InfeLens

Inference + Lens

Dokeun Oh

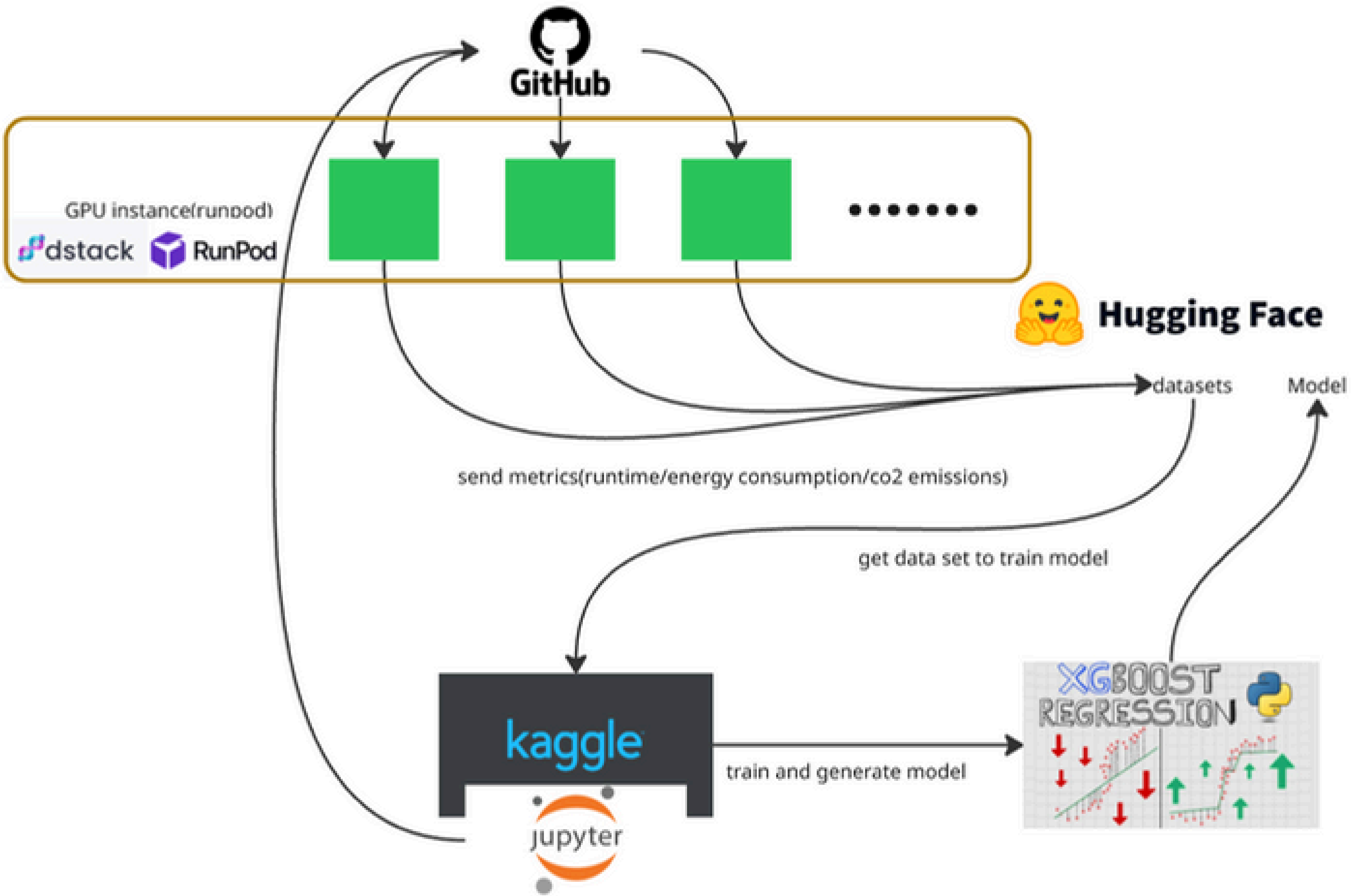
InfeLens: AI Inference Power & Runtime Estimation

Challenge 1: AI Inference Runtime & Power Estimation

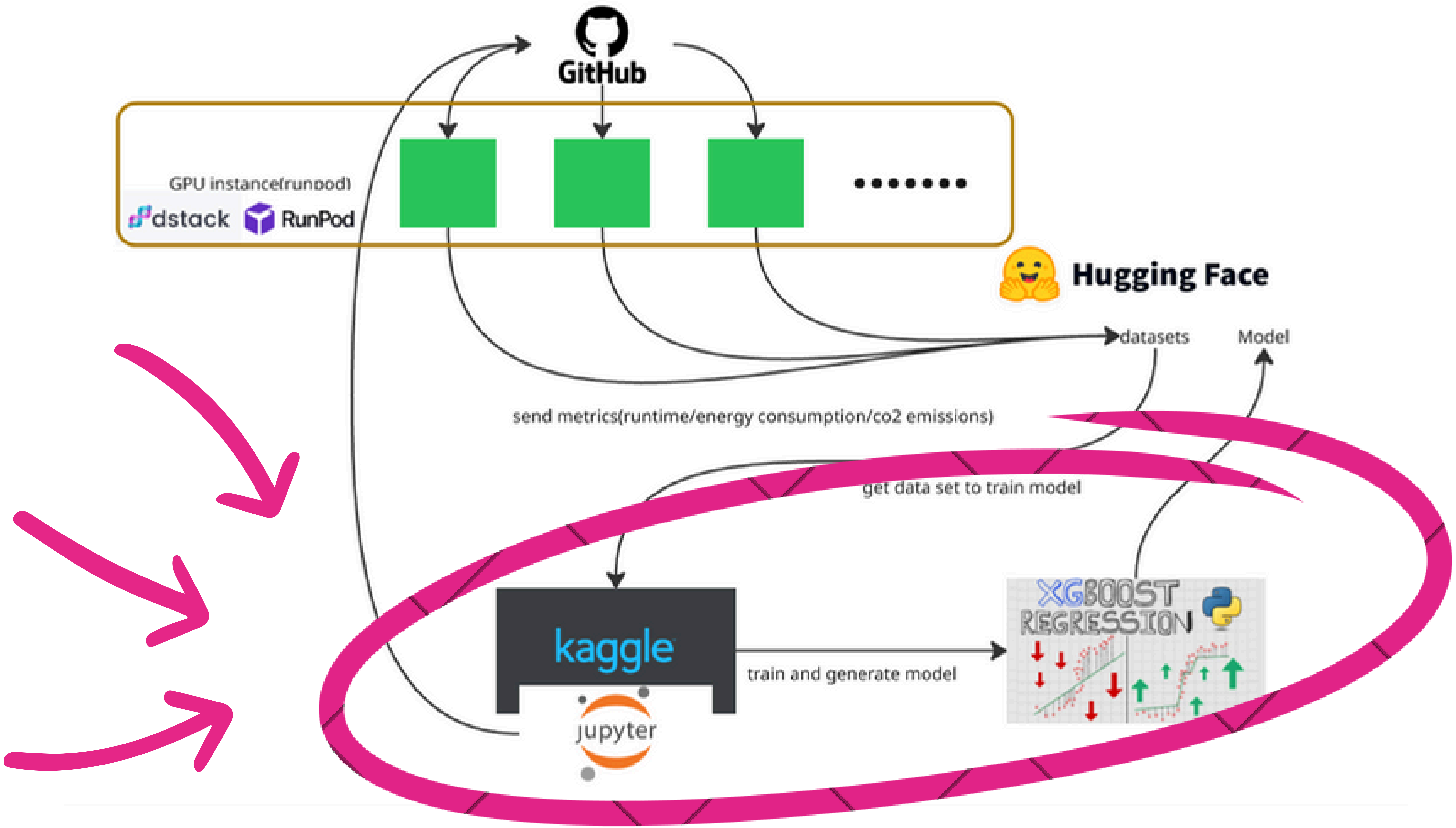
Goal: Predict inference time and power consumption of LLMs

Target Hardware: NVIDIA GPU, Opensources LLM Models

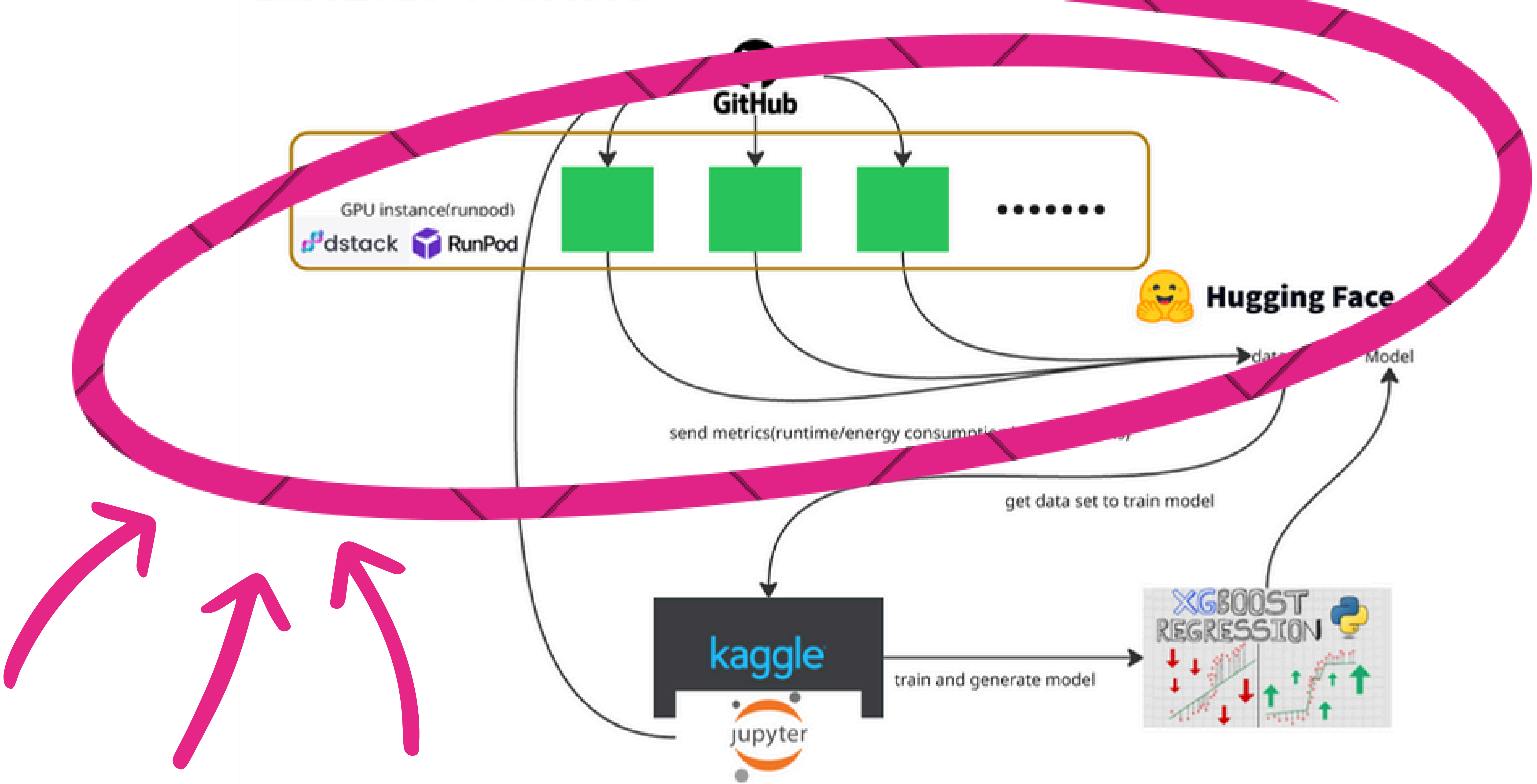
InfeLens Overview



InfeLens Overview



InfeLens Overview



Prepare Datasets with script

250 prompt
11 opensource llm models
10 available GPU in RunPod and dstack

with script

Average Runtime
Average Energy
Average Co2
Prompt Runtime
Prompt Energy
Prompt Co2

https://github.com/ohdoking/infelens

https://huggingface.co/datasets/ohdoking/energy_consumption_by_model_and_gpu

https://huggingface.co/datasets/ohdoking/gpu_spec

https://huggingface.co/datasets/ohdoking/llm_model_specs

Training Model

Dataset

number of data : 137,500

(250*5*11*10)

80 % training

20% validating

Hugging Face

jupyter

Xgboost regression

50 Optuna hyperparameter tuning

100 n_estimators(like epoch)

OPTUNA

Why Regression

Problem type Numerical prediction → Regression fits best

Why XGBoost?

Fast Training Speed

Efficient Resource Usage

Overfitting Prevention

High Performance on Small Datasets

Supports Multi-Output Regression(targets : runtime, energy, CO_2)

Model Input & Output Overview

Input Features

LLM Model Name
LLM Parameter Size
GPU Name

→

Output Features

Inference Runtime(seconds)
Energy Consumption(Joules)

UNDER CONSTRUCTION CO_2 Emission(kg CO_2)

DON'T FORGET! Predicts reliably—even on inputs it has never seen before

Model Input & Output Overview

Model Input Parameters

LLM Model Characteristics

model_name: Name of the LLM (e.g., LLaMA-7B)
huggingface_model: Hugging Face model reference
hidden_size: Size of hidden layers
num_layers: Number of layers
vocab_size: Vocabulary size
seq_length: Input sequence length
model_type: Type of model architecture
num_params_B: Total parameters (in billions)

Hardware Specifications

hardware_gpu: GPU model used
Manufacturer: GPU manufacturer (e.g., NVIDIA)
Memory (GB): GPU memory size
TDP (W): Thermal Design Power
CUDA Cores: Number of CUDA cores
FP32 TFLOPS: Floating-point performance
Architecture: GPU architecture
hardware_ram_GB: Host machine RAM size

Prompt Information

total_prompts: Number of prompts used for inference

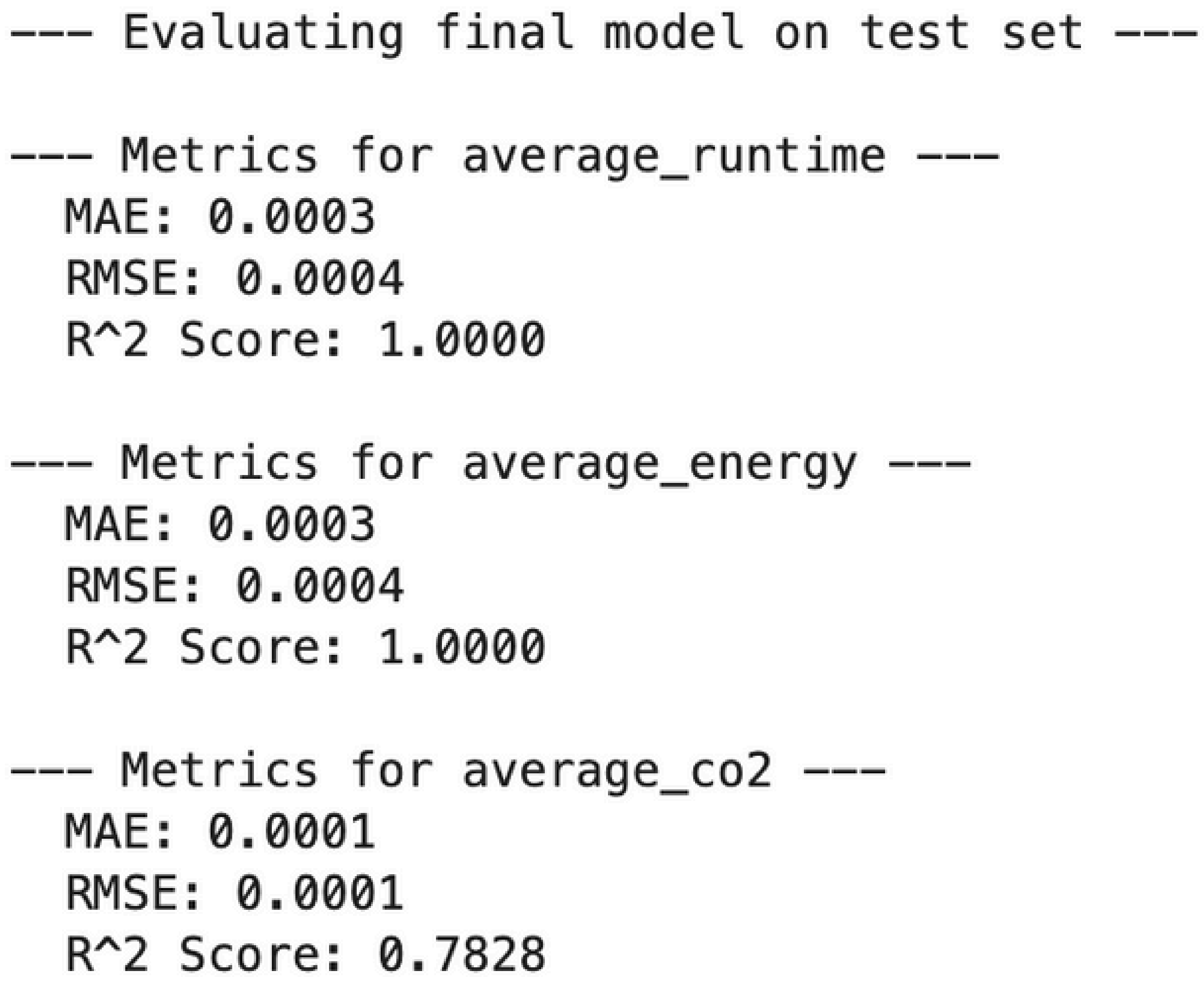
Output Features

Inference Runtime(seconds)

Energy Consumption(Joules)

UNDER CONSTRUCTION CO_2 Emission(kg CO_2)

Result



```
--- Evaluating final model on test set ---

--- Metrics for average_runtime ---
  MAE: 0.0003
  RMSE: 0.0004
  R^2 Score: 1.0000

--- Metrics for average_energy ---
  MAE: 0.0003
  RMSE: 0.0004
  R^2 Score: 1.0000

--- Metrics for average_co2 ---
  MAE: 0.0001
  RMSE: 0.0001
  R^2 Score: 0.7828
```

UNDER CONSTRUCTION

Demo

Data collecting script

Demo

Unseen Model Scenario (Untrained LLM)

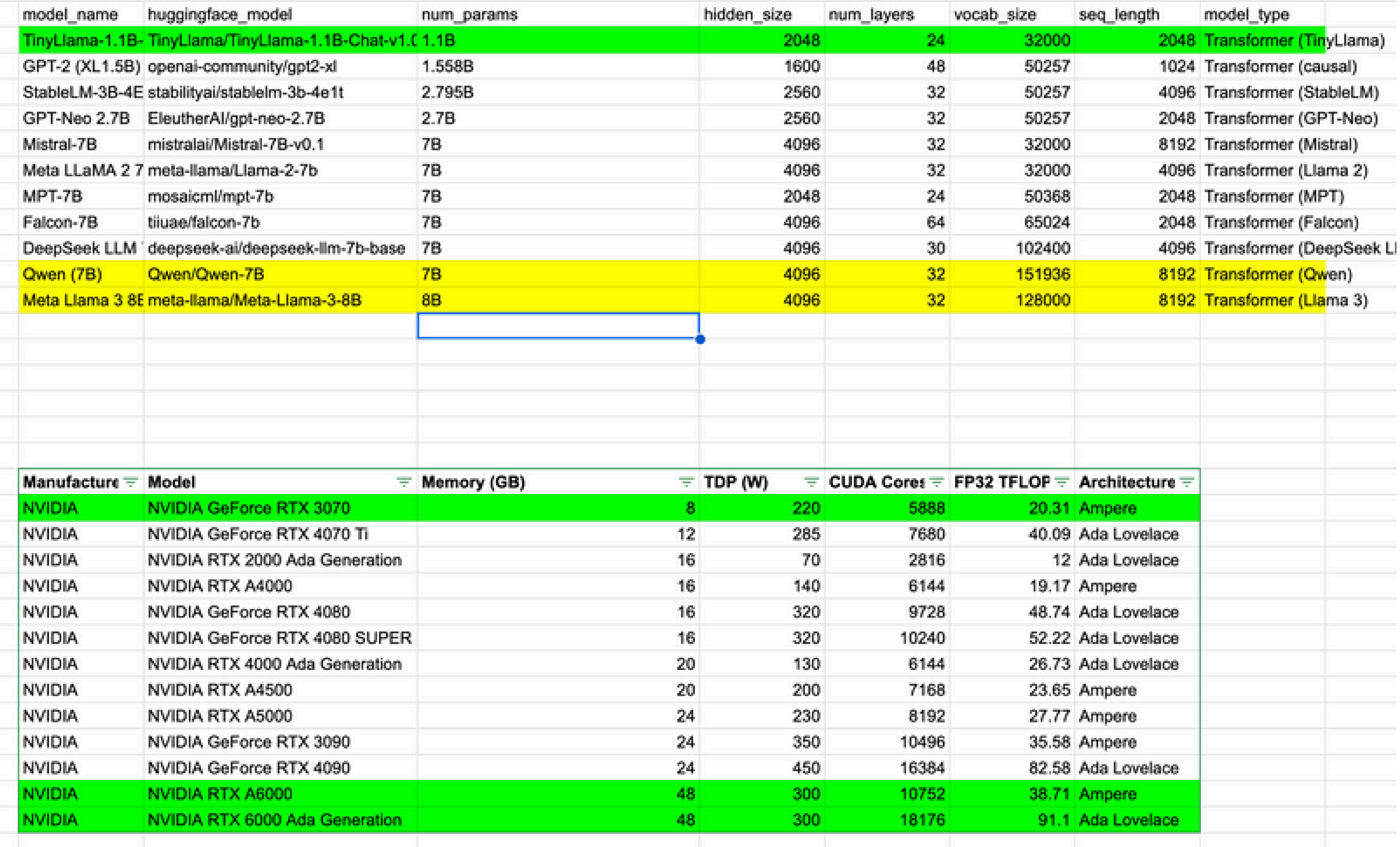
🔍 Unseen model: Qwen 7B
✅ Trained on:

- Meta Llama 3 8B (similar model)
- TinyLlama 1.1B (different model)

Unseen Hardware Scenario (Untrained GPU)

🔍 Unseen GPU: NVIDIA RTX A6000
✅ Trained on:

- NVIDIA RTX 6000 Ada Gen (similar architecture)
- NVIDIA GeForce RTX 3070 (different architecture)



model_name	huggingface_model	num_params	hidden_size	num_layers	vocab_size	seq_length	model_type
TinyLlama-1.1B-	TinyLlama/TinyLlama-1.1B-Chat-v1.(	1.1B	2048	24	32000	2048	Transformer (TinyLlama)
GPT-2 (XL1.5B)	openai-community/gpt2-xl	1.558B	1600	48	50257	1024	Transformer (causal)
StableLM-3B-4E	stabilityai/stablelm-3b-4e1t	2.795B	2560	32	50257	4096	Transformer (StableLM)
GPT-Neo 2.7B	EleutherAI/gpt-neo-2.7B	2.7B	2560	32	50257	2048	Transformer (GPT-Neo)
Mistral-7B	mistralai/Mistral-7B-v0.1	7B	4096	32	32000	8192	Transformer (Mistral)
Meta LLaMA 2 7	meta-llama/Llama-2-7b	7B	4096	32	32000	4096	Transformer (Llama 2)
MPT-7B	mosaicml/mpt-7b	7B	2048	24	50368	2048	Transformer (MPT)
Falcon-7B	tiiuae/falcon-7b	7B	4096	64	65024	2048	Transformer (Falcon)
DeepSeek LLM '	deepseek-ai/deepseek-llm-7b-base	7B	4096	30	102400	4096	Transformer (DeepSeek LI
Qwen (7B)	Qwen/Qwen-7B	7B	4096	32	151936	8192	Transformer (Qwen)
Meta Llama 3 8E	meta-llama/Meta-Llama-3-8B	8B	4096	32	128000	8192	Transformer (Llama 3)

Manufacture	Model	Memory (GB)	TDP (W)	CUDA Core:	FP32 TFLOF	Architecture
NVIDIA	NVIDIA GeForce RTX 3070	8	220	5888	20.31	Ampere
NVIDIA	NVIDIA GeForce RTX 4070 Ti	12	285	7680	40.09	Ada Lovelace
NVIDIA	NVIDIA RTX 2000 Ada Generation	16	70	2816	12	Ada Lovelace
NVIDIA	NVIDIA RTX A4000	16	140	6144	19.17	Ampere
NVIDIA	NVIDIA GeForce RTX 4080	16	320	9728	48.74	Ada Lovelace
NVIDIA	NVIDIA GeForce RTX 4080 SUPER	16	320	10240	52.22	Ada Lovelace
NVIDIA	NVIDIA RTX 4000 Ada Generation	20	130	6144	26.73	Ada Lovelace
NVIDIA	NVIDIA RTX A4500	20	200	7168	23.65	Ampere
NVIDIA	NVIDIA RTX A5000	24	230	8192	27.77	Ampere
NVIDIA	NVIDIA GeForce RTX 3090	24	350	10496	35.58	Ampere
NVIDIA	NVIDIA GeForce RTX 4090	24	450	16384	82.58	Ada Lovelace
NVIDIA	NVIDIA RTX A6000	48	300	10752	38.71	Ampere
NVIDIA	NVIDIA RTX 6000 Ada Generation	48	300	18176	91.1	Ada Lovelace

Thank you