Architektur für KI-Anwendungen

Ein Überblick

CONCISO.

Speaker

Dr. Georg Pietrek

Geschäftsführer und
Chef Architekt

Lars Winterhalder

Senior Architekt

CONCISO.

Was Euch erwartet

- Betrieb der KI
- Integration von KI in Enterprise Applications

CONCISO.

Betrieb der KI

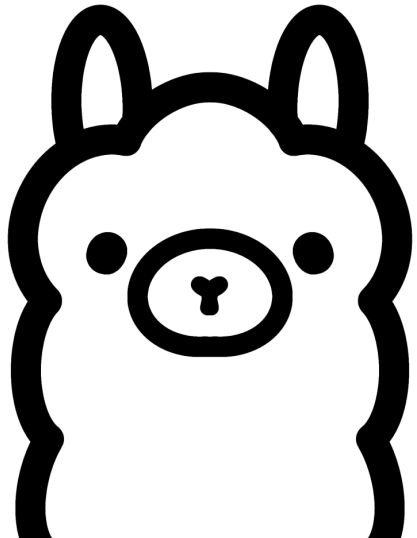
Betrieb der KI

KOMMERZIELL

- OpenAI – ChatGPT
- Google – Gemini
- Anthropic – Claude
- Meta – Meta AI
- Mistral AI
- …

PRIVAT

- Ollama – ca. 100 Modelle



CONCISO.

Kommerzieller KI-Anbieter

Betrieb in der Cloud (durch den Anbieter)

- Zugriff über REST-API
- Pay per Use (oft)

VORTEILE

- Die „besten“ Modelle nutzbar
- Sehr performant (Antwortzeiten, Skalierbarkeit)

NACHTEILE

- Datenschutz
- Evtl. Kosten

Kommer-zieller KI-Anbieter

DATENSCHUTZ

- Wo landen meine Daten?
- Wer hat Zugriff?
- Wofür werden sie genutzt?
  - Nutzt der Anbieter sie zum Trainieren seiner Modelle?
- Was gibt es an rechtlichen Vorschriften zu beachten?

Kommer-zieller KI-Anbieter

KOSTEN

- Abhängig von Anzahl der Token
  - Simple Chats: günstig
  - KI-Agent für Programmieraufgaben: Kosten gehen schnell in die Höhe

Private KI

Betrieb auf eigener Hardware

- Lokal (mein Notebook)
- Server in eigenem Rechenzentrum
- Gemieteter Server bei Hoster

VORTEILE

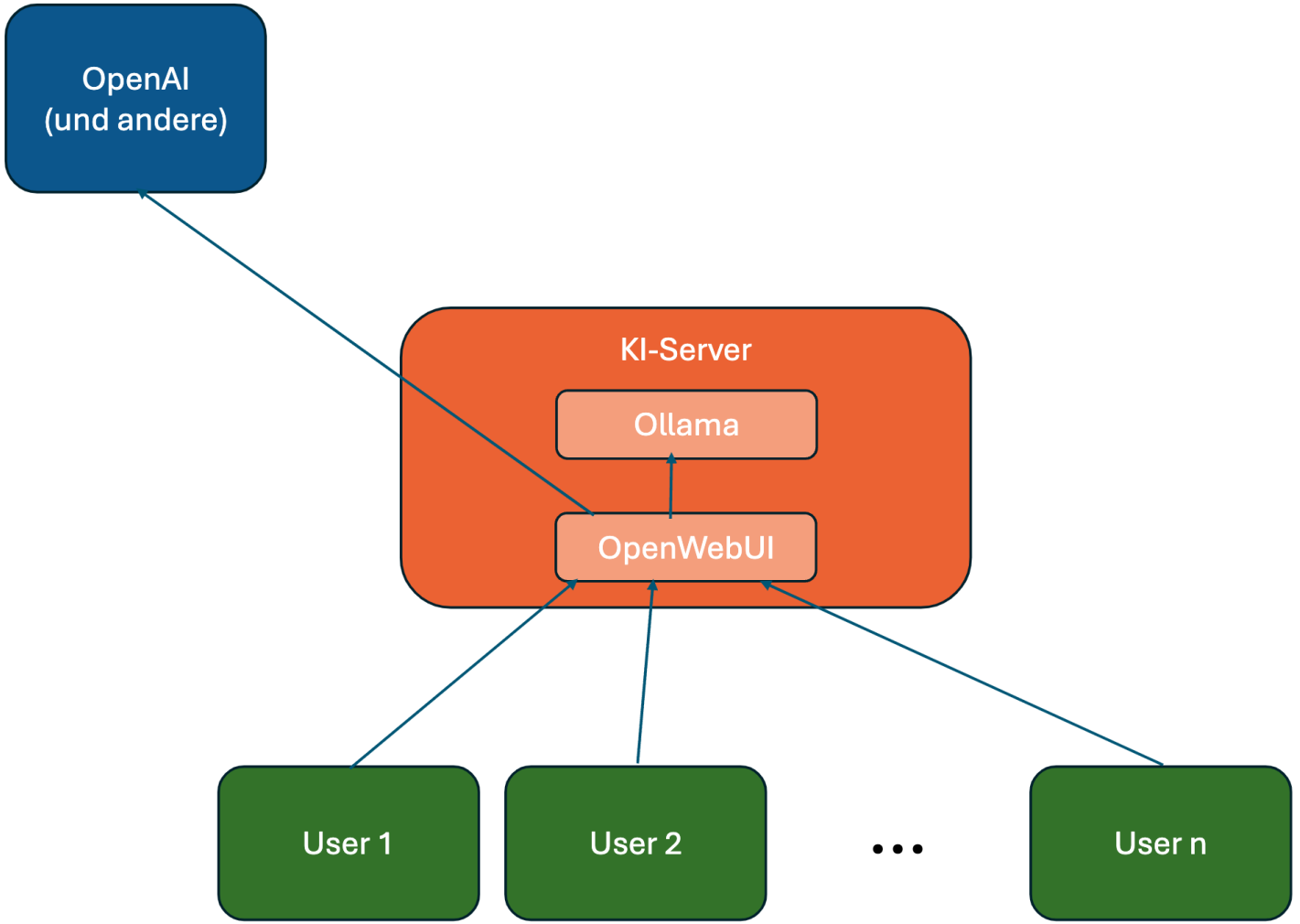
- Daten verlassen eigene Hardware nicht
- Große Auswahl an Modellen

NACHTEILE

- Nicht alle Highend-Modelle nutzbar
- Erfordert leistungsstarke Hardware

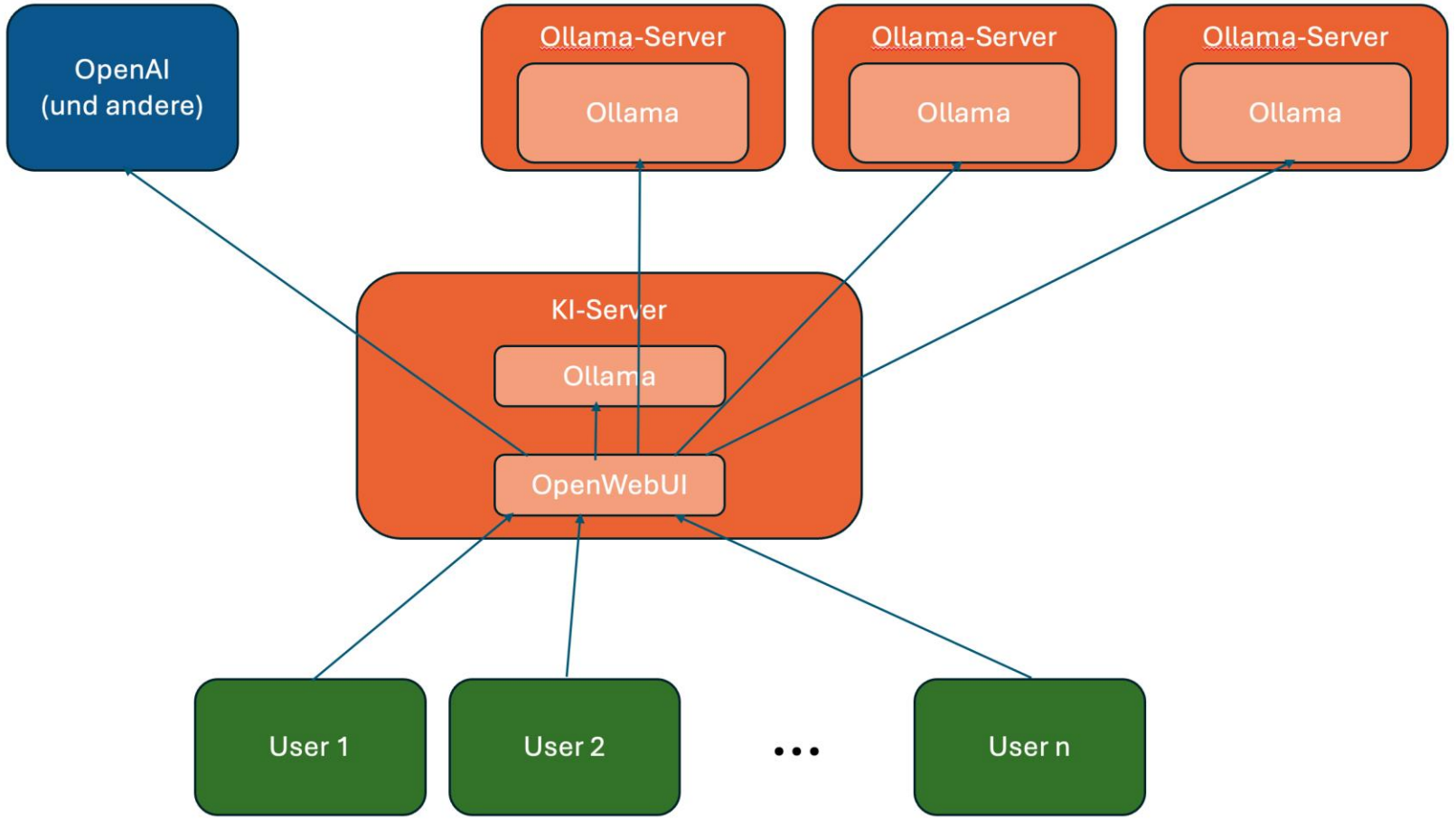
CONCISO.

KI-Play-ground



CONCISO.

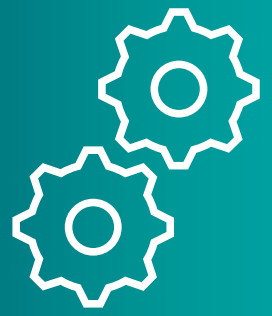
KI-Play-ground



CONCISO.

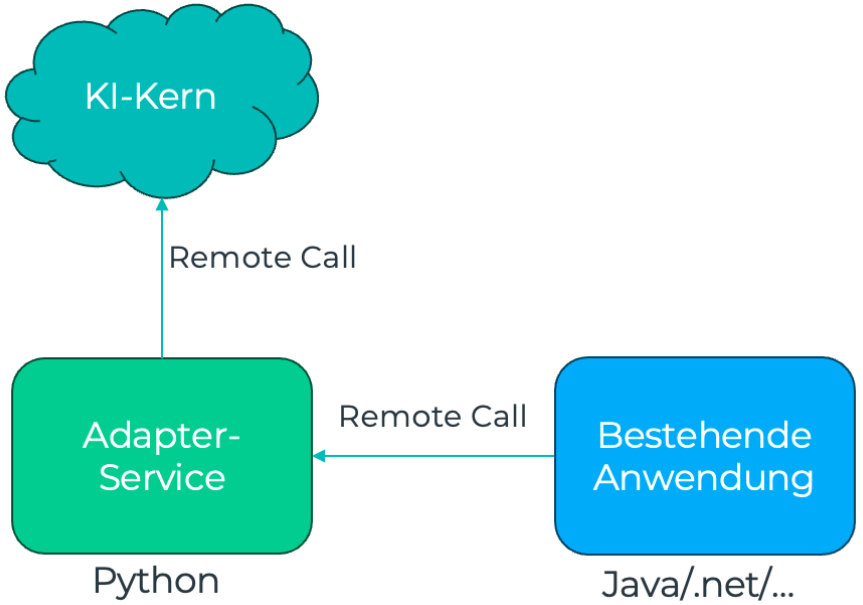
Integration von KI in Enterprise Applications

Integration von KI in Enterprise Appli-cations

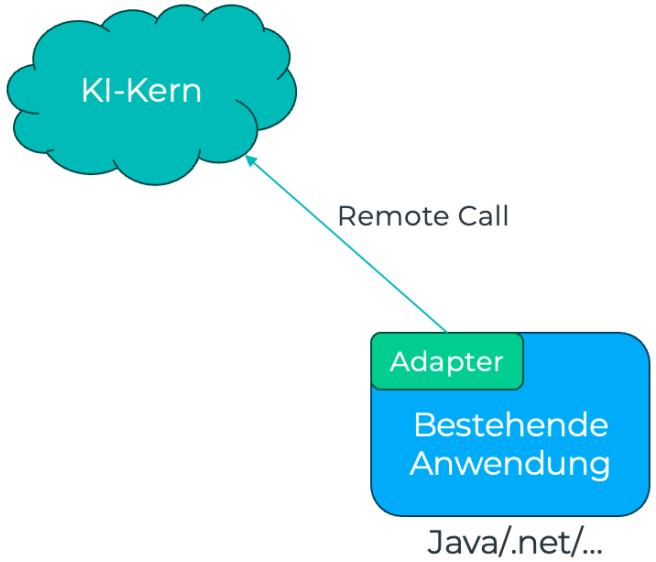


CONCISO.

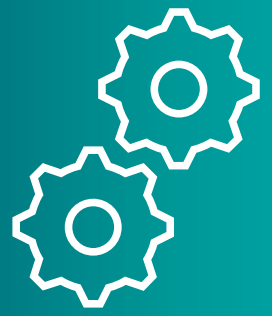
Anbindung über dedizierten Adapter



Direkte Anbindung aus der bestehenden Anwendung heraus



Integration von KI in Enterprise Appli-cations



CONCISO.

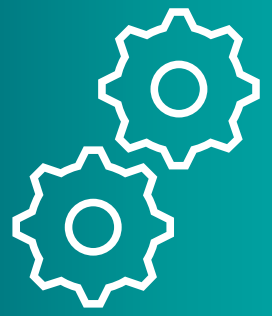
ANBINDUNG ÜBER DEDIZIERTEN ADAPTER

- Technologie kann speziell für den Usecase gewählt werden
  - Große Auswahl an bestehenden Tools/Libraries
  - Viele Beispiele und bestehende Lösungsansätze für Probleme
  - Neue Features können früher benutzt werden
- Kann (bei größeren Projekten) von eigenem Team betreut werden

DIREKTE ANBINDUNG AUS DER BESTEHENDEN ANWENDUNG HERAUS

- Bekannter Technologiestack
- Entwickler müssen keine neue Programmierspra che lernen
- Geringere Komplexität
- Stabilere APIs

Integration von KI in Enterprise Appli-cations



CONCISO.

ANBINDUNG ÜBER DEDIZIERTEN ADAPTER

- Typisch: Phyton
  - Alle Beispiele im Bereich generative KI sind in Python implementiert

DIREKTE ANBINDUNG AUS DER BESTEHENDEN ANWENDUNG HERAUS

- Typisch: Spring Boot (Spring AI)
  - Java
  - passt in „unsere“ Welt
  - Langsamer Fortschritt (seit Monaten im Milestone-Status)
  - Geringere Funktionalität

Herzlichen Dank für Eure Aufmerksamkeit!

conciso.de

CONCISO.

Dr. Georg Pietrek

M +49 151 10860 459
E georg.pietrek@conciso.de
W www.conciso.de

LinkedIn

Lars Winterhalder

M +49 170 2279357
E lars.winterhalder@conciso.de
W www.conciso.de

LinkedIn