Easy Model Deployer

None

None

None

Table of contents

1. About Easy Model Deployer

Easy Model Deployer is a lightweight tool designed to simplify the machine learning model deployment process.

1.1 Features

- Simple deployment workflow
- Support for multiple ML frameworks
- Easy configuration
- Minimal dependencies

1.2 Why Use Easy Model Deployer?

Perfect for developers who want to quickly deploy ML models without dealing with complex infrastructure setup.

1.3 Getting Started

Check our Usage Guide to start deploying your models in minutes.

2. Supported Model

ModelId	ModelSeries	ModelType	Supported Engines	Supported Instances
glm-4-9b-chat	glm4	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
internlm2_5-20b-chat-4bit-awq	internlm2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.12xlarge,g
internlm2_5-20b-chat	internlm2.5	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
internlm2_5-7b-chat	internlm2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.12xlarge,g
internlm2_5-7b-chat-4bit	internlm2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.12xlarge,g
internlm2_5-1_8b-chat	internlm2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.12xlarge,g
Qwen2.5-7B-Instruct	qwen2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.12xlarge,g
Qwen2.5-72B-Instruct-AWQ	qwen2.5	llm	vllm,tgi	g5.12xlarge,g5.24xlarge,g5.48xlarge
Qwen2.5-72B-Instruct	qwen2.5	llm	vllm	g5.48xlarge
Qwen2.5-72B-Instruct-AWQ-128k	qwen2.5	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
Qwen2.5-32B-Instruct	qwen2.5	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
Qwen2.5-0.5B-Instruct	qwen2.5	llm	vllm,tgi	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge,i
Qwen2.5-1.5B-Instruct	qwen2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge
Qwen2.5-3B-Instruct	qwen2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge
Qwen2.5-14B-Instruct-AWQ	qwen2.5	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge
Qwen2.5-14B-Instruct	qwen2.5	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
QwQ-32B-Preview	qwen reasoning model	llm	huggingface,vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
llama-3.3-70b-instruct-awq	llama	llm	tgi	g5.12xlarge,g5.24xlarge,g5.48xlarge
DeepSeek-R1-Distill-Qwen-32B	deepseek reasoning model	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
DeepSeek-R1-Distill-Qwen-14B	deepseek reasoning model	llm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge

ModelId	ModelSeries	ModelType	Supported Engines	Supported Instances
DeepSeek-R1-Distill-Qwen-7B	deepseek reasoning model	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge
DeepSeek-R1-Distill-Qwen-1.5B	deepseek reasoning model	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge
DeepSeek-R1-Distill-Llama-8B	deepseek reasoning model	llm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.16xlarge
deepseek-r1-distill-llama-70b-awq	deepseek reasoning model	llm	tgi,vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
Baichuan-M1-14B-Instruct	baichuan	llm	huggingface	g5.12xlarge,g5.24xlarge,g5.48xlarge
Qwen2-VL-72B-Instruct-AWQ	qwen2vl	vlm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
QVQ-72B-Preview-AWQ	qwen reasoning model	vlm	vllm	g5.12xlarge,g5.24xlarge,g5.48xlarge
Qwen2-VL-7B-Instruct	qwen2vl	vlm	vllm	g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.12xlarge,g
InternVL2_5-78B-AWQ	internvl2.5	vlm	lmdeploy	g5.12xlarge,g5.24xlarge,g5.48xlarge
txt2video-LTX	comfyui	video	comfyui	g5.4xlarge,g5.8xlarge,g6e.2xlarge
whisper	whisper	whisper	huggingface	g5.xlarge,g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.1
bge-base-en-v1.5	bge	embedding	vllm	g5.xlarge,g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.1
bge-m3	bge	embedding	vllm	g5.xlarge,g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.1
bge-reranker-v2-m3	bge	rerank	vllm	g5.xlarge,g5.2xlarge,g5.4xlarge,g5.8xlarge,g5.1

3. Installation

4. Usage