Self-hosting AI LLMs – a beginners guide

Nick Burch

Berlin Buzzwords 2025

Nick Burch

Director of Engineering

Nick Burch

@Gagravarr

@nick@social.earth.li

Slides, Links

github.com/Gagravarr/BBuzz2025-SelfHostedLLMs

Introduction

Models

Models

Where to get them?

Models

Size?

Parameters?

Quantisation?

Models

Evaluation?

Rankings?

Software

AKA How to run your models locally

llama.cpp

ollama

llm, from Simon W

In General

Try GitHub!

In General

It's a mixture of C and Python

Not always from Software Devs

Not always following normal build/install patterns

Problems

Prompt Injection

Incorrect Answers

aka Hallucinations

Building Solutions

Turning a Local LLM into something useful!

MCP

ollama REST

Or just copy!

Follow the interface of one of the big cloud providers

Hardware

Phones

Quick Demo

Any Questions?

github.com/Gagravarr/BBuzz2025-SelfHostedLLMs