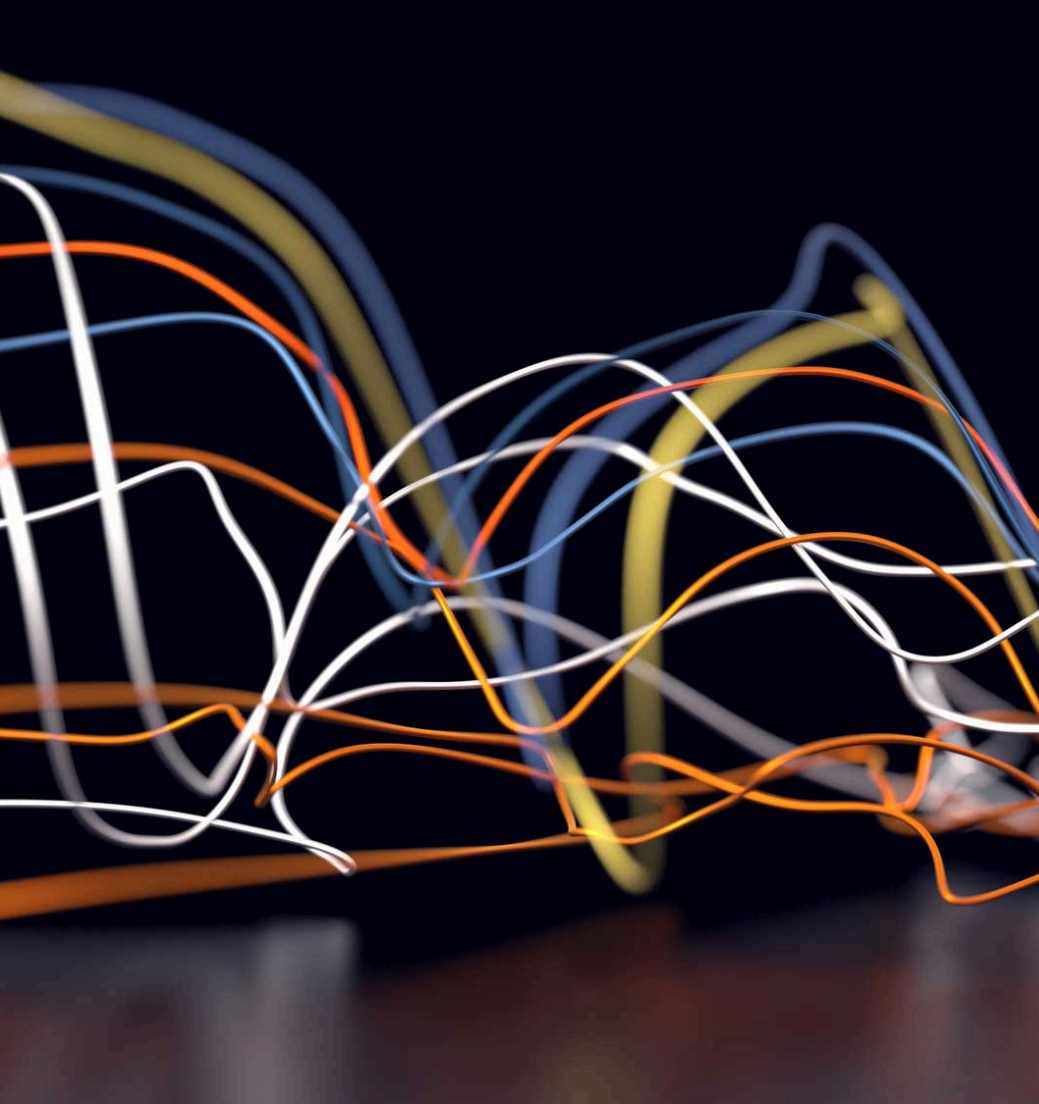




INNOQ TECHNOLOGY NIGHT // 17.09.2026

Souveräne KI betreiben

Jenseits von Claude und ChatGPT

INNOQ



NICOLAS INDEN
INNOQ



Warum nicht einfach Claude & GPT?

Vier Accounts. Zwei Anbieter.



claude.ai

Subscription



Claude Platform

API • Pay-per-Token



chatgpt.com

Subscription



OpenAI Platform

API • Pay-per-Token

📌 **Proprietär:** Selbes Modell nur beim Erstanbieter – Preis festgeschrieben, **kein Preiswettbewerb**



Open Weights: Selbes Modell bei vielen Anbietern – **Wettbewerb um denselben Token**

Was passiert mit meinen Prompts?

 **US-Frontier-Anbieter:** EU-only-Verarbeitung nur auf Anfrage • Prompts bis zu **30 Tage** gespeichert (Abuse Detection) [OpenAI ↗](#) • [Anthropic ↗](#)

 **Open Weights aus der EU:** EU Data Residency • **Zero Data Retention** • kein Training – bei EU-Providern Standard

 **Nicht das Modell, sondern der Provider zählt:** Open Weights aus US-/China-Providern bieten dieselben Garantien **nicht** automatisch

 **Herkunft ≠ Betrieb:** Ein in China trainiertes Modell **kann** in der EU datenschutzkonform laufen – die Weights „telefonieren“ nicht nach Hause

Was, wenn das Modell weg ist?



Proprietär

Beschränkung trifft *alle* Kanäle – auch Bedrock & Co.



Open Weights

Weights einmal veröffentlicht = von jedem Provider
betreibbar

Drei Warnsignale aus der Praxis



Kalkulierbarkeit • Anthropic (Juli 2025): Neue **wöchentliche Limits** mitten im laufenden Abo, gültig ab 28.08. – trifft laut Anthropic **< 5 %** der Abos (Power-User). techcrunch.com ↗



Datenschutz • OpenAI (Mai 2025): US-Gericht verpflichtet OpenAI, **alle ChatGPT-Output-Logs aufzubewahren** – auch zur Löschung vorgesehene Consumer-Chats. [Court Order](#) ↗



Resilienz • Anthropic (Juni 2025): **First-Party-Kapazität** für einen Coding-Tool-Anbieter mit **< 5 Tagen Vorlauf** gekappt – Grund: Übernahmegerüchte. techcrunch.com ↗



**Unsere Antwort:
Eine unabhängige Lösung,
basierend auf freier Software**

 WIE ALLES BEGANN

Eine Woche H200



Eine Woche Zeit

eine H200, abgekoppelt – nur zum
Ausprobieren



Was geht wirklich?

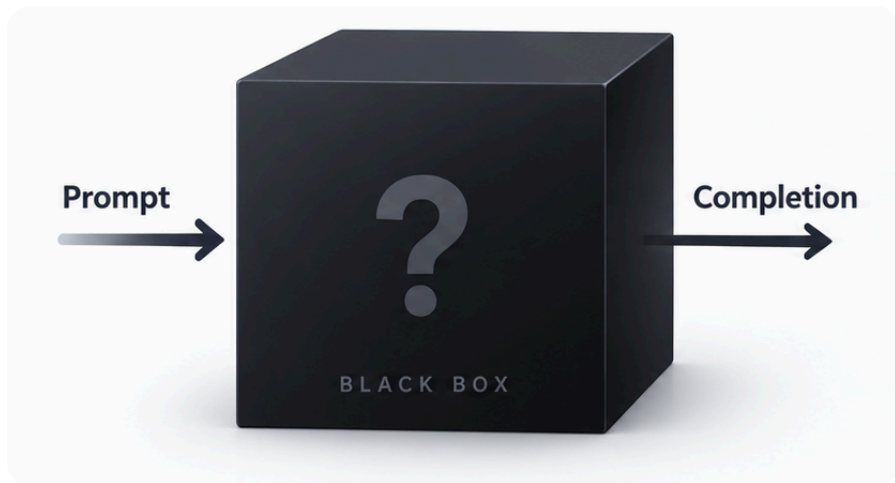
Open-Weights-Modelle für Chat,
Coding & Integration getestet



Heute produktiv

der Stack läuft im Alltag – Inferenz
per API von diversen Anbietern

Blackbox: Proprietäre LLMs

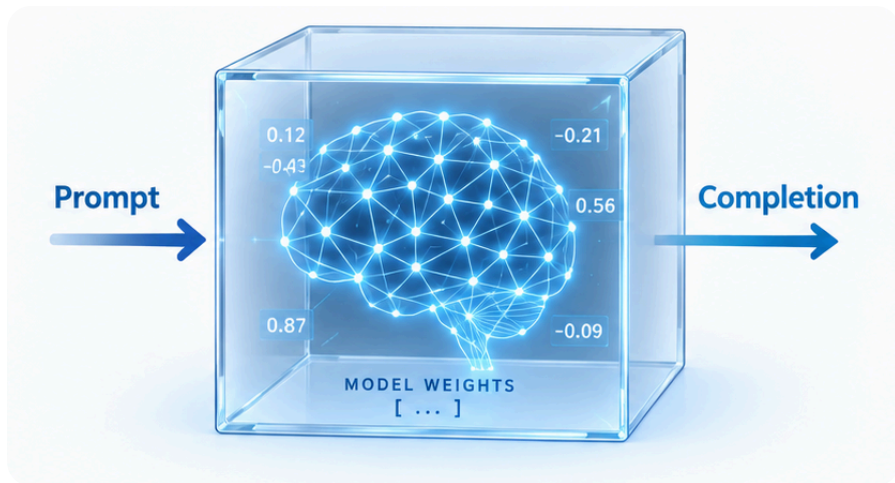


Proprietäres LLM

Geschlossen. Undurchsichtig. Nicht einsehbar.

- Modell, Weights und Trainingsdaten **nicht einsehbar**
- Funktionsweise und Entscheidungen undurchsichtig
- Kontrolle liegt beim Anbieter

Open Weights: Der Glaskasten



Open-Weights LLM

Offen. Nachvollziehbar. Frei nutzbar.

- Modellarchitektur, Weights und oft auch Trainingsdaten **einsehbar**
- Funktionsweise nachvollziehbar und überprüfbar
- Frei nutzbar, anpassbar und selbst betreibbar

MODELLQUALITÄT · 🏁 STAND HEUTE

Der Gap ist fast zu



Proprietäre Spitze

Text Arena · Overall

Claude Fable 5 · **1506**

Claude Opus 4.6 · **1505**

Claude Opus 4.7 · **1502**



Offene Spitze

Text Arena · Overall

Kimi K3 Max · **~1490**

GLM-5.3 Max · **~1485**

DeepSeek V4 Pro · **~1460**

Text Arena · Overall: Gap nur noch ~15–20 Elo-Punkte

WebDev Arena · #1 gpt-6-astra-max **1800** – bestes offenes:
qwen3.8-max **1681** (Gap ~120)

Agent Arena · noch proprietär geführt; Kimi K3 Max auf **#8** –
offen holt auf

Quelle: arena.ai/leaderboard ↗ · Stand abgerufen aus dem Live-Leaderboard · blindes A/B-Voting

DER STACK

Ein Setup, drei Szenarien



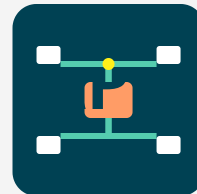
Chat & Fachexperten

OpenWebUI · RAG · MCP



Coding Agents

LiteLLM · Proxy · Budgets

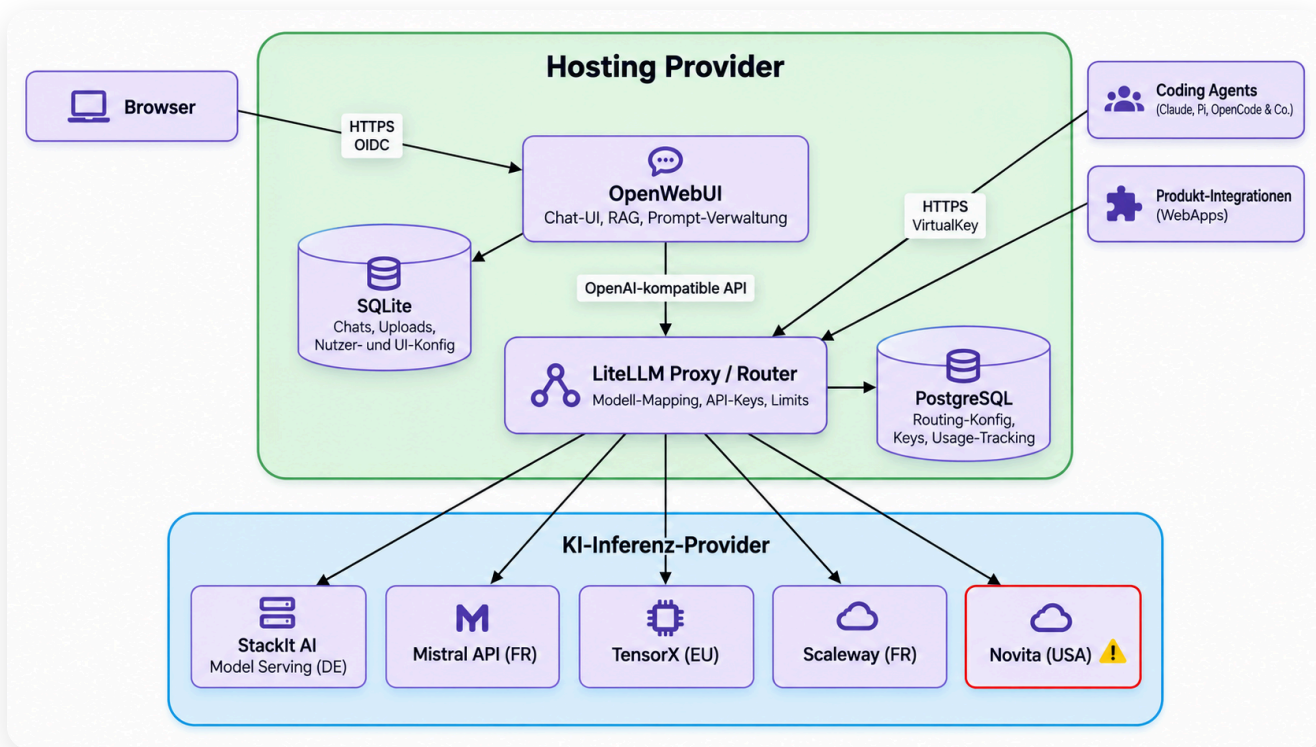


Produktintegration

LiteLLM · einheitliche API



Der Aufbau im Überblick



Was hosten wir selbst?

Selbst betreiben

Die **Zugänge** – einfach, DSGVO-nah, volle Kontrolle



OpenWebUI
LiteLLM

Extern beziehen

Die **Inferenz** – von EU-Providern per API



TensorX · Scaleway
StackIt · IONOS

 **Zugänge selbst** ·  **Inferenz extern = der Sweetspot**



Hardware-Frage

GLM-5.3 (744B): **8x H200** · Kimi
K2.7 (1T)



Verträge

AVVs mit Hosting- *und* Inferenz-
Providern



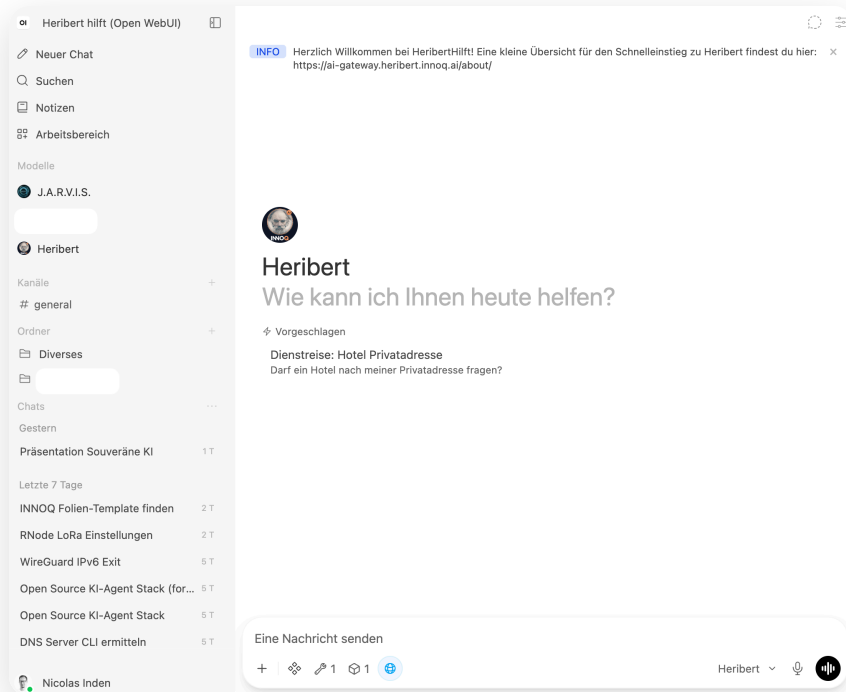
Betrieb

Updates · Preise · Keys & Budgets
pflegen

OPENWEBUI



Chat wie gewohnt – Wissen inklusive



Notizen

Text & Sprache
durchsuchbar



Wissen

Vektor-DB + RAG



MCP

Jira, Confluence & Co.



Mobile

z. B. Conduit (iOS)

BEISPIEL



Der „StackIt Terraform Expert“

● LIVE-DEMO

1



Doku-Repo klonen

terraform-provider-
stackit/docs



2



**Als Wissen
hochladen**

Arbeitsbereich → Wissen



3



**Modell + Prompt
verknüpfen**

Basis + Wissen + System-
Prompt



4



Fragen stellen

RAG liefert die Quellen
mit

LITELLM



Ein Proxy für alles



Kosten-Tracking

pro Key · Modell · Team ·
Tag



Anbieterwechsel

transparent für Clients



Budgets

VirtualKeys pro Projekt



Eine BaseURL

für alle Modelle und alle
Clients



Dashboards

Top Public Model Names

5

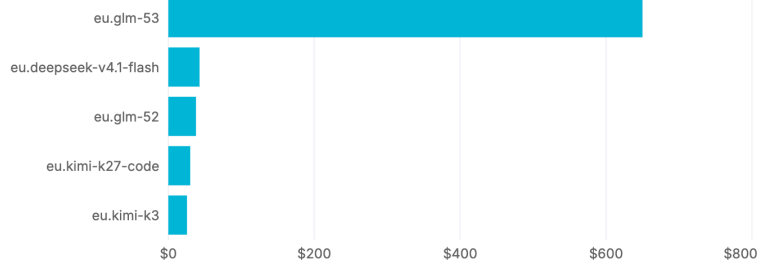
10

25

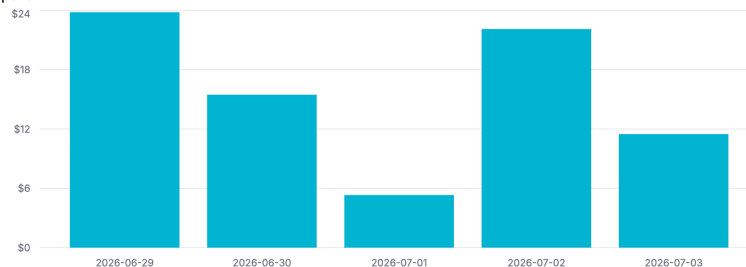
50

Public Model Name

Litellm Model Name



Daily Spend



Auswertbar pro **VirtualKey** · **Modell** · **Team** · **Tag**

LITELLM • REQUEST LOGS



Jeder Request nachvollziehbar

Request Logs													
Auto-refreshing every 15 seconds													Stop
Search logs by ID...													
Last 24 Hours Live Tail Hide Health Checks Reset Filters Filters													
Time	Type	Status	Session ID	Request ID	Cost	Duration (s)	TTFB (s)	Team Name	Key Hash	Key Alias	Model	Tokens	Internal User
Sep 17, 09:46:46	LLM	Success	5b64d5c1-8b...	95d584b684e4...	\$0.021103	2.47	2.21		768f43a473364...		deepseek/deep...	160801 (160709+92)	
Sep 17, 09:46:36	LLM	Success	7995c818-51...	ea21f36126a2...	\$0.022278	3.33	2.18		768f43a473364...		deepseek/deep...	160658 (160677+581)	
Sep 17, 09:46:29	LLM	Success	16536bc3-84...	df047056f8374...	\$0.022937	3.65	2.28		768f43a473364...		deepseek/deep...	159992 (159293+699)	
Sep 17, 09:46:25	LLM	Success	8a986256-4d...	715c821687314...	\$0.021555	3.68	3.64		768f43a473364...		deepseek/deep...	159234 (159043+191)	
Sep 17, 09:46:21	LLM	Success	a46a4758-58...	99e7f6de75564...	\$0.021995	2.80	2.34		768f43a473364...		deepseek/deep...	159001 (158727+274)	
Sep 17, 09:46:18	LLM	Success	3c9f6c84-85...	7c828358fa54...	\$0.021371	2.47	2.13		768f43a473364...		deepseek/deep...	158685 (158498+187)	
Sep 17, 09:46:15	LLM	Success	75fa1779-8a...	e7a3c487a4a2...	\$0.020899	2.36	2.19		768f43a473364...		deepseek/deep...	157551 (157459+92)	
Sep 17, 09:45:16	LLM	Success	d62b3b1d-f8...	81a3827ba1ae4...	\$0.008724	2.55	0.49		ee85813db58c...		z-align-5.3	14726 (14187+539)	
Sep 17, 09:45:15	LLM	Success	b22bbbc4-58...	328818b16ae84...	\$0.013614	0.86	0.66		ee85813db58c...		z-align-5.3	14174 (14087+777)	
Sep 17, 09:45:11	LLM	Success	337639da-45...	de14528896674...	\$0.022174	2.53	2.26		768f43a473364...		deepseek/deep...	157410 (157298+112)	
Sep 17, 09:45:11	LLM	Success	c019d876-f4...	ae9a48c5ec744...	\$0.008452	3.16	0.76		ee85813db58c...		z-align-5.3	13836 (13470+366)	
Sep 17, 09:45:08	LLM	Success	ba69289a-24...	33986aa48834...	\$0.008944	2.66	1.01		ee85813db58c...		z-align-5.3	12835 (12476+359)	
Sep 17, 09:45:08	LLM	Success	816d4752-2b...	ad04580813904...	\$0.022670	2.76	2.27		768f43a473364...		deepseek/deep...	157215 (156884+231)	
Sep 17, 09:45:06	LLM	Success	8578dd17-1d...	af262780952b4...	\$0.006553	1.37	1.14		ee85813db58c...		z-align-5.3	11091 (11010+81)	
Sep 17, 09:45:05	LLM	Success	c5f66ba4-88...	5ef7335477864...	\$0.004962	1.10	0.66		ee85813db58c...		z-align-5.3	10025 (9905+120)	
Sep 17, 09:45:03	LLM	Success	f16ec678-62...	17a376cfe884...	\$0.006988	1.84	0.65		ee85813db58c...		z-align-5.3	9893 (9823+270)	
Sep 17, 09:45:02	LLM	Success	19982d64-ea...	4775463873ba4...	\$0.015380	0.92	0.82		ee85813db58c...		z-align-5.3	8718 (8673+45)	
Sep 17, 09:45:01	LLM	Success	e8273a58-b7...	6f99c66c8c834...	\$0.004402	0.82	0.59		ee85813db58c...		z-align-5.3	8511 (8424+87)	
Sep 17, 09:44:59	LLM	Success	278e6d6e-f1...	6d84a51381824...	\$0.009840	1.78	0.75		ee85813db58c...		z-align-5.3	8243 (8020+223)	
Sep 17, 09:44:57	LLM	Success	982c24b3-d7...	c885f9e0eb04...	\$0.007320	1.13	0.84		ee85813db58c...		z-align-5.3	4024 (3923+101)	
Sep 17, 09:44:56	LLM	Success	4d276673-1a...	78082c2cc274...	\$0.001913	0.73	0.33		ee85813db58c...		z-align-5.3	2430 (2305+125)	
Sep 17, 09:44:56	LLM	Success	6b19d0b8-8c...	c809e981c78b4...	\$0.003881	0.73	0.42		ee85813db58c...		z-align-5.3	2106 (2035+71)	
Rows per page 25													
Showing 1 - 25 of 1787 Page 1 of 72													

Modell • Token • Kosten • Latenz • Status – pro Aufruf



Ohne Prompt- und Completion-Inhalt – wir loggen keine Inhalte!



**Anbieter wechseln,
ohne dass jemand es merkt**



Woher die Modelle?



TensorX

Irland



Scaleway

Frankreich



StackIt / IONOS

zuletzt kleinere/ältere Modelle

UNSER MODELL-KATALOG



Der aktuelle Stand

🇫🇷 Mistral (Frankreich) – AVV ZDR 🇪🇺 Tensorix.ai (EU) – AVV ZDR 🇫🇷 Scaleway (Frankreich) – KEIN AVV 🇺🇸 Novita (USA) – KEIN AVV 🇺🇸 Anthropic (USA) – KEIN AVV 🇩🇪 Stackit (Deutschland) – KEIN AVV 🇨🇳 Moonshot (China) – KEIN AVV						
PROVIDER:	AVV-STATUS:	SUCHE:				
Alle	Alle	Modellname...				
MODELL	PROVIDER	HOSTING	AVV	KOSTEN \$/1M	ARENA.AI SCORE	MODELL-ID
Kimi K3 NEU	TensorX	🇪🇺	JA	± \$3.00 ↑ \$15.00 ↓ \$0.75	1674 #5 x11	eu.kimi-k3
Qwen 3.8 Max NEU	TensorX	🇪🇺	JA	± \$2.50 ↑ \$6.00 ↓ \$2.50	1671 #6 x12	eu.qwen3.8-max
Qwen 3.8 Flash Next NEU	TensorX	🇪🇺	JA	± \$0.20 ↑ \$0.50 ↓ \$0.05	1635 #9 x13	eu.qwen3.8-flash-next
DeepSeek V4.1 Flash NEU	TensorX	🇪🇺	JA	± \$0.50 ↑ \$1.50 ↓ \$0.13	1614 #10 x17	eu.deepseek-v4.1-flash
GLM 5.3 NEU	TensorX	🇪🇺	JA	± \$1.75 ↑ \$4.50 ↓ \$0.44	1614 #15 x11	eu.glm-53
GLM 5.3 Flash NEU	TensorX	🇪🇺	JA	± \$0.20 ↑ \$0.50 ↓ \$0.05	1607 #17 x12	eu.glm-53-flash
GLM 5.2	TensorX	🇪🇺	JA	± \$1.50 ↑ \$4.50 ↓ \$0.38	1592 #18 x7	eu.glm-52
DeepSeek V4 Flash	TensorX	🇪🇺	JA	± \$0.25 ↑ \$0.30 ↓ \$0.06	1580 #22 x10	eu.deepseek-v4-flash
Qwen 3.7 Max	Novita	🇺🇸 USA	KEIN	± \$1.25 ↑ \$3.75 ↓ \$0.25	1517 #39 x8	us.qwen-37-max
MiniMax M3	TensorX	🇪🇺	JA	± \$0.40 ↑ \$2.00 ↓ \$0.10	1487 #48 x6	eu.minimax-m3
Kimi K2.7 Code	TensorX	🇪🇺	JA	± \$1.25 ↑ \$4.50 ↓ \$0.31	1473 #51 x10	eu.kimi-k27-code
DeepSeek V4 Pro	TensorX	🇪🇺	JA	± \$1.75 ↑ \$3.50 ↓ \$0.44	1446 #59 x8	eu.deepseek-v4-pro
Qwen 3.5 397B A17B	Scaleway	🇫🇷 Frankreich	KEIN	± \$0.60 ↑ \$3.60 ↓ \$0.60	1399 #79 x5	eu.qwen35-397b-a17b
Gemma 4 31B IT	Stackit	🇩🇪 Deutschland	JA	± \$0.45 ↑ \$0.65 ↓ \$0.45	1364 #90 x7	de.gemma-4-31b-1t
Gemma 4 26B A4B IT	Scaleway	🇫🇷 Frankreich	KEIN	± \$0.25 ↑ \$0.50 ↓ \$0.25	1361 #91 x17	eu.gemma-4-26b-a4b-1t
Mistral Medium 3.5	Mistral	🇫🇷 Frankreich	JA	± \$1.50 ↑ \$7.50 ↓ \$1.50	1265 #112 x15	eu.mistral-medium-3.5
Qwen 3.6 35B A3B	Scaleway	🇫🇷 Frankreich	KEIN	± \$0.25 ↑ \$1.50 ↓ \$0.25	1250 #115 x18	eu.qwen36-35b-a3b

📊 Arena-Scores (Code/WebDev) Stand Sep 11, 2026 - Quelle: arena.ai

Modell · Provider · Hosting · **AVV-Status** · Kosten \$/1M · Arena.ai-Score · Modell-ID



Agents an LiteLLM anbinden

```
1 // ~/.claude/settings.json – Anthropic-Subscription durchs Gateway
2 {
3   "env": {
4     "ANTHROPIC_BASE_URL": "https://url.zu.litellm.tld",
5     "ANTHROPIC_CUSTOM_HEADERS": "x-litellm-api-key: Bearer ${LITELLM_VIRTUAL_KEY}"
6   }
7 }
```



Weitere Agents, gleiches Muster

 **Pi** • `.pi/agent/models.json`

```
{
  "providers": {
    "ai-gateway": {
      "api": "openai-completions",
      "baseUrl": "https://url.zu.litellm.tld/v1",
      "apiKey": "ai-gateway",
      "models": [
        { "id": "eu.mistral-medium-3.5",
          "contextWindow": 262144,
          "input": ["text", "image"] },
        { "id": "eu.deepseek-v4-pro",
          "contextWindow": 1048576,
          "input": ["text", "image"] }
      ]
    }
  }
}
```

 **OpenCode** •

`~/.config/opencode/opencode.json`

```
{
  "provider": {
    "ai-gateway": {
      "npm": "@ai-sdk/openai-compatible",
      "name": "AI-Gateway",
      "options": {
        "baseUrl":
          "https://url.zu.litellm.tld/v1"
      },
      "models": {
        "eu.mistral-medium-3.5": {},
        "eu.deepseek-v4-pro": {},
        "eu.deepseek-v4-flash": {}
      }
    }
  }
}
```

PRODUKTINTEGRATION



Eine API für alles



OpenAI-kompatibel

/v1/chat/completions



Anthropic-Messages-kompatibel

/v1/messages

```
curl -X POST https://url.zu.litellm.tld/v1/chat/completions \  
-H "Authorization: Bearer YOUR_LITELLM_VIRTUALKEY" \  
-H "Content-Type: application/json" \  
-d '{  
  "model": "eu.mistral-medium-3.5",  
  "messages": [{ "role": "user", "content": "Why is the sky blue?" }]  
'
```



Was wir gelernt haben



Betrieb ist kein Selbstläufer



AVVs mit allen datenverarbeitenden Parteien: Hosting-Provider *und* jede:r Inferenz-Provider



Laufender Betrieb: Modell-Updates · Preisänderungen der Provider · regelmäßige Prüfung von VirtualKeys & Budgets

„Der Mehraufwand hält sich in Grenzen – aber er ist nicht null.“



Fünf Erkenntnisse



Konsolidierung wirkt

Account- & Kostenchaos
ist über einen Proxy
lösbar



Experten sind schnell

Wissen anlegen →
Experten mit
Basismodell erstellen →
Prompt



LiteLLM kam später

Start nur mit
OpenWebUI – Key-
Management & Usage-
Tracking waren
unzureichend



Token: Agents » Chat

Budgets & Tracking
zuerst bei Coding Agents



Betrieb nicht unterschätzen

Updates, Preise, AVVs –
Mehraufwand ≠ 0



Wann lohnt sich der Aufwand?



Lohnt sich bei

- Compliance-/Datenschutzanforderungen
- mehreren Projekten / Teams
- Bedarf an Kostenkontrolle
- Wunsch nach Anbieterunabhängigkeit



Weniger bei

- Einzelperson ohne Compliance-Druck
- sporadischer Nutzung

Faustregel: Je mehr Teams, Daten und Budgets – desto eher rechnet sich der eigene Stack.



Wann Claude & GPT sinnvoll bleiben



Für viele Aufgaben nach wie vor erste Wahl – wir hören nicht auf, die Frontier auszuprobieren



Hybrid-Ansatz: beide Welten durch denselben Proxy nutzbar → keine Entweder-oder-Entscheidung

Nicht „statt Claude/GPT“, sondern „nicht alles auf eine Karte“.



Kontrolle zurückgewinnen



Zugänge

selbst gehostet, ein
Gateway



Modelle

Open Weights,
austauschbar



Kosten

transparent & limitiert



Daten

EU, ZDR, kein Training

„Offene Modelle und Tools sind längst keine Spielerei mehr – sondern eine echte, betreibbare Alternative und eine gute Ergänzung.“



Vertiefung: [INNOQ-Blogartikel „Beyond Claude & GPT“ ↗](#)

AUSBLICK



Was noch kommt



LiteLLM: Agenten und MCP-Server direkt im Gateway einbinden



OpenWebUI: Skills sammeln, teilen und wiederverwenden



Betrieb: Modell-Katalog wächst, EU-Provider holen auf – Lage regelmäßig prüfen

Vielen Dank!

Fragen, Ideen oder Lust auf Austausch? Meldet euch gerne.



Nicolas Inden

nicolas.inden@innoq.com

+49 151 400 41 579

www.innoq.com/de/staff/nicolas-inden



LinkedIn



Krischerstr. 100
40789 Monheim
+49 2173 33660

Ohlauer Str. 43
10999 Berlin

Ludwigstr. 180E
63067
Offenbach

Kreuzstr. 16
80331 München

Eifflerstraße 43
22769 Hamburg

Spichernstr. 44
50672 Köln

Gewerbestr. 11
6330 Cham +41
4174 30111

Hardturmstr.
253 8005 Zürich