DropTicks AI Resource

vLLM

Developer Tool

High-throughput inference engine for serving open-weight language models.

Builder notes

Use when a system needs efficient model serving, batching, OpenAI-compatible endpoints, and better throughput for self-hosted inference.

Why use this

â¢ Use this when self-hosted inference needs higher throughput, batching, OpenAI-compatible endpoints, and efficient serving for open-weight models.

Who can use this

â¢ ML infrastructure engineers, AI platform teams, backend teams serving models at scale, and companies controlling their own inference stack.

Prerequisites

â¢ Linux/server operations, GPU deployment knowledge, model selection experience, and understanding of latency, throughput, batching, and memory tradeoffs.

System requirements

â¢ Linux server environment, Python, compatible GPU/CUDA stack for common deployments, sufficient VRAM for the selected model, and production networking/load-balancing when serving users.

Source or reference link

https://docs.vllm.ai/

Web link

https://www.dropticks.com/resources/vllm

Related links

https://github.com/vllm-project/vllm