RHEINSCHAFE

SESSION

DIE WICHTIGSTEN TREIBER DER KI-REVOLUTION: **LARGE LANGUAGE MODELS**

Mit Christian Loth, Philipp Müller und Julian Pohl

TYPO3 Association Silver Member

TCC C TCC D TCC I TCC E

DESIGN AWARD 2021

GERMAN DESIGN AWARD WINNER 2024

LOS GEHT'S

WAS SIND LLMs?

SPRACHE UND KONTEXT

Versteht Texte und kann diese deuten.

INHALTE GENERIEREN

Kann selbständig z.B. Code oder Text generieren.

"LARGE"

Basiert auf einer großen Datenbasis.

UND WIE?

WIE SETZT MAN LLMs EIN?

Fertige Software
– Integration über API
– Viele Auswahlmöglichkeiten
– ChatGPT, Gemini, Claude ...

Self Hosting
– Vortrainierte Open Source LLMs
– Finetuning
– RAG-System

UND WARUM NICHT CHATGPT?

WAS SPRICHT FÜR SELF-HOSTING?

- INVESTITIONSAUFWAND
- AUSFÜHRLICHE PLANUNG
- RESSOURCENINTENSIV
- BETREUUNG & WEITERENTWICKLUNG

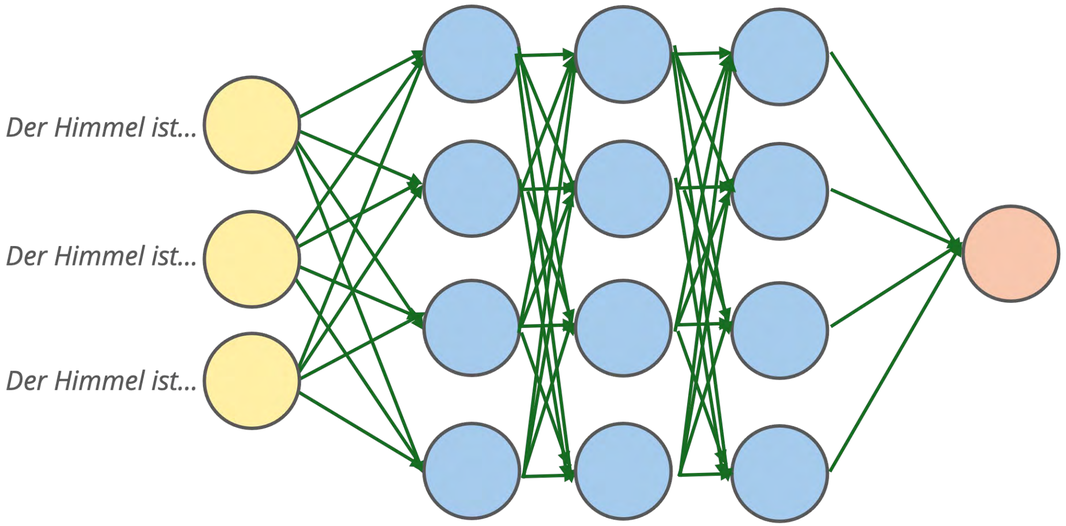
UND WARUM NICHT CHATGPT?

WAS SPRICHT FÜR SELF-HOSTING?

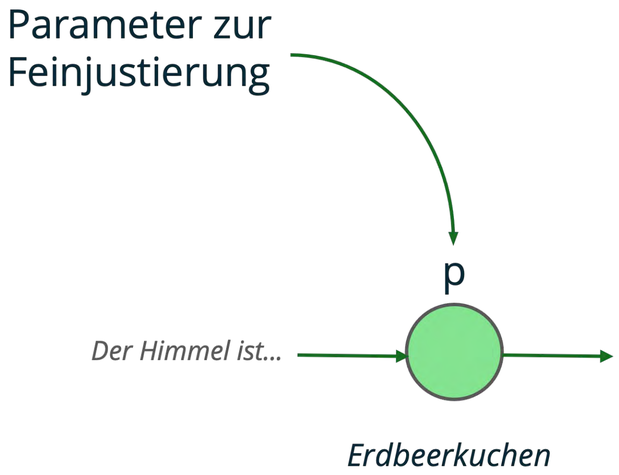
- ✔ DATENSCHUTZ
- ✔ KOSTENKONTROLLE
- ✔ FLEXIBILITÄT
- ✔ UNABHÄNGIGKEIT

NEURONALE NETZE

DAS INNERE EINER KI

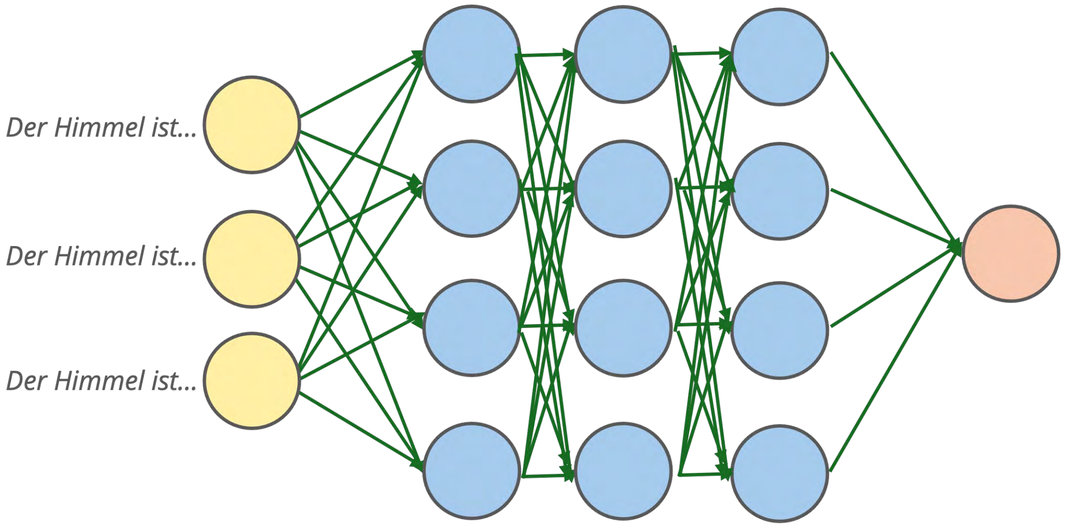


Input Layer Hidden Layers Output Layer

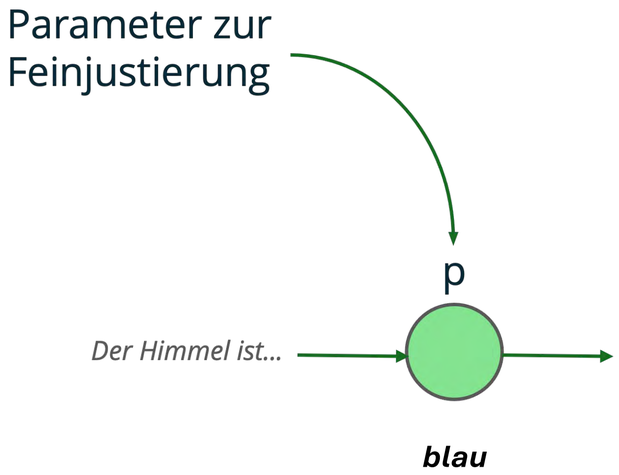


NEURONALE NETZE

DAS INNERE EINER KI



Input Layer Hidden Layers Output Layer

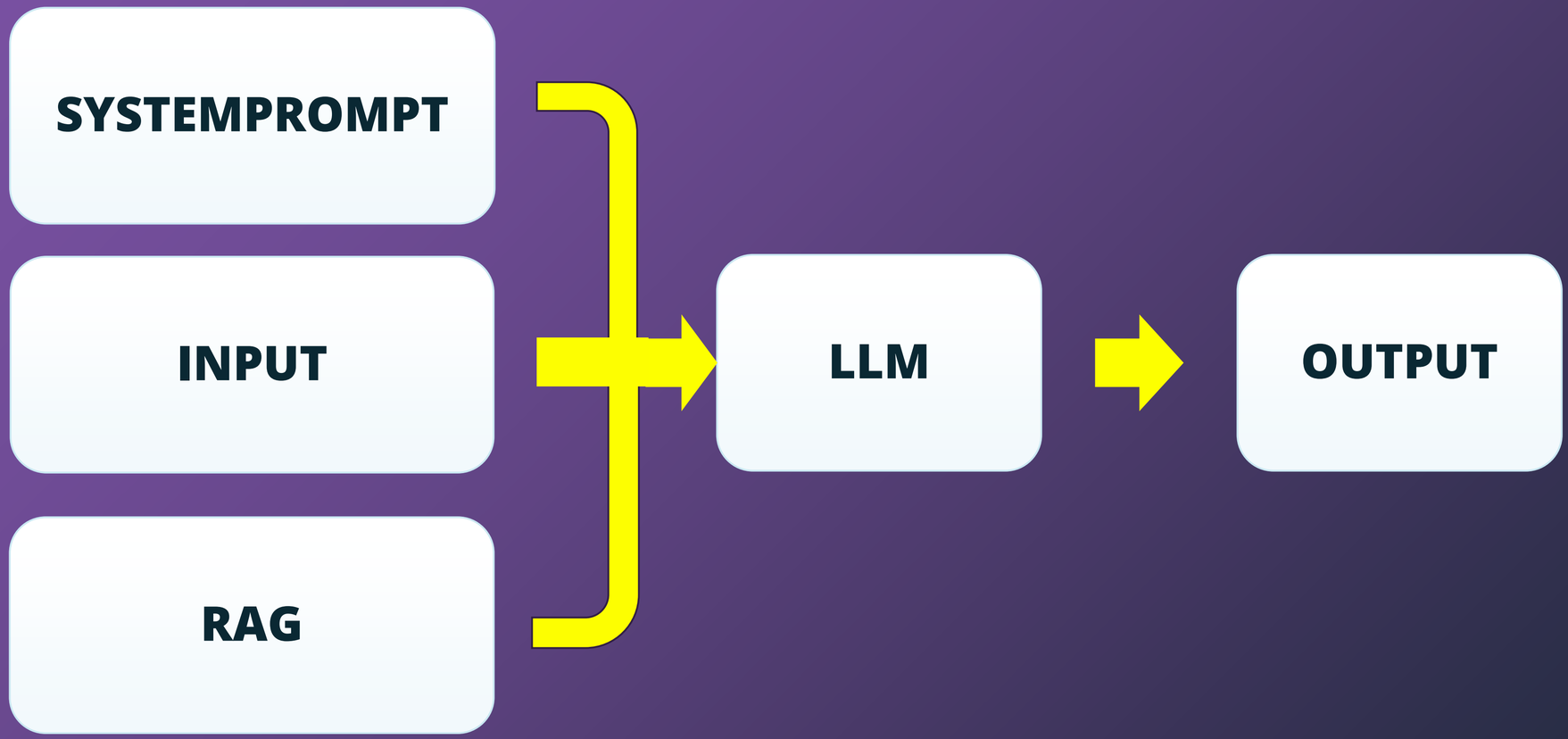


SPEZIALISIERUNG

LLMs ANPASSEN UND OPTIMIEREN

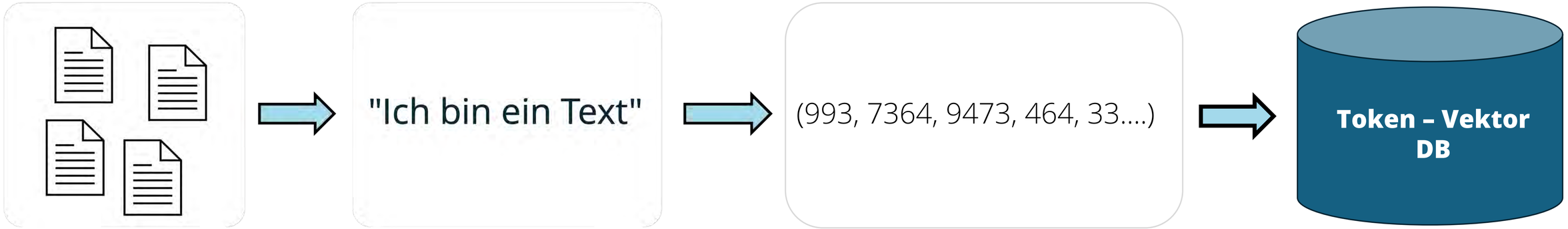
SPEZIALISIERUNG

LLMs ANPASSEN UND OPTIMIEREN



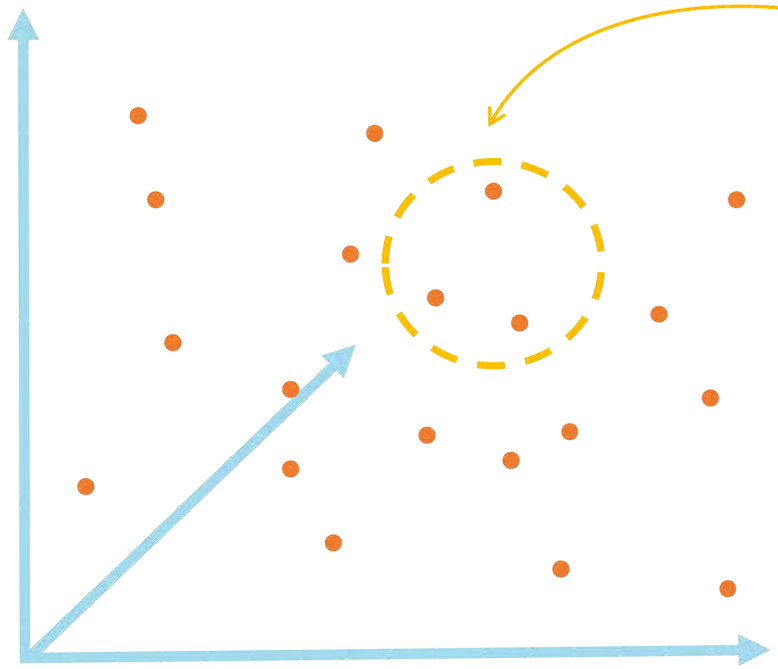
UND SO GEHT'S

WIE FUNKTIONIERT EIN RAG-SYSTEM?



UND SO GEHT'S

WIE FUNKTIONIERT EIN RAG-SYSTEM?



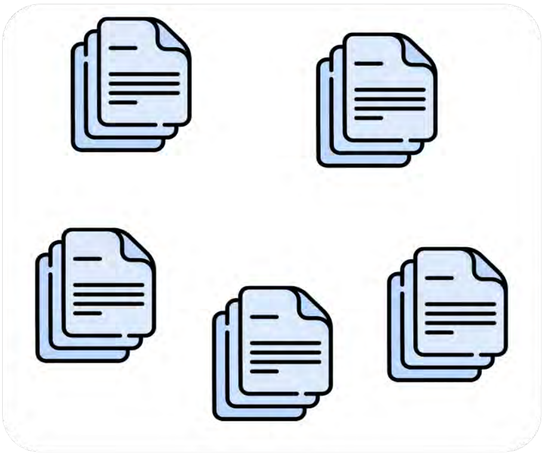
Ermittelte Quellen

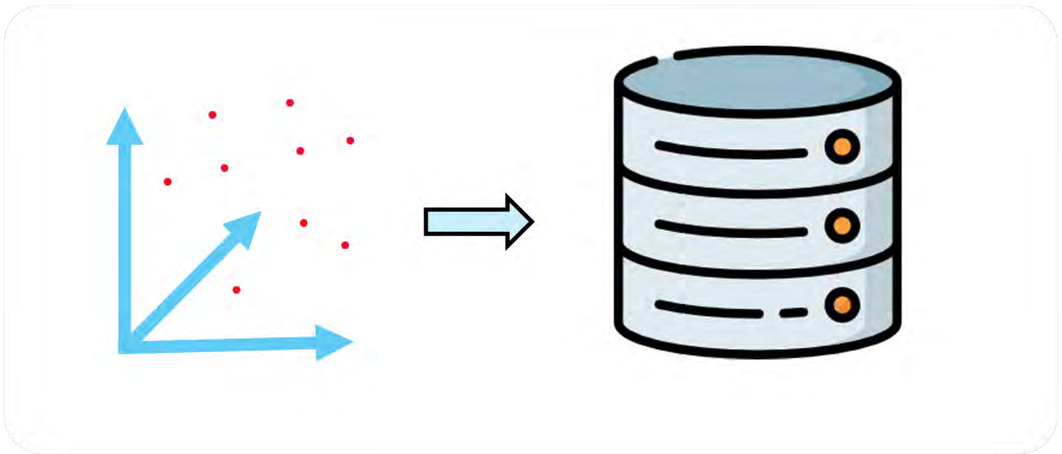
Hohe inhaltliche Überschneidung mit Prompt.

Vereinfachte Darstellung

UND SO GEHT'S

RAG-SYSTEM



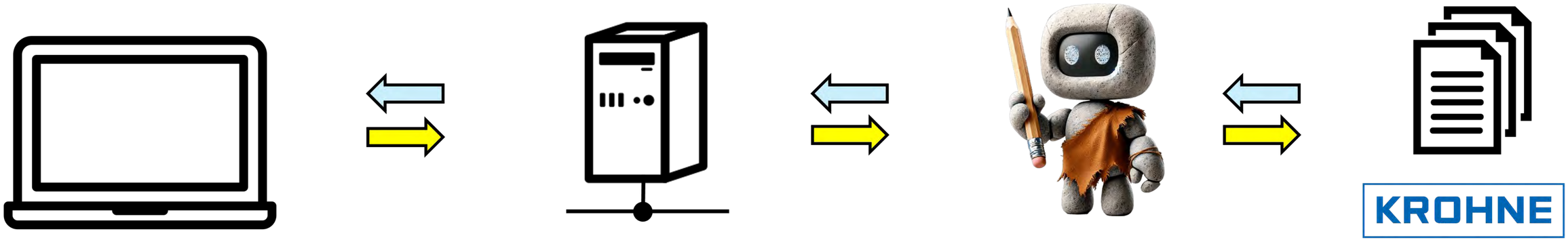


ANWENDUNG

DAFÜR KANN MAN LLMs VERWENDEN

PROJEKT

SO "EINFACH" GEHTS!



RAG-SYSTEM

MEHR ALS NUR EIN LLM

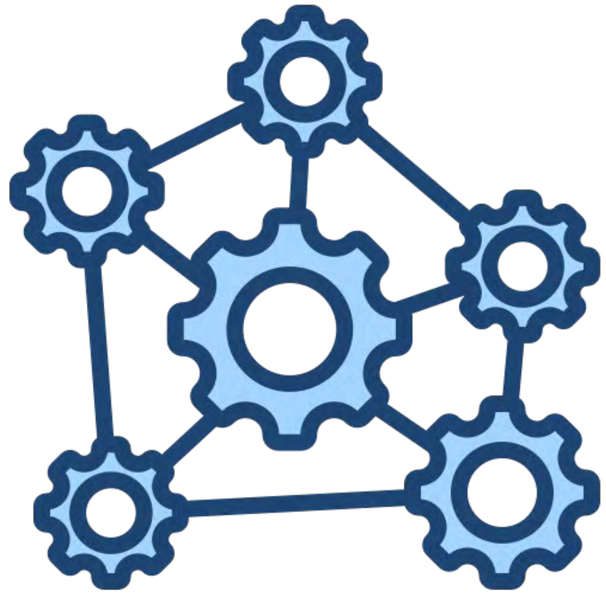
– Bausteine: Modelle

– Auswertung der Modelle ergeben Gesamtapplikation

Modelle für Knowledgebase:

– Embeddings (Token ➞ Vektoren): BAAI/bge-large-en-v1.5

Spracherkennung:
– papluca/xml-roberta-base-language-detection



RAG-SYSTEM

MEHR ALS NUR EIN LLM

Modelle für Folgeprozesse:

– Embedding des Prompts: BAAI/bge-large-en-v1.5

– Generierung der Antwort: Qwen/Qwen2-1.5B-Instruct

– Generierung Keywords: BAAI/bge-large-en-v1.5

– Scoring der Dokumente: BAAI/bge-reranker-v2-m3