Workload	Container Orchestrator	Distributed compute engine	Training & inference framework	Outcomes
Batch / offline inference Personalization, multimodal search, CLIP embeddings, LLM batch inference	**Kubernetes**-based Michelangelo Job Controller across multiple AZs (for GPU availability and scale)	**Ray** for coordinating and scaling	**PyTorch DDP**, **DeepSpeed**, Hugging Face Transformers, **vLLM** for offline scoring	GPU memory reduction enabling 2-7x larger batches leading to 2-3x throughput increase on Llama-2 70B. ✧ (blog, blog)
Online LLM inference Assistant, chat translation, voice safety	**Kubernetes** using hybrid cloud & on-prem data centers	kServe	**vLLM** is the primary LLM inference engine, Triton	Switch to vLLM delivered 2x lower latency and higher throughput. ✧ (blog)
Training pipelines Daily retraining, distributed training	**Kubernetes** using hybrid cloud & on-prem data centers with Yunikorn for queueing	Kubeflow Pipelines, **Ray** for distributed training	**PyTorch**	Platform expanded from 50 → 250 inference pipelines. ✧ (blog)