vLLM x inferact

Inside vLLM

Production Best Practices, Model integration and roadmap

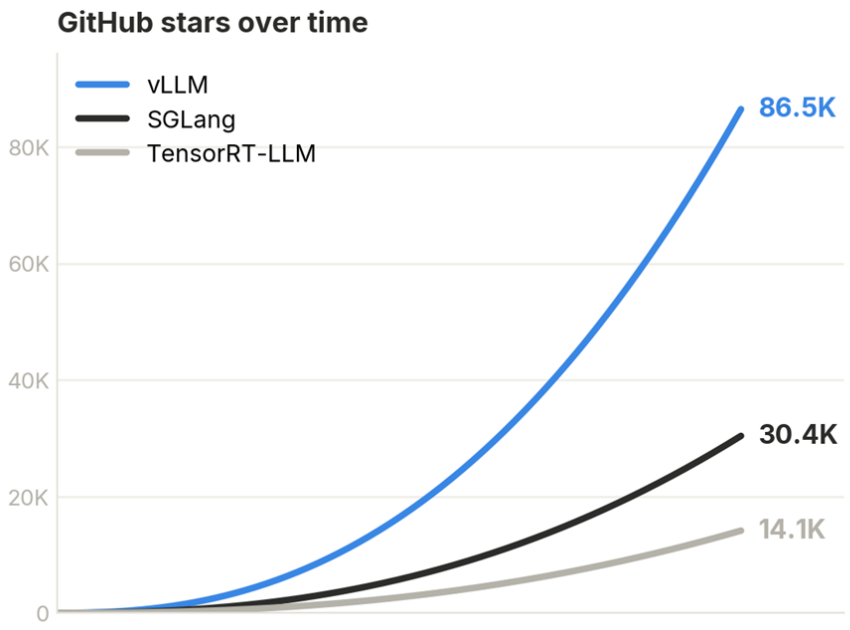
vLLM Bangkok Meetup

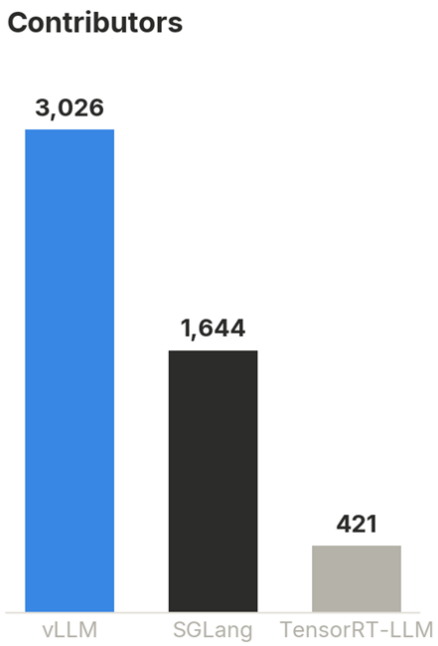
PRESENTED BY Tiezhen WANG, Head of ecosystem @ Inferact

vLLM x inferact

The state of vLLM today

The open standard for LLM inference and fastest-growing project in the ecosystem





91K

GitHub stars — #1 inference engine

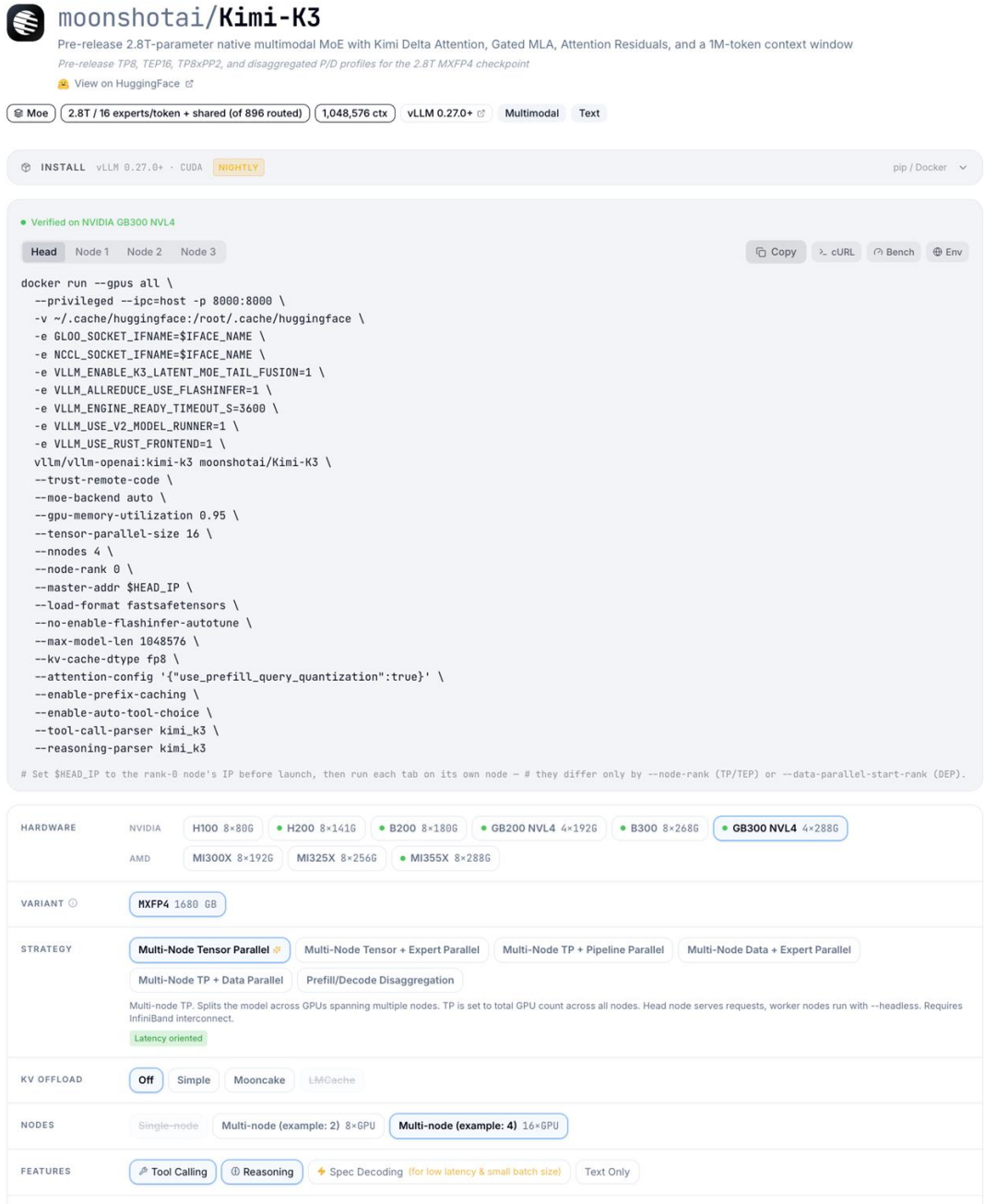
3,336

contributors and counting

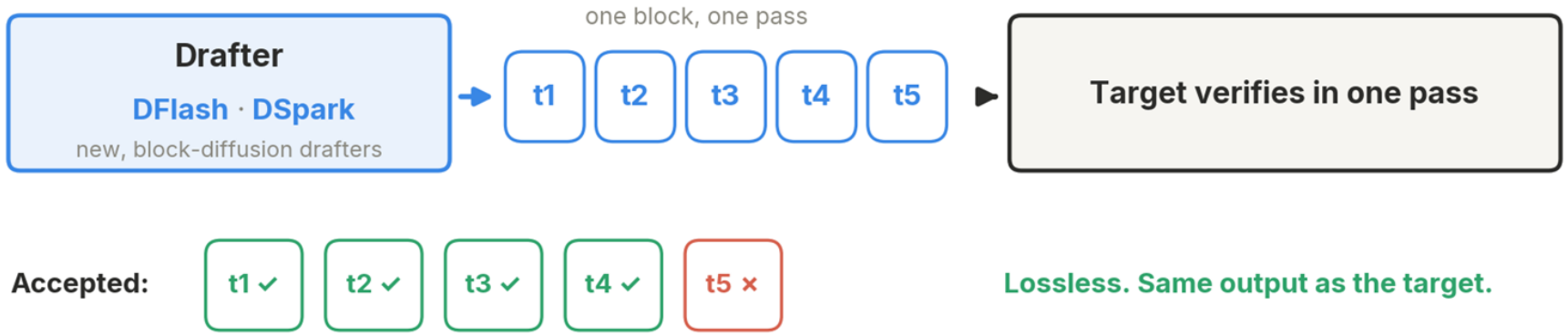
https://recipes.vllm.ai/moonshotai/Kimi-K3

--max-model-len

--max-num-seqs



Speculative decoding



EAGLE-3 has been solid for a while. DFlash and DSpark are the new, fast-moving block-diffusion drafters.

DFlash / DSpark Support

Block-diffusion drafters that draft a whole block in one pass. Example: DSpark speculators offer ~30% higher acceptance length on Qwen3 4/8/14B. Both native in vLLM.

TorchSpec and Speculators

Speculators and TorchSpec train EAGLE-3, DFlash, and DSpark drafters, with unified online and offline training on vLLM hidden states.

Spec Decode as a First-class citizen

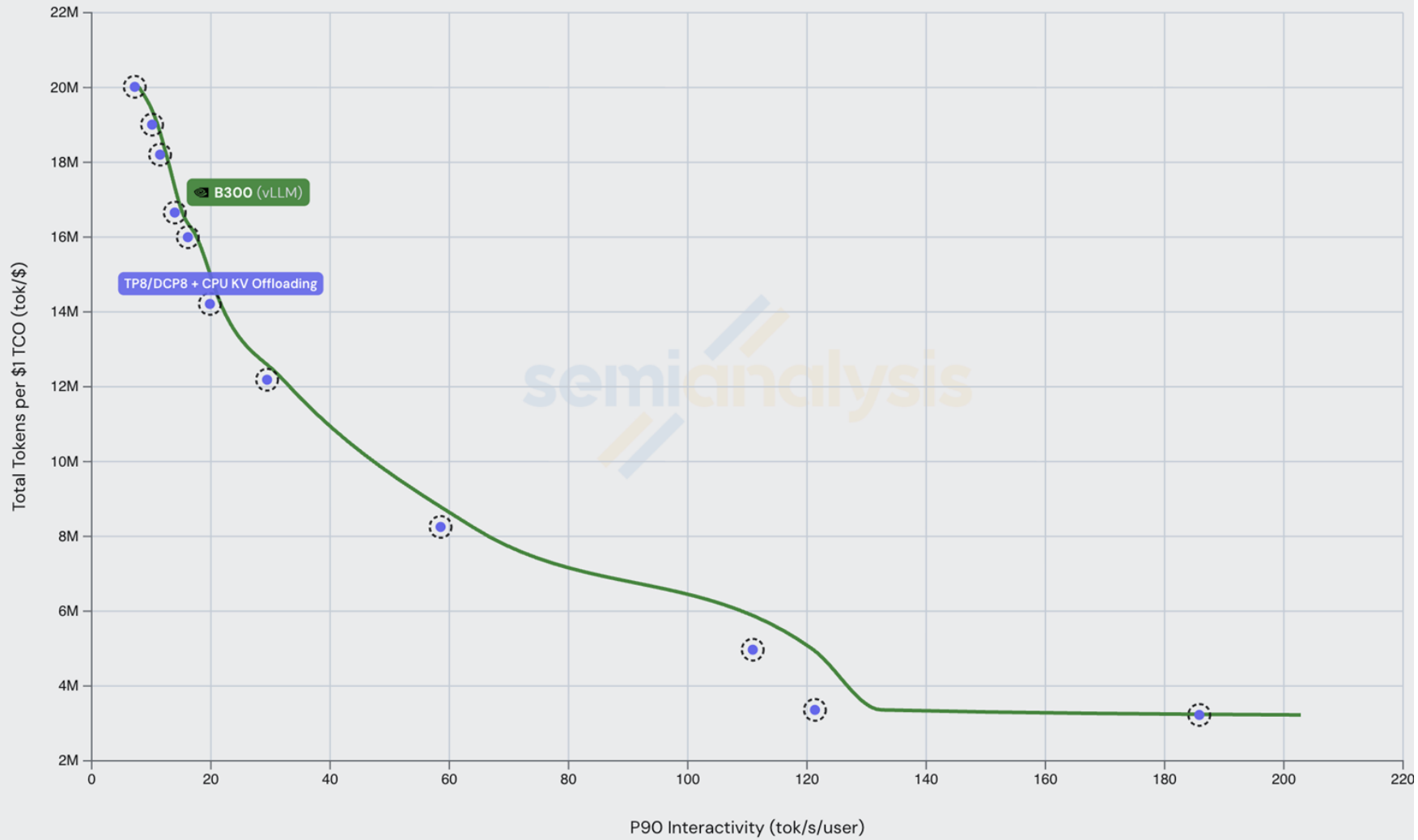
Spec decode works with the rest of vLLM: PD disaggregation, Model Runner V2, prefix caching, tool calling, etc.

Kimi K3 2.8T Agentic Total Tokens per $1 TCO vs. P90 Interactivity

Cost Tier: Owning Hyperscaler **Updated:** 2026-09-04 **Source:** SemiAnalysis InferenceX™

TCO $/chip/hr: B300: 2.26

Source: SemiAnalysis Market July 2026 Pricing Surveys & AI Cloud TCO Model



Shift+Scroll to zoom • Drag to pan • Double-click to reset • Click a point to pin tooltip

vLLM x inferact

HOW TO SUPPORT Kimi K3

The process of adding Day 0 model integration

Four stages, using Kimi K3 as a running case study — from initial contact to the integration.

01

Coordination with model vendors

Understand the basic information of the model & the timeline required for this work.

02

Implement & Verification

Work is done in a private Github repository.

03

Day 0 Docker image

Docker container to avoid the hassle of version / dependency management.

04

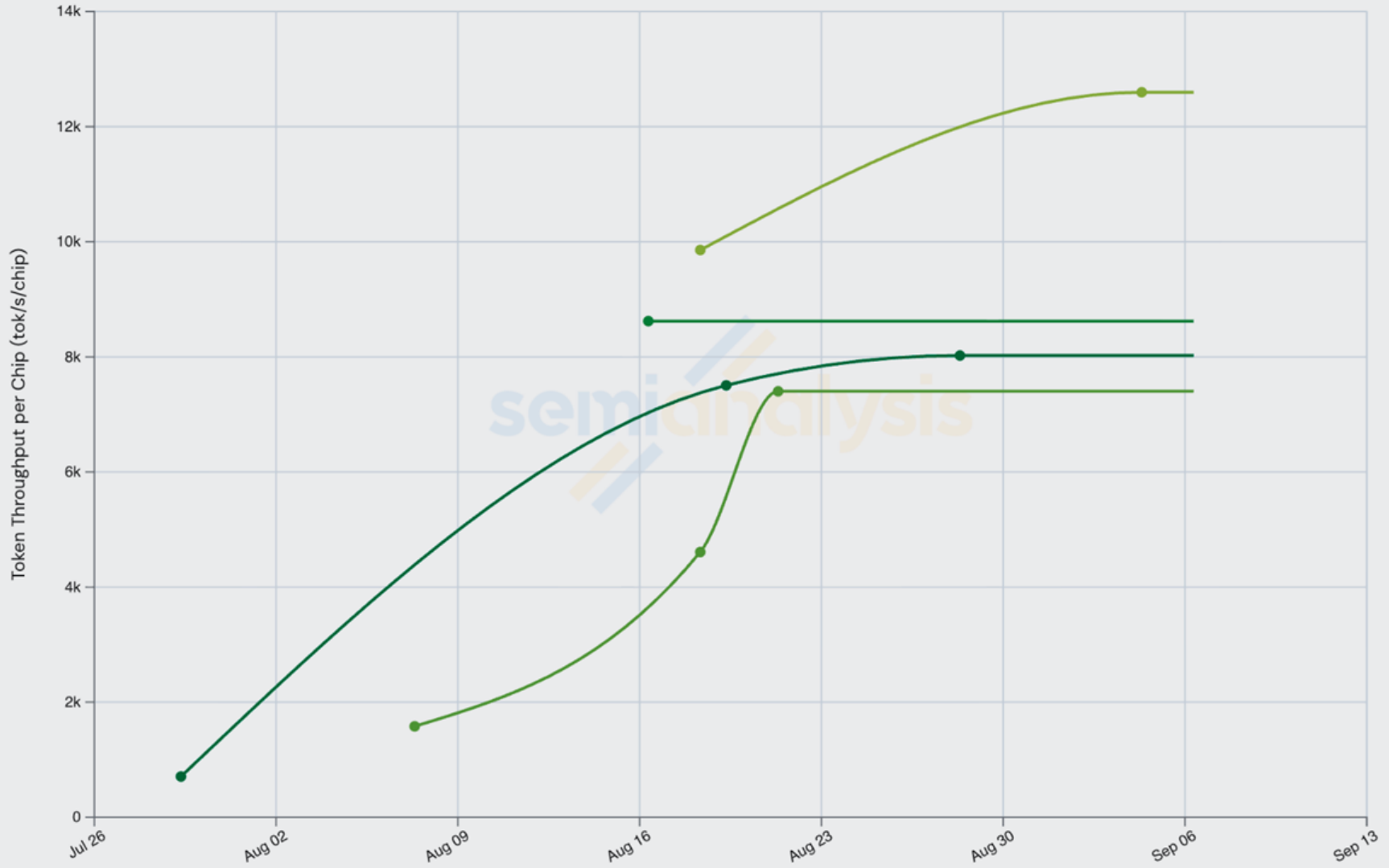
Recipes

A simple reusable recipes with required flags + Env variables.

The brutal truth: The whole process could take weeks or even months. And it doesn't stop there.

Token Throughput per Chip Over Time at 35 tok/s/user Interactivity

Model: Kimi K3 2.8T **Workload:** Agentic **Precision:** FP4 **Metric:** Token Throughput per Chip (tok/s/chip) **Target:** 35 tok/s/user **Updated:** 2026-09-04 **Source:** SemiAnalysis InferenceX™



inferact | vLLM

The Q3 roadmap

Our goal: optimize vLLM as the production agentic serving system.

PRODUCTION SERVING

Agentic workloads

Tool calling, structured output, agent-aware caching, and AgentX.

Large-scale serving

Dynamo support, disagg, and parallelism across multi-node clusters.

PERFORMANCE

Model performance

Day-0 support with leading perf; current sprints are Kimi K3 and GLM 5.2.

Speculative decoding

DSpark and DFlash, plus resilience across other frameworks.

REFINEMENT

KV cache manager refactoring

Cleaner manager and connector interfaces for offload, reuse, and external stores.

CI & release

Cut time-to-signal. Reshape CI to fit how vLLM ships.

ECOSYSTEM

Broader hardware support

Optimizations for AMD GPU and Google TPU.

RL & Omni

Weight sync, train-inference consistency, and vision models in lockstep with main.

vLLM

Building the fastest and easiest-to-use open-source LLM inference & serving engine!

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

https://slack.vllm.ai

https://docs.vllm.ai

https://twitter.com/vllm_project

https://blog.vllm.ai

https://opencollective.com/vllm

https://www.zhihu.com/people/vllm-team

vllm_project

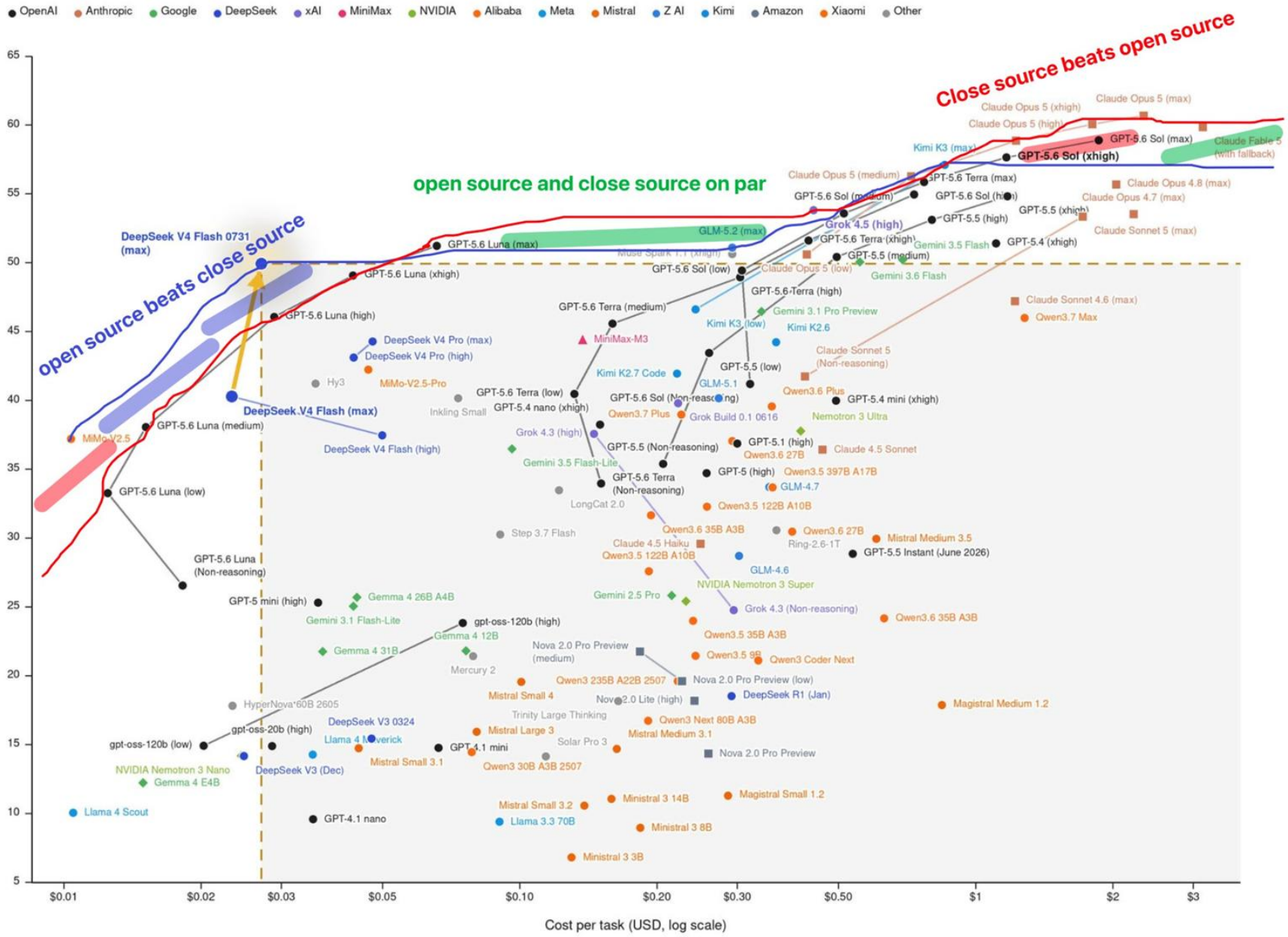
https://www.linkedin.com/company/vllm-project

collaboration@vllm.ai

Artificial Analysis Intelligence Index v4.1

Source: Artificial Analysis · extracted 2026-07-31

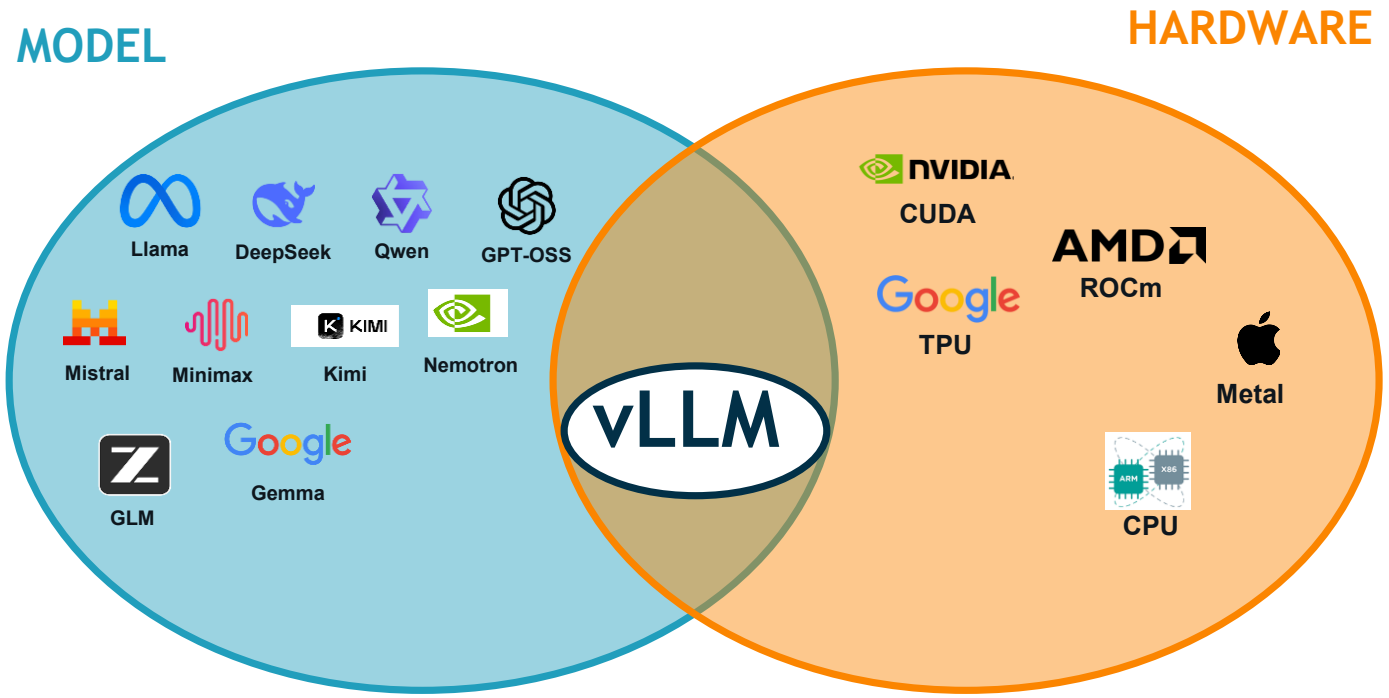
All 128 models from the supplied CSV · weighted cost per Intelligence Index task



inferact | vLLM

What does Inferact do?

Production scaling on data center environment with thousands of GPUs



THANK YOU

vLLM x inferact

Twitter: Xianbao_QIAN
Linkedin: https://www.linkedin.com/in/tiezhen
Whatsapp: +852 6206 2575