LLMs in Eigenregie

Individualisierung & Hosting von KI-Modellen

IT After Work

Marian Lambert

Drogeriemarkt dm führt firmeneigenes ChatGPT ein

- Handelsblatt, 25.08.2023

BoschGPT: Aleph Alpha will bis Jahresende KI für Bosch-Mitarbeitende einführen

- t3n.de, 14.08.2023

McKinsey & Company launches internal generative AI tool Lilli

- consultancy-me.com, 29.08.2023

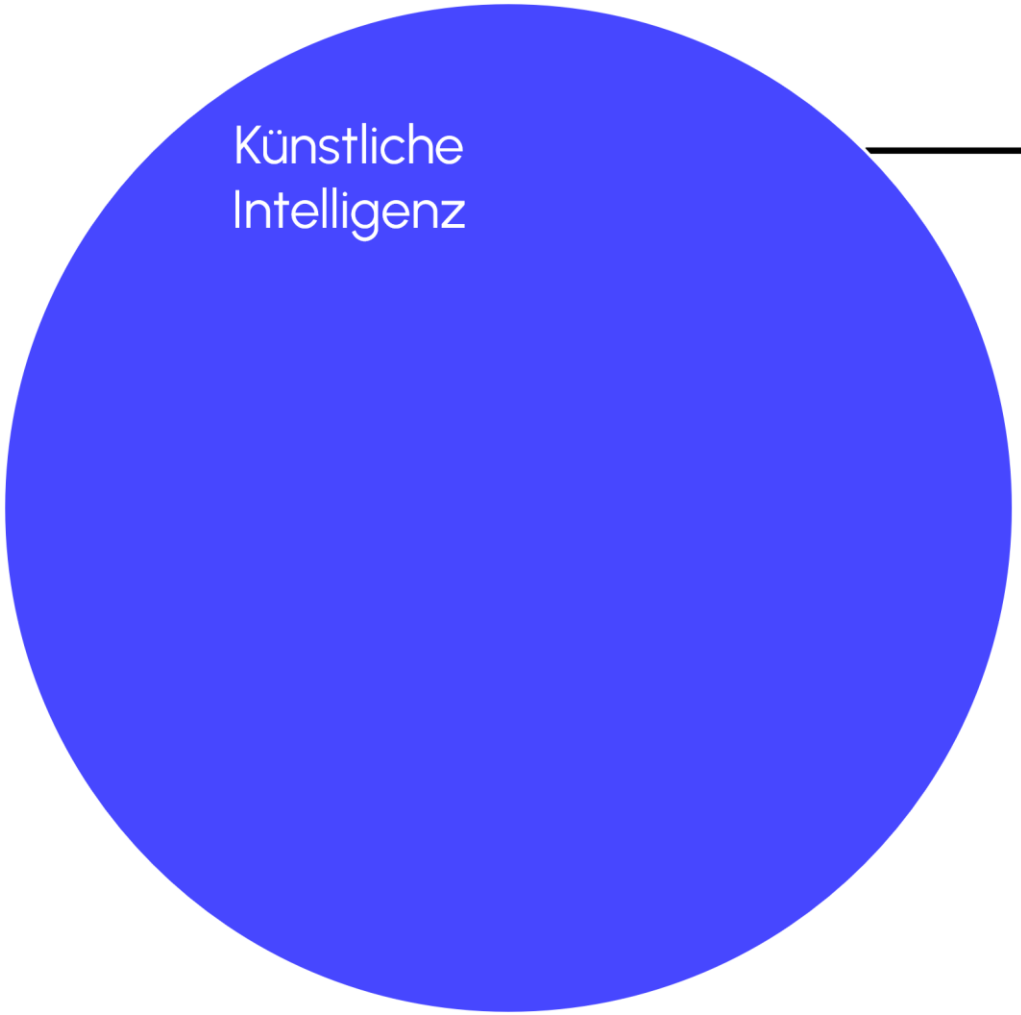
Mercedes-Benz Direct Chat: an internal ChatGPT application for employees

- businesswire.com, 30.10.2023

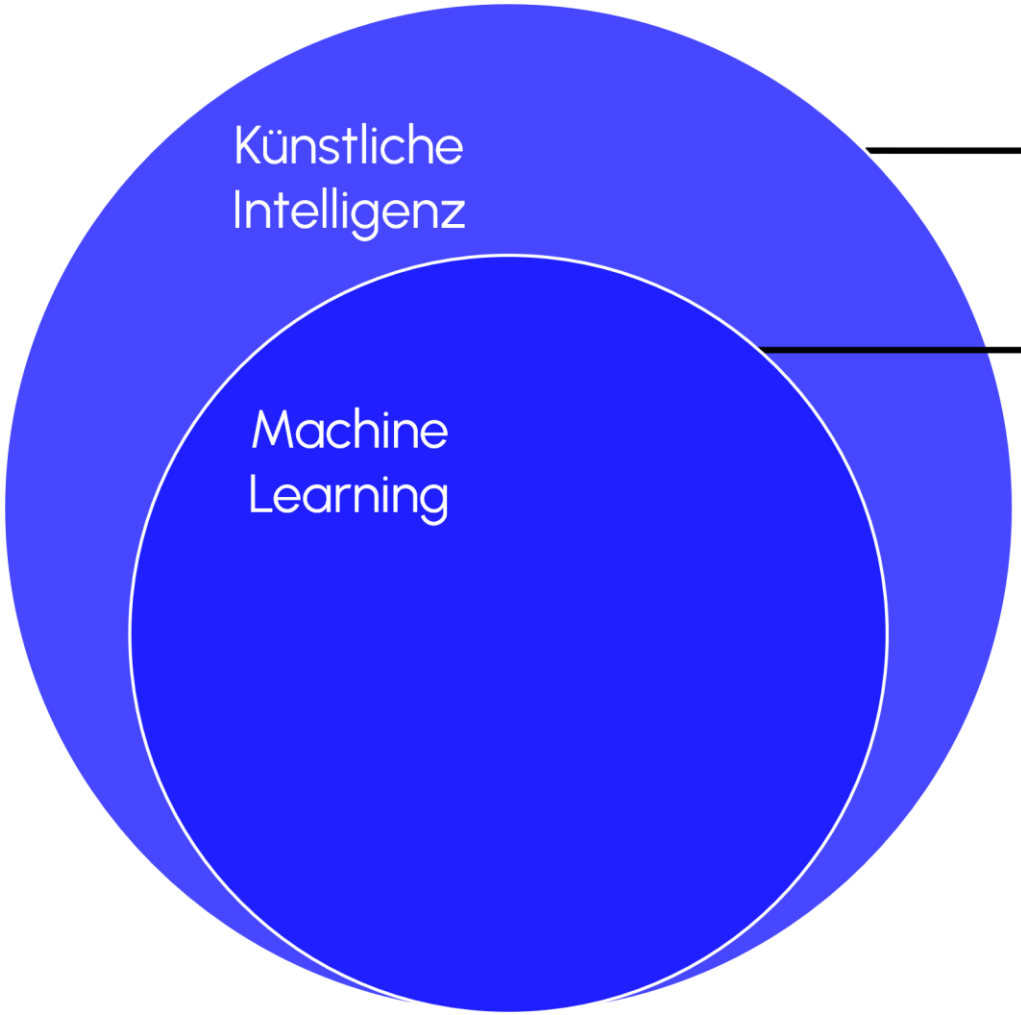
Warum denn überhaupt ein „eigenes“ LLM?

- Datenschutz und -sicherheit
- Anpassung auf unternehmensinterne Anwendungsfälle
- Leistungssteigerung durch Zugriff auf unternehmensinterne Daten
- Volle Kontrolle über Modell
- Integration in andere IT-Systeme
- Niedrigere OPEX
- …

Wie bekomme ich nun mein eigenes LLM?

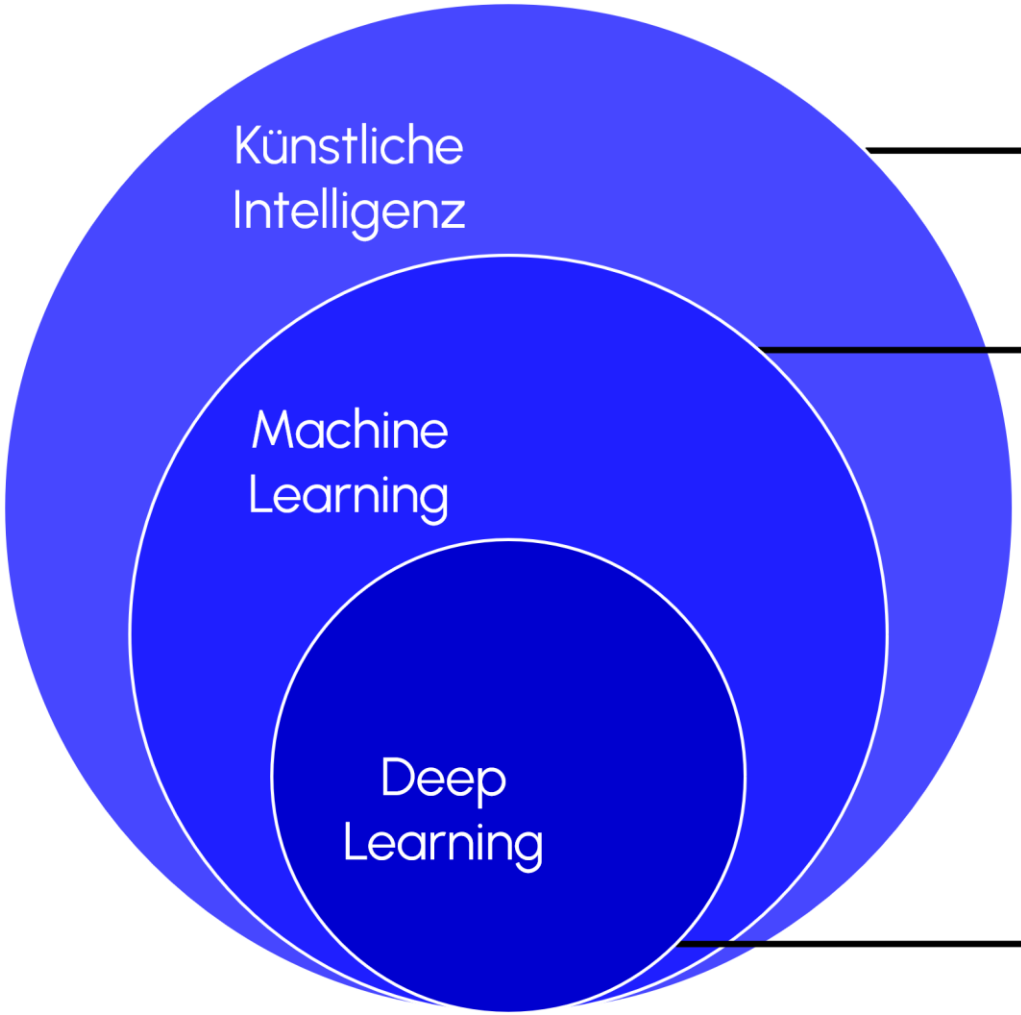


Systeme, die menschliche Intelligenz simulieren



Systeme, die menschliche Intelligenz simulieren

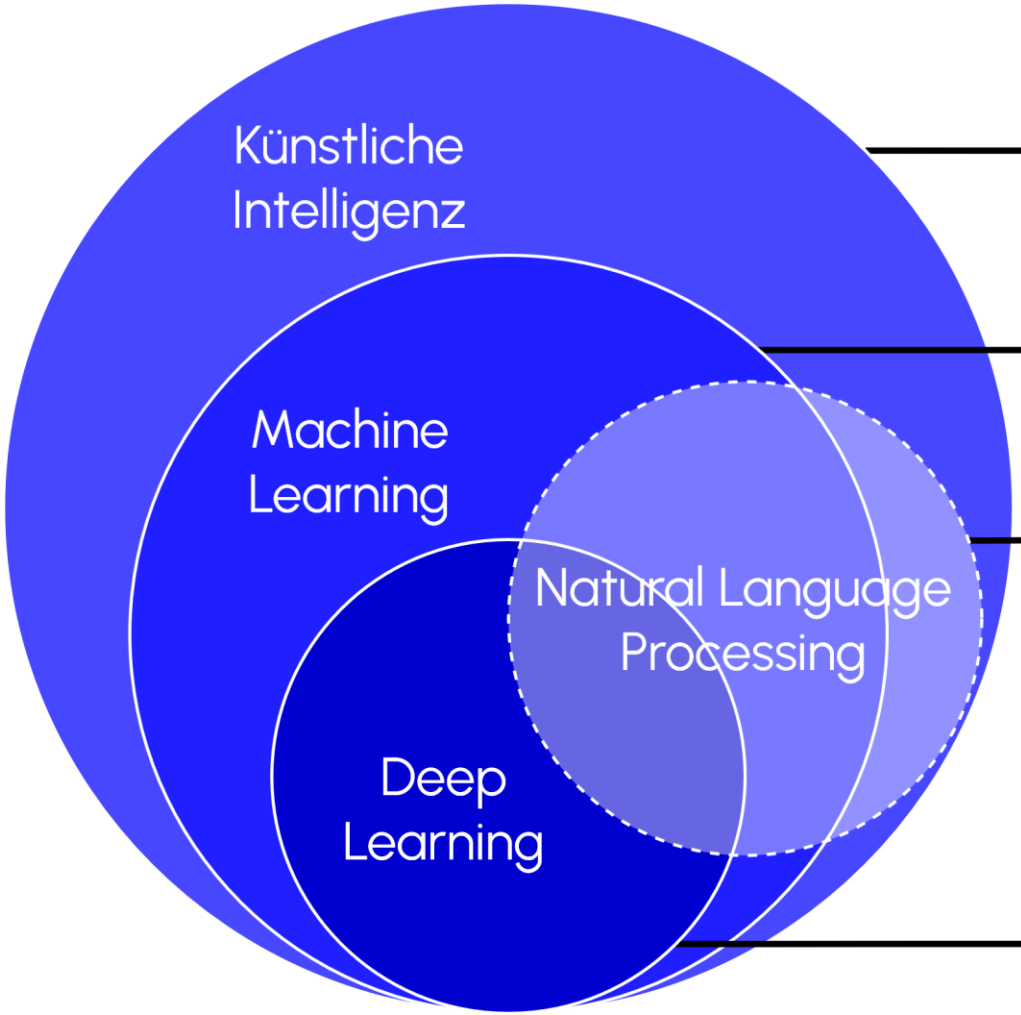
Algorithmen, die aus Daten lernen



Systeme, die menschliche Intelligenz simulieren

Algorithmen, die aus Daten lernen

Neuronale Netze, die hierarchisch aus Daten lernen

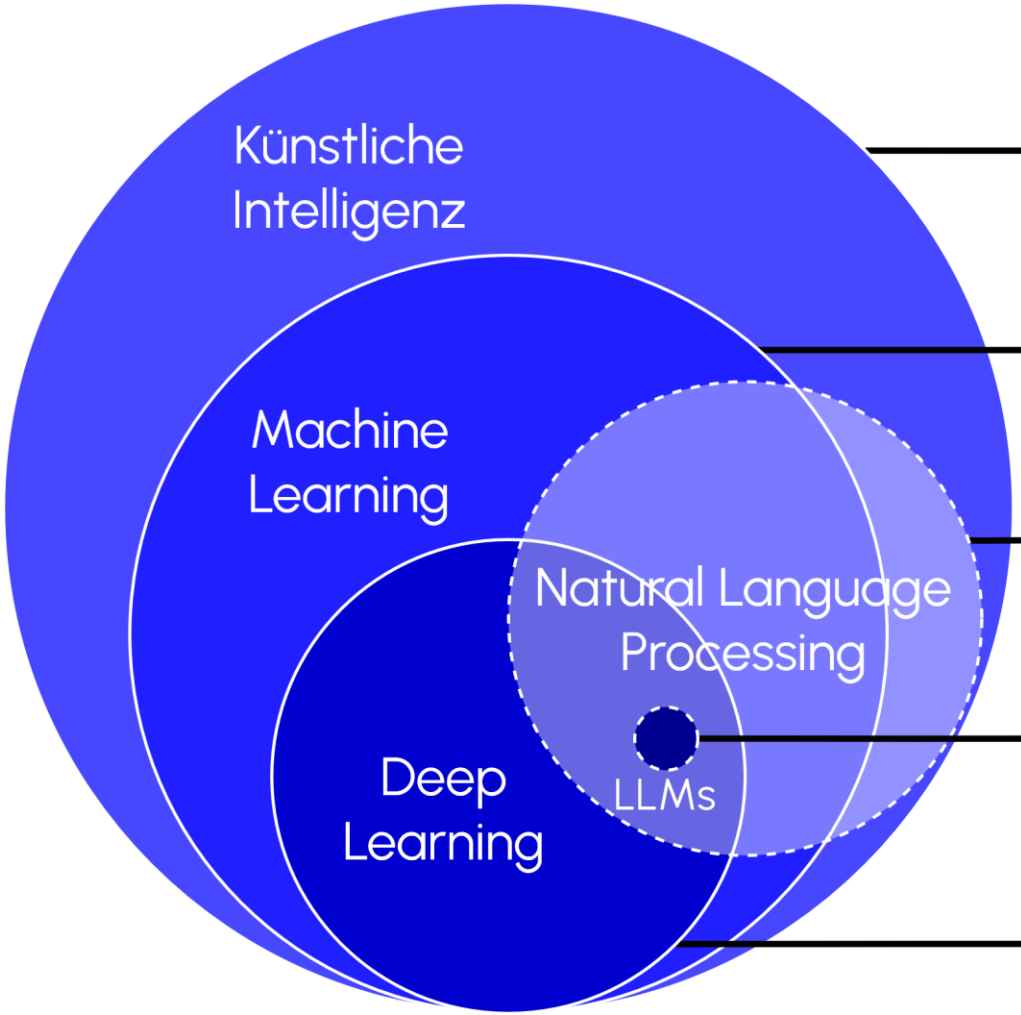


Systeme, die menschliche Intelligenz simulieren

Algorithmen, die aus Daten lernen

Systeme, die menschliche Sprache verarbeiten

Neuronale Netze, die hierarchisch aus Daten lernen



Systeme, die menschliche Intelligenz simulieren

Algorithmen, die aus Daten lernen

Systeme, die menschliche Sprache verarbeiten

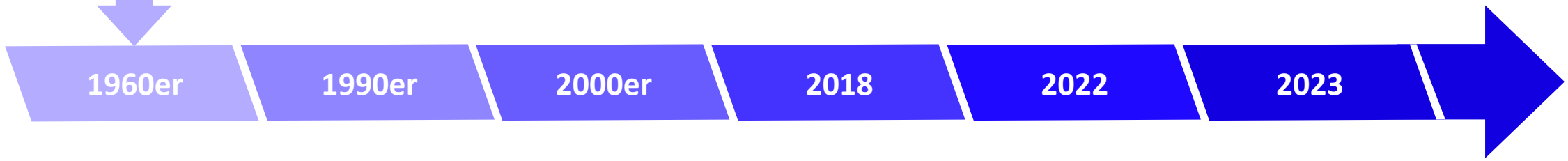
Neuronale Netze, die menschliche Sprache verstehen und generieren

Neuronale Netze, die hierarchisch aus Daten lernen

Von ELIZA zu ChatGPT und weiter…

Regelbasierte Systeme

Manuell erstellte Regeln für die maschinelle Verarbeitung und Generierung menschlicher Sprache



ELIZA

Das erste System zur maschinellen Verarbeitung und Generierung natürlicher Sprache

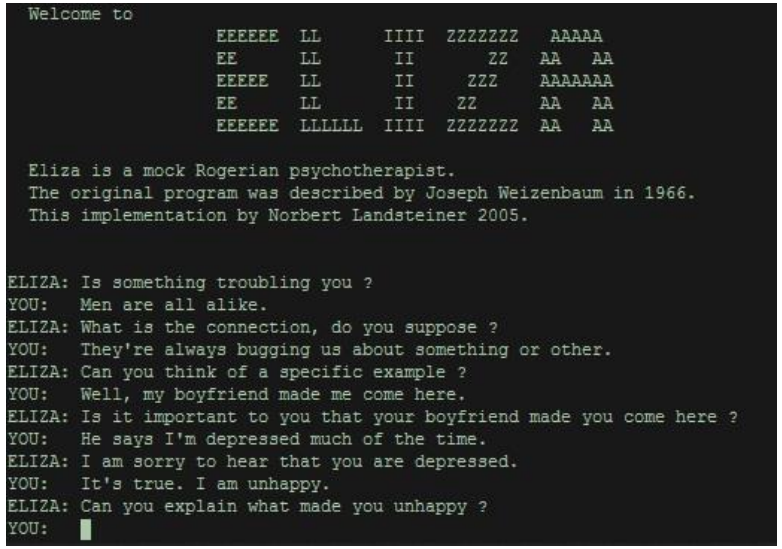
- 1966 von Joseph Weizenbaum entwickelt
- Regelbasierte Funktionsweise: Erkennung von Themengebieten und Antwort mit Standardphrasen

„Ich habe ein Problem mit meinem Vater.“

„Erzählen Sie mir mehr über Ihre Familie!“

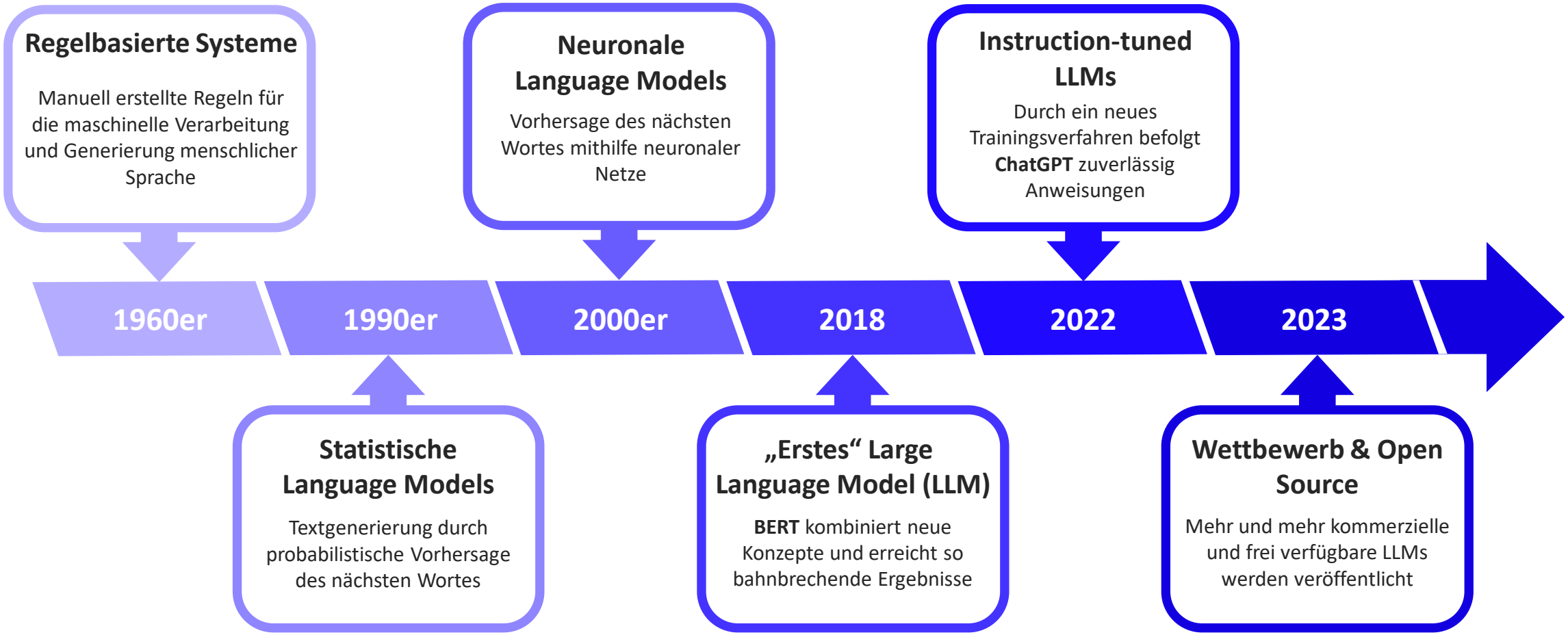
„Krieg ist der Vater aller Dinge.“

„Erzählen Sie mir mehr über Ihre Familie!“



Turing-Test: OpenAIs ChatGPT verliert gegen Sprachmodell aus den 60ern

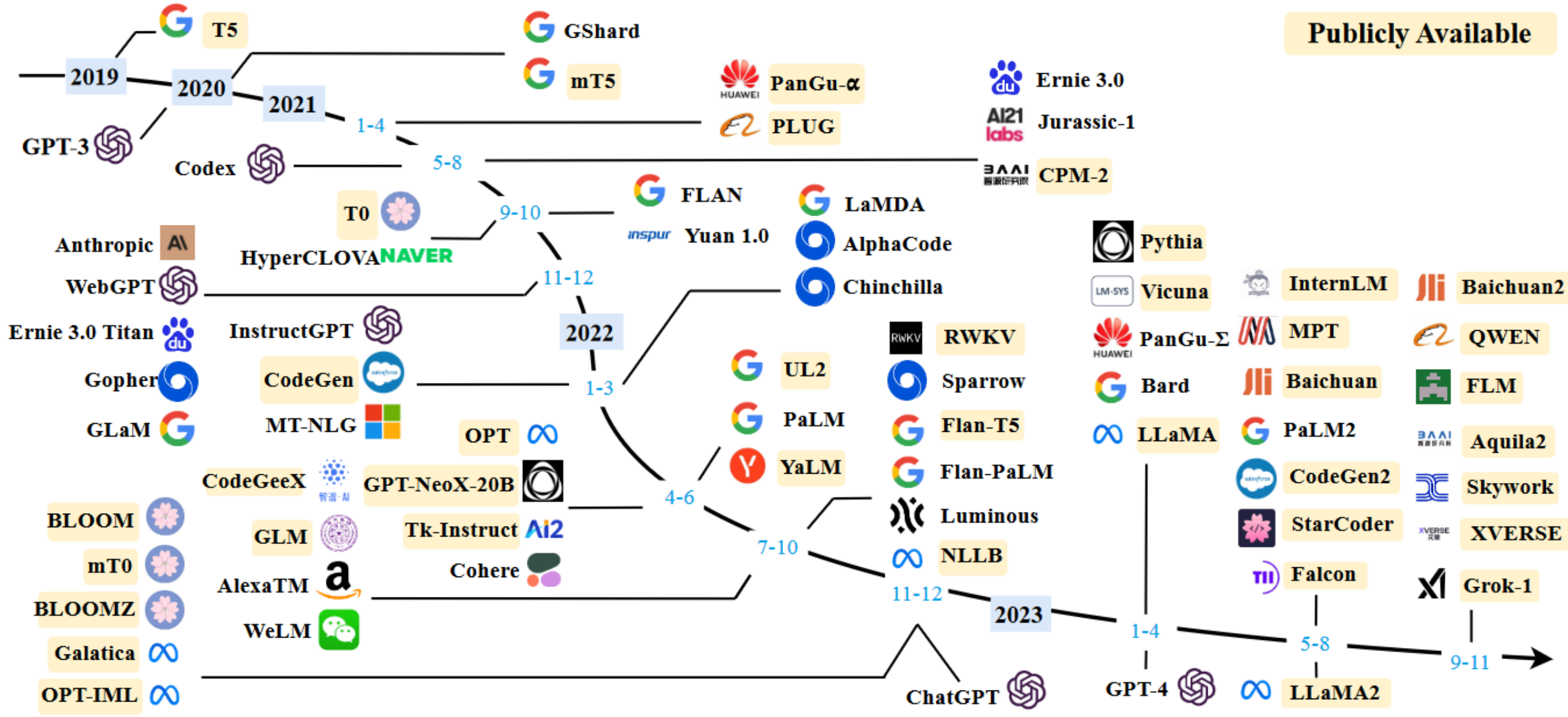
Von ELIZA zu ChatGPT und weiter…



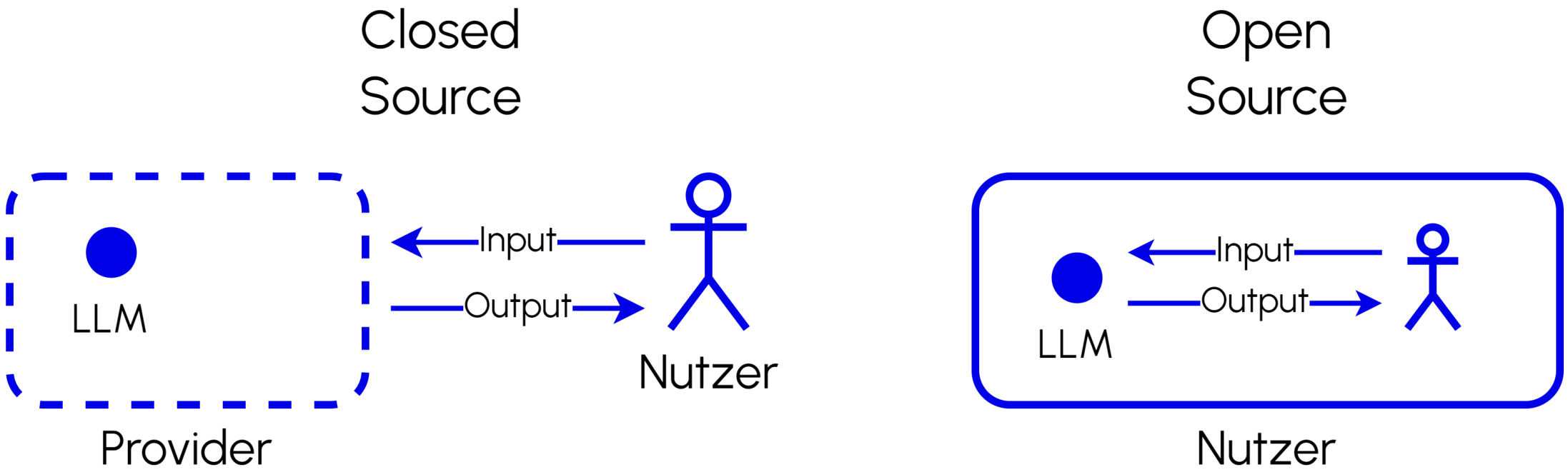
Demo

Das richtige LLM als Basis

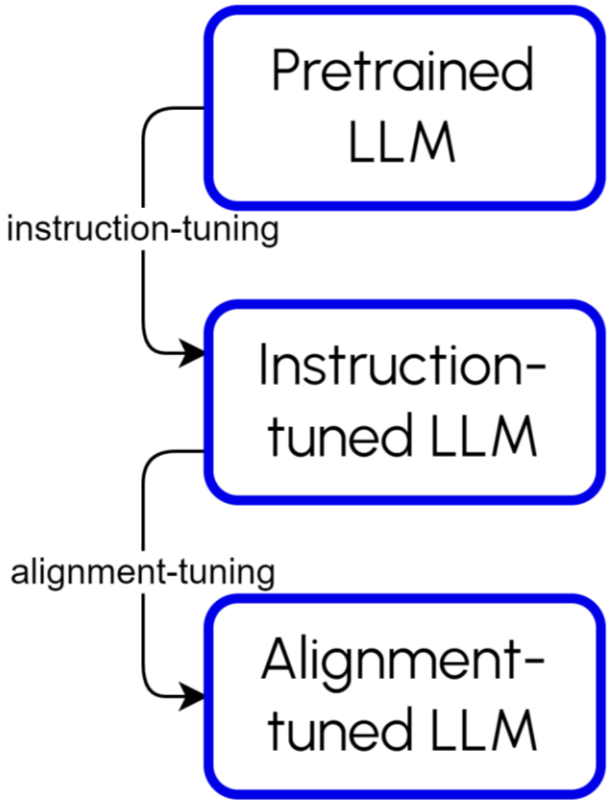
Zhao, Wayne Xin, et al. "A survey of large language models." *arXiv preprint arXiv:2303.18223* (2023).



Proprietär vs. Open Source



Pretrained vs. Instruction-tuned vs. Alignment-tuned



	Darauf trainiert ...	**Gut für bspw. ...**	**Beispiele**
Pretrained LLM	den nächsten Token vorher-zusagen	Question Answering, Autocompletion, Content Generation, ...	GPT-3, BERT, RoBERTa
Instruction-tuned LLM	Anweisungen zu befolgen	Code Generation, Extraktionsaufgaben, interne Anwendung, ...	Mistral, Zephyr
Alignment-tuned LLM	hilfreich, ehrlich und harmlos zu sein	Customer-facing Chatbots, Education, Mental Health, ...	GPT-3.5, GPT-4, LLaMA 2, Luminous

Weitere Faktoren

- **Leistungsfähigkeit** (Sprachverständnis, Reasoning, Kreativität)
- **Latenz**
- **Datenschutz**
- **Kosten**
- Kontextgröße
- Datensatz (Bias, Fairness, Transparenz)
- Rechtliche Anforderungen (bspw. AI Act)
- …

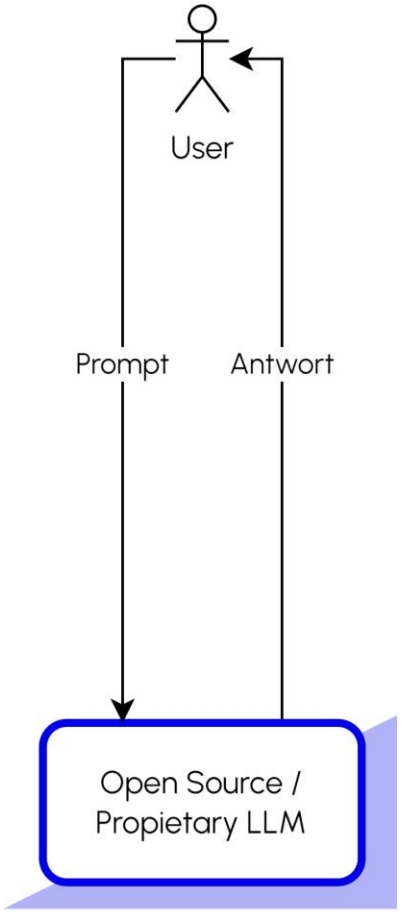
Bekannte LLMs im Überblick

Model	Organization	Typ	Context Window (Tokens)	Size (Parameters)	License
GPT-3.5-turbo	OpenAI	IT	4097, 16,385	175B	Proprietär
GPT-4	**OpenAI**	**IT**	**8192, 32768**	**1.8T (est.)**	**Proprietär**
LLaMA 2	Meta	PT/IT	4096	7B, 16B, 70B	NC Custom
StableBeluga 2	StabilityAi	IT	4096	70B	CC BY-NC-4.0
Claude 2	Anthropic	IT	200k	130B (est.)	Proprietär
Luminous	Aleph Alpha	PT/IT	2048	13B, 30B, 70B	Proprietär
PaLM 2	Google	IT	8000	340B (est.)	Proprietär
Falcon	Technology Innovation Institute	PT/IT	2048	7B, 40B, 180B	Apache-2.0
Mistral	Mistral AI	PT/IT	8000	7B	Apache-2.0
Zephyr	HuggingFace	PT/IT	8000	7B	MIT
Platypus 2	garage-bAInd	PT/IT	4096	7B, 13B, 70B	CC BY-NC-4.0
Godzilla 2	**Maya Philippines**	**IT**	**4096**	**30B, 70B**	**CC BY-NC-4.0**
Orca 2	Microsoft	IT	4096	7B, 13B	MRL
LeoLM	LAION	IT	8000	7B, 13B	NC Custom
Gemini	Google	IT	32k	?	Proprietär

PT=Pretrained; IT=Instruction/alignment-tuned

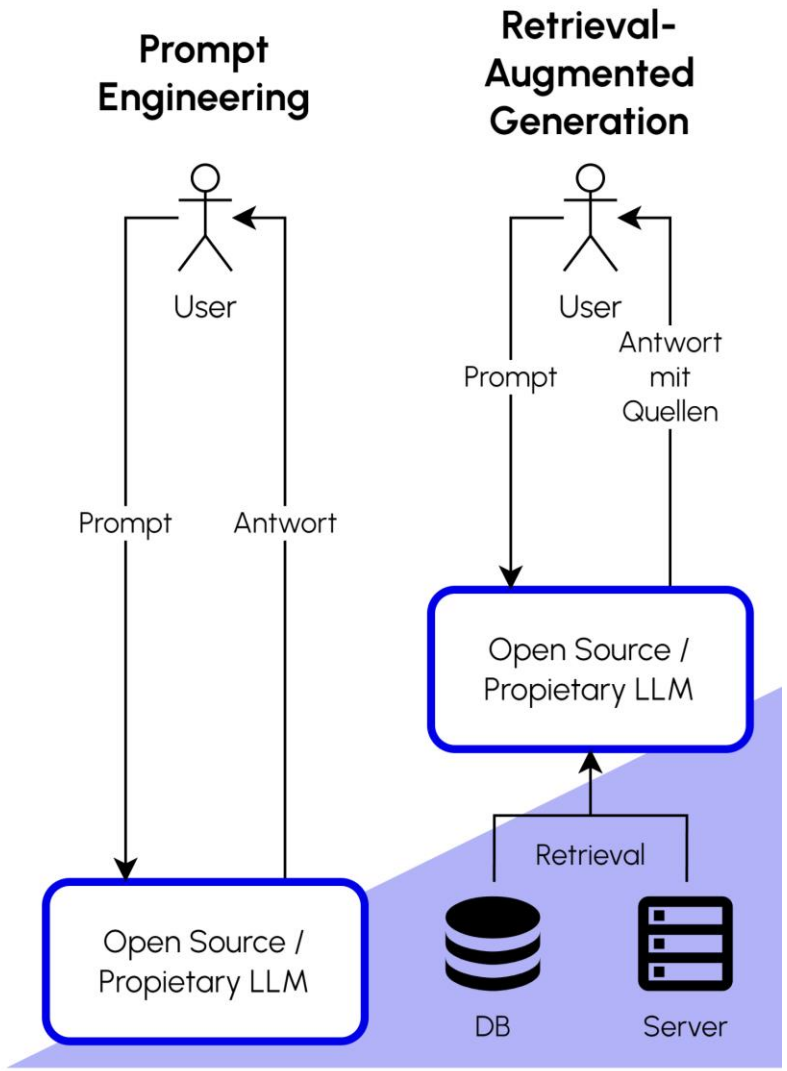
Anpassung an das Unternehmen

Prompt Engineering



Geeignet für ...

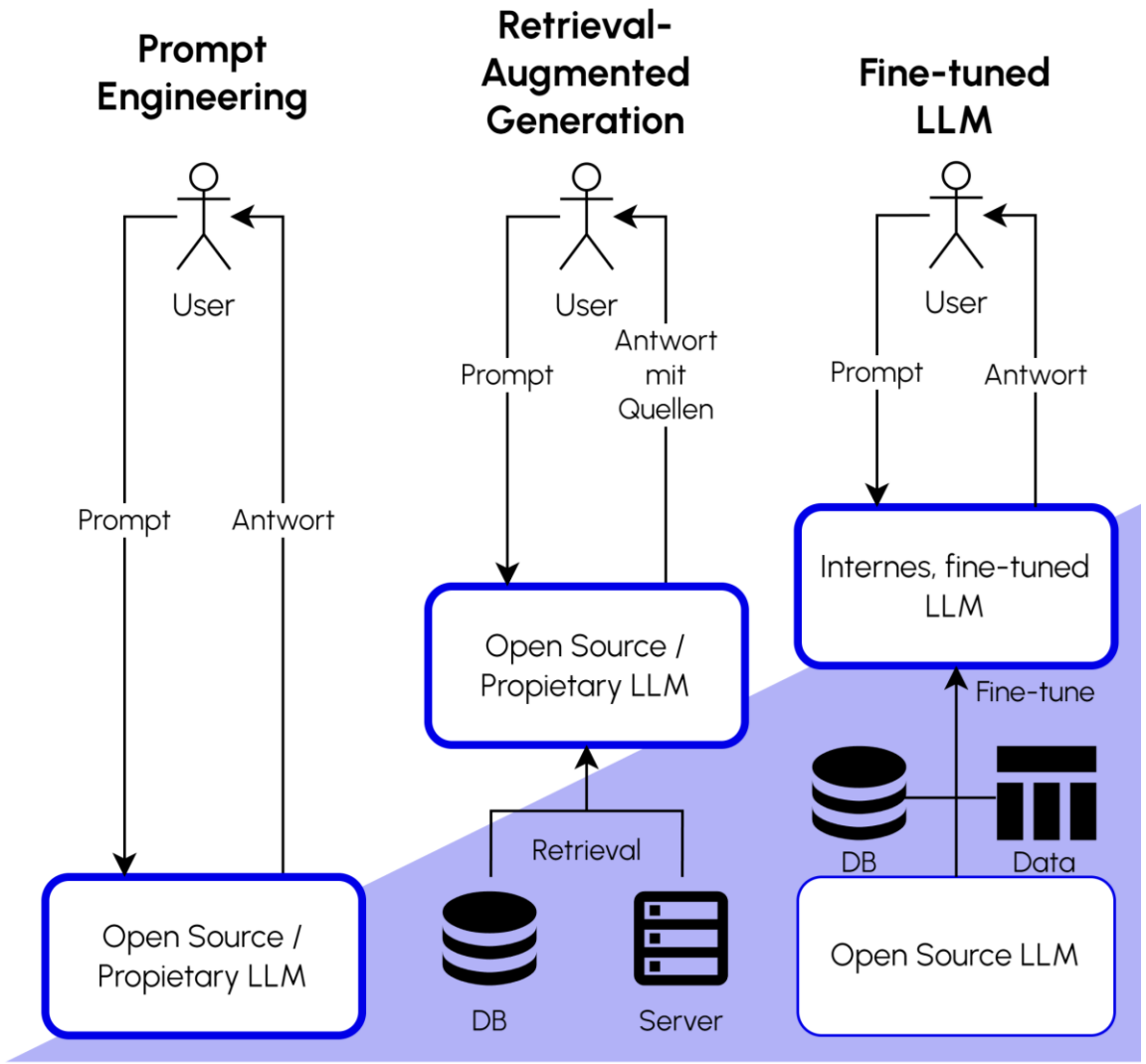
Aufgaben, die generische Fähigkeiten und Informationen erfordern



Geeignet für ...

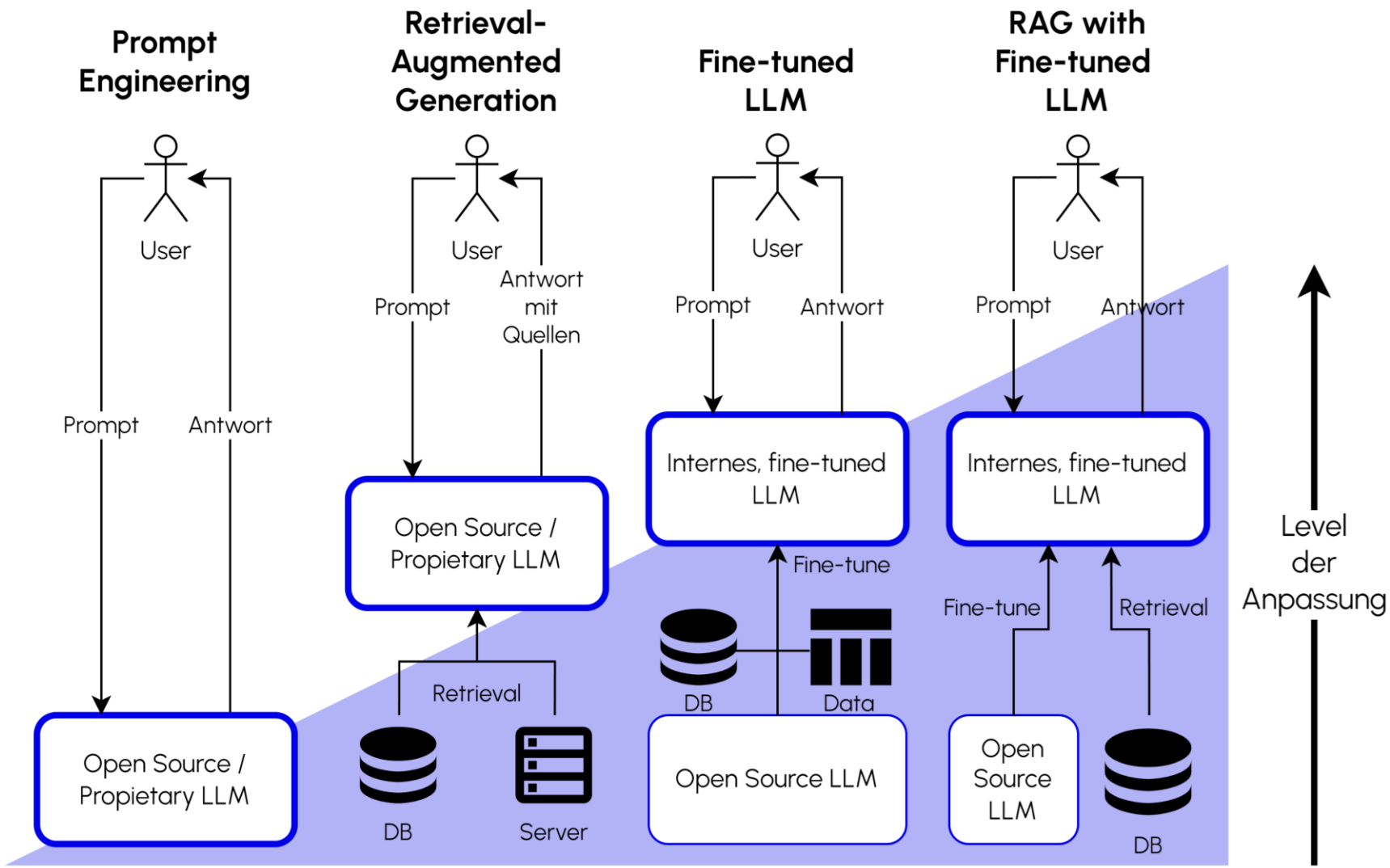
Aufgaben, die generische Fähigkeiten und Informationen erfordern

Aufgaben, die konkrete interne Informationen erfordern

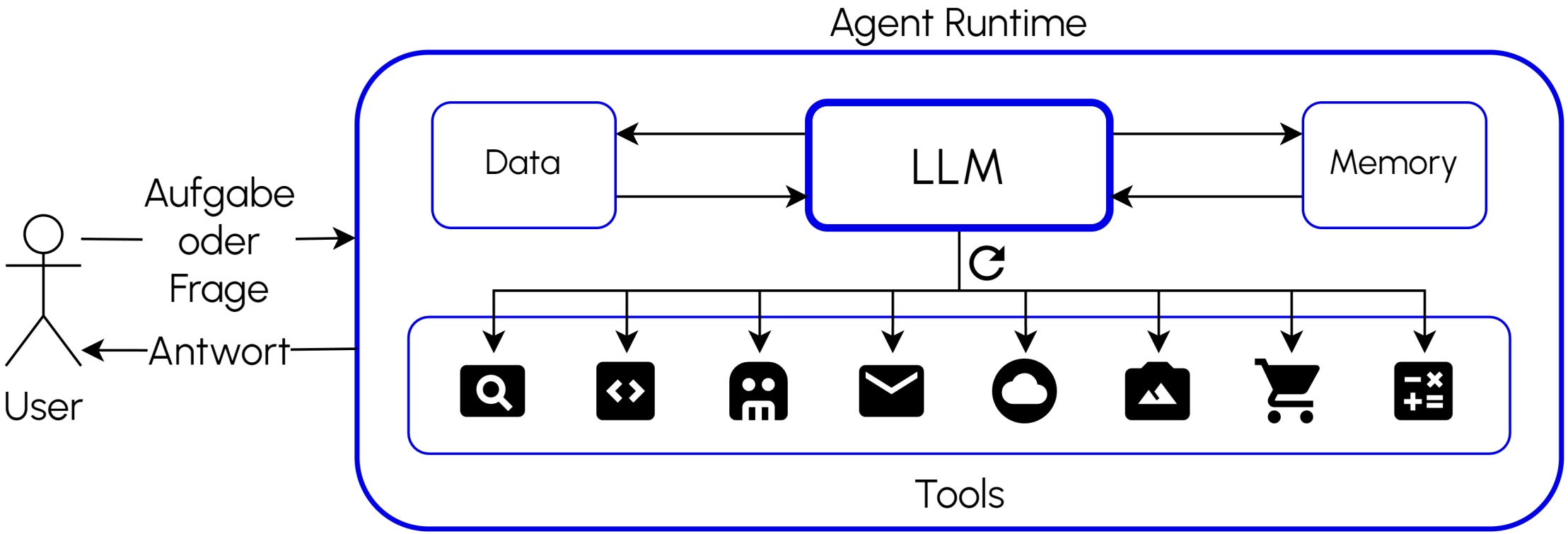


Geeignet für ...

Aufgaben, die generische Fähigkeiten und Informationen erfordern	Aufgaben, die konkrete interne Informationen erfordern	Aufgaben, die umfassendes Wissen interner Informationen, Sprache und Kontext erfordern


Prompt Engineering
User
Prompt
Antwort
Open Source / Propietary LLM
Retrieval-Augmented Generation
User
Prompt
Antwort mit Quellen
Open Source / Propietary LLM
Retrieval
DB
Server
Fine-tuned LLM
User
Prompt
Antwort
Internes, fine-tuned LLM
Fine-tune
DB
Data
Open Source LLM
RAG with Fine-tuned LLM
User
Prompt
Antwort
Internes, fine-tuned LLM
Fine-tune
Retrieval
Open Source LLM
DB
Level der Anpassung
Geeignet für ...
Aufgaben, die generische Fähigkeiten und Informationen erfordern
Aufgaben, die konkrete interne Informationen erfordern
Aufgaben, die umfassendes Wissen interner Informationen, Sprache und Kontext erfordern
Aufgaben, die umfassendes internes Wissen und einzelne Informationen erfordern

Von LLMs zu Agenten



Contact Me!

Marian Lambert

Senior Consultant AI & Software

marian.lambert@xpace.de

www.xpace.de

Danke!

Fragen?