**Evaluating Open-Source AI Models for Student Competence Analysis in Python**

**(FOSSEE Internship Screening Task)**

**SETUP**

**Steps to Setup:**

1. Set Up Python Environment

* Install Python 3.8
* Create a virtual environment (to keep dependencies clean):

python -m venv venv

venv\Scripts\activate # on Windows

2. Install Required Libraries

Install Hugging Face’s transformers and torch (PyTorch backend).

pip install torch transformers datasets

3. Load GPT-Neo Model

Use Hugging Face transformers to download and load GPT-Neo. Example (2.7B model):

from transformers import pipeline

# Load GPT-Neo model (2.7B parameters)

generator = pipeline("text-generation", model="EleutherAI/gpt-neo-2.7B")

# Quick test

result = generator("Explain Newton's First Law in simple words:", max\_length=60, do\_sample=True, temperature=0.7)

print(result[0]['generated\_text'])

4. Define Student Competence Task

Decide what *student competence analysis* means in your case:

* **Text Classification**: Evaluate student answers as “Correct”, “Partially Correct”, “Incorrect”.
* **Summarization**: Summarize long answers into key points.
* **Feedback Generation**: Give constructive feedback on student responses.

You can fine-tune GPT-Neo on a dataset of student answers, or use prompt-based evaluation.

5. Prepare Dataset

* Collect or create a small dataset

[

{"question": "What is Newton's First Law?", "answer": "An object remains at rest...", "label": "Correct"},

{"question": "Explain photosynthesis.", "answer": "Plants eat oxygen directly.", "label": "Incorrect"}

]

* You can use Hugging Face datasets library to load and preprocess it

from datasets import Dataset

data = [

{"question": "What is Newton's First Law?", "answer": "An object remains at rest...", "label": "Correct"},

{"question": "Explain photosynthesis.", "answer": "Plants eat oxygen directly.", "label": "Incorrect"}

]

dataset = Dataset.from\_list(data)

print(dataset)

6. Run Analysis

7. Evaluation Metrics

* If you fine-tune the model: use metrics like **Accuracy, F1-score** (for classification tasks).
* If you’re only prompting: measure **qualitative correctness** by comparing outputs with expected feedback.