KI bei der stepping stone AG

Niklaus Hofer

Artificial Intelligence meets Reality

5. September 2025

Übersicht

- Digitale Souveränität
- KI und GPUs
- GPUs in der stoney cloud
  - OpenStack-Integration
  - Images
  - AI on demand

Begrifflichkeiten

- stoney cloud: Unsere OpenStack basierte Cloud
- KI: Künstliche Intelligenz
- LLM: Large Language Model
- GPT: Generative Pre-trained Transformer
- Token: Kleinster Textbaustein (Wort, Satzzeiche, ...)
- Parameter: Trainierte Einstellung eines LLM
- ChatGPT basiert auf einem LLM

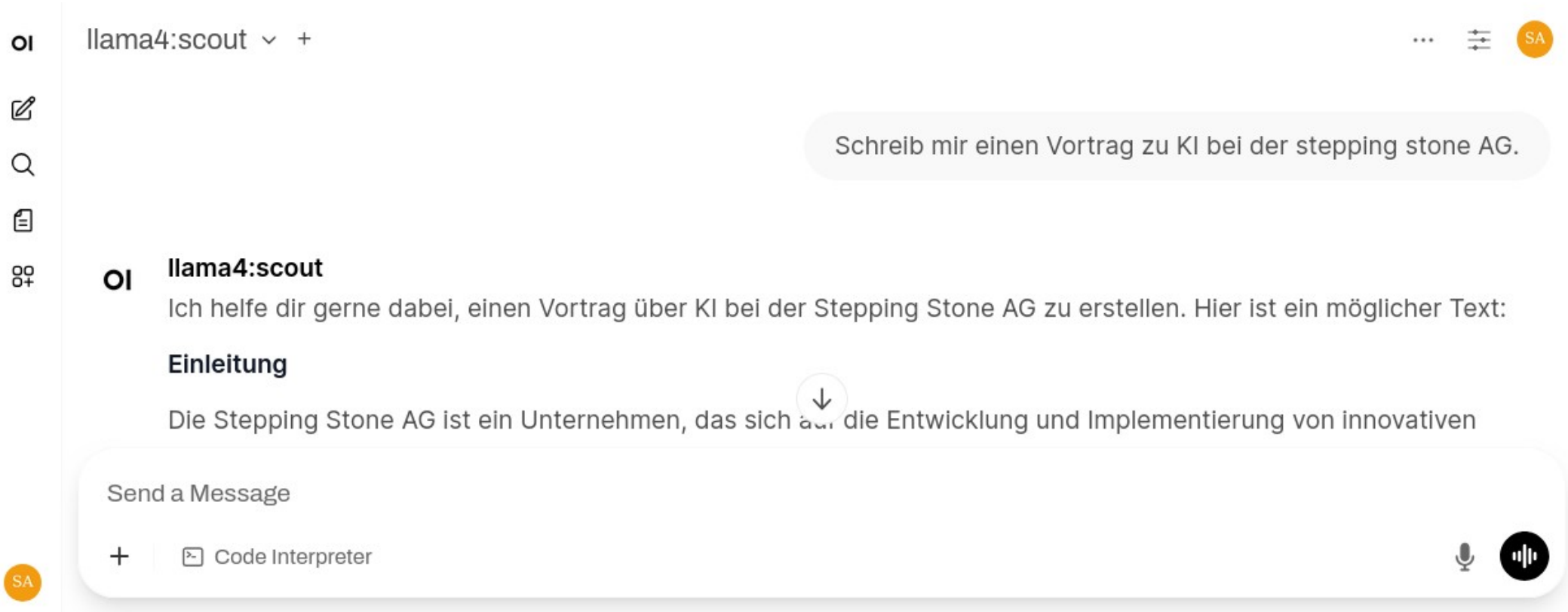
Digitale Souveränität

Digitale Souveräntität

Motivation

- Digitale Souveränität
- Offene Modelle
  - Apertus
  - Mistral
  - Llama
  - Granite
  - DeepSeek
  - Kimi
- Differenzierung von den Hyperscalern

Swissness

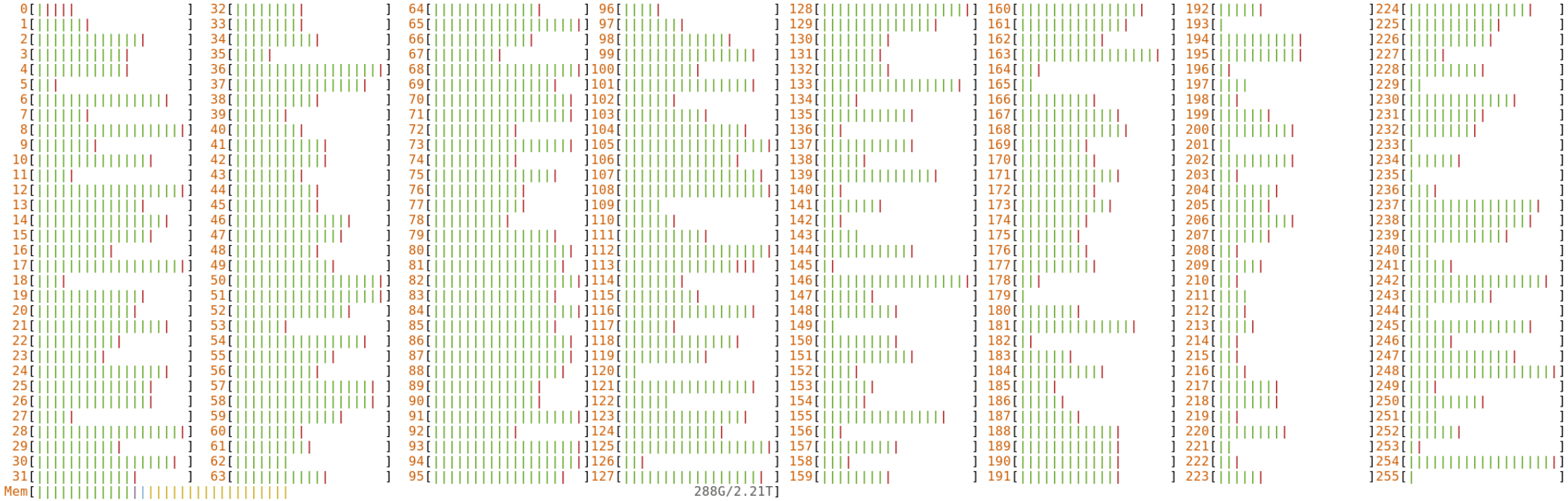

llama4:scout
SA
Schreib mir einen Vortrag zu KI bei der stepping stone AG.
llama4:scout
Ich helfe dir gerne dabei, einen Vortrag über KI bei der Stepping Stone AG zu erstellen. Hier ist ein möglicher Text:
Einleitung
Die Stepping Stone AG ist ein Unternehmen, das sich auf die Entwicklung und Implementierung von innovativen
Send a Message
Code Interpreter
SA

KI und GPUs

Hardwareanforderungen

- Modelle in unterschiedlichen Grössen
  - Je grösser, desto mächtiger
- Zur Nutzung muss das Modell im RAM liegen

LLM mit CPUs



LLM mit GPUs

stepping stone

```
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 580.65.06              Driver Version: 580.65.06      CUDA Version: 13.0     |
+-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA RTX PRO 6000 Blac...    On  |   00000000:01:00.0 Off |                    0 |
| N/A   42C    P0            176W /  600W |   78419MiB /  97887MiB |     29%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   1  NVIDIA RTX PRO 6000 Blac...    On  |   00000000:21:00.0 Off |                    0 |
| N/A   40C    P0            183W /  600W |   76827MiB /  97887MiB |     27%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   2  NVIDIA RTX PRO 6000 Blac...    On  |   00000000:C1:00.0 Off |                    0 |
| N/A   40C    P0            195W /  600W |   81925MiB /  97887MiB |     31%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   3  NVIDIA RTX PRO 6000 Blac...    On  |   00000000:E1:00.0 Off |                    0 |
| N/A   30C    P8             34W /  600W |       3MiB /  97887MiB |      0%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
```

CPU versus GPU

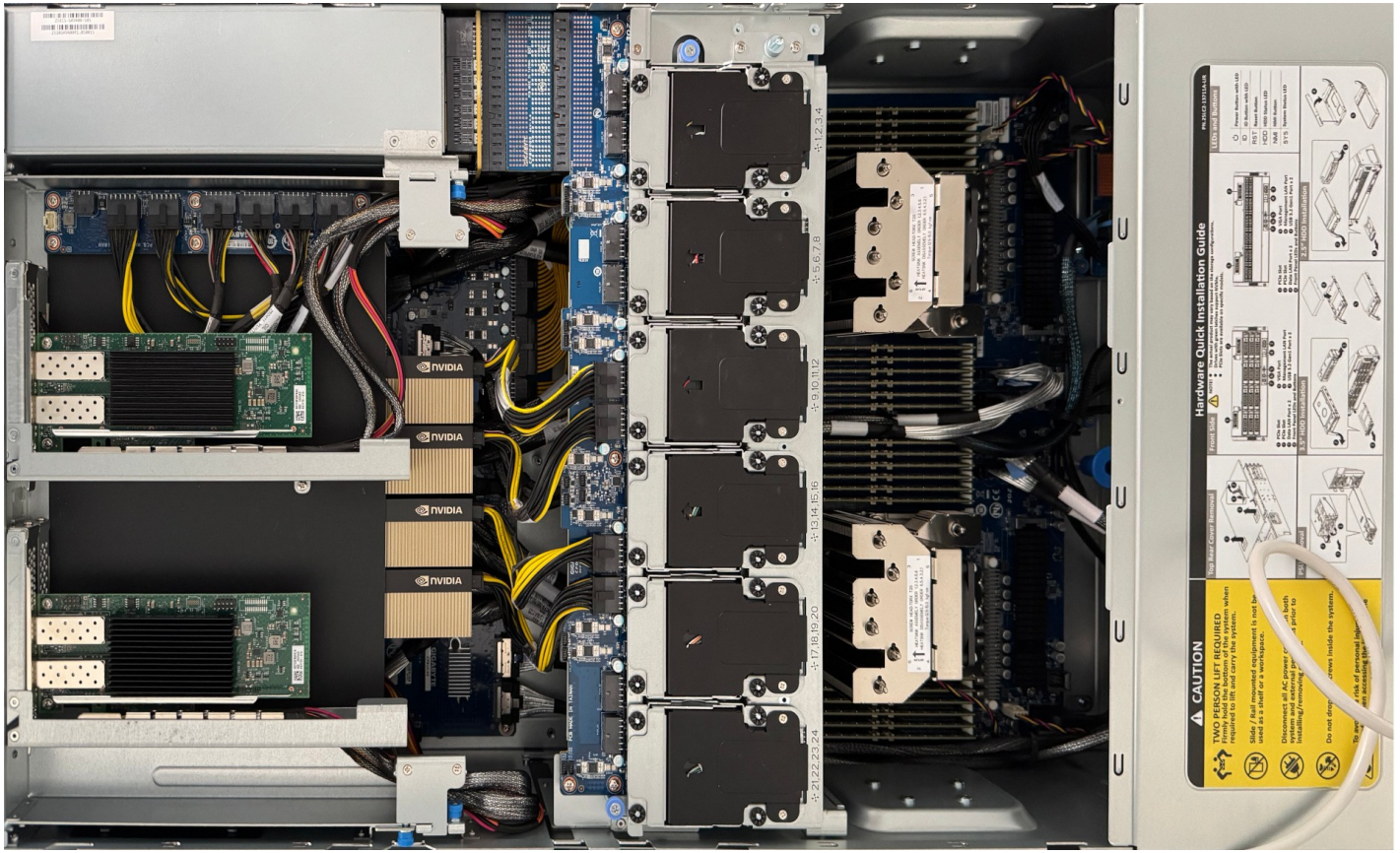
- CPU
  - 2 x AMD EPYC 9554 64-Kern Prozessoren
    - 256 Threads insgesamt
  - 2304 GiB RAM
- GPU
  - NVIDIA RTX PRO 6000 Blackwell
  - 96 GB Video-RAM
- GPU ist 10-70 mal schneller

Benchmarks

stepping stone

Model	Parameters	Size	Tokens per second		Factor (t_{GPU} / t_{CPU})
			GPU only	CPU only	
`mistral:latest`	7B	4.1 GiB	202.46	3.21	63.1
`llama4:scout`	109B	67 GiB	74.91	2.48	30.2
`llama4:maverick`	400B	245 GiB	77.12	8.14	9.5
`granite3.3:latest`	8B	4.9 GiB	165.98	2.56	64.8
`deepseek-r1:latest`	7B	4.7 GiB	161.37	2.30	70.2
Other models					
`phi4:latest`	14B	9.1 GiB	116.05	1.63	71.2
`gpt-oss:20b`	20B	14 GiB	166.85	2.18	76.5
`gemma3:1b`	1B	815 MiB	250.31	2.54	98.5

Unsere KI Server



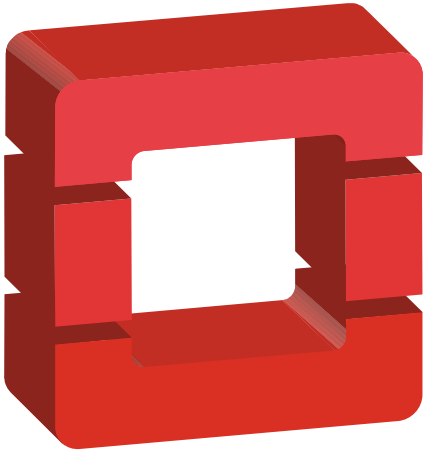
Stromverbrauch und Kühlung

GPUs in der stoney cloud

OpenStack Integration

stepping stone

- Nativ in OpenStack integriert
  - GPU Passthrough
- 1+ GPU(s) pro VM
- Verfügbar: Oktober 2025
- Vorbestellbar: Sofort!



openstack

AI ready Cloud Images

- Cloud Images
  - Mit den gängigsten Werkzeugen
    - Ollama
    - vLLM
    - Hugging Face CLI
    - Open WebUI
  - Und mit den neuesten Treibern
- Regelmässig aktualisiert
- Zusammenarbeit mit der ETHZ

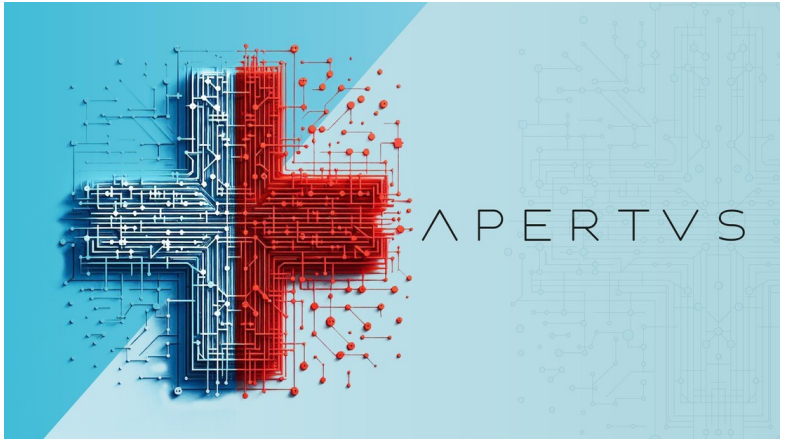
AI Cloud Images

stepping stone

Vorinstallierte Large Language Models:

- Apertus
- Mistral
- Llama
- <Dein bevorzugtes LLM?>

MISTRAL AI_



Anleitungen

- Copy & Paste
- Angepasst für und getestet mit unseren Cloud Images
- In 3 Kommandos zum KI-Prompt!

Kubernetes

- GPUs in Kubernetes
- Via Kubernets Device Plugins

AI on demand

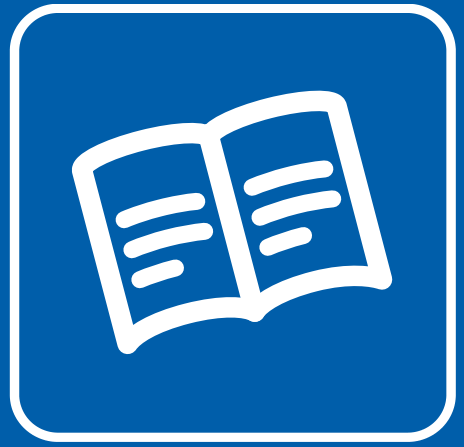
Coming soon™ to a stoney cloud near you!

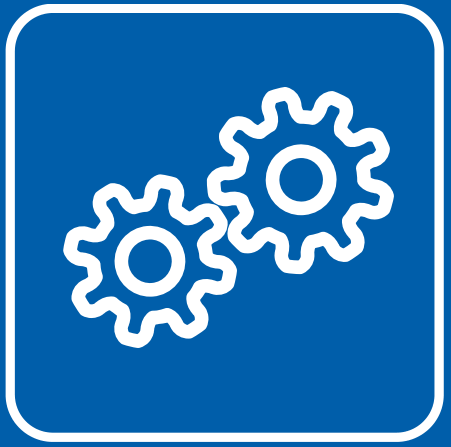
- API-Zugriff
- Abrechnung pro Token

Fragen?









Links

- https://www.stepping-stone.ch/
- https://www.stoney-backup.com/
- https://www.stoney-cloud.com/
- https://www.stoney-mail.com/
- https://www.stoney-meet.com/
- https://www.stoney-office.com/
- https://www.stoney-services.com/
- https://www.stoney-storage.com/
- https://www.stoney-wiki.com/

stepping stone AG
Wasserwerkgasse 7
CH-3011 Bern

Telefon: +41 31 332 53 63
www.stepping-stone.ch
info@stepping-stone.ch