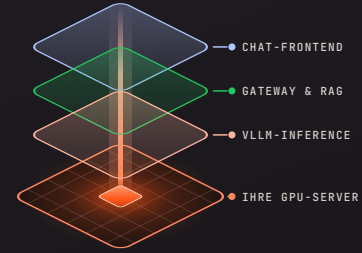


PRIVATE LLM HOSTING

Ihre KI. Ihre Daten. Ihre Kosten im Griff.

Cloud-KI heißt: steigende API-Rechnungen und Daten außer Haus. NexPatch bringt dieselbe Leistung auf Ihre Server: **compliant by design**, planbare Kosten, für Ihr Team so einfach wie ChatGPT.



Der lokale KI-Stack auf Ihrer Infrastruktur: keine Daten verlassen das Netzwerk.

0 €

EXTERNE API-ABOS

100 %

DATENKONTROLLE

5-10M

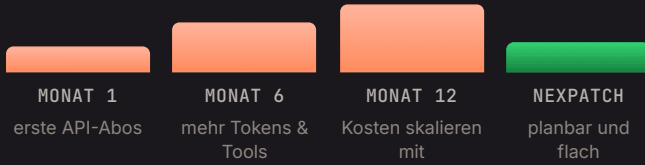
BREAK-EVEN TOKENS/MONAT

60-73 %

ZIELERSPARNIS BEI VOLUMEN

Was heute teuer wird

Der Einstieg ist leicht, doch mit echten Prozessen wachsen Rechnung und Risiko Monat für Monat. Dazu CLOUD Act, DSGVO und ab August 2026 der EU AI Act.



Was sich mit NexPatch ändert

- Private, OpenAI-kompatible KI-Schicht: vLLM, Gateway, RAG und Observability als ein System
- Daten, Prompts und Logs bleiben bei Ihnen
- Internes Wissen sicher nutzbar: klare Rechte, belegbare Quellen
- Fertige Workflows: Suche, Assistenz, Zusammenfassung, Automatisierung

Heute

API-Abos, Tokenkosten, Datentransfer

NexPatch Private AI Layer

Audit · Anbindung · RAG · Modelle · Betrieb

Danach

0 externe Abos, volle Datenkontrolle



01

Audit

Prozesse, Daten, Risiken, Sparpotenzial



02

Anbinden

ERP, CRM, DMS, Datenbanken, Rollen



03

Wissen

RAG, Embeddings, Dokumente, Rechte



04

Build

Use Cases, Modellwahl, Fine-Tuning



05

Betrieb

Monitoring, Qualität, Updates, Governance

Passt zu Ihnen, wenn

Sensible Dokumente, hohes KI-Volumen oder strenge Compliance: Finanz, Gesundheit, Recht, Fertigung, öffentlicher Sektor.

Ihr Vorteil

So einfach wie eine API, aber ohne externe Abos, ohne Datentransfer, ohne Lock-in. Modelle wechseln Sie jederzeit.

Was Sie messen

Externe API-Kosten sinken auf 0. Ab 5-10 Mio. Tokens pro Monat schlägt eigene Infrastruktur Premium-APIs.

DSGVO-KONFORM

EU AI ACT READY

ON-PREMISE ODER PRIVATE CLOUD

MADE IN GERMANY

Sehen wir uns Ihr Sparpotenzial gemeinsam an.

30 Minuten, unverbindlich: Wir prüfen Ihre Use Cases und rechnen Ihr Sparpotenzial durch.

nexpatch.ai · info@nexpatch.ai · Leipzig

Gespräch vereinbaren

calendly.com/fabian-nexpatch/30min