

Google komprimiert LLM-Cache auf 3 Bit ohne Genauigkeitsverlust

2026-03-26 10:46

Googles TurboQuant drückt den KV-Cache großer Sprachmodelle auf 3 Bit. Die Genauigkeit soll bleiben, die Geschwindigkeit sich vervielfachen.