

INNOQ TECHNOLOGY LUNCH // 04.11.2026

Souveräne KI betreiben

Jenseits von Claude und ChatGPT

INNOQ



NICOLAS INDEN
Senior Consultant
@INNOQ



Warum nicht einfach Claude & GPT?

PROBLEM 1 · 💸 KOSTEN

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**

PROBLEM 2 · 🔒 DATENSCHUTZ

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

PROBLEM 3 · ⛔ VERFÜGBARKEIT

Was, wenn das Modell weg ist?



Proprietär

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



Open Weights

Weights einmal veröffentlicht = von jedem Provider betreibbar

BELEGE · 🔗 ALLES SCHON PASSIERT

Drei Beispiele 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 Ansatz:
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

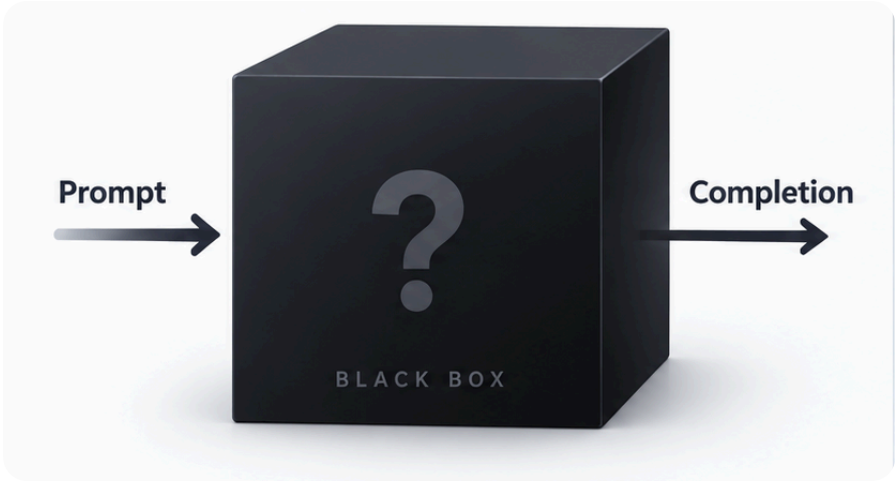


Heute produktiv

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

GRUNDLAGEN · 🧩

Blackbox: Proprietäre LLMs



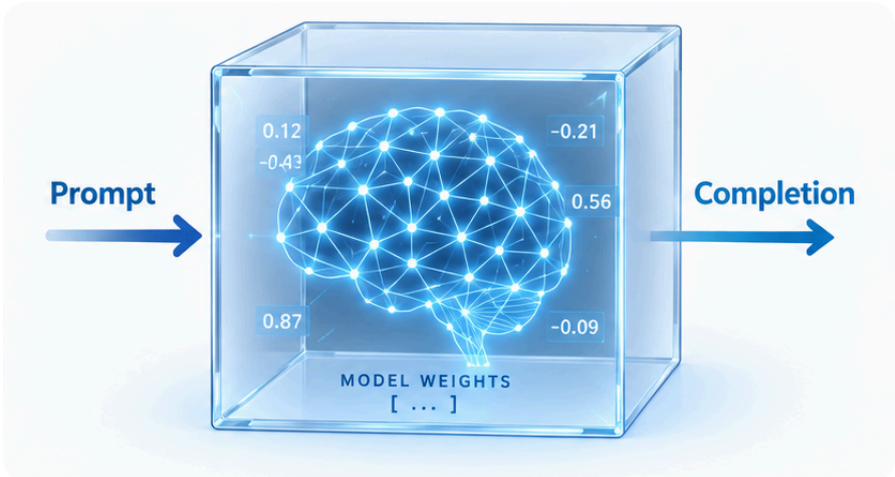
Proprietäres LLM

Geschlossen. Undurchsichtig. Nicht einsehbar.

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

GRUNDLAGEN · 🧩

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 · 🏁 MOMENTAUFNAHME · 27.10.2026

Die Sache mit dem Gap

💬 Text Arena

ELO · OVERALL

PROPRIETÄR — **1528** · #1
Gemini 4 Carbon

OFFEN — **1488** · #19
Kimi K3

GLM-5.3 1481 · MiMo-V2.6-Pro 1480

Gap ~40 Elo zur Spitze

🖥️ WebDev Arena

ELO · OVERALL

PROPRIETÄR — **1810** · #1
Claude Opus 5.5

OFFEN — **1695** · #9
K3.1 Pro

DeepSeek-V4.1-Flash 1665 · Kimi K3 1656

Gap ~115 Elo zur Spitze

🤖 Agent Arena

NET IMPROVEMENT · OVERALL

PROPRIETÄR — **+20,0 %** · #1
Claude Opus 5.5

OFFEN — **+5,1 %** · #13
K3.1 Pro

MiMo-V2.6 #16 (+2,9 %) · DeepSeek #22 (+2,3 %)

Gap ~15 pp zur Spitze

Drei Arenen, drei Abstände – und die Spitze bewegt sich.

WAS SICH SEIT DEM LETZTEN CHECK GEÄNDERT HAT

💬 Text

Führung: Claude → Gemini 4 (**1506 → 1528**)

Offen: Kimi K3 **1490 → 1488** · GLM **1485 → 1481**

🖥️ WebDev

#1: gpt-6-astra → Opus 5.5 (**1800 → 1810**)

Neu offen: K3.1 Pro **1695** (#9) · DeepSeek **1665**

🤖 Agent

Kimi K3 **#8 → #26** · K3.1 Pro **#13 (+5,1 %)**

#1 bleibt Claude Opus 5.5 **+20,0 %**

🆕 **Kimi K3.1 Pro:** Open Weights angekündigt für **04.11.** – in WebDev & Agent bereits bestes offenes Modell, im Text Arena noch ohne Score.

Quelle: arena.ai/leaderboard ↗ · Text/WebDev 27.10.2026 · Agent 03.11.2026 · blindes A/B-Voting

Momentaufnahme – Leaderboards ändern sich laufend.

DER STACK

Ein Setup, drei Szenarien



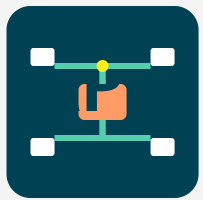
💬 Chat & Fachexperten

OpenWebUI · RAG · MCP



🤖 Coding Agents

LiteLLM · Proxy · Budgets



🔌 Produktintegration

LiteLLM · einheitliche API

ARCHITEKTURENTSCHEIDUNG

Was hosten wir selbst?

🏠 Selbst betreiben

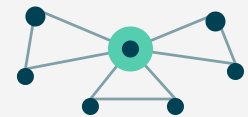
Die **Zugänge** – einfach, DSGVO-nah, volle Kontrolle



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): **8× H200** · Kimi K2.7 (1T)



Verträge

AVVs mit Hosting- *und* Inferenz-Providern

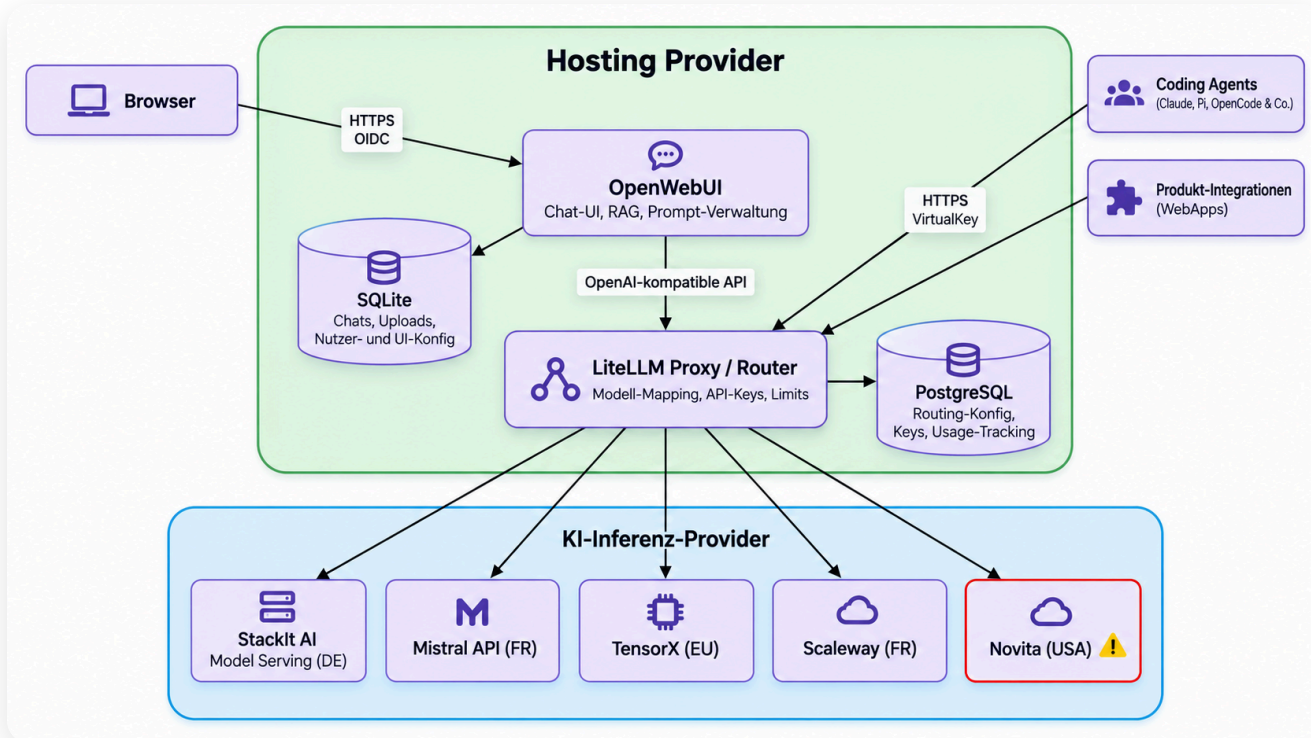


Betrieb

Updates · Preise · Keys & Budgets pflegen

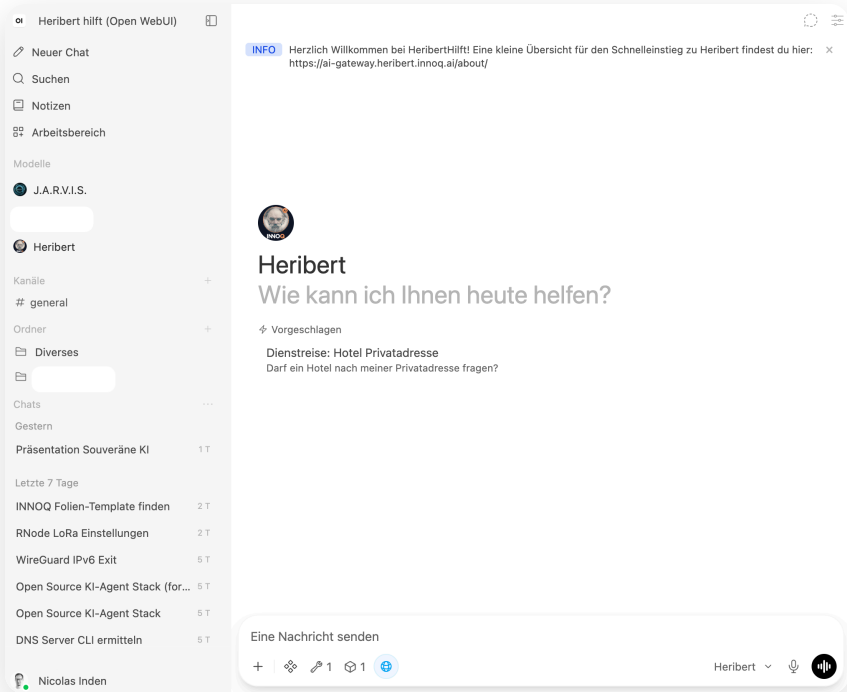
ARCHITEKTUR

Der Aufbau im Überblick



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

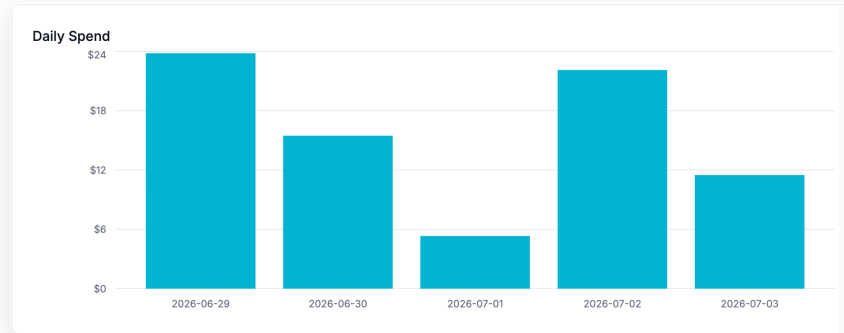
🎯 Eine BaseURL

für alle Modelle und alle Clients

LITELLM · KOSTEN IM GRIFF

📈 Dashboards

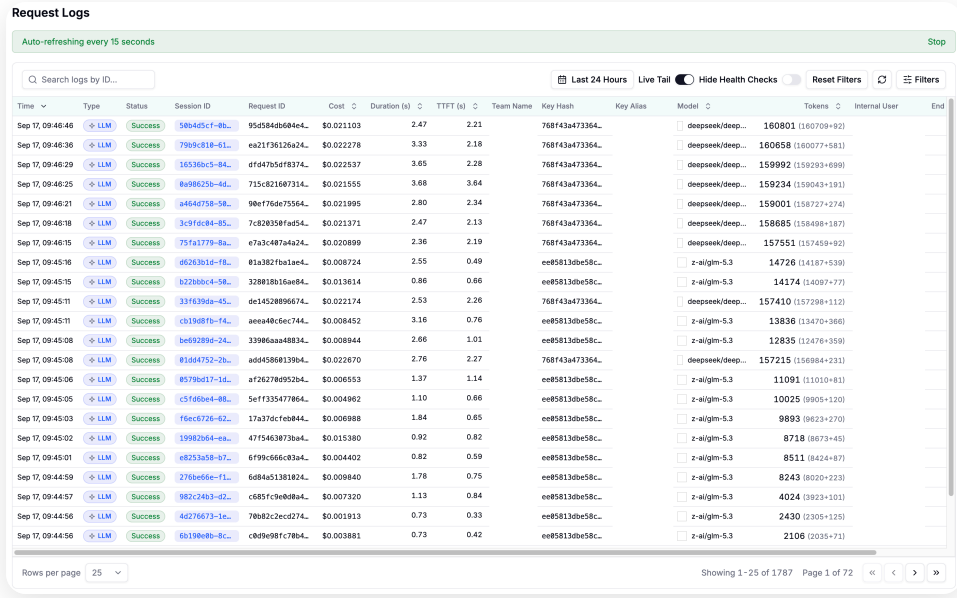
eu.glm-53 zai	$360.21	5.745 requests	10.60s avg response
eu.kimi-k3 moonshot	$286.29	1.337 requests	18.82s avg response
eu.deepseek-v4.1-flash deepseek	$259.98	23.445 requests	5.15s avg response
eu.glm-53-flash zai	$90.37	10.303 requests	16.34s avg response
eu.qwen3.8-flash-next dashscope	$20.54	4.087 requests	9.47s avg response
eu.mistral-large-4 mistral	$19.57	1.672 requests	13.71s avg response
world.clef	$17.28	13.789 requests	1.18s avg response



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

LITELLM · REQUEST LOGS

🔍 Jeder Request nachvollziehbar



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

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



Anbieter wechseln,
ohne dass jemand es merkt

EU-INFERENZ-PROVIDER

🗺️ Woher die Modelle?

🇮🇪

TensorX

Irland

🇫🇷

Scaleway

Frankreich

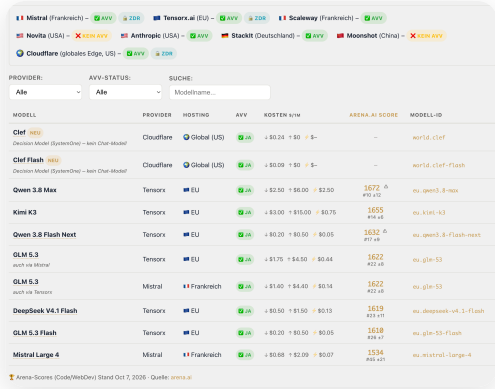
🇩🇪

StackIt / IONOS

zuletzt kleinere/ältere Modelle

UNSER MODELL-KATALOG

📋 Der aktuelle Stand



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

KONFIGURATIONSMUSTER · CLAUDE CODE

🤖 Agents an LiteLLM anbinden

```
// ~/.claude/settings.json — Anthropic-Subscription durchs Gateway
{
  "env": {
    "ANTHROPIC_BASE_URL": "https://url.zu.litellm.tld",
    "ANTHROPIC_CUSTOM_HEADERS": "x-litellm-api-key: Bearer ${LITELLM_VIRTUAL_KEY}"
  }
}
```



Was wir gelernt haben

VERANTWORTUNG & PFLEGE

 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."

ENTSCHEIDUNGSHILFE

🤷 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
- Coding mit Frontier-Modellen (dieser Punkt wackelt)

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

ABGRENZUNG · EHRLICH BLEIBEN

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".

FAZIT

🎯 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
- Smart-Routing testen

🧩 **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.

INNOQ
www.innoq.com



Nicolas Inden
nicolas.inden@innoq.com
+49 151 400 41 579
www.innoq.com/de/staff/nicolas-inden

LinkedIn

INNOQ

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