AI-FLUENCY · LLM-WIKI · HERMES AGENT

Build a Second Brain with Hermes Agent

A personal, interlinked knowledge base you grow by doing — set it up on your own machine, feed it what you learn, and let it compound.

Set up | Build | Use | Extend | Sharpen | Go

BEFORE ANYTHING — THE MINDSET

A fast-moving field. Bring a learner's mind.

Treat it as a snapshot

Models, prices and free-tier limits change monthly. Specifics go stale fast — names here may be gone by the time you read them.

Fundamentals last

Context windows, quantization, memory bandwidth, the privacy/speed/quality trade-off. Learn these deeply and specifics slot in easily.

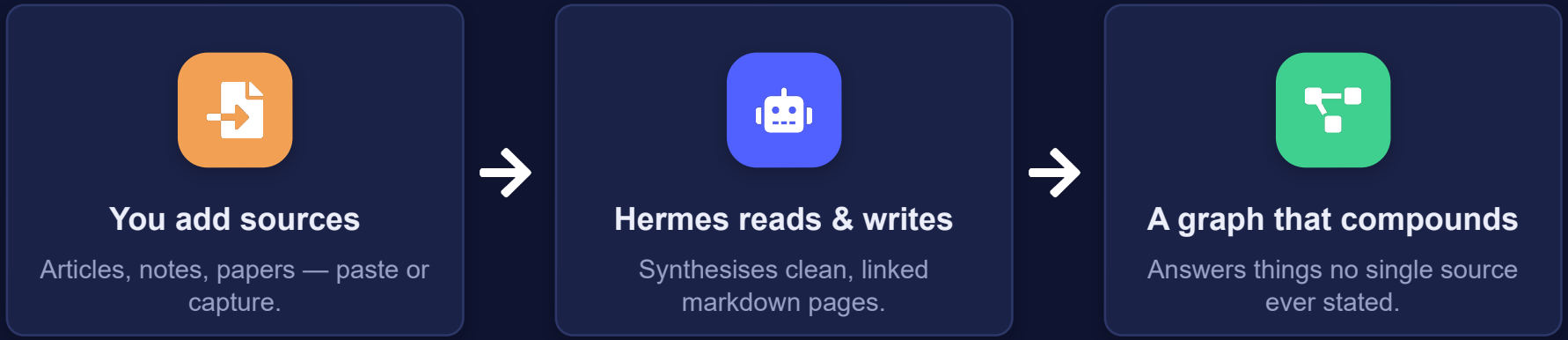
Learn by doing, daily

A working setup teaches more than any guide. Small steps compound — and a second brain is where they accumulate.

THE IDEA

What you're building

A **personal, interlinked wiki** on the topic of AI fluency — grown by an agent that reads what you feed it, writes encyclopedia-style pages, and cross-links everything. This is Karpathy's *“LLM wiki”* pattern, shipped as the built-in `llm-wiki` skill.



The guiding idea

Augment your thinking — don't offload it.

Use the wiki to revisit and connect what you've engaged with. ***Use it after the struggle, not instead of it.***

THE MAP

The journey — five phases, colour-coded

Each step is tinted by phase, so you always know where you are. The interactive mind map and guide use the same colours.

Set up
Steps 1–2

Build
Steps 3–5

Use
Steps 6–7

Extend
Steps 8–9

Sharpen
Upkeep

Go
Impact

Gold = where you are right now · green ring = done

SET UP · STEPS 1–2

Step 1 — Install Hermes Agent

Pick the way you like to work — both reach the same place. Works on Windows, macOS & Linux.

Desktop app · GUI

- Download the installer for your OS
- Click through Providers, Model, Skills
- Settings panes — no terminal needed
- (one config value via CLI later)

Terminal · CLI

- Install via the one-line script
- hermes init to set things up
- Everything by command
- Great for servers & homelabs

SET UP · STEPS 1–2

Step 2 — Connect a model (the brain)

One hard rule: Hermes needs a model with at least **64,000 tokens of context**. Most cloud models clear it; local models need their window raised.

Path A — Cloud

Easiest. Nothing to install, no hardware worries. Free tiers (Google AI Studio, OpenRouter, Groq) need no card and clear 64K easily. Start here.

Path B — Local

Private, offline, free forever — but bounded by your hardware. Run via Ollama once you want your wiki to stay on your own machine.

SET UP · STEPS 1–2

Choosing a model — free & fast first

Priority for beginners: get it working at $0, learn the workflow, optimise later. No credit card needed.

Google AI Studio

`Gemini 2.5 Flash`

~1,500 req/day · 1M context

Best first stop

OpenRouter

`28+ :free models`

20 req/min · one key

Easiest to swap

Groq

`Llama · Qwen · DeepSeek`

Hundreds of tok/s

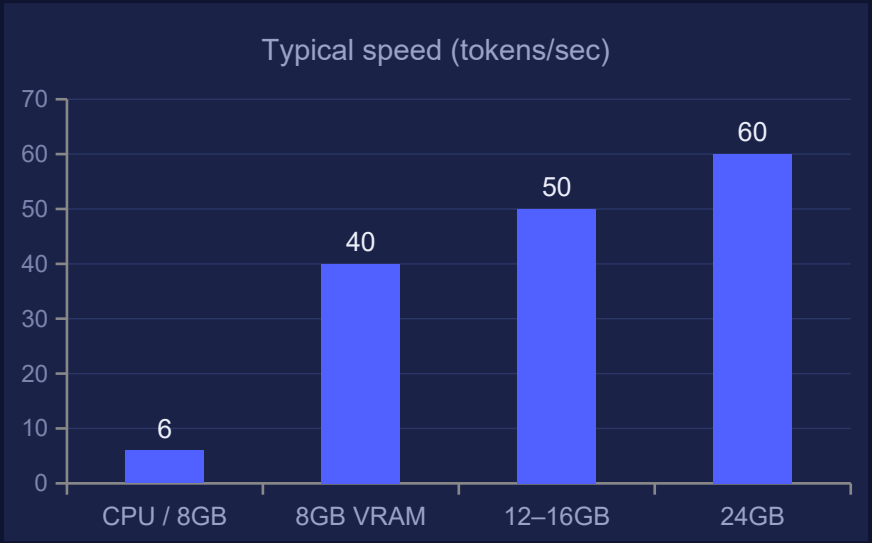
Fastest free

Privacy trade-off: free tiers are usually funded by training on your prompts. Fine for learning — keep truly private material out, or go local.

SET UP · STEPS 1–2

Local models — does your hardware fit?

Fit follows your VRAM (or Mac unified memory) at standard Q4 4-bit. Speed is set by memory bandwidth.



- **CPU / 8GB RAM**
  1–3B · testing only
- **8GB VRAM / 16GB Mac**
  7–9B · Llama 3.1 8B
- **12–16GB VRAM**
  12–14B · the sweet spot
- **24GB (3090/4090/5090)**
  27–32B · nears cloud

FUNDAMENTALS THAT DON'T GO STALE

Why GPUs win, and why quantization helps

Memory-bandwidth bound

Every token reads the whole model from memory. The maths is trivial; the limit is how fast memory feeds the chip — so high-bandwidth GPU VRAM beats system RAM, and Apple Silicon's unified memory punches above CPUs.

Quantization (Q4_K_M)

Stores weights at 4-bit: ~75% smaller and faster (less to stream). Q8 ≈ near-original but 2× size; Q2/Q3 degrade. Rule: a smaller model at Q4 beats a bigger one crushed to Q2/Q3.

THE CORE TRADE-OFF

Privacy vs speed vs quality

You can't max all three — pick your point. Choose by what your wiki holds.

Local

Privacy	**Highest**
Speed	**Hardware-bound**
Quality	**What fits**
Cost	**$0 after HW**

Free cloud

Privacy	**Lowest (trains)**
Speed	**Fast**
Quality	**Good–frontier**
Cost	**$0, capped**

Paid frontier

Privacy	**No training***
Speed	**Fast**
Quality	**Best**
Cost	**Per token**

BUILD · STEPS 3–5

Build — wire up the wiki

Three quick steps turn the agent into a working knowledge base.

3

Confirm the skill

Make sure the built-in llm-wiki skill is installed and ready.

4

Set the wiki path

Tell the skill where to keep your pages (e.g. ~/ai-fluency-wiki/).

5

Initialise

Send a first prompt that has the agent read the skill and scaffold structure.

USE · STEPS 6–7

Use — capture, then ask

The daily loop. This is where the compounding happens.

STEP 6

Add a source

Paste an article or note — or capture a URL with the strip helper. The agent reads it and writes linked pages.

STEP 7

Ask your wiki

Query across everything you've fed it. It connects ideas your individual sources never stated outright.

EXTEND · STEPS 8–9

Extend — make it ambient & automatic

Optional, but powerful. Reach your brain from anywhere and feed it on a schedule.

STEP 8

Messaging gateway

Set up a Telegram bot via BotFather and chat to your wiki from your phone — capture on the go.

STEP 9

Automated intake

Schedule a cron job (e.g. a daily arXiv pull) so fresh sources flow in and pages keep growing on their own.

SHARPEN · UPKEEP

Hard-won tips — save yourself hours

Raise the compression threshold

Premature context-compression silently evicts task instructions — looks like the model failing. Bump it (e.g. 0.5 → 0.85).

127.0.0.1, not localhost

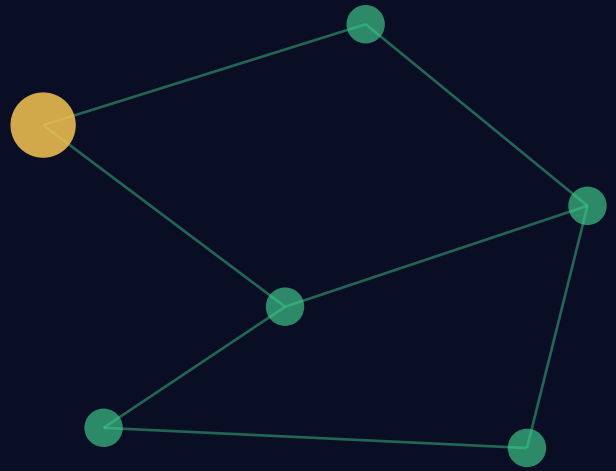
IPv4/IPv6 resolution bites gateways. Prefer explicit 127.0.0.1 over localhost.

Set cwd to the repo

Hermes' working directory must point at the real path, not '.', or file ops land in the wrong place.

Discord: privileged intents

Message Content + Server Members are the most common gateway failure point — enable them.



THE SEND-OFF

Now go build — and make it matter.

Start now — connect a free model, capture one source, ask one question today. **Keep going** — learn by doing, a little every day. **Make it matter** — use what you learn to teach someone, solve a real problem, or make something good exist that didn't before.

Knowledge that sits in a wiki is potential. Knowledge you act on is impact.