DevFest

Cloud Ho Chi Minh 2024

Deploy Gemma2 with multiple LoRA adapters with TGI DLC on GKE

Hieu Ngo, Senior AI Engineer @ MoMo

Google Developer Groups

Gemma 2

Gemma 2 is a **open-source LLM** offers best-in-class performance, runs at incredible speed across different hardware and easily integrates with other AI tools.

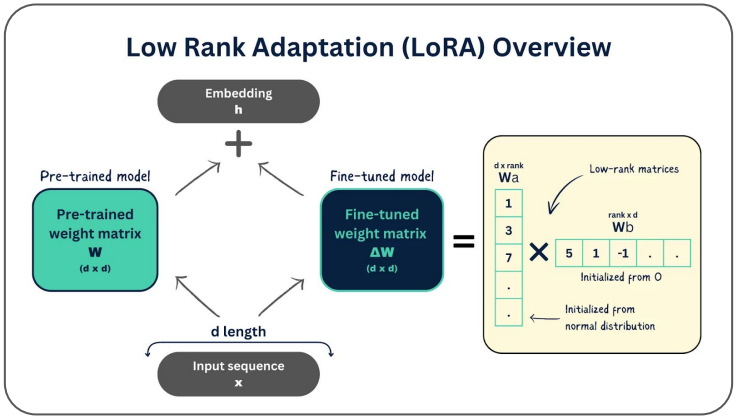
			Gemma 2		Llama 3		Grok-1
	BENCHMARK	METRIC	9B	27B	8B	70B	314B
General	MMLU	5-shot, top-1	71.3	75.2	66.6	79.5	73.0
Reasoning	BBH	3-shot, CoT	68.2	74.9	61.1	81.3	–
	HellaSwag	10-shot	81.9	86.4	82	–	–
Math	GSM8K	5-shot, maj@1	68.6	74.0	45.7	–	62.9 (8-shot)
	MATH	4-shot	36.6	42.3	–	–	23.9
Code	HumanEval	pass@1	40.2	51.8	–	–	63.2 (0-shot)

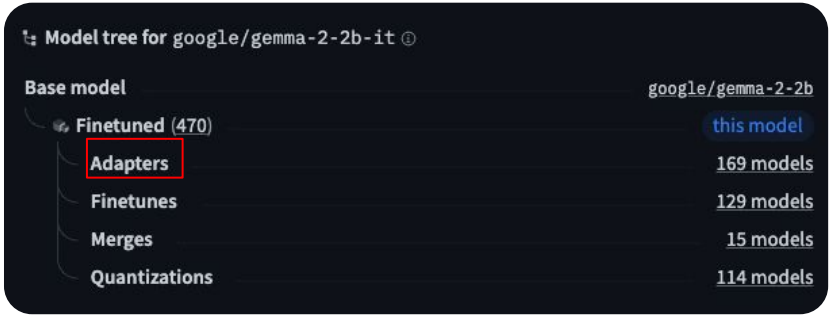
DevFest
Cloud Ho Chi Minh 2024

GDG Cloud HCM

LoRA (Low-Rank Adaptation)

- **Technique** to fine-tune models.
- The core idea is to train to specific tasks without needing to retrain the **entire model**, but only a **small set of parameters** called **adapters**.

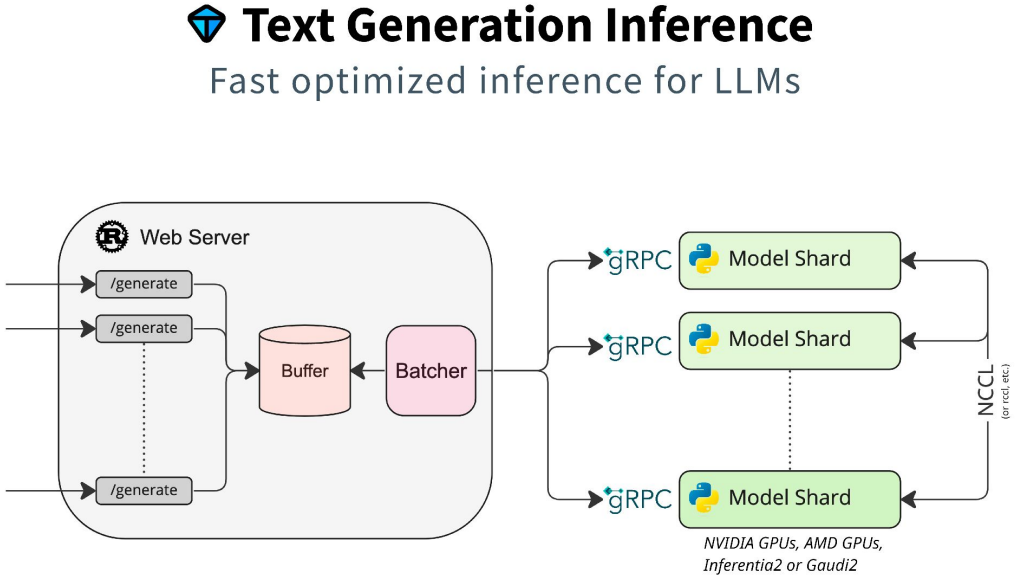




DevFest Cloud Ho Chi Minh 2024

GDG Cloud HCM

Text Generation Inference (TGI)



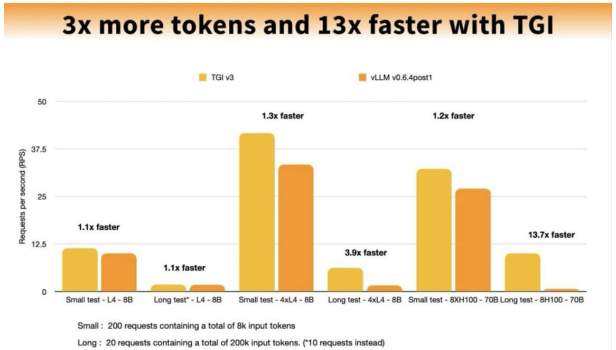
TGI

Philipp Schmid · Following
Technical Lead & LLMs at Hugging Face...
24m ·

3x more tokens and 13x faster generations than vLLM? 👀 Hugging Face TGI 3.0 released! 🎉 TGI 3.0 dramatically improves LLM inference processing by 3x more input tokens, running 13x faster than vLLM on long prompts while requiring zero configuration!

TL;DR:
🚀 Processes 3x more tokens than vLLM (30k vs 10k tokens on L4 GPU for llama 3.1-8B)
⚡ Achieves 13x faster processing on long prompts (200k+ tokens) through conversation caching
🔧 Significantly reduced memory & Zero configuration needed for models
🔬 New kernels (flash-infer, flash-decoding), optimized prefix caching, and improved VRAM efficiency
🤝 Soon available on AWS, Google Cloud, and Dell Enterprise Hub
➡️ SOON Future: special models, KV-cache retention, and multimodal models

Learn more: https://lnkd.in/e8eNnJ7E



84 2 comments • 6 reposts

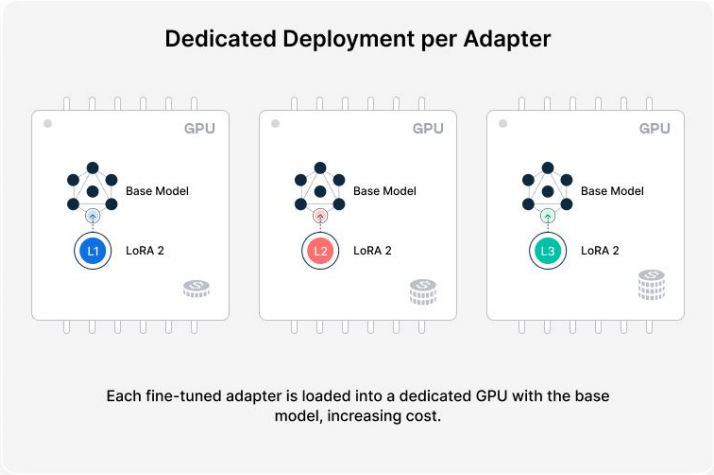
DevFest
Cloud Ho Chi Minh 2024

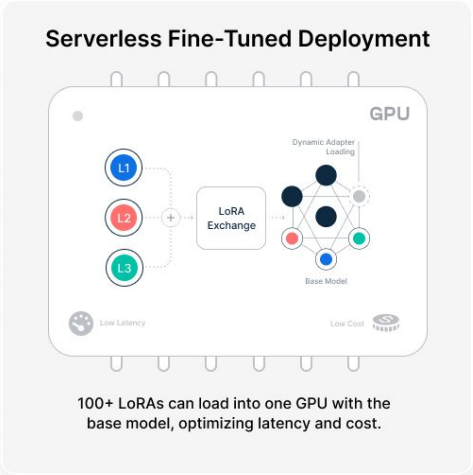
GDG Cloud HCM

TGI Multi-LoRA: Deploy Once, Serve N models

Motivation

Independence - For each task that your organization cares about, different teams can work on different fine tunes



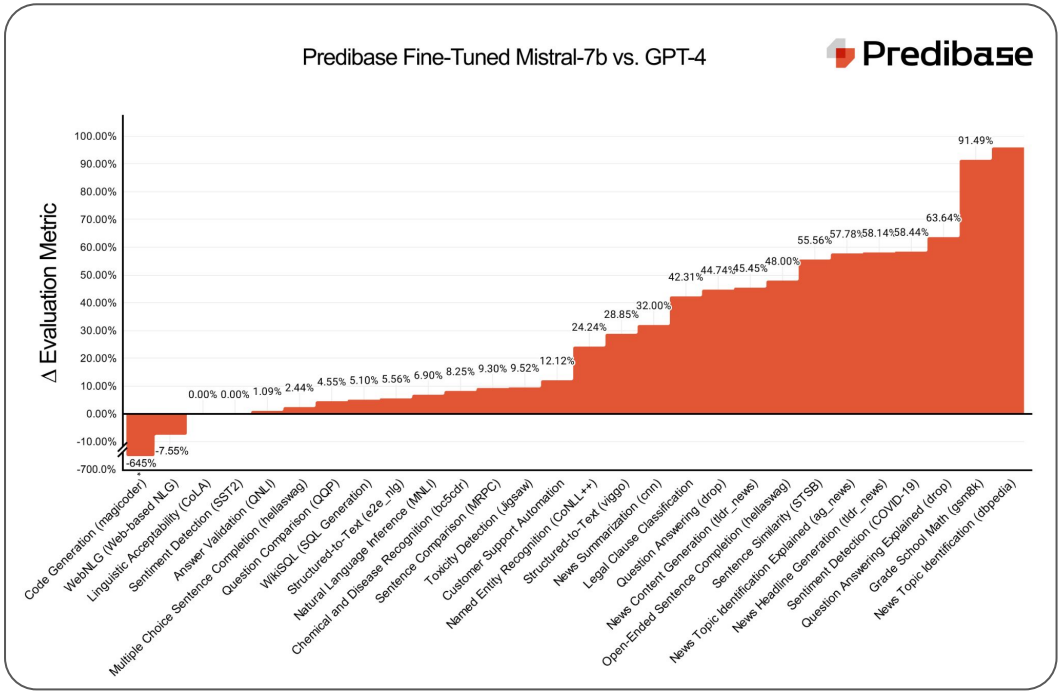


TGI Multi-LoRA: Deploy Once, Serve N models

Performance -

evidence that smaller, specialized models outperform their larger, general-purpose model.

Predibase showed better performance than GPT-4 using task-specific LoRAs with a base like mistralai/Mistral-7B-v0.1.



DevFest
Cloud Ho Chi Minh 2024

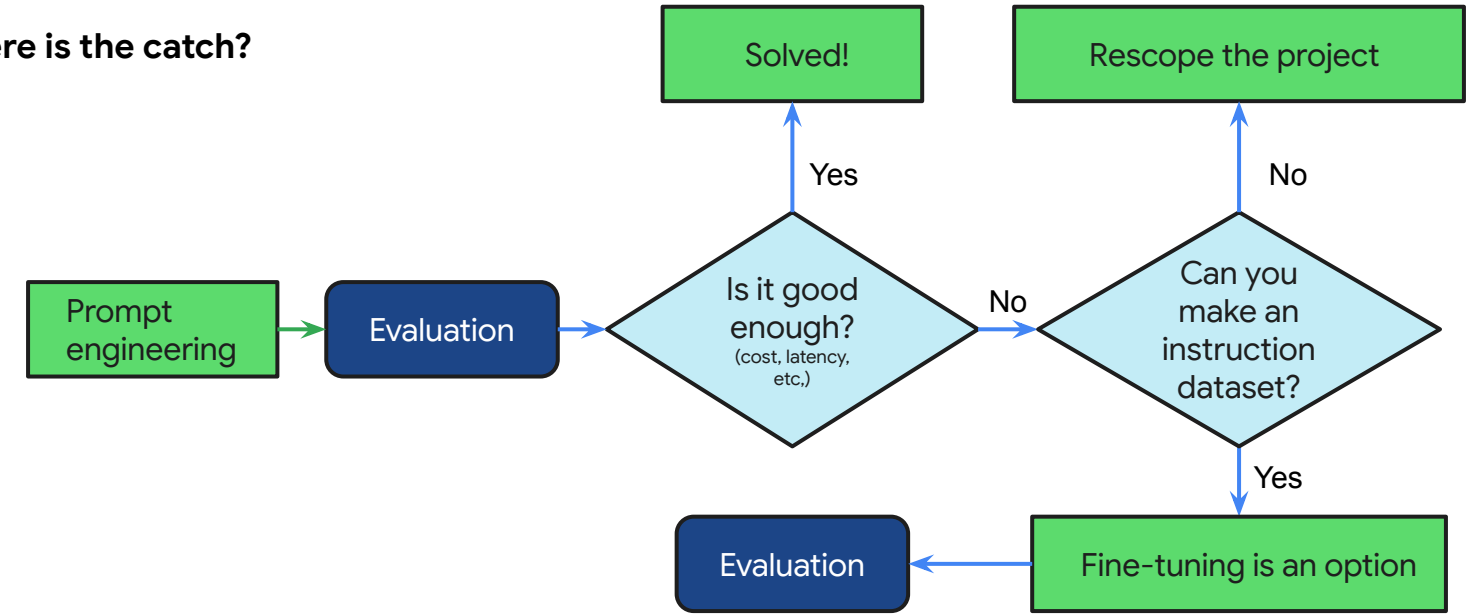
GDG Cloud HCM

TGI Multi-LoRA: Deploy Once, Serve N models

Privacy - Specialized models offer flexibility with training data segregation and access restrictions to different users based on data privacy requirements. Additionally, in cases where running models locally is important, a small model can be made highly capable for a specific task while keeping its size small enough to run on device.

TGI Multi-LoRA: Deploy Once, Serve N models

So, where is the catch?



TGI Multi-LoRA: Deploy Once, Serve N models

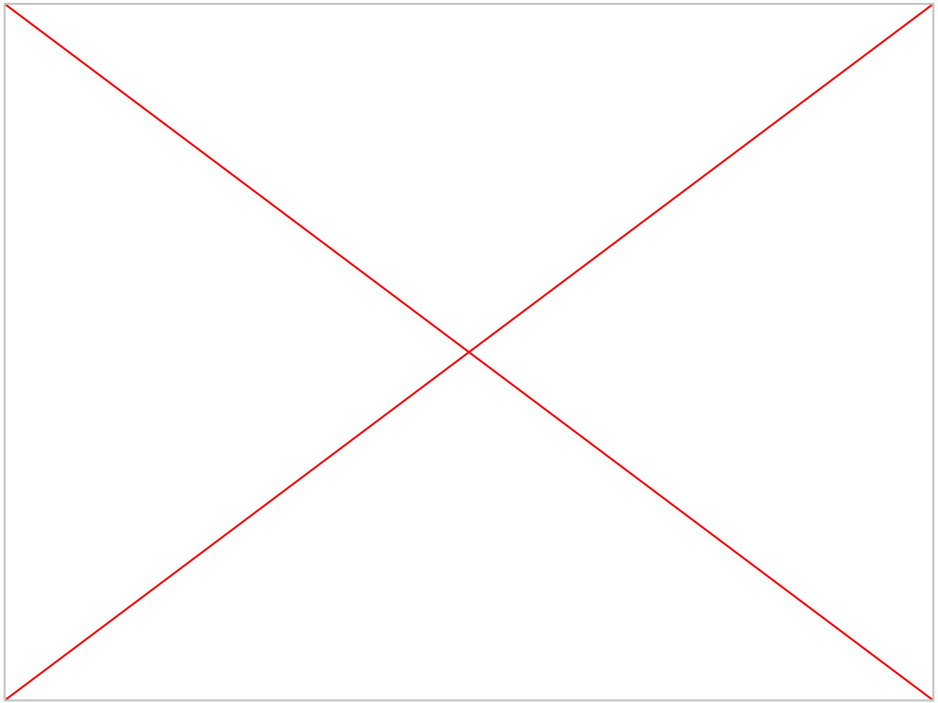
So, where is the catch?

Deploying and serving Large Language Models (LLMs) **is challenging** in many ways. **Cost and operational** complexity are key considerations when deploying a single model, let alone n models. This means that, for all its glory, fine-tuning complicates LLM deployment and serving even further.

That is why today I am super excited to introduce **TGI's feature - Multi-LoRA serving**.

Inference LORA

Multi-LoRA Serving



Multi-LoRA Serving

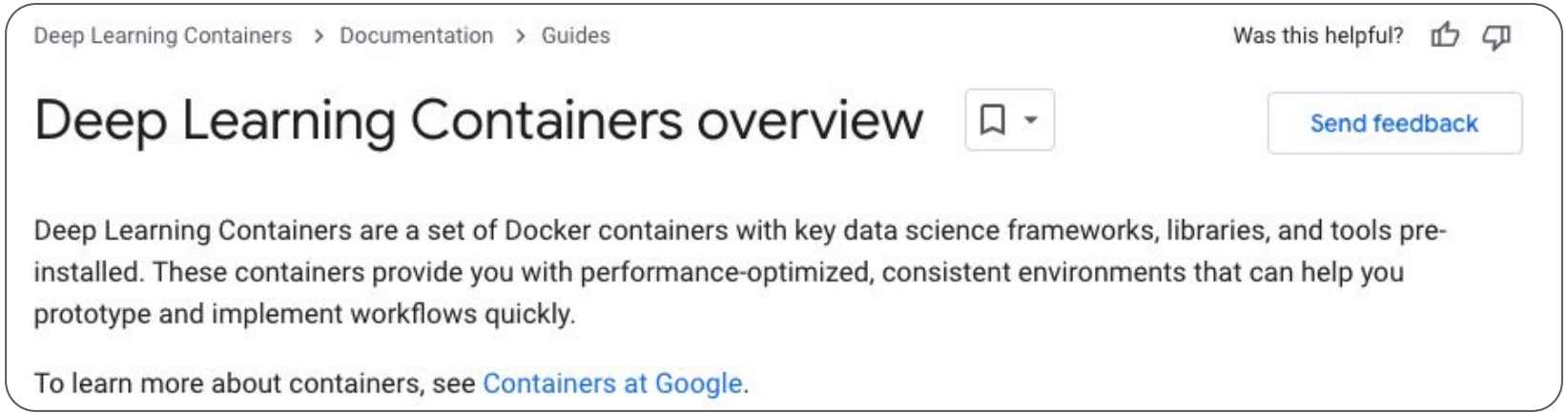
LoRAs (the adapter weights) can vary based on rank and quantization, but they are generally **quite tiny**.

Example:

predibase/magicoder is 13.6MB, which is less than 1/1000th the size of **mistralai/Mistral-7B-v0.1, which is 14.48GB.**

In relative terms, loading **30 adapters into RAM** results in only a **3% increase** in **VRAM**. Ultimately, this is not an issue for most deployments. Hence, we can have one deployment for many models.

Deep Learning Container



TGI DLC

TensorFlow

PyTorch

Hugging Face

NVIDIA
CUDA

DevFest
Cloud Ho Chi Minh 2024

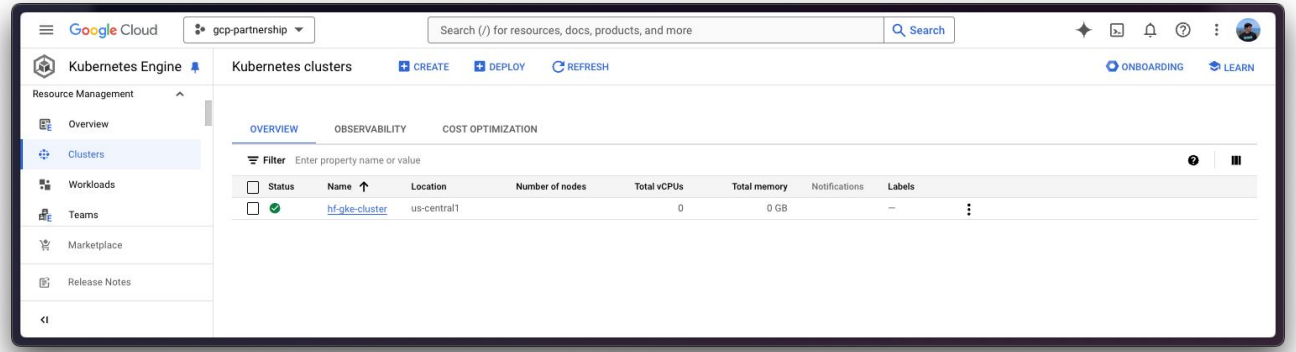
GDG Cloud HCM

Google Kubernetes Engine (GKE)

GKE, short for Google Kubernetes Engine, is a **managed Kubernetes** service provided by Google Cloud. It allows you to deploy, manage, and scale containerized applications using Kubernetes, an open-source container orchestration platform. With GKE, **Google handles much of the underlying infrastructure,** such as **provisioning, maintaining, and upgrading Kubernetes clusters,** so you can **focus on developing and running your applications.**

Create GKE Cluster

```
gcloud container clusters create-auto $CLUSTER_NAME \
    --project=$PROJECT_ID \
    --location=$LOCATION \
    --release-channel=stable \
    --cluster-version=1.29 \
    --no-autoprovisioning-enable-insecure-kubelet-readonly-port
```



DevFest
Cloud Ho Chi Minh 2024

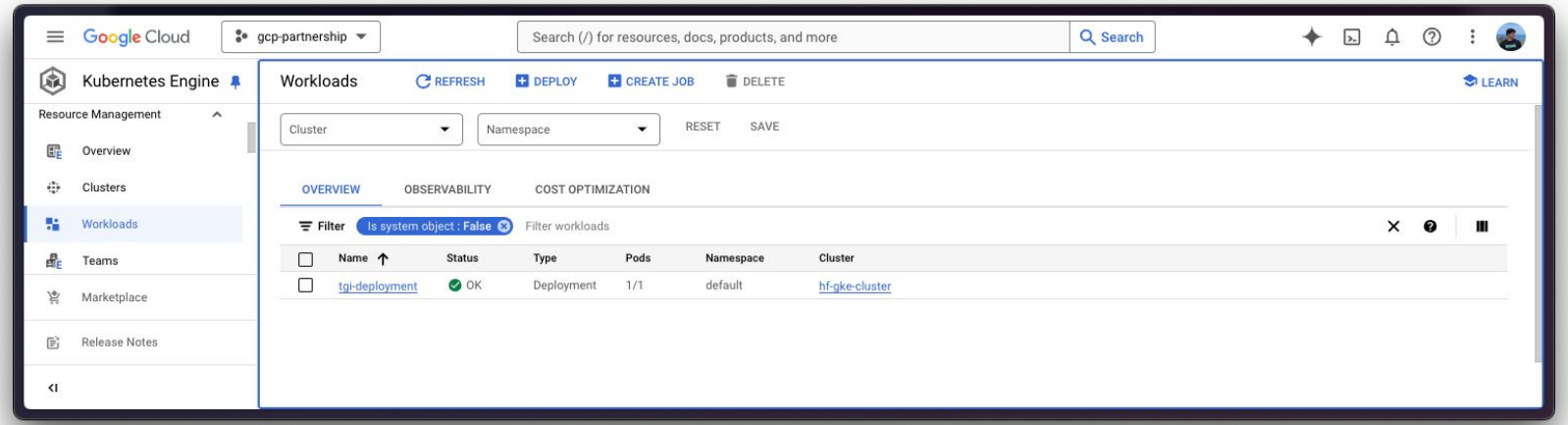
GDG Cloud HCM

Deploy TGI

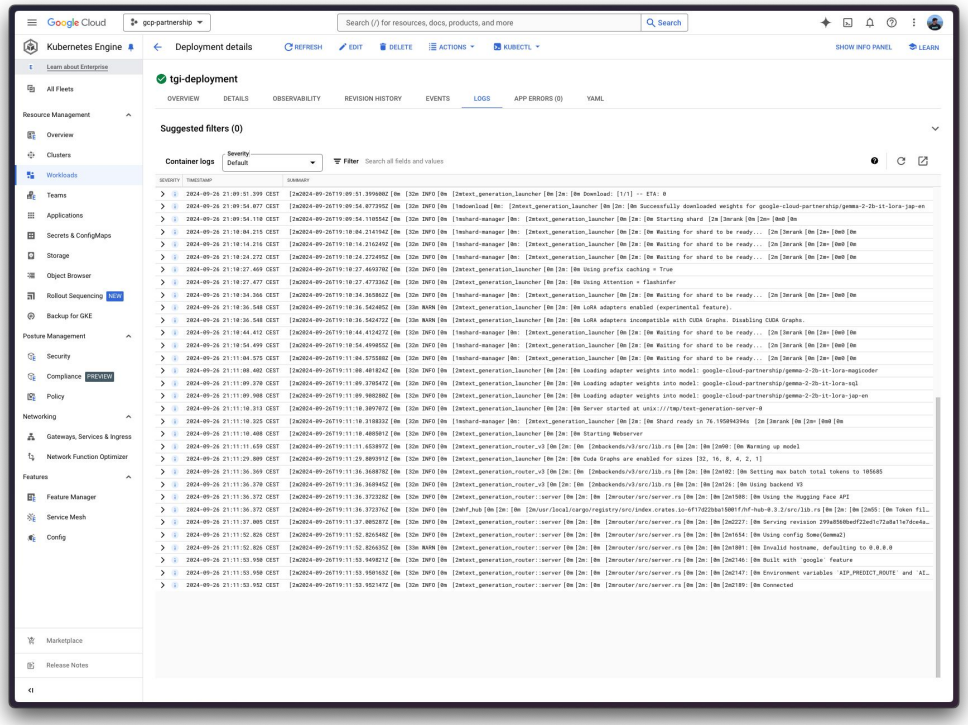
```
kubectl apply -f Google-Cloud-Containers/examples/gke/tgi-multi-lora-deployment/config
```

```
containers:
  - name: tgi-container
    image: us-docker.pkg.dev/deeplearning-platform-release/gcr.io/huggingface-text-generation-inference-cu124.2-3.ubuntu2204.py311
    resources:
      requests:
        nvidia.com/gpu: 1
    env:
      - name: MODEL_ID
        value: google/gemma-2-2b-it
      - name: LORA_ADAPTERS
        value: google-cloud-partnership/gemma-2-2b-it-lora-magicoder,google-cloud-partnership/gemma-2-2b-it-lora-sql,google-cloud-partnership/gemma-2
      - name: NUM_SHARD
        value: "1"
      - name: PORT
        value: "8080"
      - name: HUGGING_FACE_HUB_TOKEN
```

Deploy TGI


Google Cloud
gcp-partnership
Search (/) for resources, docs, products, and more
Search
Kubernetes Engine
Resource Management
Overview
Clusters
Workloads
Teams
Marketplace
Release Notes
Workloads
REFRESH
DEPLOY
CREATE JOB
DELETE
LEARN
Cluster
Namespace
RESET
SAVE
OVERVIEW
OBSERVABILITY
COST OPTIMIZATION
Filter
Is system object : False
Filter workloads
Name
Status
Type
Pods
Namespace
Cluster
tgi-deployment
OK
Deployment
1/1
default
hf-gke-cluster

Deploy TGI



DevFest
Cloud Ho Chi Minh 2024

GDG Cloud HCM

Demo and Q&A

References

- Deploy Gemma2 with multiple LoRA adapters with TGI DLC on GKE
- TGI Multi-LoRA: Deploy Once, Serve 30 models
- LoRA Land: Fine-Tuned Open-Source LLMs that Outperform GPT-4

Thank you!