

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

Example local- LLM output



[github.com/Gagravarr/BBuzz2025-
SelfHostedLLMs](https://github.com/Gagravarr/BBuzz2025-SelfHostedLLMs)

Introduction

Introduction

- Models - where to find them, picking one
- Software - what to use, where to find
- Problems - what to be aware of
- Building Solutions
- Hardware
- Demo!

Models

Models

Where to get them?

Where to get them?

Not on Github!

Where to get them?

Hugging Face

Where to get them?

Hugging Face

- Bit like Github for models
 - Discussion
 - Forks
 - Storage
 - Hosting
- Datasets for Fine-Tuning
- (And datasets for training)
- Over 1 million models!

Where to get them?

Hugging Face

- Many big models have their own hosting too
- You won't always download from Hugging Face
- But it's where to find most of the community
- Like GitHub, but for the models

Models

Size?

Parameters?

Quantisation?

Context Window? Tokens?

Model Size

Parameters

- Parameters - often measured in Billions
- Roughly - how much information it holds
- More means the model will be larger!
- So needs more memory / GPUs etc
- (Plus also disk space!)
- Languages / multimodal affects this too
- Bigger isn't always better! Especially for local
- **1B - 7B** typical for locally run models

Model Size

Quantisation

- Quantisation - fidelity of weights and activations
- Roughly - how many decimal places on the numbers
- Lower quantisation can dramatically reduce size
- So less disk space, but especially less memory!
- But - very low weights might be zero'd
- But - differences between "nearly the same" lost

Model Size

Context Window / Token Limits

- eg 32k output
- eg 16k context window
- Context Window - how long until the model forgets what went before?
- Output Context - how long a response can it generate?
- Models "forget" things out of the context window
- Summarising / re-prompting helps, but doesn't fix
- Longer is often slower and/or more memory...

Models
Evaluation?
Rankings?










Model Evaluation




- For some things, you can run a set of prompts and check answers
- As with all things AI - need some ground truth!
- But a lot of it comes down to "vibes"
- Try a bunch of stuff, and take a guess!
- Even the big providers don't have a science here...

Model Evaluation / Rankings

- Bit like benchmarks, often they pick one they win on!
- [LM Arena](#) - lots of humans testing
- [OpenRouter](#) - rankings on different aspects
- But model providers sometimes cheat on the version posted to these sites...
- Can be a good way to try a bunch of different ones out easily/cheaply!

Model Evaluation / Rankings

✍ Text 🕒 2 days ago			
Rank (UB) ↑	Model ↓	Score ↑↓	Votes ↑↓
1	 gemini-2.5-pro-preview-06-05	1470	7,343
2	 o3-2025-04-16	1447	15,210
2	 gemini-2.5-pro-preview-05-06	1446	12,351
4	 chatgpt-4o-latest-20250326	1436	19,762
4	 gpt-4.5-preview-2025-02-27	1430	15,271
5	 claude-opus-4-20250514	1420	13,850
6	 gemini-2.5-flash-preview-05-...	1418	12,614
7	 gpt-4.1-2025-04-14	1408	13,830
8	 grok-3-preview-02-24	1404	21,879

✍ WebDev View →			
Rank (UB) ↑	Model ↓	Score ↑↓	Votes ↑↓
1	 Gemini-2.5-Pro-Preview-06-05	1443	1,872
1	 Claude Opus 4 (20250514)	1412	2,466
2	 Gemini-2.5-Pro-Preview-05-06	1408	3,858
2	 Claude Sonnet 4 (20250514)	1389	2,078
5	 Claude 3.7 Sonnet (20250219)	1357	7,481
6	 Gemini-2.5-Flash-Preview-05-...	1312	2,626
7	 GPT-4.1-2025-04-14	1256	5,489
8	 Claude 3.5 Sonnet (20241022)	1238	26,338
9	 DeepSeek-V3-0324	1207	1,097

"Best Model"

- There is no "overall" best model
- You'll need to test for your own problem space
- Consider memory use, speed, context windows etc
- Consider accuracy, false positives and negatives etc
- Fine tuning can help!

Software

AKA How to run your models locally

Software - mostly llama

Facebook Llama came first!

Most other LLM models also supported

llm tool/wrapper

from Simon Wilson / Datasette

- CLI tool for interacting with LLMs
- Works with Cloud-hosted LLMs
- Works with local LLMs
- Very easy to switch between models, and cloud vs local
- Aims to be very beginner friendly
- Easy to install/setup and get started
- Plugin interface makes extensions possible

llm tool/wrapper

from Simon Wilson / Datasette

My suggestion for new users!

ollama

- Based on llama.cpp (covering shortly!)
- Pre-build binaries for most platforms
- Easy support for downloading models
- REST interface for managing and running
- Stats on most popular modals, and for what
- Common integration for many things (eg Cline)
- Fewer sharp edges, but fewer tuning points than llama.cpp

llama.cpp

- Fast CPU loading and execution
- Many GPUs supported
- But you need to enable GPU support + have libraries!
- Generally new stuff happens here first
- Lots of very cool tricks and techniques
- Lots of control over how things work

In General
Try GitHub!

In General

- It's a mixture of C and Python
- Not always from Software Devs
- (Quite often with AI helping write wrappers!)
- Not always following normal build/install patterns
- Often feels like trying to make games run in the early 2000s!

Problems

Prompt Injection

Prompt Injection

Data Exfiltration

Incorrect Answers
aka Hallucinations

Power

Electricity

Cooling, eg Water

Building Solutions

Turning a Local LLM into something useful!

MCP

Model Context Protocol

MCP

- Way to let an LLM interact with other systems
- eg send email
- eg list / search / read document
- eg write a file
- eg visit a website
- eg interact with an API

MCP

- Can be very cool and powerful
- Current standard way to integrate LLMs

See *Problems* section

And then see it again!

Multiple Agents

- Current "best practice" is to have multiple agents
- eg 1 to plan, 3 to execute, 1 to finish
- Some of these will be tool-use LLMs via eg MCP
- Can use smaller/cheaper/quicker LLMs for some parts
- Can help avoid context overload
- More LLMs → More Tokens → More Cost!

ollama REST API

- Lots of integrations can talk to the ollama REST API
- Fairly well supported pattern
- Only for trusted input / access!

Or just copy!

**Follow the interface of one of
the big cloud providers**

Hardware

Hardware matters
CPUs are slow at this!

Mac Metal

- For most in the room, easiest fast-ish LLM acceleration
- Fairly widely supported
- Often built as standard
- Quite a large speedup

Nvidia

- Big speedups possible
- Widely supported
- Subject to nation-state export controls(!)
- Laptop / Desktop / Docker
- Needs drives / libraries / permissions
- Sometimes feels like 2000s gaming...

Nvidia - Which GPU?

Which GPU for Local LLMs?

Radu Gheorghe + Rafał Kuć

2025.berlinbuzzwords.de/session/which-gpu-for-local-llms/

10.40am earlier today

Video likely to be up in about 24 hours!

AMD

- Fairly good speedups possible
- Less well supported
- Less well documented
- Usually feels like 2000s gaming...

Intel

- Some speedups possible
- Barely supported
- Documentation out-of-date
- Feels like early 2000s linux graphics drivers...
- Consider your hourly rate, and probably by a different laptop!

Huawei

- Ask your Chinese colleagues for help
- Some cool stuff going on!
- But mostly for the Chinese domestic market, for now
- Somewhat - pick your geopolitical poison!

Phones

Phones

**You can run LLMs on your
phone!**

Phones

- It'll probably run all all of *your* phones
- But not on all phones globally...
- Limited memory, storage, speed
- Use a small model, few parameters and probably quantised
- And maybe something pre-LLM!

Quick Demo

Demo output, in case of live-demo gremlins...

I am giving an introductory talk at Berlin Buzzwords 2025 on locally-hosted LLMs, and need to publicise my talk. The conference hashtag is #bbuzz . Can you suggest a few possible tweets I can use?

Demo output - Llama 3.2 3B Instruct (1.9gb)

Here are a few possible tweets you can use to publicize your talk at Berlin
Buzzwords 2025:

1. **Exciting news!** I'm honored to share my research on locally-hosted LLMs at #BBuzz2025! Don't miss my talk to learn more about the future of language models and their potential applications in a decentralized world. #LLMs #NLP
2. **Join the conversation!** I'm speaking at #BBuzz2025 about the rise of locally-hosted LLMs - a game-changer for data privacy, security, and autonomy. Mark your calendars and come learn with me! #NLP #DecentralizedLLMs
3. **The future of language models is decentralized** I'm thrilled to share my research on locally-hosted LLMs at #BBuzz2025! Don't miss this opportunity to learn about the possibilities and challenges of decentralized NLP. #LLM #Decentralization
4. **Talk to me about locally-hosted LLMs!** I'll be sharing my insights on the potential of locally-hosted LLMs at #BBuzz2025. Join the conversation and learn about the future of language models! #NLP #LocallLMs

Demo output - DeepSeek Qwen 1.5B (1gb)

Much thinking on what to write and why

Here are a few engaging tweets you can use for your talk at Berlin Buzzwords 2025, targeting the #bbuzz hashtag:

1. Impact and Cost Savings:

- *"**Locally-hosted LLMs are revolutionizing AI development. They're saving time and reducing costs while enabling faster innovation.**"
#LocallyHosted #AI #LLM #BBuzz"

2. Call to Action:







- *"**Bring your LLMs to Berlin! We're excited to announce our latest advancements in locally-hosted AI solutions. What's your take on hosting your own LLM? Let's discuss!**" **#BBuzz #AI #LocallyHosted"**

3. Versatility and Future Impact:

- *"**Your future is just around the corner with locally-hosted LLMs! They'll give you a significant edge in AI-driven challenges. Let's explore how you can get started quickly!**" **#LocallyHosted #AI #BBuzz"**

Demo output - Devstral (14gb)

Absolutely! Here are a few tweet suggestions to publicize your talk on locally-hosted LLMs at Berlin Buzzwords 2025:

1. **Engaging and Informative:** "  Excited to share my talk on locally-hosted LLMs at #bbuzz 2025! Learn how to harness the power of large language models right on your hardware. Join me to discover the benefits and best practices.  #AI #LLMs"
2. **Personal Touch:** "  Thrilled to speak at #bbuzz 2025 about locally-hosted LLMs! Dive into the future of AI with me. Let's connect, learn, and innovate together. See you there!  #BerlinBuzzwords #AICommunity"
3. **Highlighting Benefits:** "  Join me at #bbuzz 2025 for an insightful session on locally-hosted LLMs. Discover why running these models locally can offer better control, privacy, and efficiency. Can't wait to see you there! #AI #LocalLLMs"
4. **Call to Action:** "  Mark your calendars! My talk on locally-hosted LLMs is happening at #bbuzz 2025. Don't miss out on this opportunity to explore

Any Questions?

[github.com/Gagravarr/BBuzz2025-
SelfHostedLLMs](https://github.com/Gagravarr/BBuzz2025-SelfHostedLLMs)



Want to try?

Find me in the 4pm coffee break!

**I'll help you get a local model installed on your
laptop**