/// Kundentag 2024

WISSENS-MANAGAMENT MIT KI

M. Höflinger mit F. Wallner | 24.10.2024 | 15:05 Uhr

DR. WALLNER ENGINEERING

Copyright

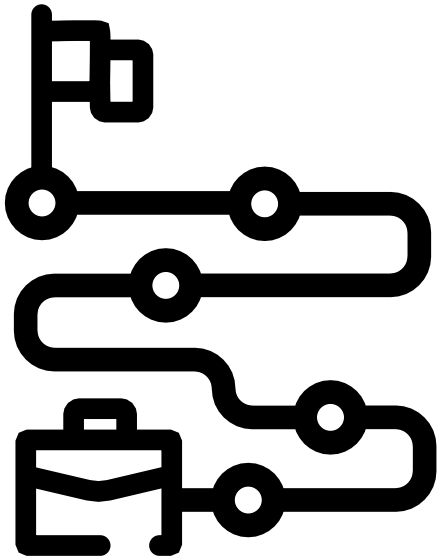
Dr. Wallner Engineering GmbH
Charles-Lindbergh-Str. 7
71034 Böblingen

Tel +49 7031 410309-0
Fax +49 7031 410309-11
Mail kontakt@drwe.de
Web www.drwe.de

/// Was ist zu erwarten?

AGENDA

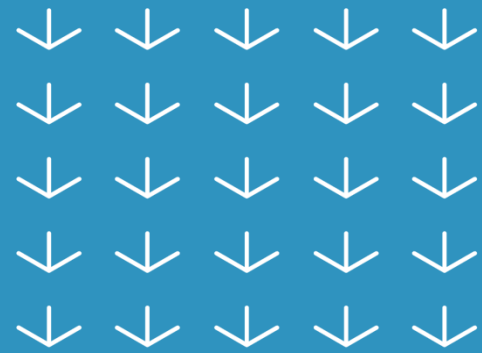
1. Warum Wissensmanagement mit KI?
2. Was ist ein LLM?
3. Warum lokal?
4. Was nutzen wir?
5. Wissen integrieren
6. Ausblick und weitere Ansätze



/// 1.

WARUM WISSENS-MANAGEMENT MIT KI?

- Userfreundlich
- Zeitsparend
- Zentralisierte Anlaufstelle über verschiedene Ressourcen hinweg



/// 2.

WAS IST EIN LLM?

Eine KI welche:

- Auf großen Mengen von Text trainiert wird
- Sprache verstehen und generieren kann
- Muster und Zusammenhänge einer Sprache versteht

/// 3.

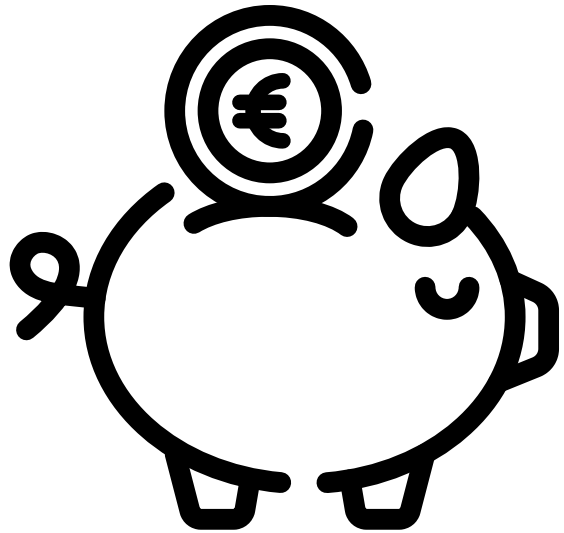
WARUM LOKAL?

- Kostenpunkt
- Privacy / Datenkontrolle

Kosten

Externe KI-Anbieter

- Chatgpt Premium → 23€ / User / Monat
- Gemini Premium (Google) → 22€ / User / Monat
- Microsoft Copilot → 22€ / User / Monat

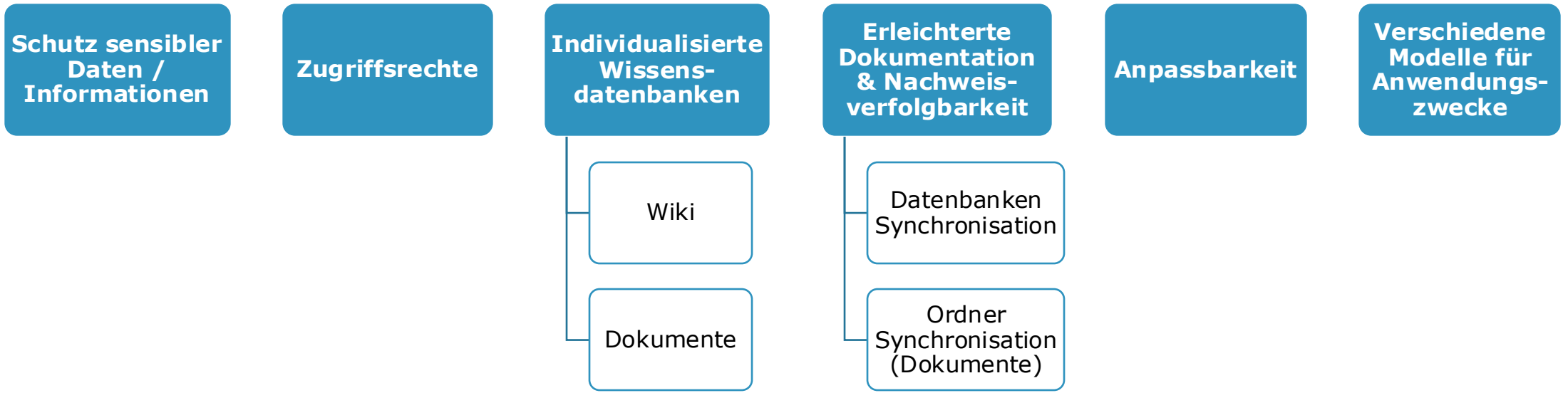


Lokale KI

Anschaffungskosten (variieren) | Beispiel:

Server + RTX 3090	3000€
Unterhalt	ca. 500€ / Jahr, bei 8h maximal Auslastung / Tag

Datenkontrolle

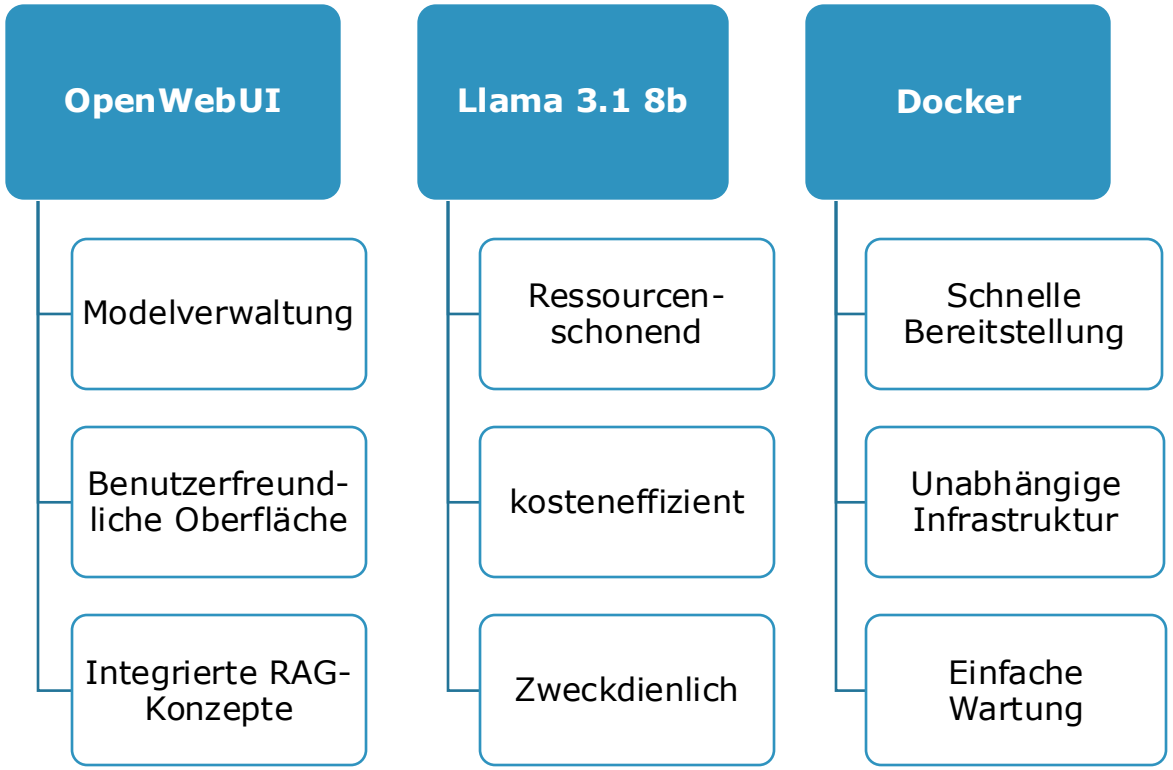


/// 4.

WAS NUTZEN WIR?

- Open Web UI
- Ollama / Llama
- Docker

Warum nutzen wir es?



/// 5.

WISSEN INTEGRIEREN

- RAG
- Training / fine-tuning

Training / fine-tuning

- Daten müssen aufbereitet werden (Textform)
- Tokenizer
- Fine-Tuning Frameworks
  - Huggingface
  - PyTorch
- Lernrate
- Batchgröße
- Epochen
- Dropout Rate
- Gewichtsinitialisierung

Vorteile

- Abfragen sind schneller
- Verbesserte Datenqualität für übergreifende Abfragen
- Antworten werden aus Wissen generiert

Nachteile

- Datenaufbereitung ist aufwendiger
- „Richtige" Initialisierung des Trainings ist komplex
- Lange Trainingszyklen
- Wissen bleibt nach Veränderung bestehen

Retrieval-Augmented-Generation

Vorteile

- Kein „Training“ benötigt
- Wissen kann in verschiedenen Medien dokumentiert sein
- Bei Änderung einer Dokumentation ist diese „sofort“ verfügbar
- KI-Modell wird „nicht“ vergrößert

Nachteile

- Datenabfrage dauert länger
- Datenabfrage kann verfälscht werden wenn mehrere Einträge ähnliche Informationen enthalten
- Antworten werden Großteils lediglich zitiert

/// 6.

AUSBLICK & WEITERE ANSÄTZE

- GPU / CPU Vergleich
- Demonstration
- Pipelines

Vergleich GPU / CPU

GPU

- Verarbeitung: parallel
- Anzahl der Kerne: (RTX3090) 10‘500
- Optimiert für Berechnungen und große Datenmengen

➡ **für datenintensive Aufgaben**

CPU

- Verarbeitung: seriell
- Anzahl der Kerne: 4 – 24
- Optimiert für Prozessverarbeitungen und Steuerungslogik

➡ **für prozessintensive Aufgaben**

VIELEN DANK!

DR. WALLNER ENGINEERING