Summary:
$0.4452 (g5.xlarge cost) x 3 (models count) x 4 (lang count) x 12 (eval time) + buffer = $90

Project Overview

In the domain of code generation, deep learning models often face challenges in generating competent outputs for various programming languages. Our project aims to address this issue by leveraging the Mixture-of-Experts (MoE) framework, which employs specialized models for each programming language. By doing so, we expect to enhance the precision and efficiency of code generation while reducing the reliance on expensive, high-end GPUs typically required for large, monolithic models.

Our proposed MoE framework comprises multiple expert models, each tailored to a specific programming language. For instance, if the input code snippet is written in C, the framework will route the task to the expert model specializing in C, thereby ensuring a more accurate and context-appropriate output. This approach intends to harness the strengths of each expert model, allowing the system to outperform traditional models that struggle to maintain consistent performance across different languages.

Resource Requirements

To implement and evaluate our MoE framework, we require EC2 compute time. Our project will involve the assessment of three distinct models across four programming languages. These models include:

1. A baseline model fine-tuned on all four languages collectively.
2. A model created by merging techniques from four different models, each fine-tuned on a specific language.
3. A model created by MoE framework consisting of four distinct models, each fine-tuned on a specific language.

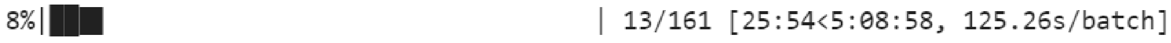
Given the time needed to evaluate one model for one language (approximately half a day) [Fig 1], we anticipate that the entire evaluation process will span 10 days. We have also allocated an additional 3-day buffer to accommodate any unforeseen challenges that may arise during the evaluation phase. Additionally, the evaluation library we are using does not support multi-GPU setups, necessitating a single instance with a minimum of 20 GB VRAM [Fig 2]. The suitable GPU for this requirement is the g5.xlarge instance type [Fig 3].

Budget and Cost Estimation

Based on the Amazon EC2 pricing calculator, we estimate that our project will require a single g5.xlarge instance with an On Demand hourly cost of $0.4452. Considering the projected compute time of 200 hours in a month, the total cost for On-Demand instances is expected to be $89.04 [Fig 4]. This budget estimation accounts for the necessary resources to conduct a comprehensive evaluation of our proposed MoE framework for code generation.

Please find the attached specifications from AWS and GPU usage, along with the detailed cost breakdown from the EC2 pricing calculator for your reference.

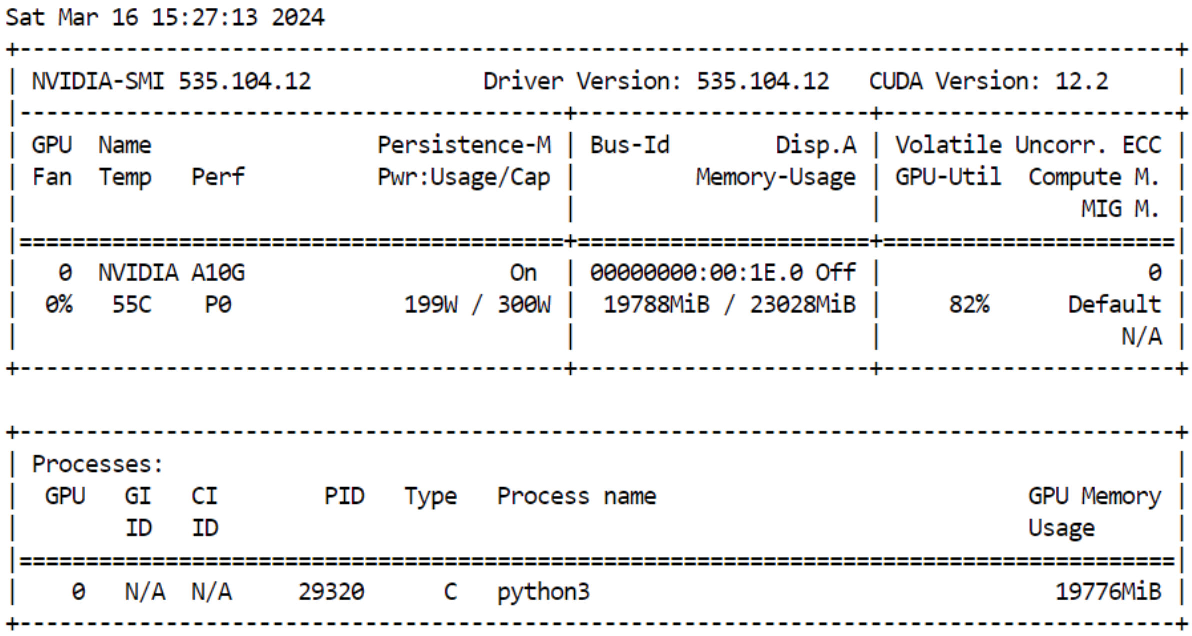
Summary:
$0.4452 (g5.xlarge cost) x 3 (models count) x 4 (lang count) x 12 (eval time) + buffer = $90



```
 8%|███                                         | 13/161 [25:54<5:08:58, 125.26s/batch]
```

Figure 1: Evaluation Script Runtime and Pre/Post-Processing Time

This figure captures the time taken to complete the evaluation script for our MoE framework, as measured during the script's execution. Additionally, it does not account for the approximate six-hour duration of pre and post-processing of results, which, when combined, extends the total evaluation time to approximately half a day per model and language.

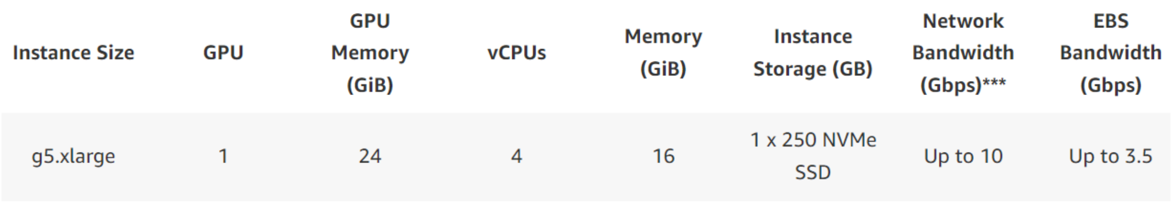


```
Sat Mar 16 15:27:13 2024
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 535.104.12             Driver Version: 535.104.12   CUDA Version: 12.2     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  NVIDIA A10G                    On  | 00000000:00:1E.0 Off |                    0 |
|  0%   55C    P0             199W / 300W |  19788MiB / 23028MiB |     82%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
|    0   N/A  N/A     29320      C   python3                                   19776MiB |
+---------------------------------------------------------------------------------------+
```

Figure 2: GPU VRAM Usage for MoE Evaluation using the nvidia-smi command

This figure illustrates the GPU VRAM usage during the evaluation of our Mixture-of-Experts (MoE) framework, as monitored through the nvidia-smi command.

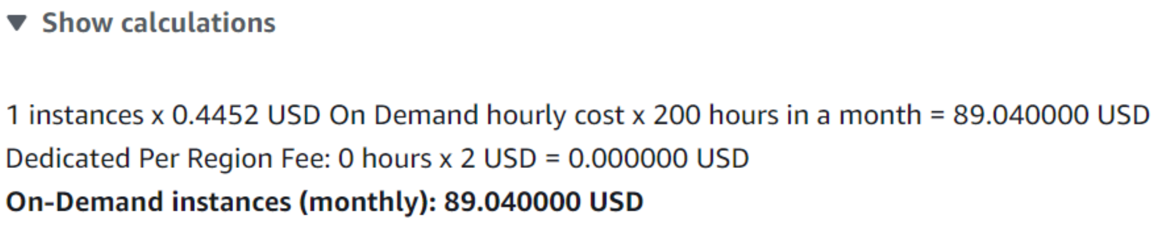


Instance Size	GPU	GPU Memory (GiB)	vCPUs	Memory (GiB)	Instance Storage (GB)	Network Bandwidth (Gbps)***	EBS Bandwidth (Gbps)
g5.xlarge	1	24	4	16	1 x 250 NVMe SSD	Up to 10	Up to 3.5

Figure 3: EC2 Instance Type Specification from AWS

This figure presents the specifications of the g5.xlarge EC2 instance type required for the evaluation of our MoE framework, as obtained from Amazon Web Services (AWS).

Summary:
$0.4452 (g5.xlarge cost) x 3 (models count) x 4 (lang count) x 12 (eval time) + buffer = $90



▼ Show calculations

1 instances x 0.4452 USD On Demand hourly cost x 200 hours in a month = 89.040000 USD

Dedicated Per Region Fee: 0 hours x 2 USD = 0.000000 USD

On-Demand instances (monthly): 89.040000 USD

Figure 4: Cost Estimation from AWS EC2 Pricing Calculator

This figure provides a detailed cost estimation for the On-Demand g5.xlarge EC2 instance required, as generated by the Amazon EC2 Pricing Calculator.