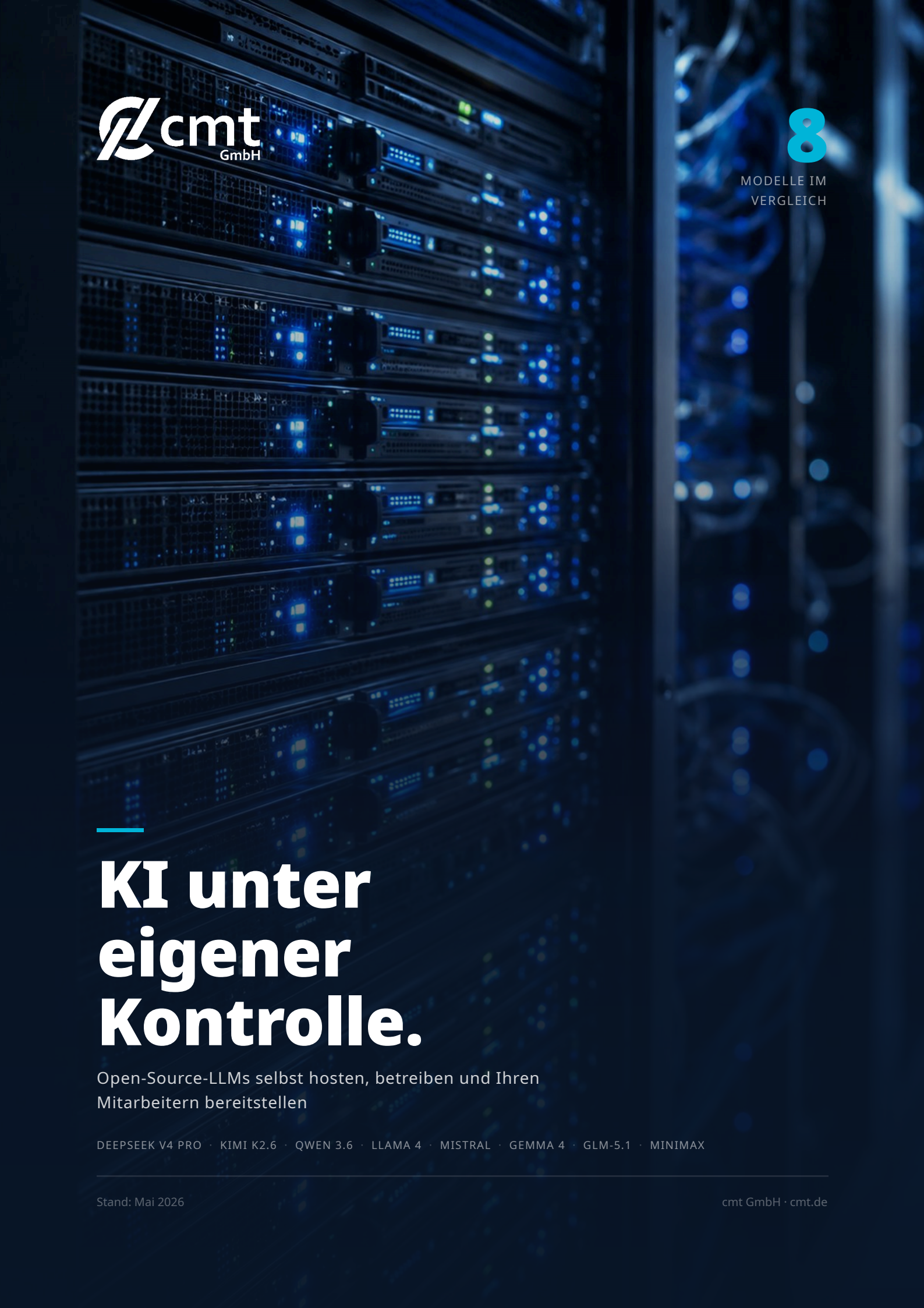


cmt GmbH

8
MODELLE IM VERGLEICH

# KI unter eigener Kontrolle.

Open-Source-LLMs selbst hosten, betreiben und Ihren Mitarbeitern bereitstellen

DEEPSEEK V4 PRO · KIMI K2.6 · QWEN 3.6 · LLAMA 4 · MISTRAL · GEMMA 4 · GLM-5.1 · MINIMAX

Stand: Mai 2026

cmt GmbH · cmt.de

# Inhalt



**8**
Open-Source-Modelle im Detail verglichen

**30**
Seiten Praxis-Wissen für IT-Entscheider

**120+**
passende cmt-Seminare zu KI & Infrastruktur

# Vorwort

**"Wir warten noch ab." Diesen Satz höre ich seit zwei Jahren - in Vorstandsetagen, in IT-Abteilungen, in Amtsstuben. Er klingt vernünftig. Er ist es nicht mehr.**

Ich verstehe die Vorsicht. Wer Verantwortung für IT-Infrastruktur, Personaldaten oder Bürgeranliegen trägt, darf nicht jedem Hype hinterherlaufen. Skepsis gegenüber KI ist begründet. Gescheiterte Digitalisierungsprojekte, unausgereifte Piloten und leere Versprechen haben berechtigtes Misstrauen hinterlassen.

Aber 2026 hat sich die technische Realität grundlegend verändert. Open-Source-Modelle wie DeepSeek, Kimi und Qwen liefern Ergebnisse auf dem Niveau der besten proprietären Systeme. Der entscheidende Unterschied: Sie laufen auf Ihrer Hardware, in Ihrem Rechenzentrum, unter Ihrer Kontrolle. Keine Daten verlassen das Haus. Keine Auftragsverarbeitung mit Drittanbietern. Keine Abhängigkeit von externen Cloud-Diensten.

Das ist besonders relevant für alle, die mit sensiblen Daten arbeiten - ob Personaldaten nach Art. 9 DSGVO, Sozialdaten oder vertrauliche Geschäftsinformationen. Self-Hosted KI macht Schluss mit dem Dilemma zwischen Innovation und Compliance.

Dabei geht es nicht um vollautomatisierte Entscheidungen. Es geht um intelligente Assistenz: Wissensmanagement, Dokumentenanalyse, Code-Reviews, Barrierefreiheit. Immer mit dem Menschen als letzter Instanz. Denn KI-Projekte scheitern selten an der Technik. Sie scheitern an fehlendem oder unzureichendem Change-Management, an unklaren Zielen und an schlechter Datenqualität.

Gleichzeitig hat sich der Markt für Open-Source-Modelle in einer Geschwindigkeit entwickelt, die selbst Experten überrascht. Modelle mit wenigen Milliarden aktiven Parametern lösen heute Aufgaben, für die vor einem Jahr noch große GPU-Cluster nötig waren. Mit Tools wie LM Studio oder Ollama ist ein Modell in wenigen Minuten einsatzbereit.

Dieses Whitepaper ist ein technisch fundierter Leitfaden. Wir vergleichen acht Modelle, rechnen Kosten durch, benennen Grenzen und zeigen einen konkreten Fahrplan vom ersten Test bis zum produktiven Rollout. Für Unternehmen und Behörden, die KI nicht verbieten, sondern professionell einführen wollen.

**Yves Hoppe**
Fachbereichsleiter KI & Open Source, cmt GmbH

# Warum jetzt?

## Die Ausgangslage in deutschen Unternehmen

**Ihre Mitarbeiter nutzen bereits ChatGPT. Die Frage ist nur, ob Sie es wissen.**

In nahezu jedem Unternehmen ist generative KI angekommen - oft ohne Wissen der IT-Abteilung. Mitarbeiter laden vertrauliche Dokumente in ChatGPT hoch, lassen Verträge zusammenfassen und nutzen KI-Assistenten für E-Mails und Code. Diese sogenannte Shadow AI ist eines der größten unkontrollierten Risiken der aktuellen IT-Landschaft.


### Die drei Risiken unkontrollierter KI-Nutzung

1. **Datenabfluss**
   Vertrauliche Informationen landen auf Servern außerhalb Ihrer Kontrolle. Einmal gesendet, lässt sich nichts mehr zurückholen. Trainingsdaten externer Anbieter sind intransparent.

2. **Compliance-Verstoß**
   DSGVO Art. 28 verlangt eine Auftragsverarbeitungsvereinbarung. Schrems II schränkt Drittlandtransfers ein. Ohne Kontrolle drohen Bußgelder und Reputationsschäden.

3. **Kontrollverlust**
   Keine Nachvollziehbarkeit, keine Qualitätssicherung, keine zentrale Steuerung. Die KI-Nutzung fragmentiert in Dutzende unkontrollierte Einzellösungen.

**Die Lösung**

Open-Source-Modelle auf eigener Infrastruktur bieten dieselbe Funktionalität wie ChatGPT und Co. - ohne Kontrollverlust, ohne Datenabfluss, ohne Drittanbieter-Abhängigkeit.

# Der Open-Source-Durchbruch

## 2026: Open Source hat gleichgezogen

**Noch 2024 lagen Open-Source-Modelle weit hinter proprietären Anbietern. Heute erreichen DeepSeek, GLM und Kimi bei Coding und Reasoning Frontier-Niveau - bei MIT-Lizenz.**

**80,6%**
SWE-bench Verified (DeepSeek V4 Pro)

**58,4**
SWE-Bench Pro (#1 GLM-5.1)

**88,4%**
GPQA Diamond (Qwen 3.6, Bestwert OS)

### Die Top-8 Open-Source-Modelle im Überblick

| Modell | Aktive Params | Lizenz | Stärke | Index |
|---|---|---|---|---|
| **DeepSeek V4 Pro** | 49B (MoE) | MIT | Coding, Reasoning | 52 |
| **GLM-5.1** | ~40B (MoE) | MIT | SWE-Bench Pro #1 | - |
| **Qwen 3.6** | 17B (MoE) | Apache 2.0 | GPQA, Mathematik | - |
| **Kimi K2.6** | 32B (MoE) | Mod. MIT | Coding, Mathe, #1 OW | 54 |
| **Llama 4 Maverick** | 17B (MoE) | Meta CL | Ökosystem, Speed | - |
| **Mistral Small 4** | 6,5B (MoE) | Apache 2.0 | EU, Speed | 28 |
| **Gemma 4** | 31B (Dense) | Apache 2.0 | Preis/Leistung | - |
| **MiniMax-M2.7** | 45,9B (MoE) | Apache 2.0 | 1M Kontext, Mathe | - |

### MIT und Apache 2.0: Kommerziell frei nutzbar

Die Mehrheit der Top-Modelle steht unter Lizenzen, die kommerzielle Nutzung ohne Einschränkungen erlauben. Kein Vendor-Lock-in, keine Lizenzgebühren, volle Kontrolle.

# DeepSeek V4 Pro

## Das leistungsstärkste Open-Source-Modell

**Hersteller** DeepSeek (China)
**Parameter** 1,6T gesamt / 49B aktiv (MoE)
**Geschw.** ~31 Tokens/s
**Lizenz** MIT
**Kontext** 1.000.000 Tokens
**Index** 52 (#3 gesamt)

### Benchmarks

| Benchmark | Score | Einordnung |
|---|---|---|
| MMLU-Pro | **87,5** | Frontier-Niveau |
| GPQA Diamond | **90,1** | Top 3 aller Modelle |
| SWE-bench Verified | **80,6** | #1 Open Source |
| LiveCodeBench | **93,5** | Bestwert |
| AIME 2026 | 85,0 | Starkes Reasoning |

### Stärken

- Frontier-Niveau bei Coding und Reasoning
- Günstigste Frontier-API ($1,74 / $3,48 pro 1M Tokens)
- 1 Million Tokens Kontext
- MIT-Lizenz: volle kommerzielle Freiheit

### Schwächen

- Sehr groß für Self-Hosting (Multi-GPU-Cluster erforderlich)
- MoE-Architektur braucht viel RAM

### Am besten für

Coding, Reasoning, komplexe Analyse - wenn maximale Leistung gefragt ist und die Infrastruktur vorhanden ist. Ideal als API-Backend oder auf dedizierten GPU-Clustern.

**cmt-Seminar:** DeepSeek: Einsatz und Self-Hosting (KKC_0221)

# GLM-5.1

## Spitzenreiter im SWE-Bench Pro

**Hersteller** Z.ai / Zhipu AI (China)
**Parameter** 754B gesamt / ~40B aktiv (MoE)
**Geschw.** ~51 Tokens/s

**Lizenz** MIT Open Weights
**Kontext** 200.000 Tokens
**Architektur** Mixture of Experts

### Benchmarks

| Benchmark | Score | Einordnung |
|---|---|---|
| GPQA Diamond | **86,2** | Stark |
| SWE-bench Verified | **77,8** | Top 3 Open Source |
| SWE-Bench Pro | **58,4** | #1 zum Release |
| LiveCodeBench | 78,0 | Stark |

### Stärken

- SWE-Bench-Pro-Führender zum Release
- MIT-Lizenz: vollständig offen
- Starkes Coding und technische Analyse
- Kompakter als DeepSeek bei guter Leistung

### Schwächen

- Kleinerer Kontext (200K vs. 1M)
- Weniger Community und Ökosystem als Llama/DeepSeek

### Am besten für

Softwareentwicklung, Code-Reviews und technische Analyse. Besonders stark bei komplexen Engineering-Aufgaben (SWE-Bench Pro). Gute Wahl für Teams, die einen leistungsstarken Coding-Assistenten suchen.

**cmt-Seminar:** GLM-5.1: Open Source LLM produktiv nutzen

# Qwen 3.6

## Die breiteste Modellpalette im Open-Source-Bereich

| | | | |
|---|---|---|---|
| **Hersteller** | Alibaba Cloud (China) | **Lizenz** | Apache 2.0 |
| **Parameter** | Qwen 3.6: bis 397B-A17B (MoE), Qwen 3.6: 35B-A3B | **Kontext** | 262K - 1M Tokens |
| **Geschw.** | ~35 Tokens/s (3.6 Plus) | **Varianten** | 0,8B bis 397B - für jede Hardwareklasse |

### Benchmarks (Qwen 3.6)

| Benchmark | Score | Einordnung |
|---|---|---|
| GPQA Diamond | **88,4** | Bestwert Open Source |
| AIME 2026 | **91,3** | Exzellent |
| LiveCodeBench | 83,6 | Stark |
| MMLU-Pro | 84,0 | Gut |

### Stärken

- Höchster GPQA-Score aller OS-Modelle
- Apache 2.0: keine Einschränkungen
- Breiteste Palette: 0,8B bis 397B
- Starkes Mathematik-Reasoning

### Schwächen

- Große Variante braucht Multi-GPU
- Kleinere Modelle schwächer beim Coding

### Am besten für

Wissenschaftliches Reasoning, Mathematik und breite Einsatzszenarien dank Modellvielfalt. Ideal für Unternehmen, die verschiedene Hardwareklassen bedienen müssen - vom Laptop bis zum GPU-Server.

**cmt-Seminar:** Qwen und weitere Alibaba-Modelle: Was lohnt sich? (KKC_0222)

# Llama 4

## Das größte Ökosystem und 10 Millionen Tokens Kontext

| | | | |
|---|---|---|---|
| **Hersteller** | Meta (USA) | **Lizenz** | Meta Community License |
| **Parameter** | Scout: 109B/17B aktiv, Maverick: 400B+/17B aktiv | **Kontext** | Scout: 10M, Maverick: 1M Tokens |
| **Geschw.** | ~112 Tokens/s | **API-Preis** | $0,15 / $0,60 pro 1M Tokens |

### Benchmarks (Maverick)

| Benchmark | Score | Einordnung |
|---|---|---|
| MMLU-Pro | 80,5 | Gut |
| GPQA Diamond | 69,8 | Unterer Bereich |
| LiveCodeBench | 43,4 | Schwächer |
| Kontext (Scout) | **10M** | Größter verfügbarer Kontext |

### Stärken

- Großes Ökosystem (Tools, Anleitungen, Community)
- 10M Kontext: ganze Codebases analysieren
- Sehr schnell (~112 tok/s)
- Extrem günstig als API ($0,15 Input)

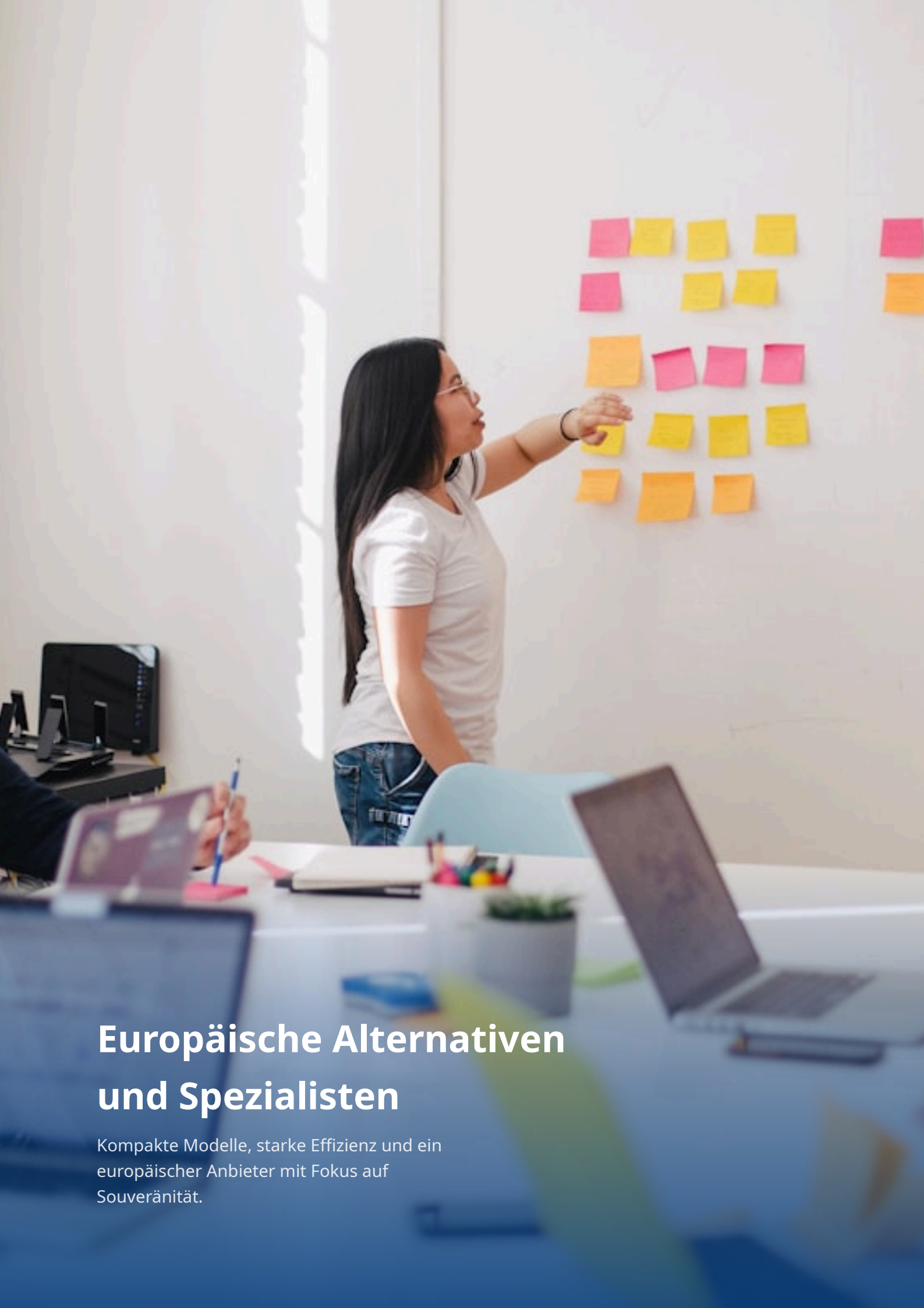
### Schwächen

- Benchmarks hinter DeepSeek/Qwen/GLM
- Meta Community License (nicht vollständig offen, kommerziell bis 700M MAU)
- Schwächer bei Coding-Aufgaben

### Am besten für

Hoher Durchsatz, lange Dokumente und kostenoptimierte Szenarien. Das breite Ökosystem macht den Einstieg besonders einfach. Ideal als erstes Self-Hosting-Projekt dank hervorragender Dokumentation und Community.

**cmt-Seminar:** Llama-Modelle im Vergleich: Nutzen statt Hype (KKC_0220)



# Europäische Alternativen und Spezialisten

Kompakte Modelle, starke Effizienz und ein europäischer Anbieter mit Fokus auf Souveränität.

# Mistral Small 4

## Die europäische Alternative - Souveränität aus Frankreich

**Hersteller** Mistral AI (Frankreich / Europa)
**Parameter** 119B gesamt / 6,5B aktiv (128 Experten)
**Geschw.** ~148 Tokens/s
**Lizenz** Apache 2.0
**Kontext** 256.000 Tokens
**Index** 28 (#7)

### Benchmarks

| Benchmark | Score | Einordnung |
|---|---|---|
| GPQA Diamond | 71,2 | Mittelfeld |
| Intelligence Index | 28 | #7 gesamt |
| Speed | **148 tok/s** | Schnellstes Modell |
| Aktive Parameter | **6,5B** | Geringster Hardware-Bedarf |

### Stärken

- Europäischer Anbieter: volle EU-Souveränität
- Apache 2.0: keine Einschränkungen
- Extrem schnell (148 tok/s)
- Nur 6,5B aktiv: läuft auf Consumer-Hardware
- Starke Mehrsprachigkeit

### Schwächen

- Schwächer bei Coding als asiatische Frontier-Modelle
- Reasoning-Leistung hinter DeepSeek/Kimi
- Weniger Benchmark-Erfolge

### Am besten für

Unternehmen, die EU-Souveränität priorisieren. Ideal für hohen Durchsatz auf kleiner Hardware, multilingualen Einsatz und Szenarien, in denen die Herkunft des Modells regulatorisch relevant ist.

**cmt-Seminar:** LLM Self-Hosting und Deployment / Open-Source-LLMs lokal betreiben

# Gemma 4

## Starke Leistung in kompaktem Format

**Hersteller** Google DeepMind (USA)
**Parameter** 31B (Dense) / 26B-A4B (MoE)
**Multimodal** Ja (Text + Bild)
**Lizenz** Apache 2.0
**Kontext** 256.000 Tokens
**Architektur** Dense + MoE-Variante

### Benchmarks

| Benchmark | Score | Einordnung |
|---|---|---|
| MMLU Multilingual | **85,2** | Sehr stark für 31B |
| GPQA Diamond | **84,3** | Schlägt viele 70B+ Modelle |
| AIME 2026 | **89,2** | Exzellentes Reasoning |
| LiveCodeBench | **80,0** | Stark |

### Stärken

- Hervorragend für seine Größe (schlägt 70B+-Modelle)
- Apache 2.0: volle Freiheit
- Starke multimodale Variante
- Läuft auf einer einzelnen GPU

### Schwächen

- Kleineres Ökosystem als Llama
- Keine Frontier-Größen verfügbar
- Google-Abhängigkeit bei Updates

### Am besten für

Edge- und On-Device-Szenarien, ressourcenbeschränkte Umgebungen und Unternehmen, die maximale Leistung pro GPU-Dollar suchen. Sehr gutes Preis-Leistungs-Verhältnis im Open-Source-Bereich.

**cmt-Seminar:** Open-Source-LLMs lokal betreiben

# Kimi K2.6

## #1 Open Weights – maximale Leistung bei kompakter Größe

| | | | |
|---|---|---|---|
| **Hersteller** | Moonshot AI (China) | **Lizenz** | **Modified MIT** (Namensnennung ab 100M MAU) |
| **Parameter** | 1T gesamt / 32B aktiv (384 Experten) | **Kontext** | 256.000 Tokens |
| **Geschw.** | Variabel (abhängig von Quantisierung) | **Index** | 54 (#1 Open Weights) |

### Benchmarks

| Benchmark | Score | Einordnung |
|---|---|---|
| GPQA Diamond | **90,5** | Bestwert Open Weights |
| SWE-bench Verified | **80,2** | Top 2 |
| SWE-Bench Pro | **58,6** | Bestwert |
| AIME 2026 | **96,4** | Bestwert |
| LiveCodeBench | **89,6** | Exzellent |

### Stärken

- #1 Open Weights Intelligence Index (54)
- Spitze bei Coding, Mathematik und Reasoning
- HLE-Full 54,0 – Bestwert aller Modelle
- Nur 32B aktiv – effizient zu betreiben

### Schwächen

- Modified MIT (nicht rein MIT)
- Jüngeres Ökosystem

### Am besten für

Coding, Mathematik und wissenschaftliches Reasoning. Bestes Leistung/Größe-Verhältnis aller Open-Weights-Modelle. Ideal, wenn Spitzenleistung bei moderatem Hardware-Aufwand gefragt ist.

**cmt-Seminar:** Kimi K2.6: Open Source LLM sofort produktiv

# MiniMax-M2.7

## Hybrid-Attention, 1 Million Tokens und Rekord-Trainingseffizienz

**Hersteller** MiniMax (China)
**Parameter** 456B gesamt / 45,9B aktiv (MoE)
**Training** 534.700 USD (Rekord-Effizienz)
**Lizenz** Apache 2.0
**Kontext** 1.000.000 Tokens
**Architektur** Hybrid-Attention + MoE

### Benchmarks

| Benchmark | Score | Einordnung |
|---|---|---|
| MMLU-Pro | 81,1 | Gut |
| MATH-500 | **96,8** | Spitzenwert |
| SWE-bench Verified | 56,0 | Mittelfeld |
| LiveCodeBench | 65,0 | Solide |

### Stärken

- Erstes Hybrid-Attention-Reasoning-Modell
- 1 Million Tokens Kontext
- Training nur 534.700 USD (Bruchteil üblicher Kosten)
- Apache 2.0: volle Freiheit
- Exzellent bei Mathematik (96,8 MATH-500)

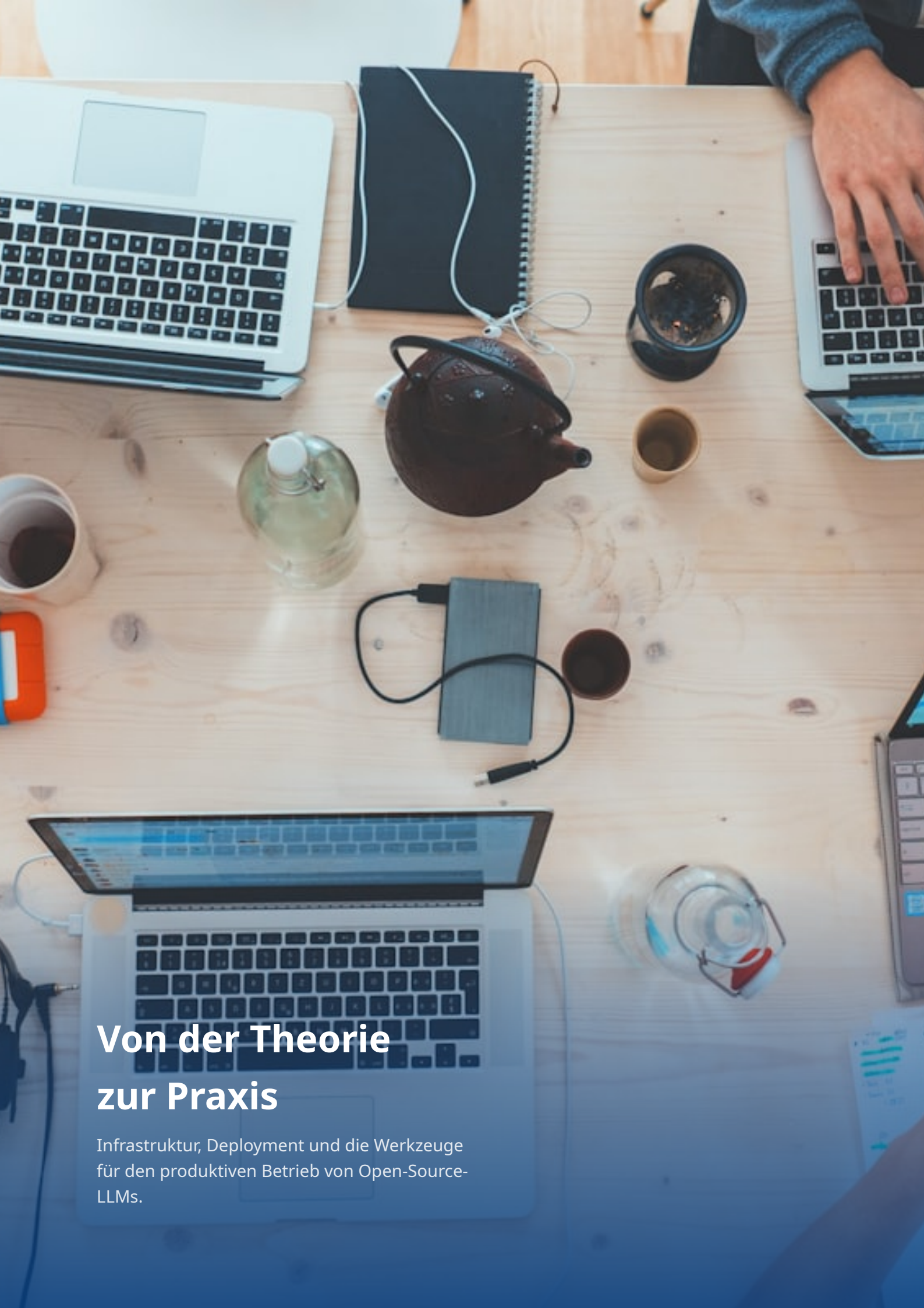
### Schwächen

- Schwächer bei Coding als DeepSeek/Kimi/GLM
- Jüngeres Ökosystem
- Weniger Community-Support

### Am besten für

Lange Dokumente, Mathematik und kosteneffizientes Training/Fine-Tuning. Die Trainingskosten von 534.700 USD liegen weit unter den üblichen Millionenbeträgen für Frontier-Modelle.

**cmt-Seminar:** MiniMax M2.7 Training (KKC_0223)



# Von der Theorie zur Praxis

Infrastruktur, Deployment und die Werkzeuge für den produktiven Betrieb von Open-Source-LLMs.

# Hardware und Infrastruktur

## Was Sie wirklich brauchen

**Die gute Nachricht: Sie brauchen keinen GPU-Cluster. Viele leistungsstarke Modelle laufen auf einer einzelnen Grafikkarte - quantisiert sogar auf Consumer-Hardware.**

### Hardware-Matrix nach Modellgröße

| Klasse | Parameter | VRAM | Beispiel-GPU | Beispiel-Modell |
|---|---|---|---|---|
| **Klein** | 1-4B | 2-4 GB | RTX 4060, Mac M2 | Qwen 3.6 3B, Gemma 4B |
| **Mittel** | 7-14B | 8-16 GB | RTX 4070, Mac M3 Pro | Qwen 3.6 14B, Mistral Small |
| **Groß** | 26-35B | 16-24 GB | RTX 5090, Mac M4 Max | Gemma 4 31B, Kimi K2.6 (Q4) |
| **Sehr groß** | 70B+ | 48-80 GB | 2x RTX 5090, Mac M3 Ultra | Llama 4 Scout, Qwen 3.6 72B |
| **Frontier MoE** | 400B+ | Multi-GPU Cluster | H100/B200 Cluster | DeepSeek V4 Pro, GLM-5.1 |

### Sweet Spots für LLM-Inference

**NVIDIA RTX 5090:** 32 GB VRAM, Blackwell-Architektur. Zwei Karten nähern sich der Inference-Leistung einer H100 bei deutlich geringeren Kosten. Aktuell ab ca. 3.700 EUR.

**Apple Mac M4 Max / M3 Ultra:** Unified Memory (bis 192 GB beim Ultra) macht Macs zu einer leisen, energieeffizienten Alternative. Ein M4 Max mit 128 GB kann 70B-Modelle betreiben - ideal als Entwickler-Workstation oder kleiner Team-Server. LM Studio und Ollama laufen nativ.

**AMD Ryzen AI 395+:** AMDs Strix-Halo-Plattform bringt bis zu 128 GB Unified Memory auf x86. Eine interessante Option für Linux-basierte Inference-Server ohne dedizierte GPU.

### Quantisierung: 75% VRAM sparen bei 95% Qualität

4-Bit-Quantisierung (GGUF, AWQ, GPTQ) reduziert den Speicherbedarf eines Modells um 75% bei nur ~5% Qualitätsverlust. Ein 32B-Modell schrumpft von ~64 GB auf ~16 GB VRAM. Für die meisten Anwendungsfälle ist der Unterschied nicht spürbar.

#### Einstieg ab ca. 3.500 EUR

Ein Mac Mini M4 Pro mit 48 GB (ab ca. 2.200 EUR) betreibt quantisierte 32B-Modelle flüsterleise. Alternativ: ein RTX-5090-Server (ab ca. 5.500 EUR) mit deutlich höherem Durchsatz für mehrere gleichzeitige Nutzer.

# Software-Stack

## Vom Modell zum Nutzer in unter 10 Minuten

**Der Software-Stack für Self-Hosted LLMs ist 2026 ausgereift. Drei Schritte genügen: Modell herunterladen, Inference-Server starten, Frontend bereitstellen.**

### Inference-Engines im Vergleich

| Engine | Setup | Stärke | Throughput* |
|---|---|---|---|
| **LM Studio** | <1 Min. | Desktop-GUI, kein Terminal nötig | ~400 tok/s |
| **Ollama** | <2 Min. | CLI-Einstieg, breites Ökosystem | 484 tok/s |
| **vLLM** | ~10 Min. | Production-Grade | **8.033 tok/s** |
| **TGI** | ~10 Min. | HuggingFace-Integration | ~5.000 tok/s |
| **SGLang** | ~15 Min. | Structured Output | ~7.000 tok/s |

*Gemessen mit Llama 3 70B auf A100, 128 gleichzeitige Anfragen. LM Studio bietet den einfachsten Einstieg: Modell auswählen, herunterladen, chatten - alles über eine grafische Oberfläche ohne Terminal-Kenntnisse. Ollama ist ideal für Entwickler und kleine Teams. Für produktiven Betrieb mit mehreren gleichzeitigen Nutzern ist vLLM die Referenz - 16,6x schneller als Ollama unter Last.

### Frontends: ChatGPT-Erfahrung im eigenen Netz

#### Open WebUI

- ChatGPT-ähnliche Oberfläche
- Integriertes RAG (Dokumenten-Upload)
- Rollenbasierte Zugriffskontrolle (RBAC)
- Self-hosted, Docker-ready

#### LibreChat

- Multi-Provider (Ollama + API)
- Audit-Logging für Compliance
- Plugin-System
- Enterprise-Features

### IDE-Integration: On-Premise Copilot

#### Tabby (23k Stars)

On-Premise-Alternative zu GitHub Copilot. Code-Vervollständigung, Chat und Code-Review - alles auf eigenem Server.

#### Continue (20k Stars)

Open-Source-IDE-Extension für VS Code und JetBrains. Verbindet sich mit jedem lokalen oder Remote-LLM.

#### OpenCode

Terminal-basierter KI-Coding-Agent. Arbeitet direkt im Projektverzeichnis, liest und editiert Dateien, führt Befehle aus. Ideal für Self-Hosted-Backends via OpenAI-kompatible API.

#### LM Studio

Desktop-App mit integriertem Server-Modus. Modelle lokal laden und als OpenAI-kompatible API bereitstellen - für alle IDE-Tools nutzbar.

# Souveränität und Datenschutz

## Warum Self-Hosting die DSGVO-Frage löst

**Jede Nutzung externer KI-APIs wirft datenschutzrechtliche Fragen auf. Self-Hosting eliminiert die kritischsten Risiken: keine Auftragsverarbeitung, keine Drittlandtransfers, volle Kontrolle.**

### Das Problem mit Cloud-APIs

| | | |
|---|---|---|
| ✕ Daten verlassen das Unternehmen | » | ✓ Daten bleiben im eigenen Netz |
| ✕ Art. 28 DSGVO: AVV erforderlich | » | ✓ Keine Auftragsverarbeitung nötig |
| ✕ Schrems II: Drittland-Transfer | » | ✓ Kein Transfer in Drittländer |
| ✕ EDPB: Hochrisiko-Einstufung | » | ✓ Volle Transparenz und Kontrolle |

### DSK-Orientierungshilfe: Die deutsche Aufsicht gibt Leitplanken

Die Datenschutzkonferenz (DSK) hat in drei Veröffentlichungen (Mai 2024, Juni 2025 mit 7 Gewährleistungszielen, Oktober 2025 zu RAG) klare Anforderungen formuliert. Self-Hosting erfüllt die meisten dieser Anforderungen automatisch: Transparenz über Datenverarbeitung, keine unkontrollierten Abflüsse, nachvollziehbares Logging.

### Deutsche Cloud-Infrastruktur

1. **Hetzner**
   GPU-Server ab 184 EUR/Monat (GEX44). Deutsche Rechenzentren, faire Preise, etablierter Anbieter.

2. **Deutsche Telekom Industrial AI Cloud**
   10.000+ Blackwell-GPUs in deutschen Rechenzentren. Enterprise-Grade mit vollem Compliance-Paket.

3. **IONOS**
   GPU-as-a-Service aus Deutschland. ISO 27001, BSI C5, DSGVO-konform.

# EU AI Act und Compliance

## Was ab wann gilt - und warum Open Source Vorteile hat

**Der EU AI Act ist in Kraft. Die Fristen laufen. Wer Open-Source-Modelle selbst hostet, profitiert von einer wichtigen Ausnahme - muss aber die Deployer-Pflichten kennen.**

### Timeline: Was wann gilt


### Open-Source-Ausnahme: Art. 53(2)

Open-Source-Modelle unter freien Lizenzen (MIT, Apache 2.0) sind von den meisten Provider-Pflichten des AI Acts befreit. Wer ein Modell nur herunterlädt und betreibt, gilt als Deployer - mit deutlich geringeren Anforderungen als ein Provider, der Modelle trainiert und anbietet.

### Deployer vs. Provider: Ihre Pflichten

| Pflicht | Provider | Deployer (Sie) |
|---|---|---|
| Risikobewertung | ✓ | ✓ |
| Technische Dokumentation | ✓ | - |
| Qualitätsmanagement | ✓ | - |
| Logging und Monitoring | ✓ | ✓ |
| Transparenz gegenüber Nutzern | ✓ | ✓ |
| Menschliche Aufsicht | ✓ | ✓ |

### Weitere Regulierung

Das KRITIS-Dachgesetz (ab März 2026), BSI-Empfehlungen und BaFin-Leitlinien für den Finanzsektor stellen zusätzliche Anforderungen an KI-Systeme in kritischen Infrastrukturen. Self-Hosting erleichtert die Einhaltung, da volle Kontrolle über Logging, Zugriffsrechte und Datenflüsse besteht.

**cmt-Seminare:** EU AI Act · KI Compliance · Risikomanagement KI

# Sicherheit

## Risiken kennen, systematisch absichern

**Auch selbst gehostete LLMs sind nicht automatisch sicher. Die OWASP Top 10 for LLMs zeigen die realen Angriffsvektoren - und wie Sie sich schützen.**

### Die häufigsten Angriffsvektoren

1. **Prompt Injection**
   Angreifer manipulieren das Modell über geschickt formulierte Eingaben. Besonders kritisch bei Agenten-Systemen mit Tool-Zugriff.

2. **Datenexfiltration**
   Sensible Daten aus dem Kontext oder RAG-System werden durch gezielte Anfragen extrahiert.

3. **Model Poisoning**
   Kompromittierte Modell-Weights aus nicht vertrauenswürdigen Quellen enthalten Backdoors.

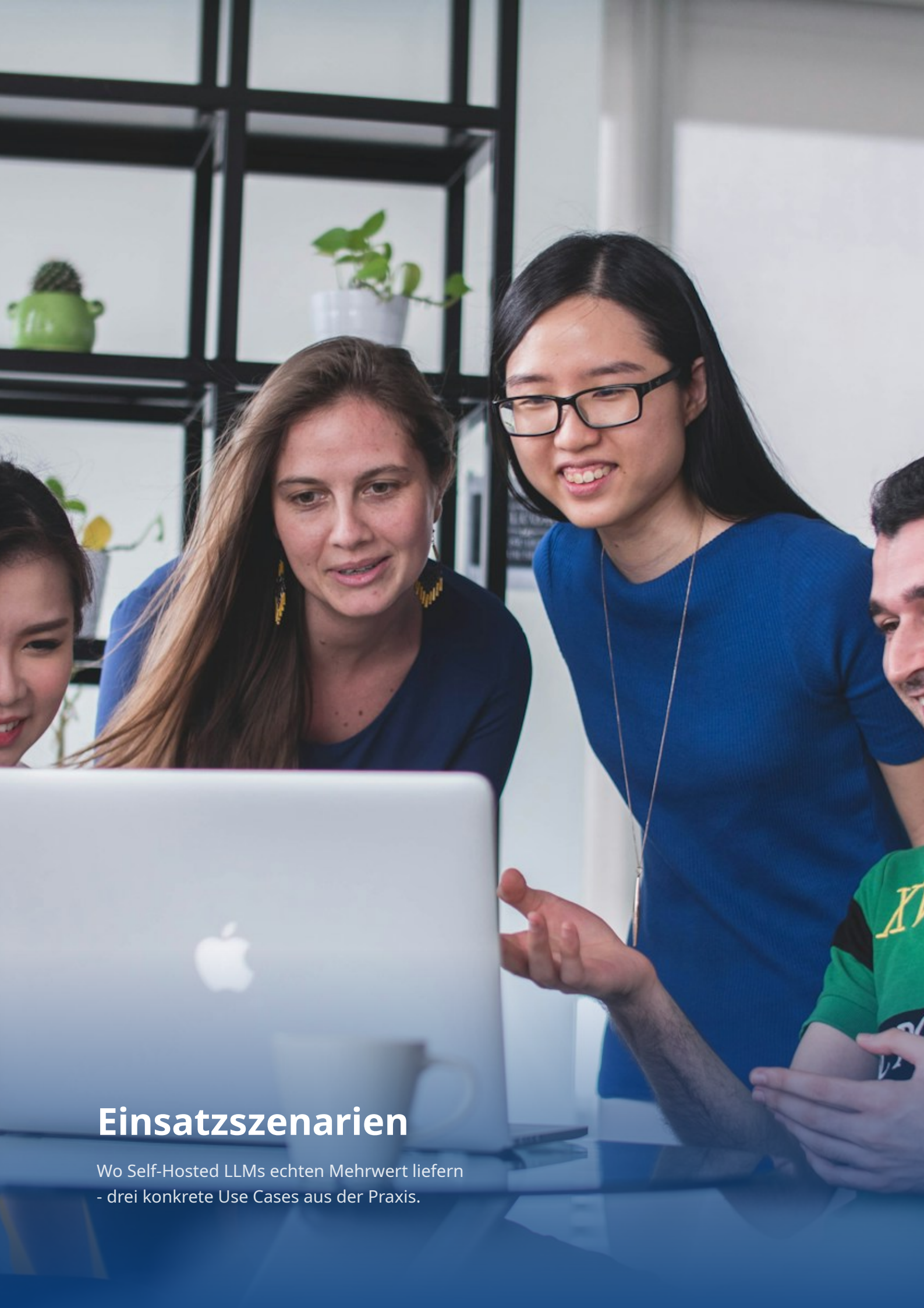
4. **Shadow AI**
   Unkontrollierte Nutzung externer KI-Dienste durch Mitarbeiter neben der offiziellen Lösung.

### Von unkontrolliert zu abgesichert

| | | |
|---|---|---|
| ✕ Unkontrollierte Nutzung | » | ✓ Zentral verwaltet mit RBAC |
| ✕ Beliebige Modellquellen | » | ✓ Trusted Registries |
| ✕ Keine Ein-/Ausgabe-Filter | » | ✓ Guardrails und Monitoring |

Weitere Maßnahmen: Least-Privilege-Zugriff für Agenten, Air-Gapped-Betrieb für hochsensible Daten, regelmäßige Security-Audits und Penetrationstests auf LLM-Endpoints.

**cmt-Seminare:** LLM Security · KI Security · Shadow AI stoppen



# Einsatzszenarien

Wo Self-Hosted LLMs echten Mehrwert liefern
- drei konkrete Use Cases aus der Praxis.

# Use Case: Coding und Entwicklung

## GitHub Copilot - aber auf eigenem Server

**Coding-Assistenten gehören zu den produktivsten KI-Anwendungen. Mit Open-Source-Modellen und Tools wie Tabby lässt sich ein vollwertiger On-Premise-Copilot aufbauen - ohne dass Code das Unternehmensnetz verlässt.**

### Leistung: Open Source auf Augenhöhe

| 80,6% | 89,6 | 58,6 |
|---|---|---|
| SWE-bench (DeepSeek V4 Pro) | LiveCodeBench (Kimi K2.6) | SWE-Bench Pro (Kimi K2.6) |

### Empfohlene Modelle für Coding

| Aufgabe | Empfohlenes Modell | Warum |
|---|---|---|
| Inline-Completion | Qwen 3.6 14B / Gemma 4 | Schnell, klein, hohe Trefferquote |
| Code-Review & Chat | Kimi K2.6 / DeepSeek V4 | Bestes Reasoning bei Code |
| Agentic Coding | Kimi K2.6 / GLM-5.1 | SWE-bench Pro Spitzenwerte |
| Legacy-Code verstehen | DeepSeek V4 Pro | 1M Kontext, starke Analyse |

Toolchain: Tabby oder Continue als IDE-Extension, OpenCode für Terminal-Workflows, LM Studio oder Ollama als lokaler Server (Details siehe S. 17).

### ROI-Rechnung

| Position | GitHub Copilot Enterprise | Self-Hosted (Tabby) |
|---|---|---|
| Kosten pro Entwickler/Monat | 240 EUR | ~0 EUR (Server amortisiert) |
| 10 Entwickler, 12 Monate | 28.800 EUR | ~5.500 EUR einmalig |
| Code verlässt Netzwerk | Ja (Microsoft Cloud) | **Nein** |
| Anpassbar / Fine-Tuning | Nein | **Ja** |

# Use Case: Agentic AI

## Autonome KI-Agenten auf eigener Infrastruktur

**Gartner prognostiziert: 40% aller Enterprise-Anwendungen werden bis Ende 2026 KI-Agenten integrieren. Mit Open-Source-Frameworks lassen sich diese Agenten vollständig selbst betreiben.**

### Frameworks und Standards

| Framework | GitHub Stars | Stärke |
|---|---|---|
| **LangGraph** | 24.800 | Komplexe Workflows mit Zustandsmaschine |
| **CrewAI** | 44.300 | Multi-Agent-Teams mit Rollenverteilung |
| **MS Agent Framework** | - | Enterprise-Integration (Azure, M365) |
| **MCP (Anthropic)** | 97 Mio. Downloads/Mo. | Universelles Tool-Protokoll für LLMs |

### Vier Einsatzszenarien

1. **Dokumentenverarbeitung**
   Eingehende Rechnungen, Verträge und Angebote automatisch klassifizieren, extrahieren und ins ERP übertragen.

2. **Wissensmanagement**
   Internes Wiki, Confluence und SharePoint durchsuchen und Antworten mit Quellenangabe generieren.

3. **Workflow-Automation**
   Wiederkehrende Aufgaben (Ticket-Routing, E-Mail-Triage, Report-Erstellung) durch Agenten automatisieren.

4. **IT-Operations**
   Log-Analyse, Incident-Triage und Runbook-Ausführung durch KI-Agenten mit Zugriff auf Monitoring-Tools.

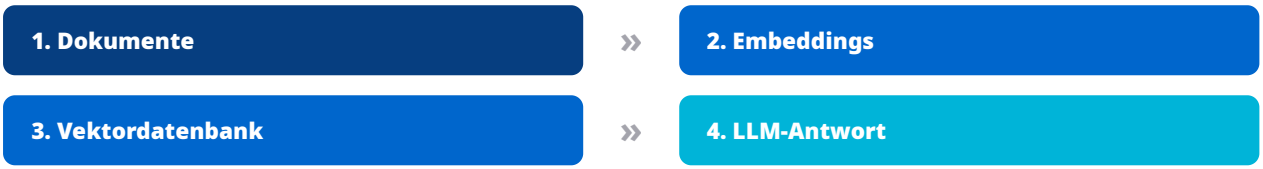
**cmt-Seminare:** KI-Agenten · n8n Workflow · Custom KI-Agents

# Use Case: RAG und Wissensbasis

## Ihr Unternehmenswissen, durchsuchbar per KI

**Retrieval-Augmented Generation (RAG) verbindet LLMs mit Ihren eigenen Dokumenten. Statt Halluzination: fundierte Antworten mit Quellenangabe - vollständig auf eigener Infrastruktur.**

### Architektur




Dokumente werden in Chunks aufgeteilt, vektorisiert und in einer Datenbank gespeichert. Bei einer Anfrage werden die relevantesten Chunks abgerufen und dem LLM als Kontext mitgegeben. Das Ergebnis: Antworten, die auf Ihren tatsächlichen Daten basieren.

### Vektordatenbanken im Vergleich

| Datenbank | Typ | Stärke |
|---|---|---|
| **pgvector** | PostgreSQL-Extension | Kein neues System nötig, wenn PostgreSQL vorhanden |
| **Qdrant** | Dediziert | Höchste Performance, Rust-basiert, Self-Hosted |
| **Milvus** | Dediziert | Skalierung auf Milliarden Vektoren |

### Praxisbeispiel: Open WebUI + RAG

Open WebUI bringt RAG-Funktionalität von Haus aus mit. Nutzer laden Dokumente per Drag-and-Drop hoch, die automatisch vektorisiert werden. Fragen werden gegen die Dokumentenbasis beantwortet - inklusive Quellenangabe. Für größere Deployments empfiehlt sich die Anbindung an Qdrant als externe Vektordatenbank.

#### DSK-konforme RAG

Die DSK hat im Oktober 2025 spezifische Anforderungen an RAG-Systeme formuliert. Self-Hosted RAG erfüllt die Kernforderung automatisch: keine Weitergabe von Dokumenten an Dritte.

**cmt-Seminare:** RAG & Vektordatenbanken · LLM-Dokumentenanalyse

# Ehrliche Bilanz

## Wo Open Source (noch) an Grenzen stößt

**Open-Source-LLMs haben deutlich aufgeholt, aber proprietäre Modelle sind in einigen Bereichen noch leistungsstärker. Eine nüchterne Einordnung hilft bei der richtigen Entscheidung.**

### Wo Open Source noch schwächer ist

| Bereich | Open Source | Proprietär |
|---|---|---|
| Halluzinationsrate | Je nach Modell und Aufgabe höher | Niedriger bei Frontier-Modellen |
| Instruction Following | Gut, aber inkonsistenter | Sehr konsistent |
| Multimodal (Bild/Video) | Aufholend (Gemma 4) | Führend |
| Lange Kontexte (>100K) | Wachsend, aber Qualität sinkt | Stabiler |

### Betriebsaufwand nicht unterschätzen

Self-Hosting bedeutet Verantwortung: Updates, Monitoring, Troubleshooting, Security-Patches. Rechnen Sie mit 10-20 Stunden pro Monat für einen produktiven LLM-Service. Ohne dediziertes Personal wird das Projekt zum Risiko.

### Unsere Empfehlung: Hybrid

Nicht alles muss self-hosted sein. Der pragmatische Ansatz:

- **Self-Hosted:** Standard-Aufgaben (Chat, Zusammenfassung, Code-Completion, RAG), sensible Daten, hohe Volumen
- **Proprietäre API:** Komplexes Reasoning, kreative Aufgaben, seltene Sprachen, Aufgaben mit geringer Fehlertoleranz

#### Wann eine API die bessere Wahl ist

**Unter 2M Tokens/Tag:** Die API ist wahrscheinlich günstiger als dedizierte Hardware.
**Geringe Fehlertoleranz:** Wenn jede Antwort stimmen muss, sind proprietäre Modelle (noch) zuverlässiger.
**Kein DevOps-Team:** Ohne Personal für Betrieb und Wartung wird Self-Hosting zur Belastung.

# Kostenvergleich

## Ab wann sich Self-Hosting rechnet

**Self-Hosting lohnt sich nicht für jeden. Die TCO-Rechnung hängt von Nutzungsvolumen, Modellwahl und vorhandener Infrastruktur ab. Hier die konkreten Zahlen.**

### TCO-Vergleich: Hetzner GEX44 vs. API

| Position | Self-Hosted (Hetzner) | API (Frontier) | API (Günstig) |
|---|---|---|---|
| Monatliche Kosten | 184 EUR fix | Variabel | Variabel |
| 5M Tokens/Tag | 184 EUR/Mo. | ~1.350 EUR/Mo. | ~120 EUR/Mo. |
| 20M Tokens/Tag | 184 EUR/Mo. | ~5.400 EUR/Mo. | ~480 EUR/Mo. |
| 50M Tokens/Tag | 184 EUR/Mo. | ~13.500 EUR/Mo. | ~1.200 EUR/Mo. |

### Break-even-Punkte

**2-5M**
Tokens/Tag vs. Frontier-APIs (Opus, GPT-5)

**50M+**
Tokens/Tag vs. günstige APIs (DeepSeek, Llama)

**3-6 Mo.**
Amortisation eigener Hardware

### Versteckte Kosten nicht vergessen

- **Personal:** 10-20 Std./Monat DevOps (ca. 800-1.600 EUR anteilig)
- **Strom:** GPU-Server verbrauchen 200-600W, ca. 20-60 EUR/Monat
- **Amortisation:** Hardware-Kauf über 3 Jahre abschreiben
- **Updates:** Neue Modelle erfordern gelegentlich Hardware-Upgrades

### Beispielrechnung: 50 Nutzer, 5M Tokens/Tag

**Self-Hosted (Hetzner GEX44):** 184 EUR/Mo. Server + ~1.000 EUR/Mo. anteiliges Personal = **1.184 EUR/Mo.**
**API (Opus 4.7):** 5M Tokens x 30 Tage x $15/1M = **~2.100 EUR/Mo.**
**Ersparnis:** ~900 EUR/Mo. bzw. ~10.800 EUR/Jahr - plus volle Datenkontrolle.

# Der Einstiegs-Fahrplan

## In drei Phasen zum produktiven LLM-Betrieb

**Sie müssen nicht alles auf einmal machen. Ein schrittweiser Ansatz minimiert Risiken und liefert schnell erste Ergebnisse.**

### Phase 1: Evaluation

**Woche 1-2**
Modelle testen (Ollama + Open WebUI). Use Cases priorisieren. Hardware-Bedarf ermitteln. Quick Win: internes Chat-Tool für IT-Team.

### Phase 2: Pilot

**Monat 1-2**
Produktions-Setup (vLLM + Frontend). RAG mit Pilotdaten. Security-Konzept. 10-20 Power-User als Testgruppe.

### Phase 3: Rollout

**Monat 3-6**
Alle Mitarbeiter onboarden. IDE-Integration (Tabby/Continue). Agenten für erste Automatisierungen. Monitoring etablieren.

## Trainings nach Rolle

| Rolle | Fokus | Empfohlenes cmt-Seminar |
|---|---|---|
| **IT-Leitung / CTO** | Strategie, Make-or-Buy, Compliance | EU AI Act, KI Compliance |
| **DevOps / Infra** | Deployment, Monitoring, Skalierung | LLM Self-Hosting, KI-Apps containerisieren |
| **Entwickler** | IDE-Integration, Agenten, RAG | KI-Agenten, RAG & Vektordatenbanken |
| **Security** | LLM-Absicherung, OWASP LLM Top 10 | LLM Security, Shadow AI stoppen |
| **Compliance** | AI Act, DSGVO, Dokumentation | KI Compliance, Risikomanagement KI |
| **Alle Mitarbeiter** | Grundlagen, Prompt-Engineering | ChatGPT Grundlagen, KI im Büroalltag |

### Quick Win in unter einer Stunde

Ollama installieren, Gemma 4 herunterladen, Open WebUI starten. Drei Befehle, ein internes Chat-Tool. Damit überzeugen Sie Ihr Management, bevor Sie Budget beantragen.

# Seminar-Übersicht

## Über 120 KI-relevante Seminare bei cmt.de

**Von der Grundlagen-Schulung bis zum spezialisierten Modell-Training: cmt.de bietet eines der umfassendsten Seminarangebote für KI und LLMs im deutschsprachigen Raum.**

### KI und LLM

- DeepSeek: Einsatz und Self-Hosting (KKC_0221)
- GLM-5.1: Open Source LLM produktiv nutzen
- Qwen und weitere Alibaba-Modelle: Was lohnt sich? (KKC_0222)
- Llama-Modelle im Vergleich: Nutzen statt Hype (KKC_0220)
- Kimi K2.6: Open Source LLM sofort produktiv
- MiniMax M2.7 Training (KKC_0223)
- LLM Self-Hosting und Deployment
- Open-Source-LLMs lokal betreiben
- LLM Fine-Tuning und Anpassung
- RAG & Vektordatenbanken
- LLM-Dokumentenanalyse

### Infrastruktur und DevOps

- KI-Apps containerisieren (Docker, Kubernetes)
- KI Cloud sicher betreiben
- GPU-Cluster aufbauen und verwalten
- Monitoring und Observability für KI-Systeme

### Security und Compliance

- LLM Security (OWASP Top 10 for LLMs)
- KI Security: Angriff und Verteidigung
- Shadow AI stoppen: Kontrolle zurückgewinnen
- EU AI Act: Anforderungen und Umsetzung
- KI Compliance und Governance
- Risikomanagement KI
- DSGVO-Workshop für KI-Anwendungen

### Coding und Automatisierung

- KI-Agenten bauen (LangGraph, CrewAI, MCP)
- n8n Workflow-Automation mit KI
- Custom KI-Agents für Unternehmen
- ChatGPT für Entwickler
- GitHub Copilot produktiv nutzen
- AI-Driven Operations

### Alle Seminare auch als Inhouse-Training

Jedes Seminar kann maßgeschneidert für Ihr Team durchgeführt werden - vor Ort oder remote. Inklusive Anpassung an Ihre Infrastruktur und Use Cases.

# Warum cmt?

## Ihr Partner für KI-Kompetenz

**Seit über 25 Jahren entwickeln wir Fach- und Führungskräfte im deutschsprachigen Raum. Mit über 120 KI-spezifischen Seminaren sind wir einer der umfassendsten Trainingsanbieter für generative KI und LLMs in Deutschland.**

| 2.100+ | 25+ | 120+ |
|---|---|---|
| Seminare im Programm | Jahre Erfahrung | KI-Seminare |

### Was uns auszeichnet

✓ **Praxisorientierte Trainer**
Alle KI-Trainer arbeiten selbst aktiv mit LLMs, deployen Modelle und bauen Agenten. Keine Folienvorträge, sondern echte Hands-on-Erfahrung aus Produktivumgebungen.

✓ **Modell-spezifische Seminare**
Als einer der wenigen Anbieter bieten wir dedizierte Trainings für einzelne Modelle: DeepSeek, Qwen, Llama, Kimi, GLM und mehr. Nicht generisch, sondern passgenau.

✓ **Offene Seminare und Inhouse**
Wählen Sie zwischen offenen Terminen oder maßgeschneiderten Inhouse-Trainings. Beide Formate auch remote verfügbar. Inhouse-Trainings werden auf Ihre Infrastruktur und Use Cases angepasst.

✓ **Individuelle Beratung**
Wir helfen Ihnen, den optimalen Einstieg in Self-Hosted KI zu planen - von der Modellwahl bis zum Schulungsplan. Das Erstgespräch ist kostenlos.

### Kostenlos beraten lassen

Telefon: 0800 71 20 000 (Mo-Fr 8-17 Uhr)
E-Mail: info@cmt.de
Oder direkt online buchen auf www.cmt.de

## Bereit für KI unter eigener Kontrolle?

**Yves Hoppe**

KI und Technologie

089 / 68 08 97 3-0 · hoppe@cmt.de

cmt GmbH · Telefon: 0800 71 20 000 · info@cmt.de

**www.cmt.de/kategorien/kuenstliche-intelligenz/**