



Prepared by **Annanahmed Shaikh, Rohan Pratap Reddy Ravula, GROUP 5**

A Novel Fine-Tuning Approach Using Backward Attention Mechanism

Enhancing Math Reasoning in GPT-2 with a Novel Final Attention Head: A Fine-Tuning
Approach on
OpenMathInstruct-2

02 April, 2025

Course Name: **Adv Topics in Large Lang Model**
Subject Code: **DATA 6300**



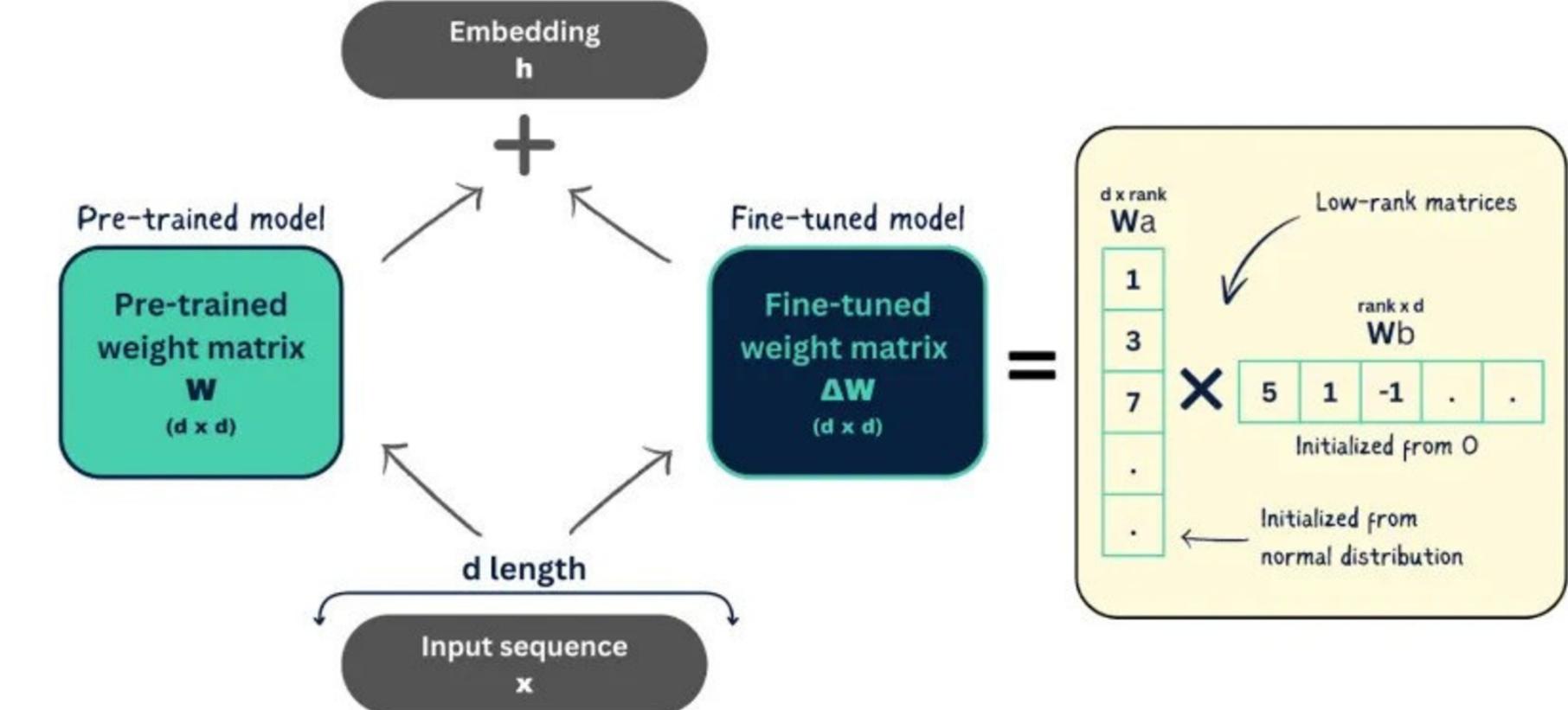
WENTWORTH INSTITUTE OF TECHNOLOGY
Professor Salem Othman



Motivation & Problem Statement

- Traditional token selection methods (greedy, top-k) lack robustness in math-based reasoning.
- Manual tuning (temperature, top-k) required for quality generation.
- Full model fine-tuning is computationally expensive.
- Our Goal: Enhance GPT-2's reasoning capability by replacing the final linear head with a Backward Attention module.
- Inspired by LoRA and QLoRA to maintain efficiency.

Low Rank Adaptation (LoRA) Overview



Dataset Overview

- Dataset Used: [OpenMathInstruct-2 \(by NVIDIA\)](#).
- Type: Math reasoning dataset
- Size: 12.6 GB (generated using Llama3.1-405B-Instruct)
- Input: "problem"
- Output: "generated solution" + "expected answer" (merged)
- Format: Supervised Q&A style fine-tuning



Screenshot of the Datasets page on the NVIDIA website showing the OpenMathInstruct-2 dataset.

The dataset has 1M rows and is split into 4 parts. The columns are:

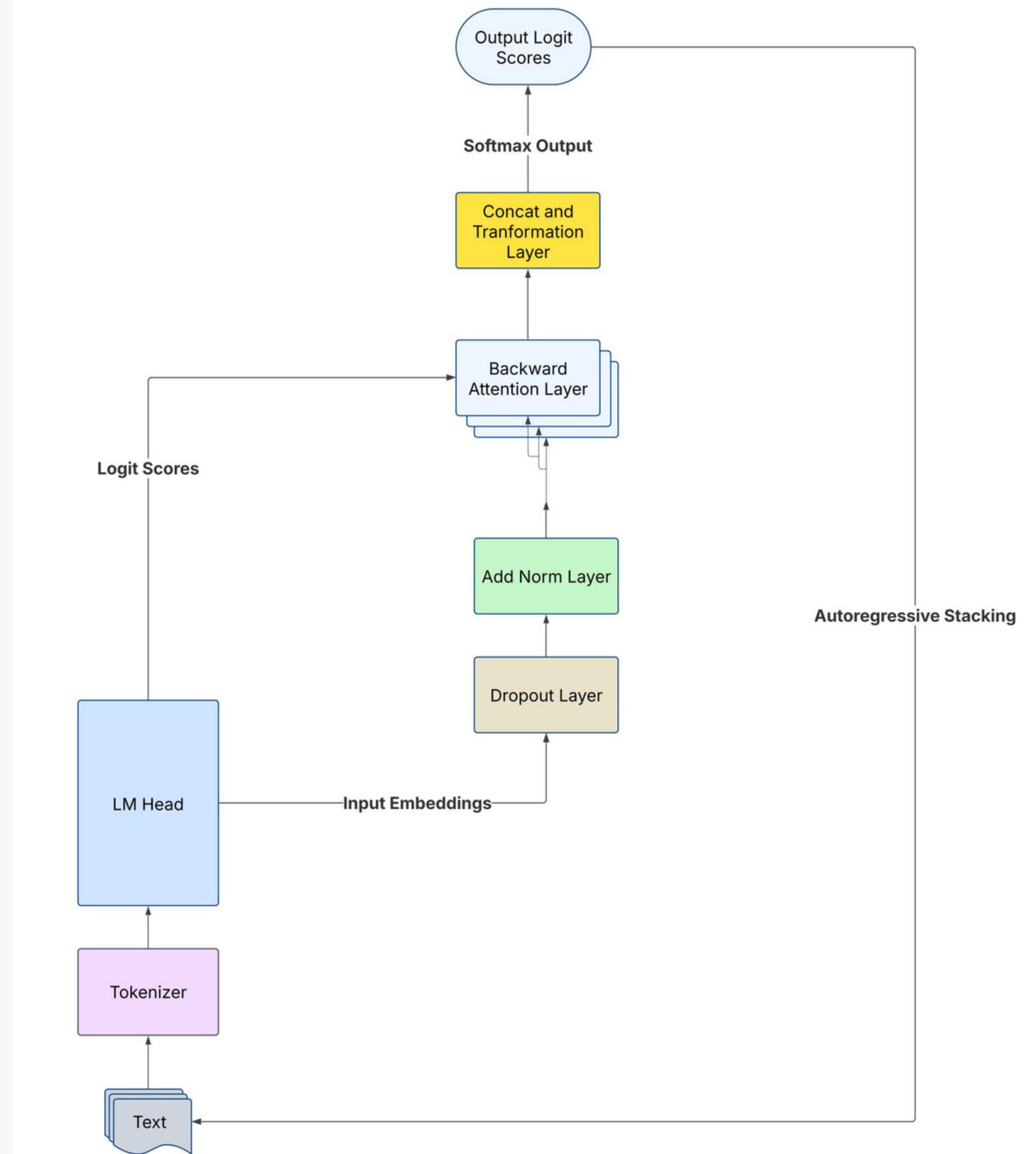
- problem: string · lengths
- generated_solution: string · lengths
- expected_answer: string · lengths
- problem_source: string · classes

The table displays several rows of math problems and their solutions. The first few rows are:

problem	generated_solution	expected_answer	problem_source
Solve for y : $\frac{y^2 - 3y + 2}{y - 2} = y + 1$	Start by multiplying both sides by $y - 2$ to eliminate the denominator: $(y^2 - 3y + 2) = (y + 1)(y - 2)$	2	augmented_math
Given a circle centered at the origin, a point A is translated along the circle to a new point B . The rotation matrix for rotating a point (x, y) by an angle θ counterclockwise is $\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$.	The rotation matrix for rotating a point (x, y) by an angle θ counterclockwise is $\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$.	(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})	augmented_math
The repeating decimal $0.\overline{099}$ can be expressed as a fraction $\frac{b}{100}$ in lowest terms. Let m and n be two positive integers such that their greatest common divisor is 18 and $m > n$.	To express $0.\overline{099}$ as a fraction, let $x = 0.\overline{099}$. Multiply both sides by 1000 to get $1000x = 99.999\ldots$. Then, subtract x from $1000x$ to get $999x = 99$, or $x = \frac{99}{999} = \frac{1}{11}$.	224	augmented_math
Let (x, y) be a solution to the system of equations: $\begin{cases} x^2 + \lfloor y \rfloor = 25 \\ \lfloor x \rfloor + y^2 = 25 \end{cases}$	## Step 1: Recall the relationship between GCD and LCM of two numbers. For any two positive integers a and b , $\text{GCD}(a, b) \cdot \text{LCM}(a, b) = a \cdot b$.	4	augmented_math
Emma visited a local farm and saw 120 baby chicks in the coop. She asked the farmer if she could take some home. The farmer allowed Emma to take one-third of the 120 chicks.	We have two equations: $x^2 + \lfloor y \rfloor = 25$ and $\lfloor x \rfloor + y^2 = 25$.	9	augmented_math
In a group of 10 friends, each person has a different favorite sports team. At a party, each friend can only talk about their own favorite team. How many different ways can they talk about their favorite teams?	Let's break down the problem step by step. The farmer allowed Emma to take one-third of the 120 chicks.	14	augmented_gsm8k
A group of friends want to share some pizzas that come in three different sizes: small, medium, and large. They want to have $\frac{1}{3}$ of a pizza left over, so they need to order 12 pizzas.	We have 10 friends, with 5 supporting teams from one league and 5 from another. First, let's find the number of ways to choose 5 teams from 10.	21	augmented_math
The difference between the cube of a number and twice the number is equal to the square of the number.	The group has 12 people and they want to have $\frac{1}{3}$ of a pizza left over, so they need to order 12 pizzas.	3	augmented_math
Eva is planning a 240-mile road trip. She decides to drive 20 miles for 10 days. How many miles does she have left to drive?	Let x be the number. According to the problem, we have the equation: $x^3 - 2x = x^2 - 3x$.	0	augmented_math
A right triangle with one side of length 10 cm and a hypotenuse of length 26 cm is rotated around its hypotenuse.	Eva needs to drive a total of 240 miles. She drives 20 miles a day for 10 days which amounts to 200 miles.	40	augmented_gsm8k
Find the smallest positive integer that is divisible by both 5 and 7, and has exactly three digits.	To find the volume of the double cone, we need to first find the radius and height of each cone.	70	augmented_math
Consider two lines: line p parametrized as $\begin{cases} x = 3 + 2t \\ y = -2 + 5t \end{cases}$	To find the smallest positive integer that is divisible by both 5 and 7, and has exactly three digits, we need to find the least common multiple of 5 and 7.	15/7 \begin{pmatrix} 1 & 2 \\ 3 & 5 \end{pmatrix}	augmented_math

Model Architecture

- GPT-2 (Frozen)
- Additional Backward Attention Layer
- Flow: Text → Tokenizer → Embeddings → K, V, Logits
Weighted V → Query → Dot(K,Q) → Softmax → Output
- Output: Refined probabilities for next-token prediction



Backward Attention Explained



-
- Step 1: Normalize input embeddings (RMSNorm)
 - Step 2: Project to latent space generate Key (K) and Value (V)
 - Step 3: Weight V using LM logits Sum Normalize Query (Q)
 - Step 4: Dot(Q, K) Softmax Output Probabilities
 - Removes need for temperature/top-k tuning
 - Adds randomness for robustness in token selection
-



Pseudo Code

- **Input:**

- IE Input Embeddings (Vocab Size × Embedding Dimension)
- Logits Output logits from frozen GPT-2
- Config Model configuration (dims, flags, device)

- **Procedure:**

- 1. Normalize IE using RMS normalization.
- K IE × WK
- V IE × WV

Compute weighted vector:

- Weighted-V Logits × V // Scale value vectors
- Dot Sum (Weighted-V) / Vocab Size

Project dot product:

- Q Dot × WQ // Query vector

Attention score:

- Attention Q × Transpose(K)

Output:

- Output SoftMax (Attention)
- Output token probabilities

Training Strategy



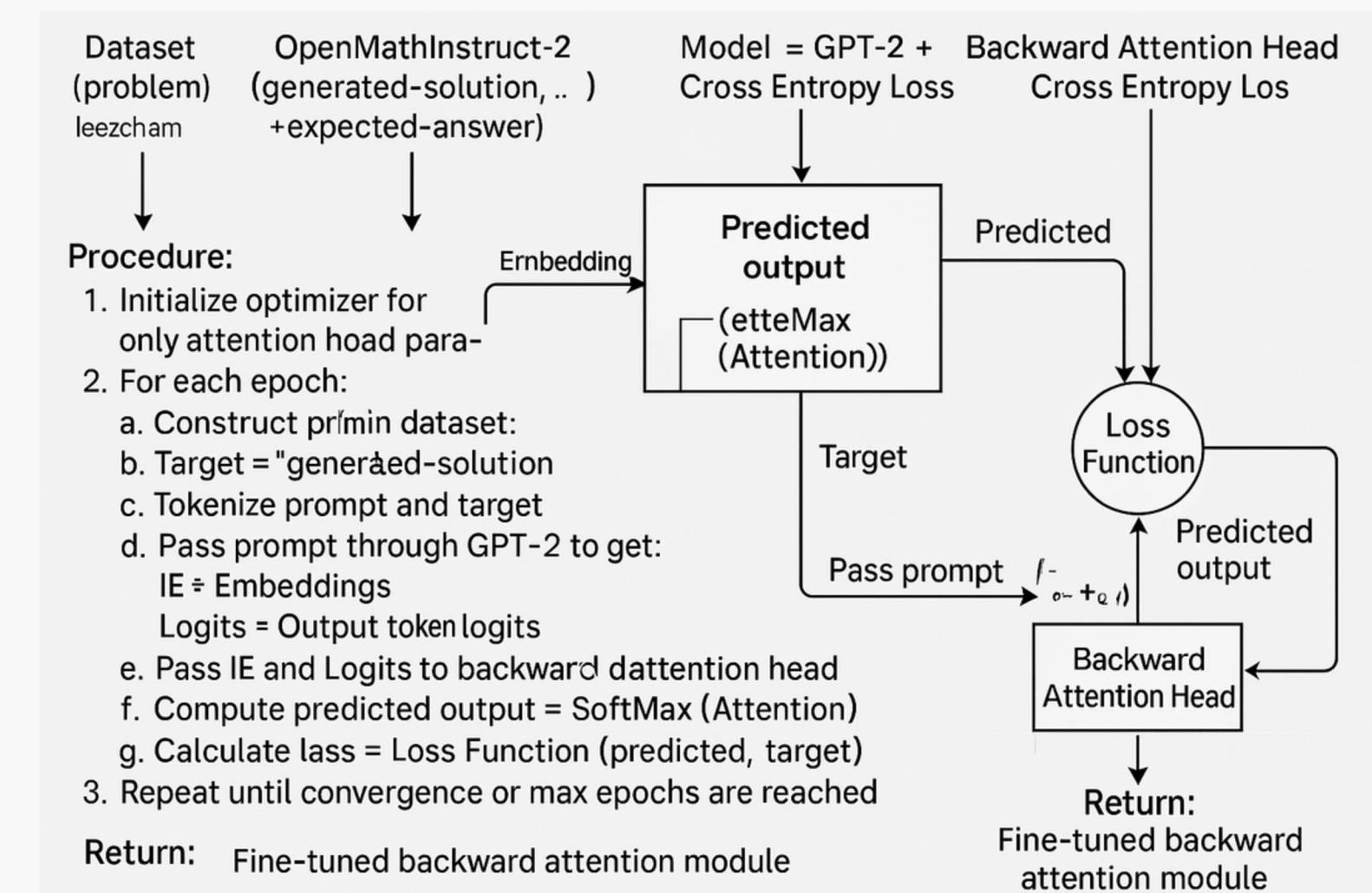
- Approach: Supervised Fine-Tuning with Teacher Forcing
- Loss Function: Cross-Entropy
- Metrics:
 - Perplexity
 - MSE, RMSE (for numeric answers)
 - Accuracy, Precision, Recall, F1
- Limitation: Training not completed due to hardware constraints
- But full training pipeline + modular architecture is implemented

Due to lack of system hardware, we couldn't train our model on time.



Evaluation Strategy

- Baselines:
 - GPT-2 fine-tuned on academic topics
 - Llama3.1-405B-Instruct (original dataset generator)
 - Expected Gains:
 - Improved token precision in math contexts
 - Better generalization through probabilistic attention
 - Metrics Focus: Quantitative (accuracy), Qualitative (reasoning clarity)



Future Work



- Multi-Head/Layered Backward Attention: Better contextual depth
- Dynamic Chunking: Optimize memory + performance
- Non-linear transformations: Add complexity to token interactions
- Transfer Learning: Adapt to small domain-specific datasets
- Real-world deployment: Modular LLM assistants, math tutoring bots
- Error analysis & new metrics: Refine contextual token evaluation
- **Modularity:** This model can be called multiple times for different task without disturbing the GPT Model, Hence They can act as the assistant doing different tasks.





Thank you

