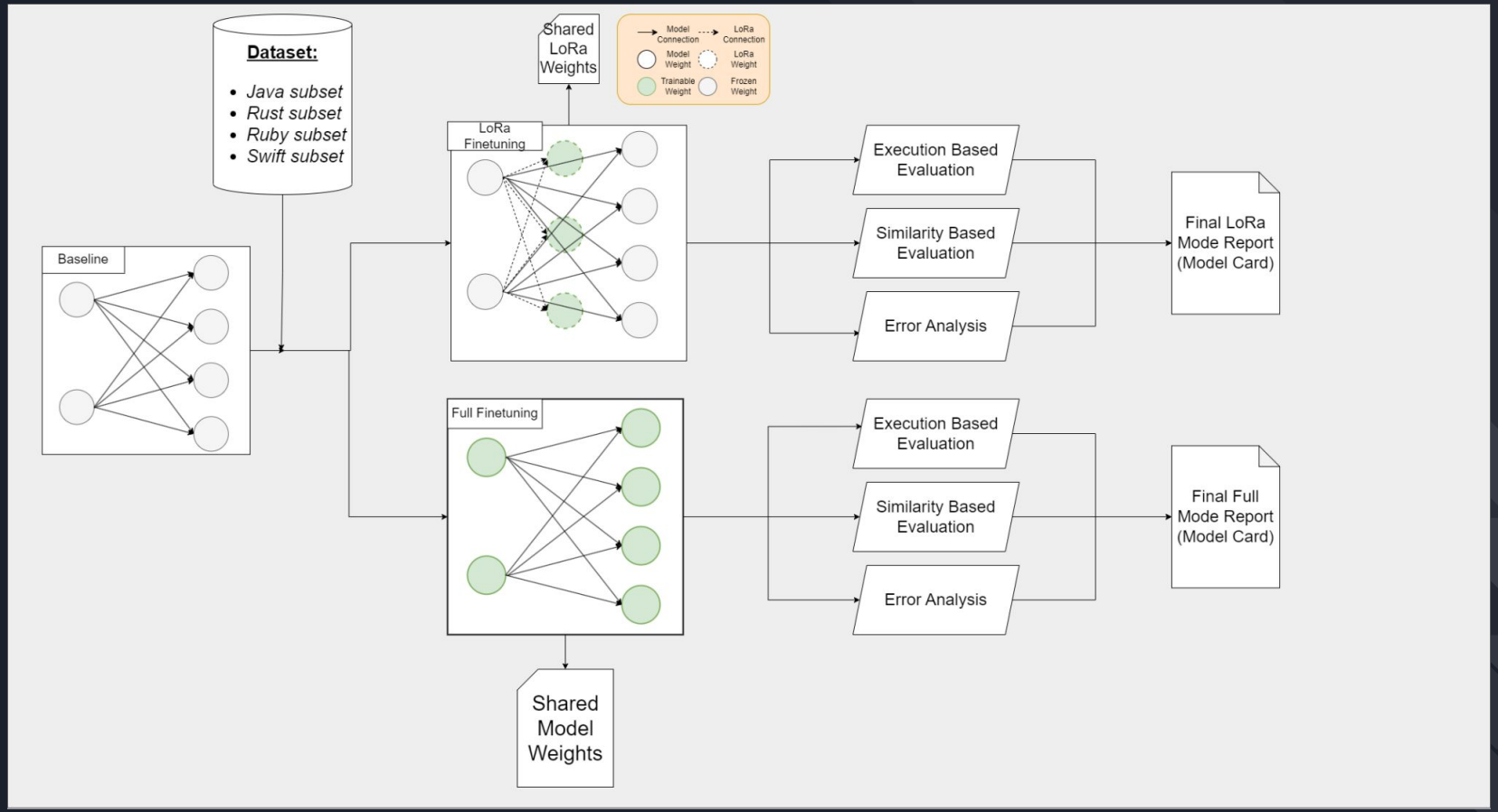
Scaling Down Multilingual Code Language Models

Objectives

- Sharing models weights is not enough to enable equitable access to Code LLMs because of the **Resource Limitations** and the **Knowledge Gap**.
- This project aims to:
    a. Provide a set of small code language models that can perform code completion in multiple programming languages (resource limitations)
    b. Provide a framework that practitioners can follow to fine-tune code LLMs to their own needs and according to their own constraints (knowledge gap)

Methodology



Baseline & Dataset

Dataset: 1 Million files of target programming languages, cleaned and filtered.

Baseline: CodeGen-350M-Mono Autoregressive Code LLM pretrained on Python.

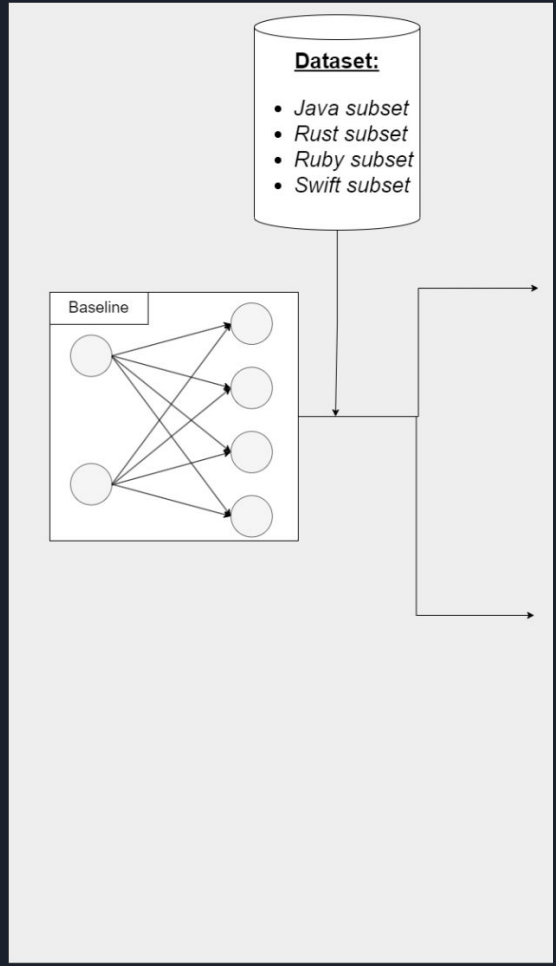
Task: Casual Language Modelling

$$L_{CLM}^{(x)} = -\frac{1}{|x|}\sum_{i=1}^{|x|} logP(x_i/x_{<i})$$

where:

$x = \{x_1, x_2, x_3, \ldots, x_{|x|}\}$ represents a sequence

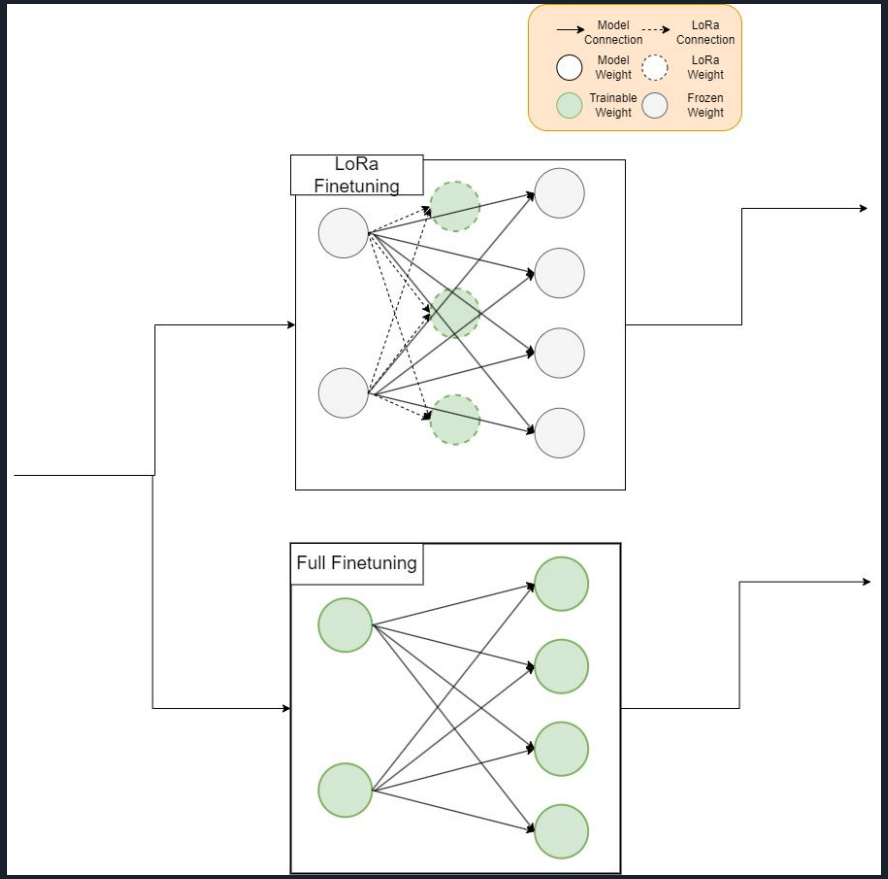
$x_{<i} = x_1, x_2, x_3, ..x_{i-1}$



Finetune Methods

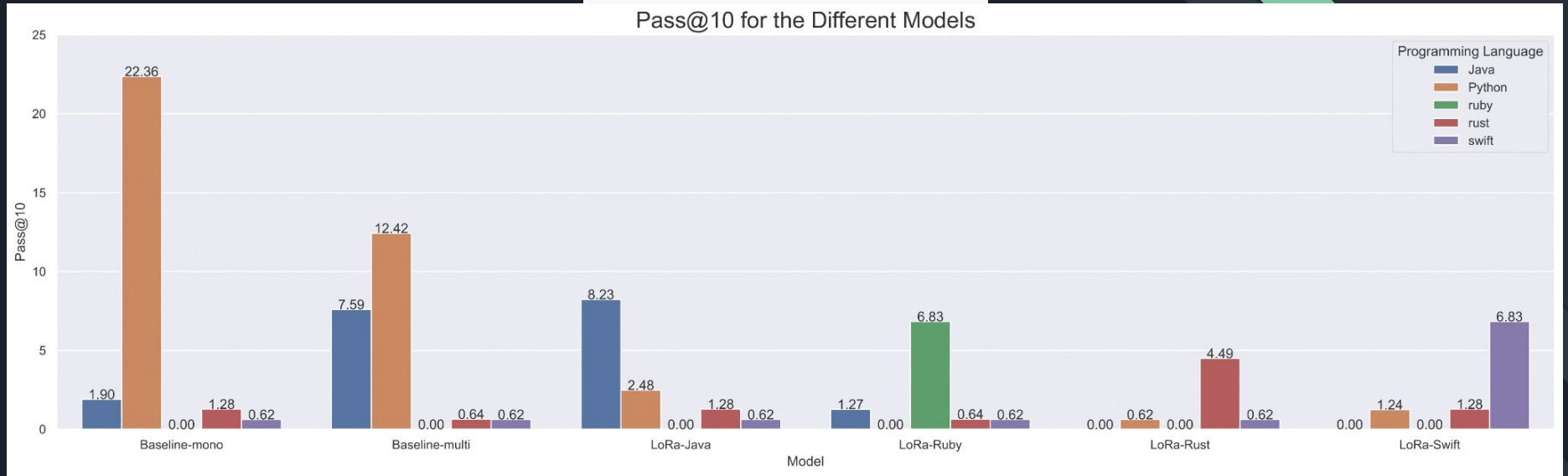
Full: All the parameters of the model are trainable.

Low Rank Adaptation (LoRa): Small set of Parameters (inserted at specific layers) are trainable

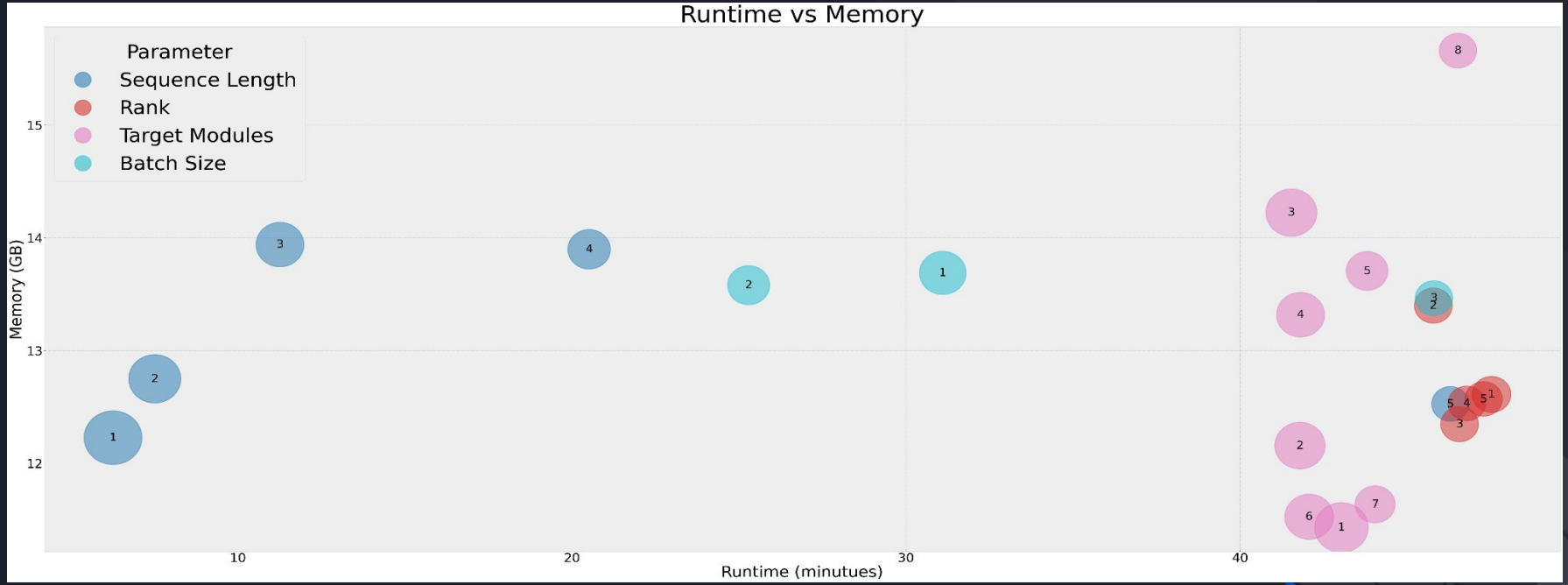


Results

```python
from typing import List

def median(l: List[int]) -> float:
    """Return median of elements in the list l.
    >>> median([3, 1, 2, 4, 5])
    3
    >>> median([-10, 4, 6, 1000, 10, 20])
    15.0
    """


def check(candidate):
    assert candidate([3, 1, 2, 4, 5]) == 3
    assert candidate([-10, 4, 6, 1000, 10, 20]) == 8.0
    assert candidate([5]) == 5
    assert candidate([6, 5]) == 5.5
    assert candidate([8, 1, 3, 9, 9, 2, 7]) == 7

def test_check():
    check(median)

test_check()
```

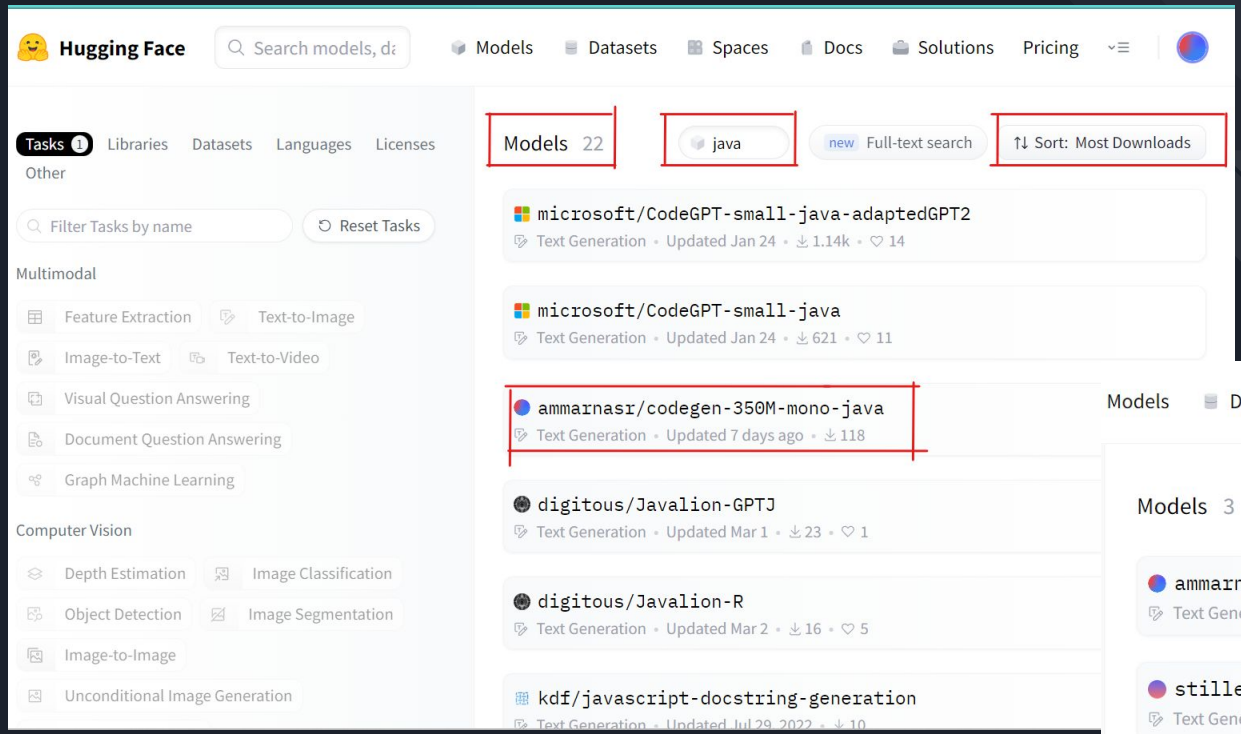


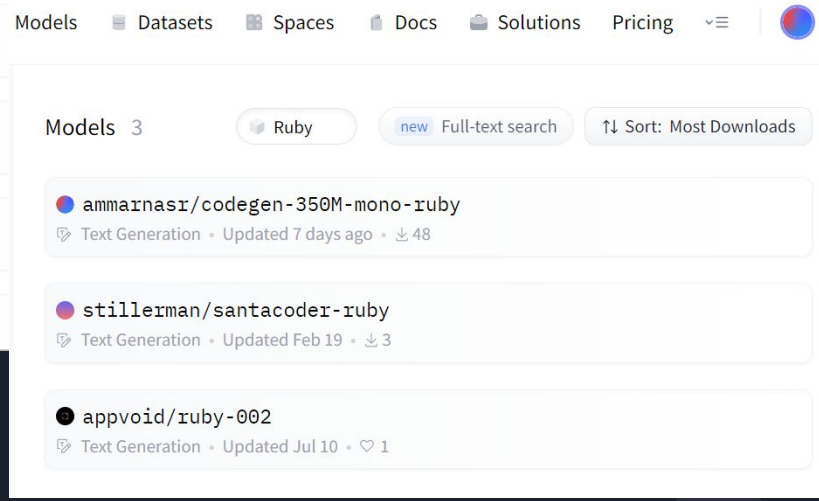
Trade-Off Analysis



Variable	Initial Value	Range
LoRa Rank	64	[8, 16, 32, 64, 128]
Batch Size	2	[1, 2, 4]
Sequence Length	1024	[128, 256, 512, 1024, 2048]
Learning Rate	5e-5	[5e-6, 1e-6, 5e-5, 5e-4]
Target Modules	Conf-6	[Attention Modules, Language Feature Modules, Fully Connected Module]

Open-Source (Inference Demo)





Thank You!

Questions ?