

KI-Ressourcen & Links

Interessante Links und Videos zu KI- und LLM-Themen

Daniel Miessler

danielmiessler.com

Daniel Miessler schreibt regelmäßig über Themen rund um Informationssicherheit, künstliche Intelligenz, Denkmodelle und persönliche Wissenssysteme. Viele seiner Texte beschäftigen sich mit der praktischen Nutzung von KI im Alltag und in professionellen Kontexten.

Fabric (GitHub)

github.com/danielmiessler/Fabric

Fabric ist ein Open-Source-Framework zur Erweiterung menschlicher Fähigkeiten mithilfe von KI. Es stellt ein modulares System bereit, um konkrete Probleme mit Hilfe spezialisierter, gemeinschaftlich entwickelter KI-Prompts zu lösen. Diese Prompts sind flexibel einsetzbar und nicht an eine bestimmte Plattform gebunden.

Der Fokus liegt weniger auf einem einzelnen Modell, sondern auf wiederverwendbaren Denk- und Arbeitsmustern für den praktischen Einsatz von LLMs.

Glukhov

glukhov.org

Eine hochtechnische Ressource und ein Blog, der sich auf die Optimierung und Analyse von Large Language Models (LLMs) spezialisiert hat. Die Seite bietet tiefgehende Einblicke in Themen wie Quantisierung, Modell-Architekturen und die Effizienzsteigerung von KI-Modellen. Besonders wertvoll für Nutzer, die verstehen wollen, wie Modelle "unter der Haube" funktionieren und wie sie für die lokale Ausführung optimiert werden können.

NVIDIA Nemotron

[NVIDIA Nemotron Developer Repository](#)

Open and efficient models for agentic AI. Training recipes, deployment guides, and use-case examples for the Nemotron family.

- 24 Mar 2026: [NVIDIA Nemotron 3 Ultra Release](#)
 - Nemotron 3 Nano (4B): Es gibt auch ein Nemotron 3 Nano in einer 4B-Version, welches nur 2.84GB groß ist. [HuggingFace Link](#)
-

LLMs

- [Llama RPC RCE](#)
- [Distributed Inference llama.cpp via RPC](#)
- [llama.cpp Docker Documentation](#)
- [Google Gemma 4 31B IT](#)
- [Self-hosted LLM Leaderboard](#)

RAG

- [RAG Chunking Strategies: The 2026 Benchmark Guide](#)
 - [Open WebUI RAG Troubleshooting](#)
 - [Optimizing RAG Chunk Size Guide](#)
 - [Gemini Chat](#)
 - [Processing Large Documents 128k Limit](#)
-

Revision #7

Created 2025-11-28 07:47:45 UTC by Carsten

Updated 2026-07-05 20:28:10 UTC by Carsten