

heise+ | KI-Benchmark: Inferenz-Engines vLLM, SGLang und NIM im Vergleich

2026-03-19 11:00

Inferenz-Engine entscheidet über Durchsatz, Latenz und Stabilität großer Sprachmodelle unter Last. Ein praxisnaher Benchmark vergleicht die drei Werkzeuge.