20250307-092440 --zone us-central1-c
⏳ Time taken: 4.85 seconds

🚀 Executing: 'gcloud compute instances start instance-20250307-092440 --zone us-central1-c'
```

instance-2025... EDIT RESET CREATE MACHINE IMAGI

DETAILS OBSERVABILITY OS INFO SCREENSHOT

SSH CONNECT TO SERIAL CONSOLE

Connecting to serial ports is disabled

Logs

Logging
Serial port 1 (console)

SHOW MORE

Basic information

Name	instance-20250307-092440
Instance Id	8301981463392748941
Description	None
Type	Instance
Status	Staging
Creation time	Mar 7, 2025, 2:54:51 PM UTC+05:30
Location	us-central1-c
Instance template	None
In use by	None
Physical host	None
Maintenance status	—
Reservations	Automatically choose
Labels	goog-ops-a... : v2-x86-tem...
Tags	—

Windows PowerShell

```
PS C:\Users\spt\Desktop\Infinitum-Hackathon> python .\test.py
🔑 Authenticating with Google Cloud...
✅ Authentication successful!


📝 Enter a command (or type 'exit' to quit): start an already existing instance with name instance-20250307-092440 with zone us-central1-c

🔍 Raw LLaMA Output: 'gcloud compute instances start instance-20250307-092440 --zone us-central1-c'

🖥️ Generated Command: gcloud compute instances start instance-20250307-092440 --zone us-central1-c
⏳ Time taken: 4.85 seconds

🚀 Executing: 'gcloud compute instances start instance-20250307-092440 --zone us-central1-c'

✅ Command executed successfully!


📝 Enter a command (or type 'exit' to quit): exit
```

instance-2025... EDIT RESET CREATE MACHINE IMAGE C

DETAILS OBSERVABILITY OS INFO SCREENSHOT

SSH CONNECT TO SERIAL CONSOLE

Connecting to serial ports is disabled

Logs

Logging
Serial port 1 (console)

SHOW MORE

Basic information

Name	instance-20250307-092440
Instance Id	8301981463392748941
Description	None
Type	Instance
Status	Running
Creation time	Mar 7, 2025, 2:54:51 PM UTC+05:30
Location	us-central1-c
Instance template	None
In use by	None
Physical host	None
Maintenance status	—
Reservations	Automatically choose
Labels	goog-ops-a... : v2-x86-tem...
Tags	—

Windows PowerShell

```
PS C:\Users\spt\Desktop\Infinitum-Hackathon> python .\test.py
🔑 Authenticating with Google Cloud...
✅ Authentication successful!


📝 Enter a command (or type 'exit' to quit): start an already existing instance with name instance-20250307-092440 with zone us-central1-c

🔍 Raw LLaMA Output: 'gcloud compute instances start instance-20250307-092440 --zone us-central1-c'

🖥️ Generated Command: gcloud compute instances start instance-20250307-092440 --zone us-central1-c
⏳ Time taken: 4.85 seconds

🚀 Executing: 'gcloud compute instances start instance-20250307-092440 --zone us-central1-c'

✅ Command executed successfully!



📝 Enter a command (or type 'exit' to quit): exit
👋 Exiting CLI tool.
PS C:\Users\spt\Desktop\Infinitum-Hackathon>
```