LLAMA.cpp and RAG Resources To Read

LLAMA.cpp

- https://retr0.blog/blog/llama-rpc-rce
- https://tekkix.com/articles/ai/2024/09/distributed-inference-llamacpp-via-rpc
- https://github.com/ggml-org/llama.cpp/blob/master/docs/docker.md
- https://huggingface.co/google/gemma-4-31B-it
- https://onyx.app/self-hosted-llm-leaderboard

RAG

- https://blog.premai.io/rag-chunking-strategies-the-2026-benchmark-guide/
- https://docs.openwebui.com/troubleshooting/rag/
- https://machinelearningplus.com/gen-ai/optimizing-rag-chunk-size-your-definitive-guide-to-better-retrieval-accuracy/
- https://gemini.google.com/app/97b148ccb6e03fc1
- https://community.openai.com/t/processing-large-documents-128k-limit/620347/9
-

Revision #2

Created 2026-04-10 12:00:50 UTC by Carsten

Updated 2026-04-10 12:12:30 UTC by Carsten