on-premise LLM-Vergleich

RAG = Retrieval-Augmented Generation

KI generiert Antworten auf Anfragen, indem das Modell nicht nur auf sein internes Wissen zurückgreift, sondern auch Informationen aus externen Dokumenten oder Datenbanken abruft.

Modell	Quantisierung & Speichergröße (ggf.)	RAM-Bedarf (ungefähr)	Eignung für Dokumentensuche & RAG	Quelle
Mistral 7B (Q4_K_M)	Q4_K_M (4-bit), ca. 4,37 GB	ca. 6,87 GB RAM	Gut geeignet, ausgewogen zwischen Größe, Performance und Qualität; effizientes Fine-Tuning und gute Inferenzzeiten lokal möglich	https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.1-GGUF
Phi-2 7B	Q4_0 (4-bit), ca. ~4 GB	ca. 6-7 GB RAM	Sehr gut geeignet, speziell trainiert auf Code und kontextuelle Aufgaben	https://huggingface.co/microsoft/phi-2
LLaMA 3 7B	ca. 6-7 GB (unquantisiert); quantisierte Varianten existieren	ca. 8-10 GB RAM	Sehr gute allgemeine Leistung, tendenziell höherer Ressourcenbedarf; große Community (womöglich Lizenz benötigt)	https://www.llama.com/models/llama-3/
Alpaca 7B	ca. 6-7 GB (meist fine-tuned LLaMA 7B)	ca. 8-10 GB RAM	Gut für prototypische RAG-Lösungen, basierend auf LLaMA; kleinere Community, aber viel Open-Source-Support	https://github.com/tatsu-lab/stanford_alpaca