Local AI

Einführung, Erfahrungen,
persönlicher Stack

Local AI - Warum?

https://www.businessinsider.com/jensen-huang-500k-engineers-250k-ai-tokens-nvidia-compute-2026-3

AI

Jensen Huang says he would be 'deeply alarmed' if his $500,000 engineer did not consume at least $250,000 of tokens

Quelle: https://www.golem.de/news/ki-angeblich-halbe-milliarde-dollar-in-tokens-verbrannt-2605-209231.html

KI

Angeblich halbe Milliarde Dollar in Tokens verbrannt

Ein Konzern soll rund 500 Millionen Dollar für Claude ausgegeben haben, weil niemand Nutzungslimits für Mitarbeiter eingerichtet hatte.

31. Mai 2026 um 12:11 Uhr / Michael Linden

Quelle: https://www.heise.de/news/US-Regierung-erzwingt-Abschaltung-von-Anthropics-KI-Fable-5-und-Mythos-5-11331129.html

US-Regierung erzwingt Abschaltung von Anthropics KI Fable 5 und Mythos 5

Eine Exportdirektive der US-Regierung zwingt Anthropic, Fable 5 und Mythos 5 abzuschalten. Das Unternehmen spricht von einem Missverständnis.

Local AI - Wie?

Local AI - Wie?

Voraussetzungen

Hardware:

- GPUs (Nvidia, AMD)
- (V)RAM
- Apple Silicon (Shared Memory)

Local AI - Wie?

Welche Modelle sind mit meiner Hardware kompatibel?

https://github.com/AlexsJones/llmfit

https://huggingface.co/models

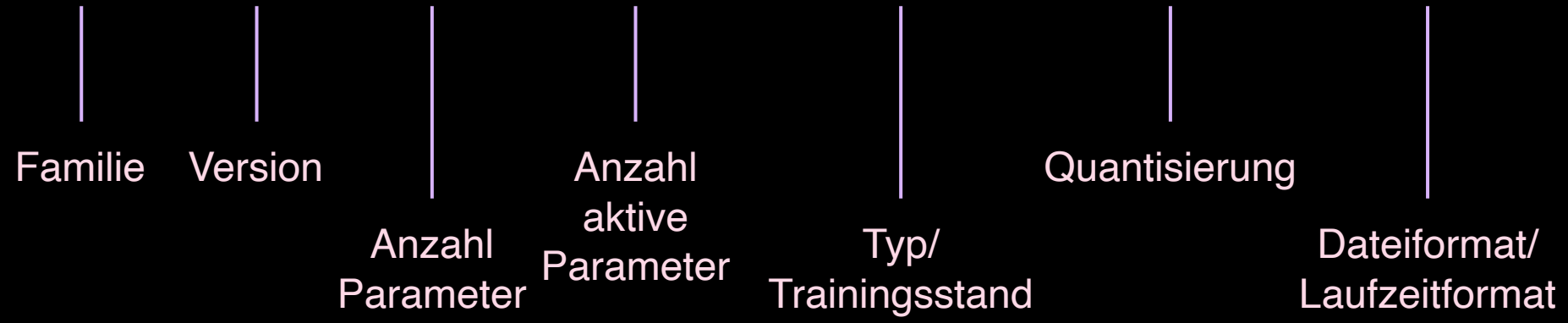
Login -> Settings -> Hardware

https://artificialanalysis.ai

Local AI - Wie?

Welches Modell wofür?
-> Open Weight Bezeichnungen verstehen

qwen3.6:35b-a3b-coding-nvfp4 (MLX)



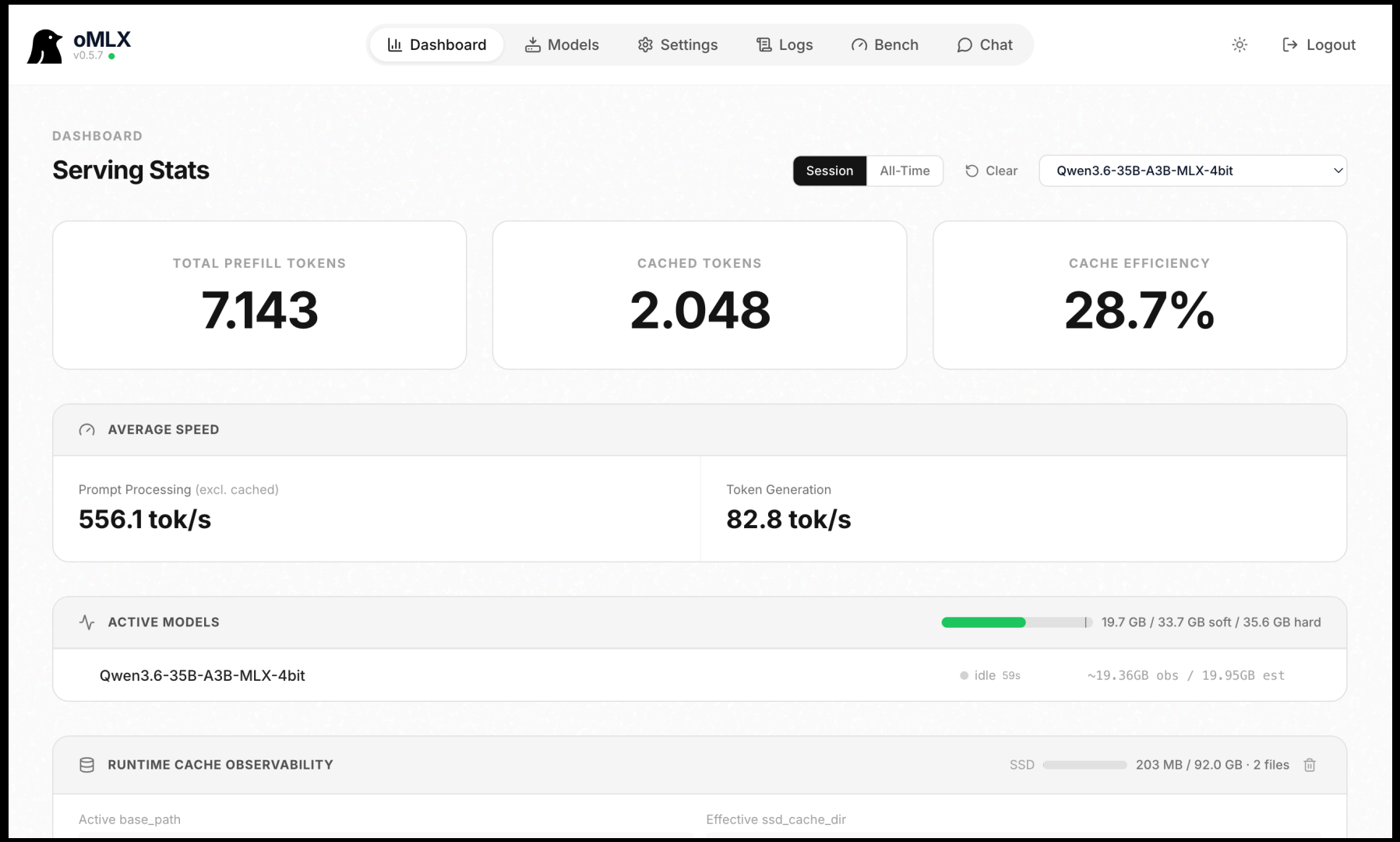
Local AI - Wie?

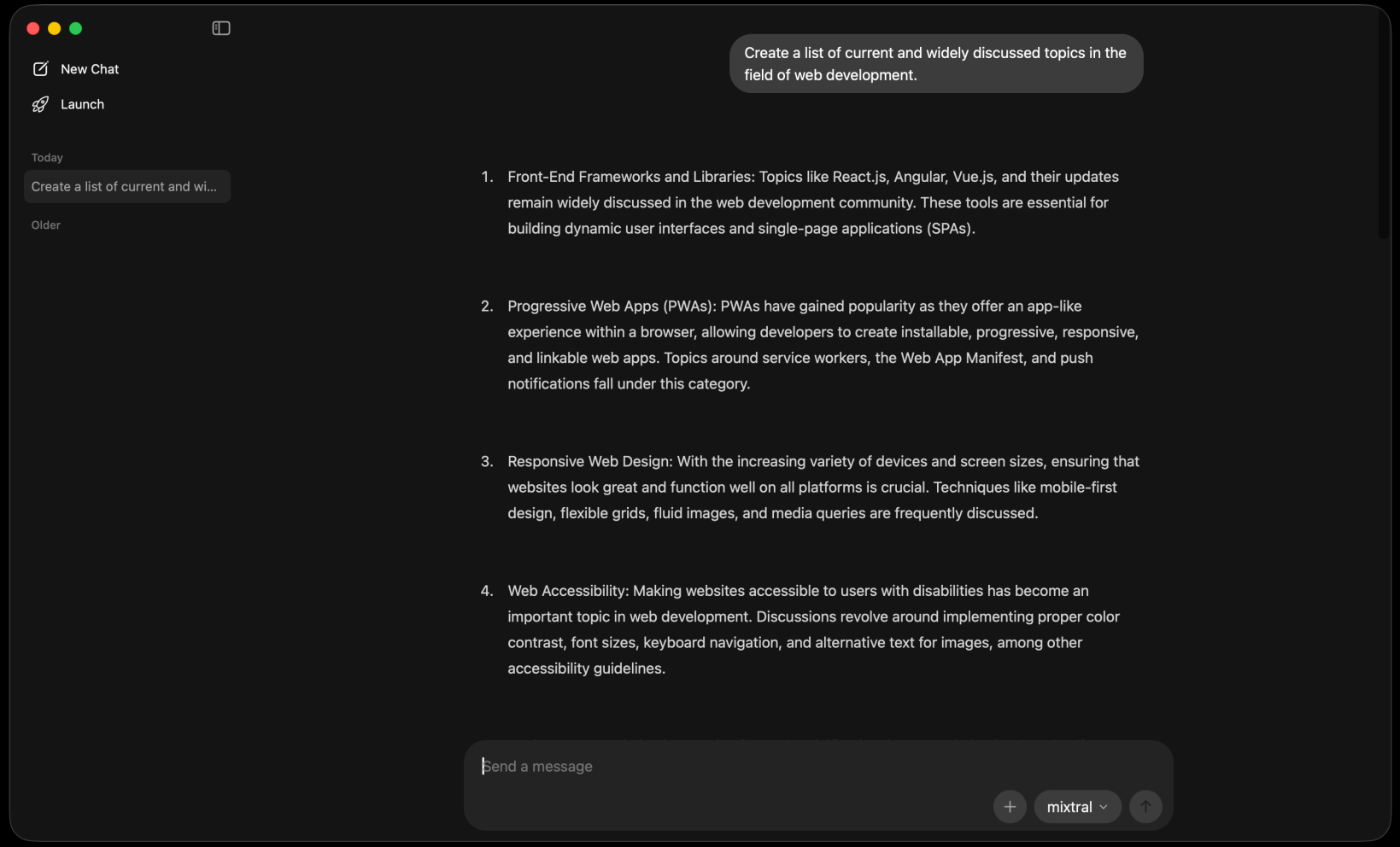
Wie kann ich Modelle lokal installieren und ausführen (Inference)?

-> Beispiel: macOS

https://omlx.ai/

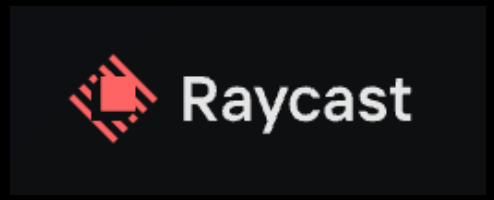
https://ollama.com/


oMLX
v0.5.7
Dashboard
Models
Settings
Logs
Bench
Chat
Logout
DASHBOARD
Serving Stats
Session
All-Time
Clear
Qwen3.6-35B-A3B-MLX-4bit
TOTAL PREFILL TOKENS
7.143
CACHED TOKENS
2.048
CACHE EFFICIENCY
28.7%
AVERAGE SPEED
Prompt Processing (excl. cached)
556.1 tok/s
Token Generation
82.8 tok/s
ACTIVE MODELS
19.7 GB / 33.7 GB soft / 35.6 GB hard
Qwen3.6-35B-A3B-MLX-4bit
idle 59s
~19.36GB obs / 19.95GB est
RUNTIME CACHE OBSERVABILITY
SSD
203 MB / 92.0 GB · 2 files
Active base_path
Effective ssd_cache_dir


New Chat
Launch
Today
Create a list of current and wi...
Older
Create a list of current and widely discussed topics in the field of web development.
1. Front-End Frameworks and Libraries: Topics like React.js, Angular, Vue.js, and their updates remain widely discussed in the web development community. These tools are essential for building dynamic user interfaces and single-page applications (SPAs).
2. Progressive Web Apps (PWAs): PWAs have gained popularity as they offer an app-like experience within a browser, allowing developers to create installable, progressive, responsive, and linkable web apps. Topics around service workers, the Web App Manifest, and push notifications fall under this category.
3. Responsive Web Design: With the increasing variety of devices and screen sizes, ensuring that websites look great and function well on all platforms is crucial. Techniques like mobile-first design, flexible grids, fluid images, and media queries are frequently discussed.
4. Web Accessibility: Making websites accessible to users with disabilities has become an important topic in web development. Discussions revolve around implementing proper color contrast, font sizes, keyboard navigation, and alternative text for images, among other accessibility guidelines.
Send a message
mixtral

Local AI - Wofür?

Für welche Aufgaben kann ich Local AI anwenden?



Danke

Fragen bitte an: Martin Wilmer <martin.wilmer@robole.de>