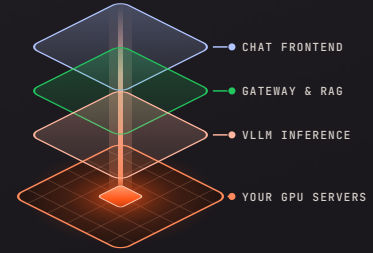


## PRIVATE LLM HOSTING

# Your AI. Your data. Your costs under control.

Cloud AI means rising API bills and data leaving the house. NexPatch brings the same power onto your servers: **compliant by design**, predictable costs, as easy for your team as ChatGPT.



The local AI stack on your own infrastructure: no data ever leaves your network.

**0 €**

EXTERNAL API  
SUBSCRIPTIONS

**100 %**

DATA CONTROL

**5-10M**

BREAK-EVEN TOKENS/MONTH

**60-73 %**

TARGET SAVINGS AT VOLUME

## What gets expensive today

Getting started is easy, but with real processes the bill and the risk grow month after month. Add CLOUD Act, GDPR and, from August 2026, the EU AI Act.



## What changes with NexPatch

- Private, OpenAI-compatible AI layer: vLLM, gateway, RAG and observability as one system
- Data, prompts and logs stay with you
- Internal knowledge safely usable: clear permissions, verifiable sources
- Ready-made workflows: search, assistance, summarisation, automation

### Today

API subscriptions, token costs, data transfer

### NexPatch Private AI Layer

Audit · Integration · RAG · Models · Operations

### After

0 external subscriptions, full data control



01

#### Audit

processes, data, risks, savings potential



02

#### Integrate

ERP, CRM, DMS, databases, roles



03

#### Knowledge

RAG, embeddings, documents, permissions



04

#### Build

use cases, model choice, fine-tuning



05

#### Operate

monitoring, quality, updates, governance

## A fit for you if

Sensitive documents, high AI volume or strict compliance: finance, healthcare, legal, manufacturing, public sector.

## Your advantage

As easy as an API, but without external subscriptions, data transfer or lock-in. Swap models whenever you want.

## What you measure

External API costs drop to 0. From 5-10M tokens per month, your own infrastructure beats premium APIs.

● GDPR COMPLIANT

● EU AI ACT READY

● ON-PREMISE OR PRIVATE CLOUD

● MADE IN GERMANY

## Let's look at your savings potential together.

30 minutes, no strings attached: we review your use cases and run your savings numbers.

nexpatch.ai · info@nexpatch.ai · Leipzig, Germany

### Book a call

calendly.com/fabian-nexpatch/30min