KI-Rechner

Der PING KI-Rechner heißt cogito. Zur Hardware gibt es eine eigene Wiki-Seite.
Hier geht es jetzt um Software und Dienste und Interoperabilität.

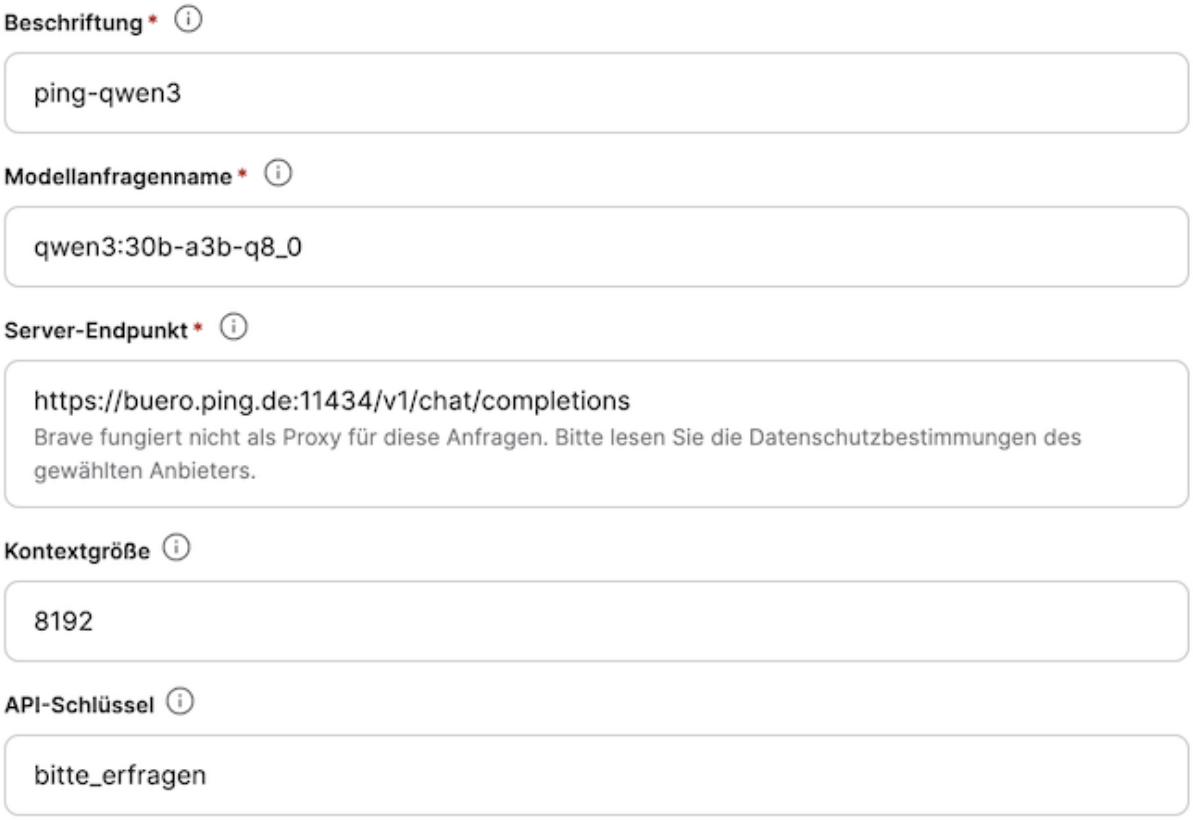
- Brave-Leo-AI mit PING KI
- Visual Studio Code und OpenCode
- Inferenz Benchmarks
- Zugriff auf vLLM via OpenAI-kompatibler API

Brave-Leo-AI mit PING KI

Konfigurieren von Leo AI im Brave Browser mit dem PING AI Server

Auf cogito (KI-Server) läuft ein Ollama-server, der einen OpenAI-API kompatiblen Endpunkt bereitstellt. Er ist unter https://ki.ping.de:8000/ erreichbar, man benötigt für den Zugriff ein Bearer Token.

1. Besorgt euch das Bearer Token, siehe API-Token Seite (für Mitglieder)
2. Im Brave Browser oben rechts auf das "Leo AI" Icon  klicken.
3. Klickt oben rechts auf die 3 Punkte übereinander " ⋮ "
4. Klickt ganz unten auf "Erweiterte Einstellungen" (mit dem Zahnrad). Ihr landet dann auf `brave://settings/leo-ai`
5. Unter "Bringen Sie Ihr eigenes Modell mit" klickt auf "Neues Modell hinzufügen"
6. Macht folgende Einstellungen:
   - Beschriftung: `ping-qwen3`
   - Modellanfragename: `qwen38-27b` (diesen Namen seht ihr so auch im open-webui oder s.u.)
   - Server-Endpunkt: `https://ki.ping.de:8000/v1/chat/completions`
   - Kontext-Größe: `65536` (mehr geht auch, hängt vom LLM und freien VRAM ab).
   - API-Schlüssel: Siehe Punkt 1. Ohne "Bearer" davor eingeben.
     ⚠ ACHTUNG, Screenshot **veraltet**:



7. Klickt auf "Modell speichern"
8. Stellt das "Standardmodell für neue Unterhaltungen" auf `ping-qwen3`

Fertig. Wenn ihr jetzt auf das Leo-AI-Icon klickt startet eine neue Unterhaltung mit dem LLM auf dem PING Server cogito. Wenn ihr auf das "Seitenleiste anzeigen" Icon daneben klickt teilt sich das Browserfenster und ihr seht neben der Webseite das Leo AI Chatinterface, dort könnt ihr dann das LLM zur gerade aktiven Webseite befragen (zusammenfassen etc.).

Die KI von der Brave Search läuft davon unabhängig in der Cloud von Brave.

Verfügbare Modelle auflisten

Wenn ihr eine Liste aller installierten Modelle sehen möchtet, dann könnt ihr das entweder in open-webui oder es geht über die Ollama API wie folgt (ihr benötigt die Befehle `curl` und `jq`):

```
BEARER_TOKEN=siehe_oben
curl -sH "Authorization: Bearer $BEARER_TOKEN" https://ki.ping.de:8000/v1/models | jq
```

Visual Studio Code und OpenCode

Wie nutzt ihr Visual Studio Code und OpenCode mit dem PING KI Server?

Bearer Token / API key / API token

Zunächst braucht ihr das Bearer Token. Es heißt manchmal auch API Token.

PING Mitglieder finden es unter API Token.

Ermittelt welche Modelle der Server anbietet, die euch fürs Programmieren interessieren. Unter `max_model_len` seht ihr die maximale Kontext-Länge.

```
BEARER_TOKEN=siehe_oben
curl -sH "Authorization: Bearer $BEARER_TOKEN" https://ki.ping.de:8000/v1/models|jq
```

Aktuell (2026-08) gibt es nur ein Model: `qwen38-27b` mit Kontext-Länge 230.000. Es ist gut und schnell und bietet Tool calling, ist multi-modal etc. □□

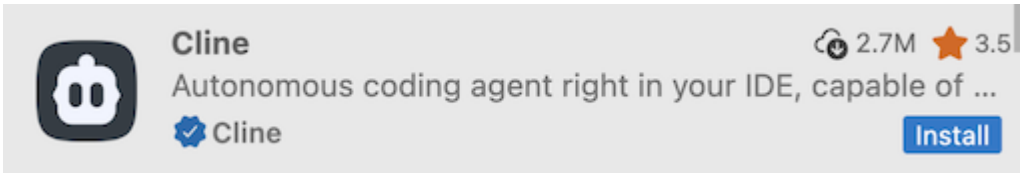
Dann geht es wie folgt:

Visual Studio Code

Für VS Code gibt es mehrere KI-Plugins, die in Frage kommen. Hier findet ihr für einige exemplarische Plugins die benötigten Schritte zur Konfiguration:

Cline Plugin

1. In VS Code das Cline Plugin installieren. Achtet darauf dass es das blaue Checkmark hat:



2. *How will you use Cline?* "**Bring my own API key**"
3. API Provider:**OpenAI**
4. Custom base URL: `https://ki.ping.de:8000`
5. API key: **siehe_oben** (nutzt das Bearer Token)
6. Model: `qwen38-27b`

Fertig!

Kilo Code Plugin

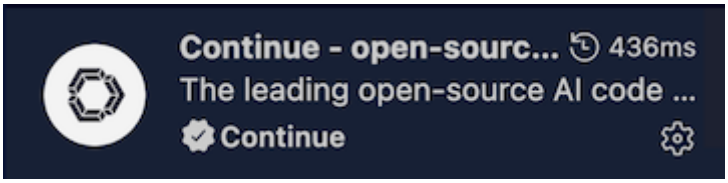
1. In VS Code das Kilo Code AI Agent Plugin installieren. Achtet darauf dass ihr kein falsches erwischt. Aktuell sieht es etwa so aus:


2. Wählt "Use your own API key"
3. API Provider: OpenAI compatible
4. Base URL: `https://ki.ping.de:8000`
5. API Key: **siehe_oben** (nehmt das Bearer Token)
6. Model Name: `qwen38-27b`

Fertig!

Continue Plugin

Installiert das Continue Plugin in VS Code.



Nehmt folgende Config Datei (ungeprüft). Tragt dort den API Key ein in der letzten Zeile. (Doku)

```
name: My Config
version: 0.0.1
schema: v1

models:
  - name: qwen38-ping
    provider: openai
    model: qwen38-27b
    apiBase: https://ki.ping.de:8000/v1
    apiKey: siehe_oben
```

Fertig!

OpenCode

Für OpenCode müsst ihr eine `opencode.json` Datei anlegen z.B. im Verzeichnis `~/.config/opencode/`

Tragt das/die Modelle in die opencode.json Datei ein, hier die Variante bei der Ihr die Menge an Reasoning (denken) auch noch steuern könnt:

```
{
  "$schema": "https://opencode.ai/config.json",
  "provider": {
      "ki.ping.de": {
      "npm": "@ai-sdk/openai-compatible",
      "options": {
        "baseURL": "https://ki.ping.de:8000/v1",
        "apiKey": "HIER_API_KEY_EINFUEGEN"
      },
      "models": {
        "qwen38-27b": {
          "name": "Qwen 3.8 (ping)",
          "limit": {
            "context": 135000,
            "output": 32000
```

```
},
"compatibility": {
  "reasoningField": "reasoning_content"
},
"modalities": {
  "input": ["image", "text"], "output": ["text"]
},
"body": {
  "reasoning_effort": "medium",
  "preserve_thinking": true
},
"variants": {
  "xhigh": {
    "name": "Max Reasoning (xhigh)",
    "body": {
      "reasoning_effort": "xhigh",
      "preserve_thinking": true
    }
  },
  "med": {
    "name": "Balanced (medium)",
    "body": {
      "reasoning_effort": "medium",
      "preserve_thinking": true
    }
  },
  "low": {
    "name": "Fast (low)",
    "body": {
      "reasoning_effort": "low",
      "preserve_thinking": true
    }
  },
  "off": {
    "name": "Thinking Disabled",
    "body": {
      "chat_template_kwargs": {
        "enable_thinking": false
      },
      "preserve_thinking": false
```

```
            }
          }
        }
      }
    }
  }
}
```

Wenn ihr das nächste Mal OpenCode startet und /connect tippt könnt ihr nach "PING" suchen und solltet den Eintrag "Qwen 3.8 (ping)" angezeigt bekommen. Falls euch dann das OpenCode nach dem API Key fragt, gebt das Bearer Token (s.o.) ein.

Inferenz Benchmarks

2026-09-21 vLLM mit Qwen 3.8 27b INT8

(config in /opt/vllm-qwen38)

Benchmark-Aufruf auf milla:

```
uvx llama-benchy --base-url $OPENAI_API_URL --api-key $OPENAI_API_KEY --pp 32000 --no-cache --concurrency 4 --format md --model Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8 --served-model-name qwen38-27b  --runs 2
```

model	test	t/s (total)	t/s (req)	peak t/s	peak t/s (req)	ttfr (ms)	est_ppt (ms)	e2e_ttft (ms)
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp32000 (c4)	1927.26 ± 3.08	1006.52 ± 564.21			41470.01 ± 18590.88	41467.98 ± 18590.88	41470.01 ± 18590.88
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c4)	2.46 ± 0.00	20.87 ± 30.50	32.00 ± 0.00	23.31 ± 30.47			

llama-benchy (0.4.0) date: 2026-09-21 19:36:48 | latency mode: api

2026-08-30 vLLM mit Qwen 3.8 27b INT8

Benchmark-Aufruf auf milla:

```
uvx llama-benchy --base-url $OPENAI_API_URL --api-key $OPENAI_API_KEY --pp 2048 32000 64000 128000 \
--no-cache --concurrency 1 2 4 8 --format md --model Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8 \
--served-model-name qwen38-27b
```

model	test	t/s (total)	t/s (req)	peak t/s	peak t/s (req)	ttfr (ms)	est_ppt (ms)	e2e_ttft (ms)

Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp2048 (c1)	2195.92 ± 88.95	2195.92 ± 88.95			936.52 ± 37.23	934.75 ± 37.23	936.52 ± 37.23
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c1)	66.82 ± 4.55	66.82 ± 4.55	68.97 ± 4.70	68.97 ± 4.70			
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp2048 (c2)	2175.03 ± 13.52	1620.96 ± 532.74			1418.79 ± 465.48	1417.02 ± 465.48	1418.79 ± 465.48
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c2)	40.12 ± 0.82	36.44 ± 14.22	61.00 ± 0.82	40.91 ± 11.42			
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp2048 (c4)	2204.24 ± 14.81	1153.90 ± 648.27			2319.89 ± 1041.14	2318.12 ± 1041.14	2319.89 ± 1041.14
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c4)	36.73 ± 0.51	25.62 ± 19.11	104.67 ± 1.25	33.27 ± 15.75			
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp2048 (c8)	2176.39 ± 2.29	738.87 ± 592.25			4226.90 ± 2154.08	4225.13 ± 2154.08	4226.90 ± 2154.08
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c8)	34.38 ± 0.40	15.08 ± 14.38	132.00 ± 7.87	20.90 ± 13.77			

Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp32000 (c1)	1990.71 ± 1.41	1990.71 ± 1.41			16076.91 ± 11.35	16075.14 ± 11.35	16076.91 ± 11.35
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c1)	74.29 ± 9.37	74.29 ± 9.37	76.68 ± 9.67	76.68 ± 9.67			
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp32000 (c2)	1977.61 ± 4.66	1487.13 ± 498.28			24241.74 ± 8121.94	24239.97 ± 8121.94	24241.74 ± 8121.94
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c2)	3.71 ± 0.02	36.44 ± 34.03	32.00 ± 0.00	38.99 ± 33.78			
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp32000 (c4)	1956.14 ± 1.19	1018.54 ± 568.61			40902.14 ± 18286.44	40900.37 ± 18286.44	40902.14 ± 18286.44
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c4)	2.47 ± 0.04	17.15 ± 23.79	32.00 ± 0.00	20.06 ± 24.93			
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp32000 (c8)	1731.80 ± 307.50	521.02 ± 460.47			94488.30 ± 48079.06	94486.53 ± 48079.06	94488.36 ± 48079.10
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c8)	2.05 ± 0.06	7.90 ± 15.94	23.33 ± 12.26	10.79 ± 17.22			

Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp64000 (c1)	1489.29 ± 52.36	1489.29 ± 52.36			43028.05 ± 1484.93	43026.28 ± 1484.93	43029.51 ± 1482.93
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c1)	49.42 ± 4.98	49.42 ± 4.98	63.96 ± 11.91	63.96 ± 11.91			
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp64000 (c2)	1367.74 ± 336.29	966.10 ± 442.01			81524.79 ± 35740.88	81523.02 ± 35740.88	81524.79 ± 35740.88
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c2)	1.49 ± 0.02	27.45 ± 24.44	32.00 ± 0.00	32.43 ± 28.16			
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp64000 (c4)	1620.36 ± 21.02	814.26 ± 432.98			100069.83 ± 43246.41	100068.06 ± 43246.41	100069.83 ± 43246.41
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c4)	0.99 ± 0.03	18.92 ± 27.29	32.00 ± 0.00	23.43 ± 28.51			
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp64000 (c8)	1691.43 ± 2.78	567.24 ± 451.32			170771.27 ± 86373.88	170769.50 ± 86373.88	170771.27 ± 86373.88
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c8)	0.93 ± 0.00	9.75 ± 17.10	32.00 ± 0.00	13.02 ± 19.26			

Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp128000 (c1)	1333.24 ± 16.29	1333.24 ± 16.29			96023.90 ± 1182.87	96022.12 ± 1182.87	96023.90 ± 1182.87
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c1)	62.31 ± 9.03	62.31 ± 9.03	110.27 ± 47.76	110.27 ± 47.76			
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp128000 (c2)	1361.46 ± 3.35	1020.37 ± 339.68			141079.37 ± 46959.48	141077.60 ± 46959.48	141079.37 ± 46959.48
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c2)	0.59 ± 0.01	63.79 ± 8.14	32.00 ± 0.00	73.59 ± 7.25			
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp128000 (c4)	1359.38 ± 1.06	707.95 ± 395.29			235394.83 ± 105268.48	235393.06 ± 105268.48	235394.83 ± 105268.48
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c4)	0.39 ± 0.00	65.37 ± 11.10	32.00 ± 0.00	76.58 ± 13.26			
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	pp128000 (c8)	1358.85 ± 1.57	462.11 ± 373.76			423654.57 ± 215887.54	423652.80 ± 215887.54	423654.57 ± 215887.54
Freaksterz/Qwen3.8-27B-SmoothQuant-W8A8-INT8	tg32 (c8)	0.33 ± 0.00	58.32 ± 14.31	32.00 ± 0.00	68.63 ± 14.86			

llama-benchy (0.4.0) date: 2026-08-30 15:13:03 | latency mode: api

2026-04-18 vLLM mit cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit

vLLM optionen:

```
--model cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit
--tensor-parallel-size 2
--max-model-len 65536
--gpu-memory-utilization 0.85
--enable-prefix-caching
--reasoning-parser qwen3
--enable-auto-tool-choice
--tool-call-parser qwen3_coder
--max-num-seqs 32
--speculative-config '{"method":"qwen3_next_mtp","num_speculative_tokens":2}'
```

Benchmark mit `uvx llama-benchy --base-url http://cogito.buero.ping.de:8000/v1 --depth 2000 32768 63000`

model	test	t/s	peak t/s	ttfr (ms)	est_ppt (ms)	e2e_ttft (ms)
cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit	pp2048 @ d2000	5463.38 ± 111.87		748.82 ± 14.93	741.48 ± 14.93	748.93 ± 14.93
cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit	tg32 @ d2000	103.13 ± 22.06	112.49 ± 24.41			
cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit	pp2048 @ d32768	5178.25 ± 25.55		6731.33 ± 33.06	6724.00 ± 33.06	6731.41 ± 33.05
cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit	tg32 @ d32768	25.65 ± 1.43	27.93 ± 1.52			
cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit	pp2048 @ d63000	4534.72 ± 42.10		14353.15 ± 133.93	14345.82 ± 133.93	14353.26 ± 133.94
cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit	tg32 @ d63000	12.85 ± 3.50	14.45 ± 3.21			

Plan: P2P einschalten, da geht noch mehr...

Zugriff auf vLLM via OpenAI-kompatibler API

Für viele Tools benötigt ihr lediglich den Zugriff auf den OpenAI-kompatiblen API Endpunkt, den unser vLLM bereitstellt.

Hier findet ihr die nötigen Daten:

Base URL: `https://ki.ping.de:8000/v1`

API-Key: siehe API Token (für Mitglieder)