

KI-Systeme für das Wissensmanagement aufbauen: CustomGPTs, RAG und Infrastrukturentscheidungen

ITSM Summit, Referent: Alexander Poschke
Hamburg, 24.09.2025





Quelle: Pexels, <https://www.pexels.com/>

LLM sind Lernmodelle, die auf riesigen Datenmengen trainiert werden, um Sprache zu verstehen und zu generieren.

Grenzen:

- LLM haben eine Wissensgrenze (ab dem Zeitpunkt des Trainings) können keine Echtzeitdaten abgerufen werden.
- LLM verstehen nicht wirklich Sprache, sondern arbeiten auf Basis von Mustererkennung und Wahrscheinlichkeiten.
- Die Modelle können Verzerrungen aus den Trainingsdaten übernehmen.
- Es werden teilweise fehlerhafte oder widersprüchliche Informationen generiert, da keine Logiken hinterlegt sind.



Quelle: Pexels, <https://www.pexels.com/>

- Ein CustomGPT ist ein speziell auf die Bedürfnisse eines Unternehmens zugeschnittene Version eines GPT-Modells.
- Häufig spricht man auch von „Finetuning eines Sprachmodells“
- Es wird auf spezifische Datenquellen oder branchenspezifische Inhalte trainiert, um spezialisiertes Wissen und maßgeschneiderte Antworten zu liefern.

CustomGPT

vs.

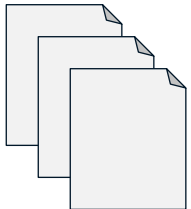
RAG-Ansatz

CustomGPT

vs.

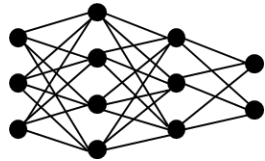
RAG-Ansatz

Eigene Dokumente
und Daten



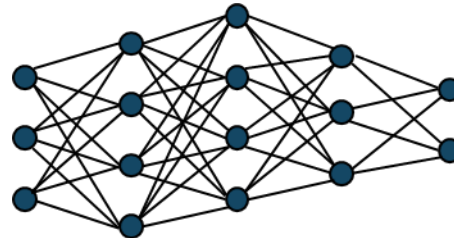
Vortrainiertes
Sprachmodell

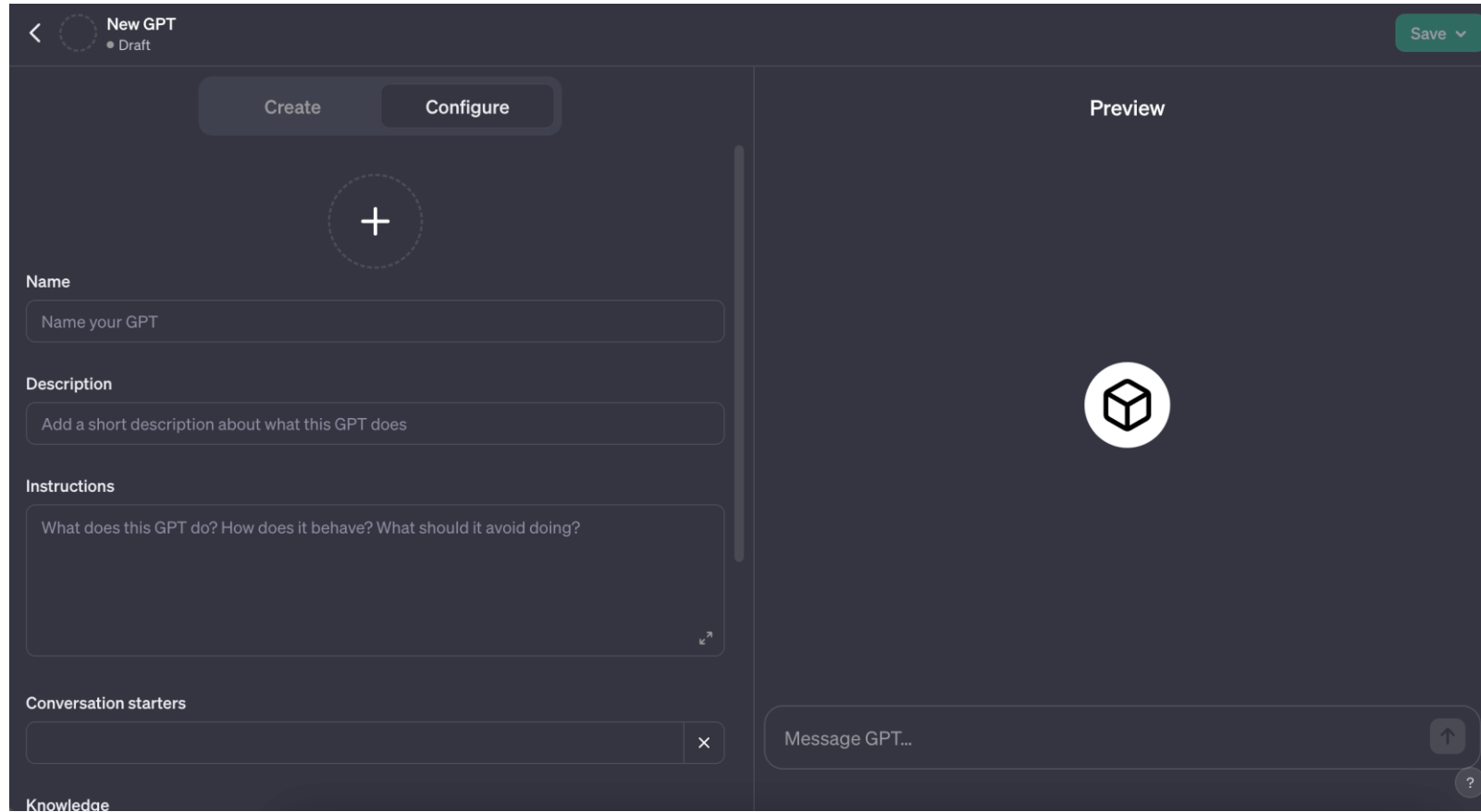
+



=

Auf Basis der eigenen Daten
trainiertes Sprachmodell





The screenshot shows the 'New GPT' configuration interface. The top bar includes a back arrow, the title 'New GPT' with a 'Draft' status, and a 'Save' button. Below the title are 'Create' and 'Configure' tabs, with 'Configure' being the active tab. A large dashed circle with a plus sign is centered in the main configuration area. The configuration fields on the left include:

- Name:** A text input field with the placeholder 'Name your GPT'.
- Description:** A text input field with the placeholder 'Add a short description about what this GPT does'.
- Instructions:** A larger text input field with the placeholder 'What does this GPT do? How does it behave? What should it avoid doing?'.
- Conversation starters:** A text input field with a clear button (X).
- Knowledge:** A section header at the bottom of the configuration panel.

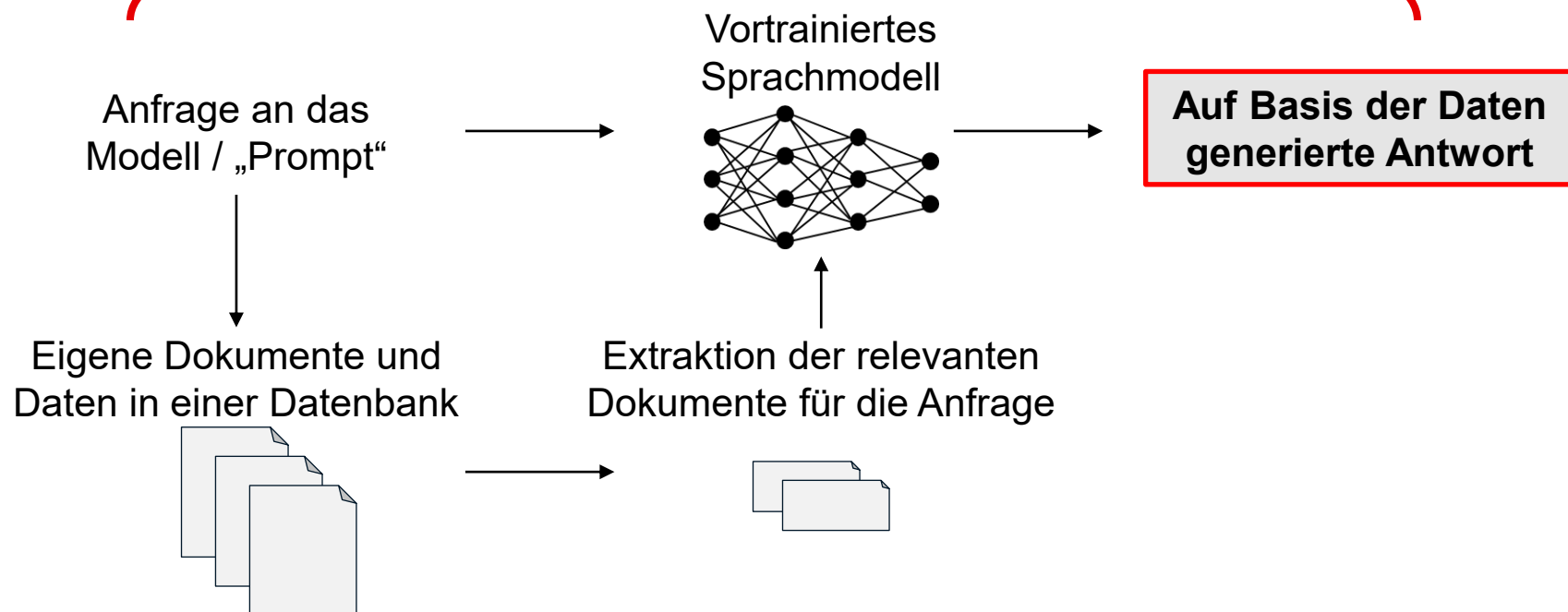
The right side of the interface is a 'Preview' section. It features a large circular icon with a cube inside, representing the GPT's avatar. At the bottom of the preview section is a text input field labeled 'Message GPT...' with a send button (upward arrow) and a help button (question mark).

- RAG steht für Retrieval Augmented Generation und kombiniert die Fähigkeit eines Modells, relevante Informationen aus einer externen Wissensbasis abzurufen (Retrieval), mit der Fähigkeit eines generativen Modells (wie GPT), auf Grundlage dieser Informationen neue Texte zu erstellen.
- Durch den Zugriff auf eine externe Wissensquelle kann RAG spezifischere und genauere Antworten liefern, besonders bei Anfragen, **die detailliertes oder aktualisiertes Wissen** erfordern.
- RAG ist besonders nützlich in Bereichen, die hochgradig spezialisiertes Wissen erfordern, indem es das generative Modell mit den relevanten Informationen unterstützt, die nicht im Modell selbst vortrainiert sind.
- **Der RAG-Ansatz ermöglicht es, das generative Modell dynamisch mit aktuellen oder spezifischen Daten zu versorgen, wodurch es anpassungsfähiger und vielseitiger wird, ohne dass das Modell selbst umfangreich nachtrainiert werden muss.**

CustomGPT

vs.

RAG-Ansatz



CustomGPT

Es ist besser kontrollierbar, was in das Modell fließt.

Wenn Daten hinzukommen, muss das Modell komplett neutrainiert werden.

Keine Nachverfolgung, woher die Informationen herkommen.

RAG-Ansatz

Es besteht hier die Gefahr, dass bei Übernahme kompletter bestehender Datenbanken viele schlechte Daten in der Datenbank landen.

Wenn Daten hinzukommen, muss das Modell nicht neutrainiert werden. Die Datenbank wird um die neuen Daten ergänzt.

Gute Nachverfolgung durch den Zugriff auf entsprechende Dokumente.

CustomGPT

vs.

RAG-Ansatz

lokal

vs.

online

	CustomGPT	RAG-Ansatz
online		
lokal		



+

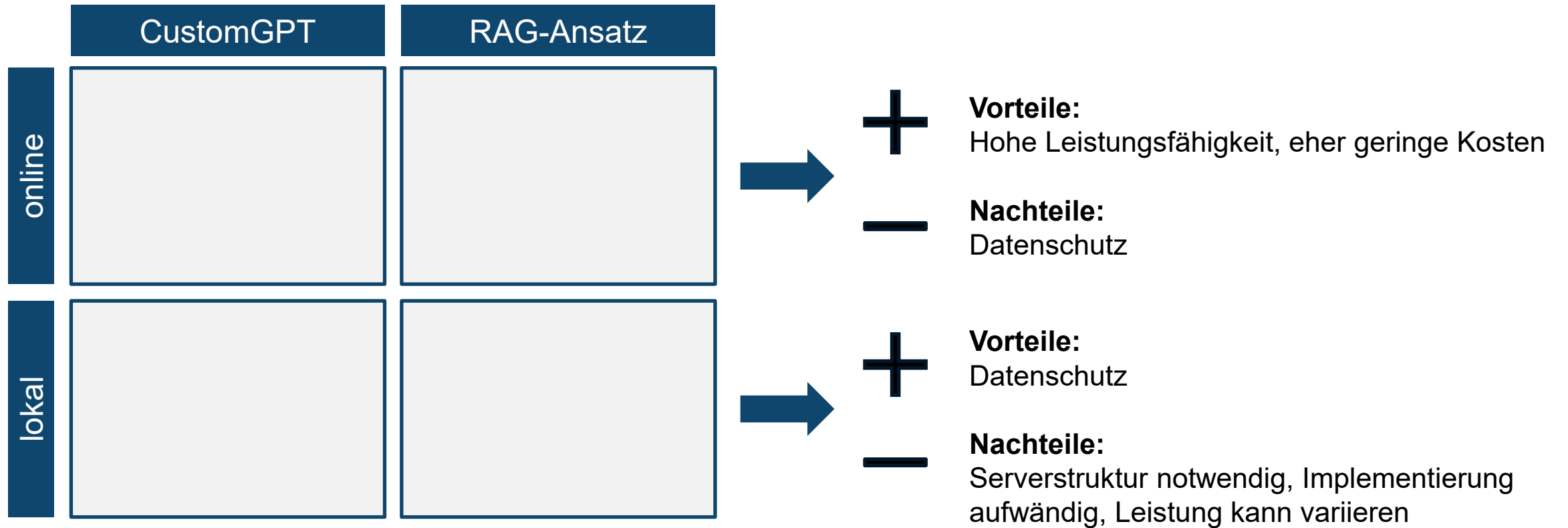
Vorteile:

Hohe Leistungsfähigkeit, eher geringe Kosten

—

Nachteile:

Datenschutz



DSGVO - Datenschutz-
grundverordnung

TDDDG - Telekommunikations-
Telemetrie-Datenschutz-Gesetz

KI-VO

UrhG - Urheberrecht

AGG – Allgemeines
Gleichbehandlungsgesetz

- KI-Systeme nutzen eine große Menge an **sensiblen Informationen und Daten**, um Muster zu erkennen und Prognosen zu treffen.
- Trainingsdaten können personenbezogene Daten beinhalten.
- Regeln der Datenschutzgrundverordnung (DSGVO) anwendbar

Vorsicht: Werden die Regeln des DSGVO nicht eingehalten, drohen **Geldstrafen von bis zu 10 Millionen Euro** oder 4 % des weltweiten Jahresumsatzes

DSGVO -
Datenschutz-
grundverordnung

TDDDG –
Telekommunika-
tions-Telemetrie-
Datenschutz-
Gesetz

KI-VO

UrhG -
Urheberrecht

AGG –
Allgemeines
Gleichbehand-
lungsgesetz

- Regelung von Datenschutz bei digitalen Diensten und Telekommunikation
- setzt die ePrivacy-Richtlinie um
- Schutz von Grundrechten und Privatsphäre

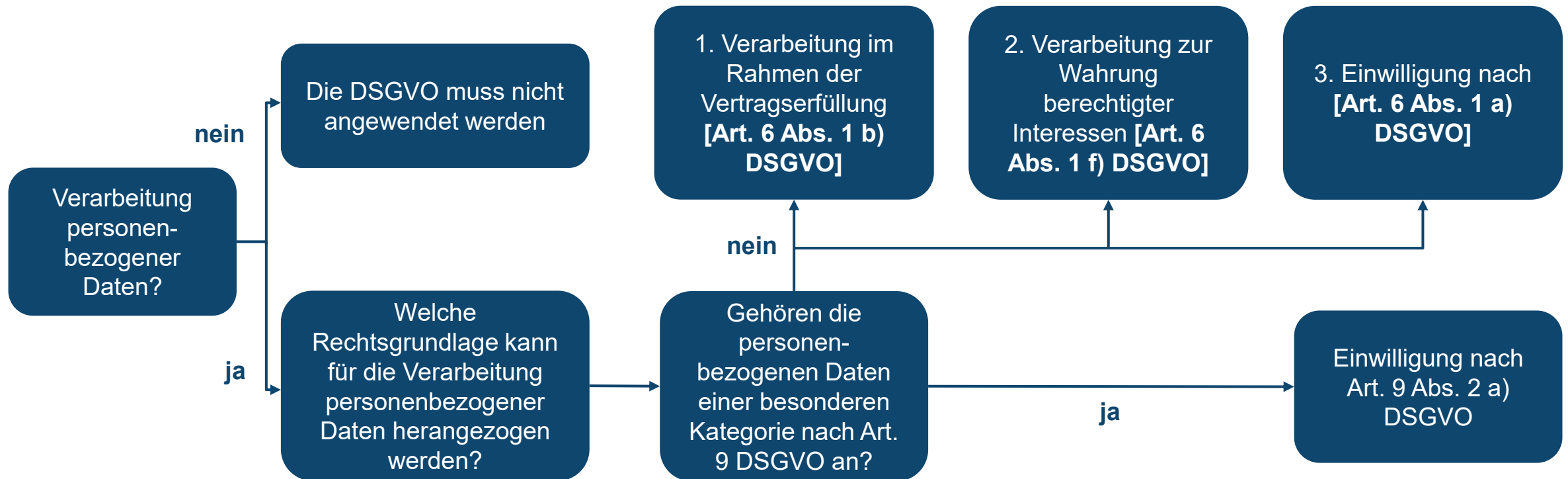
Beispiel: Ein KI-Chatbot erhebt Nutzungsdaten zur Echtzeit-Personalisierung.



„Im Sinne dieser Verordnung bezeichnet der Ausdruck:

1. "personenbezogene Daten" alle Informationen, die sich auf eine **identifizierte oder identifizierbare natürliche Person** (im Folgenden "betroffene Person") beziehen; als identifizierbar wird eine natürliche Person angesehen, die direkt oder indirekt, insbesondere mittels **Zuordnung zu einer Kennung** wie einem **Namen**, zu einer **Kennnummer**, zu **Standortdaten**, zu einer **Online-Kennung** oder zu einem oder mehreren besonderen **Merkmale**n, die Ausdruck der **physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität** dieser natürlichen Person sind, identifiziert werden kann;
2. [...]"

(1) Die Verarbeitung personenbezogener Daten, aus denen die **rassische und ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen oder die Gewerkschaftszugehörigkeit** hervorgehen, sowie die Verarbeitung von **genetischen Daten, biometrischen Daten** zur eindeutigen Identifizierung einer natürlichen Person, **Gesundheitsdaten** oder **Daten zum Sexualleben** oder der **sexuellen Orientierung** einer natürlichen Person ist untersagt.



Quelle: <https://digitalzentrum-chemnitz.de/wissen/ki-anwendungen-und-dsgvo/>

Rechtmäßigkeit, Verarbeitung nach Treu und Glauben, Transparenz

- Daten müssen rechtmäßig, fair und transparent für die betroffene Person verarbeitet werden

Zweckbindung

- Daten nur für spezifische, legitime und ausdrücklich angegeben Zwecke erheben und verwenden

Datenminimierung

- Nur absolut erforderliche Daten erheben

Richtigkeit

- Daten korrekt und aktuell

Speicherbegrenzung

- Aufbewahrung von Daten nicht länger als erforderlich

Integrität und Vertraulichkeit

- Personenbezogene Daten müssen sicher verarbeitet werden, um unbefugten oder unrechtmäßigen Zugriff, Verlust oder Beschädigung der Daten zu verhindern

- vor der Nutzung von personenbezogenen Daten: Datenschutz-Folgeabschätzung (Art. 35 DSGVO) anfertigen und datenschutzbeauftragte Person kontaktieren
- schriftliche Einwilligung vor der Nutzung von personenbezogenen Daten einholen

Beispiel: Ein KI-Tool wertet Metadaten aus E-Mails, Kalendern und Microsoft Teams aus, um Stressindikatoren zu prognostizieren.

	Lokale Nutzung	Online-Nutzung
Datenfluss	Eingaben und Ausgaben bleiben auf Gerät oder internen Server	Eingaben werden an Server des KI-Anbieters geschickt, dort verarbeitet und Antwort zurückgesendet
Risiken	- keine Übertragung an Dritte → gering - Verantwortung für Datensicherheit verbleibt im Unternehmen (z. B. Verschlüsselung, Zugriffsrechte)	- Daten verlassen das System - Risiko von Zugriffen durch Dritte oder unklaren Speicher-/Verarbeitungsarten
DSGVO-Relevanz	- Kein externer Auftragsverarbeiter im Spiel → es findet in der Regel keine „Übermittlung an Dritte“ statt - Unternehmen allein Verantwortlicher im Sinne der DSGVO - Wenn sensible oder personenbezogene Daten verarbeitet werden, muss das Unternehmen selbst technische und organisatorische Maßnahmen nach Art. 32 DSGVO sicherstellen	- Auftragsverarbeitungsvertrag (AVV) nach Art. 28 DSGVO erforderlich - Zusätzliche Anforderungen, falls Daten außerhalb der EU verarbeitet werden - Transparenzpflichten: Du musst deine Nutzer/Kunden darüber informieren, dass und wie ihre Daten an einen KI-Dienst übermittelt werden.

Rechtsleitfaden KI im Unternehmen

von Niklas Mühleis, Nick Akinci

ISBN: 978-3-367-10098-9

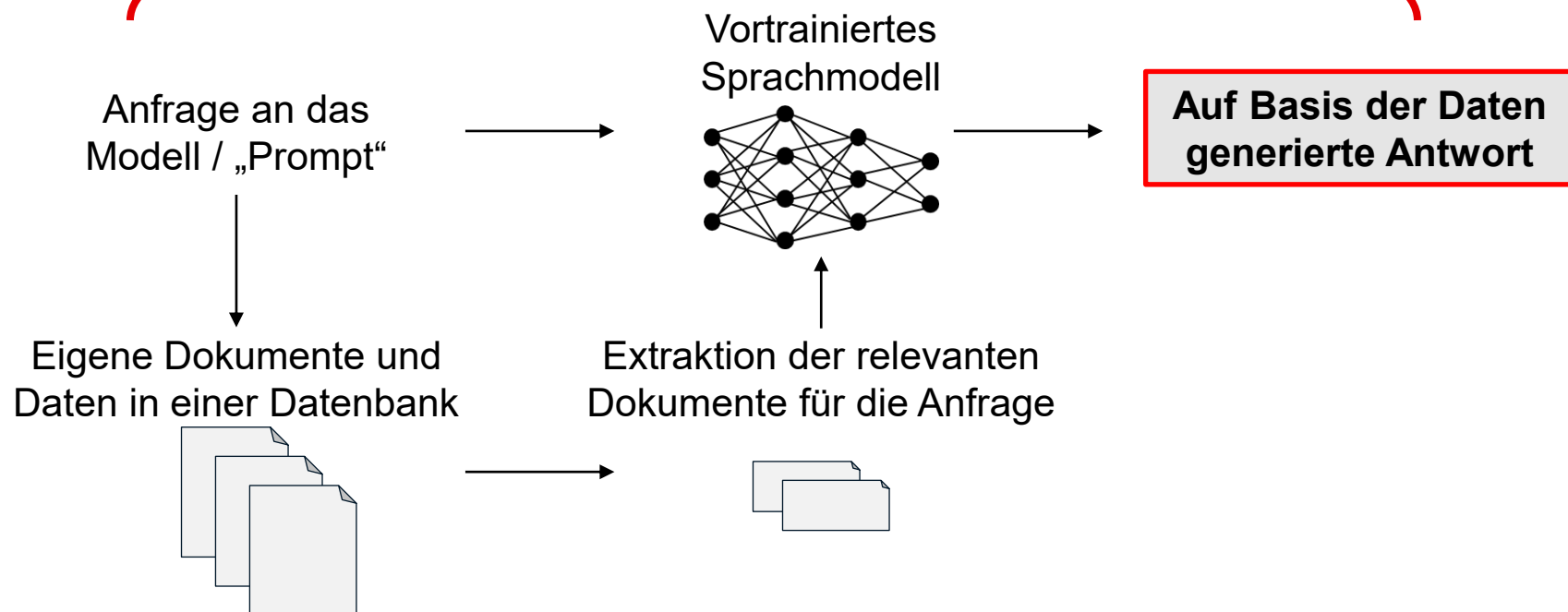
<https://www.rheinwerk-verlag.de/rechtsleitfaden-ki-im-unternehmen/?srsltid=AfmBOoptxmu5XE3t9tiAeOFjGE-vgzKKQKtVM8vJfD-fh0x-7H1iK-mE>

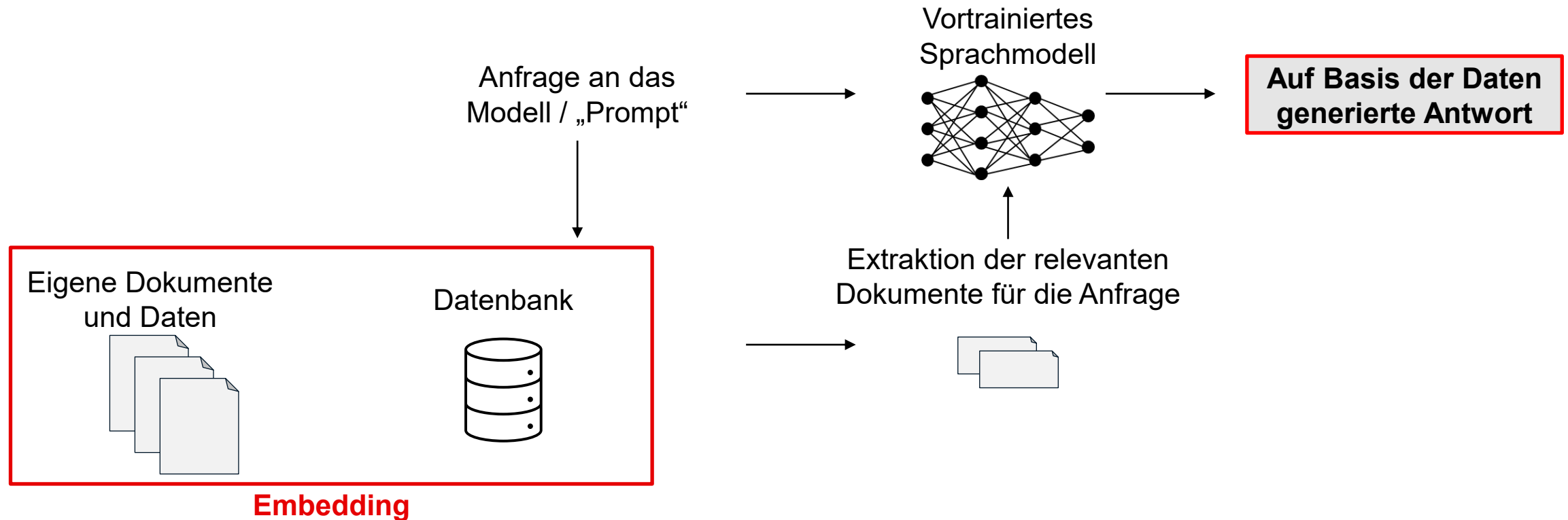


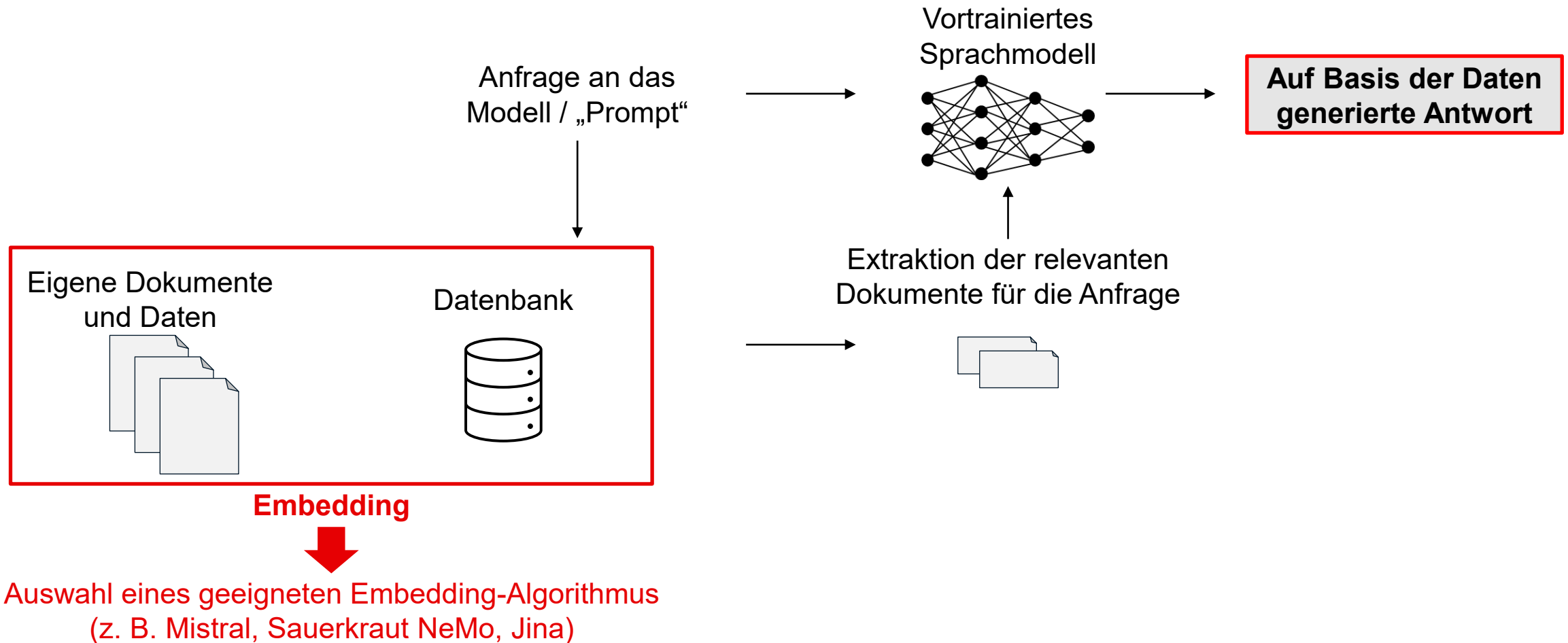
CustomGPT

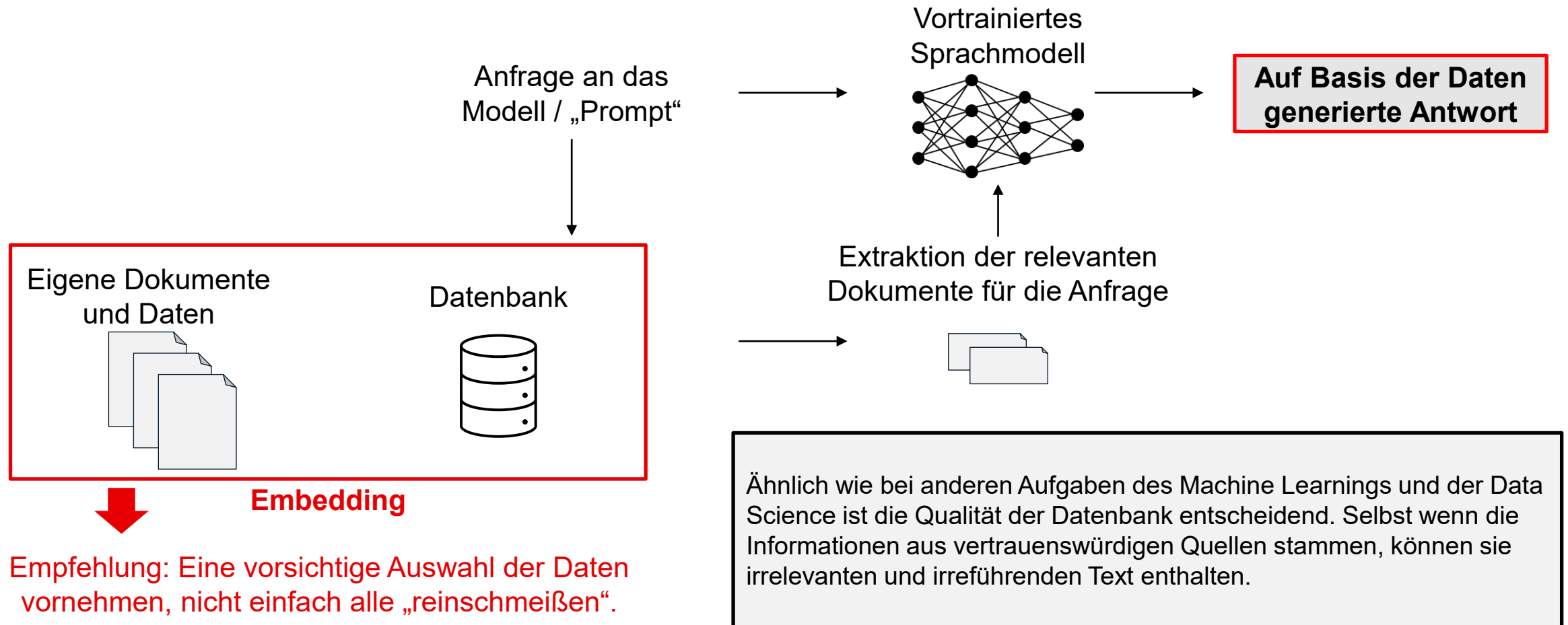
vs.

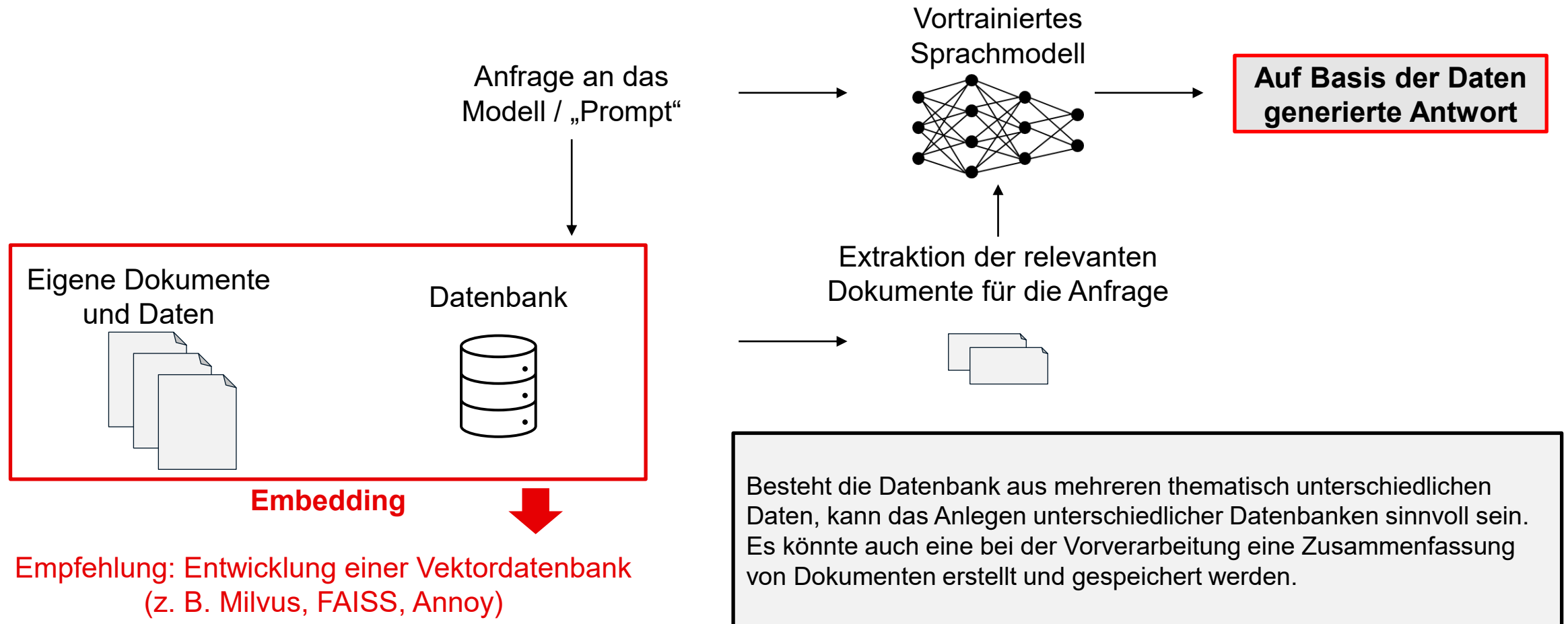
RAG-Ansatz

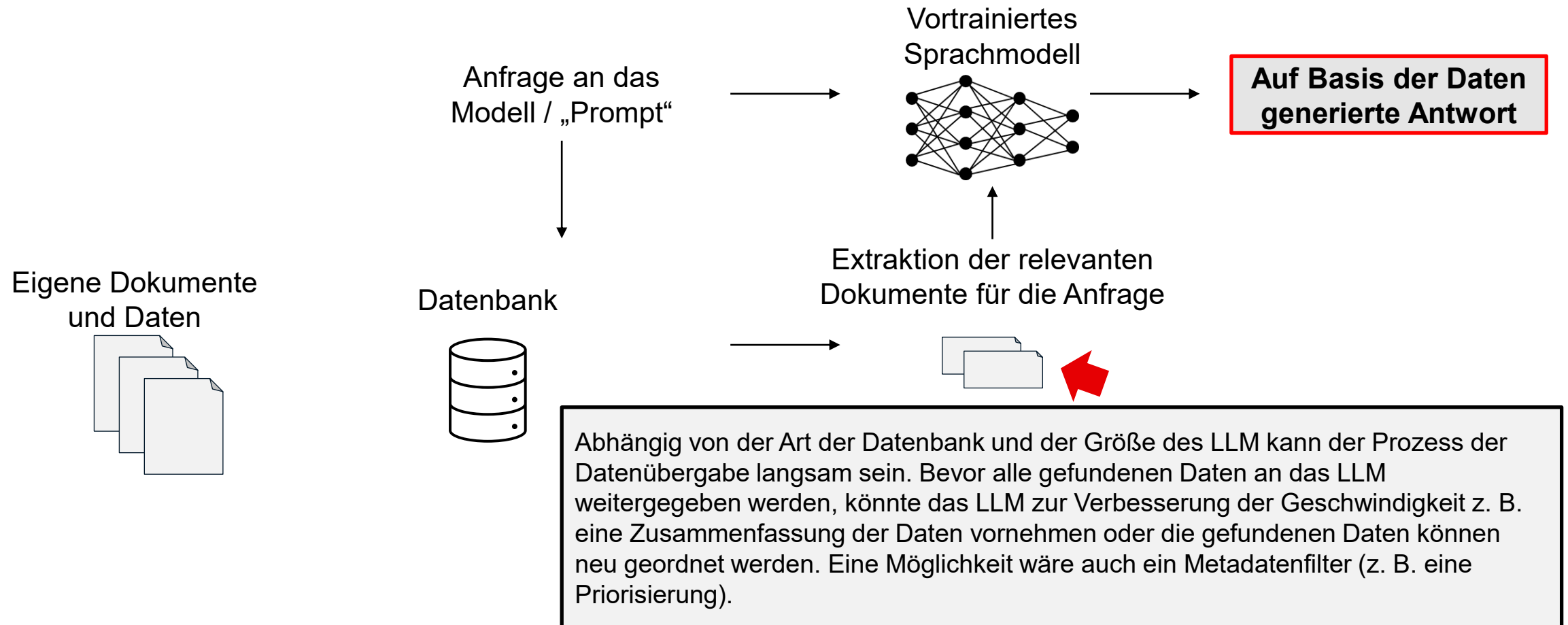


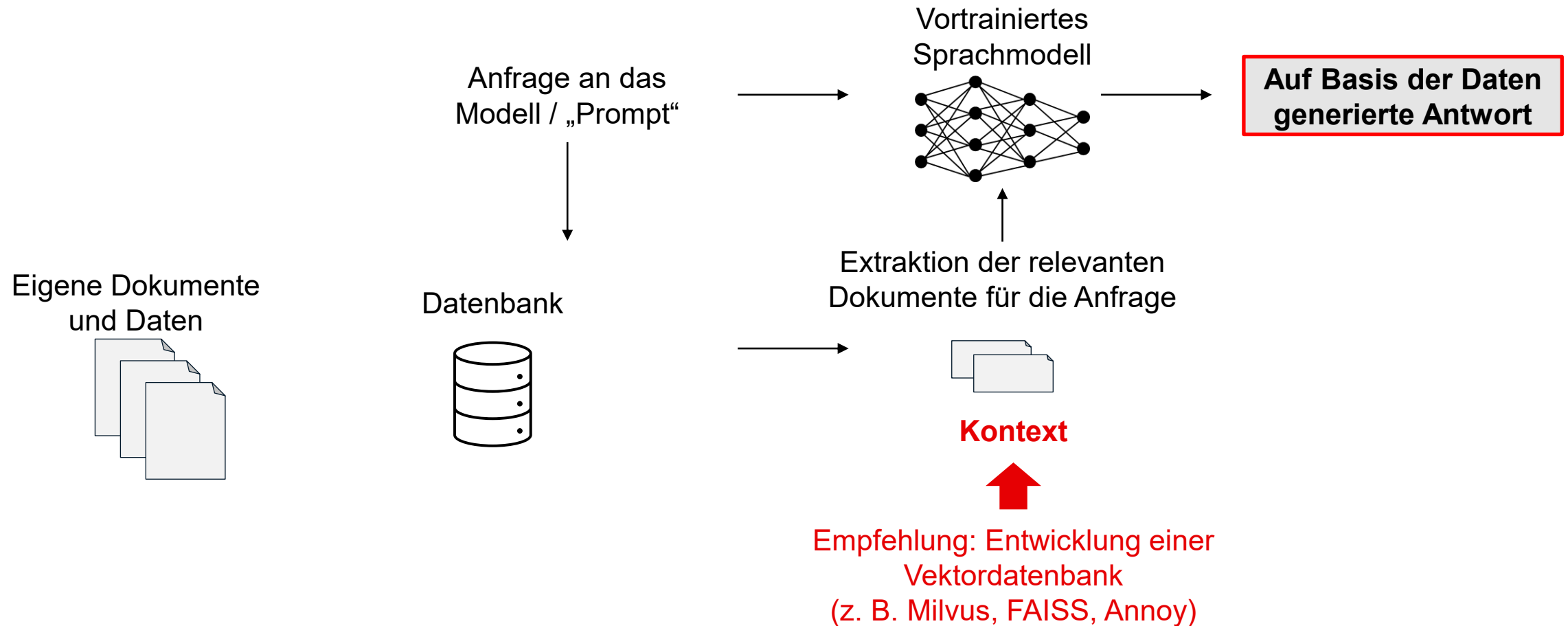












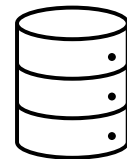
Empfehlung: Vorverarbeitung
der Anfrage, indem sie in
thematisch getrennte
Teilanfragen zerlegt wird



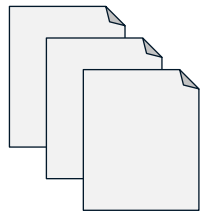
Anfrage an das
Modell / „Prompt“



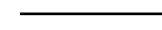
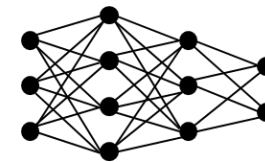
Datenbank



Eigene Dokumente
und Daten



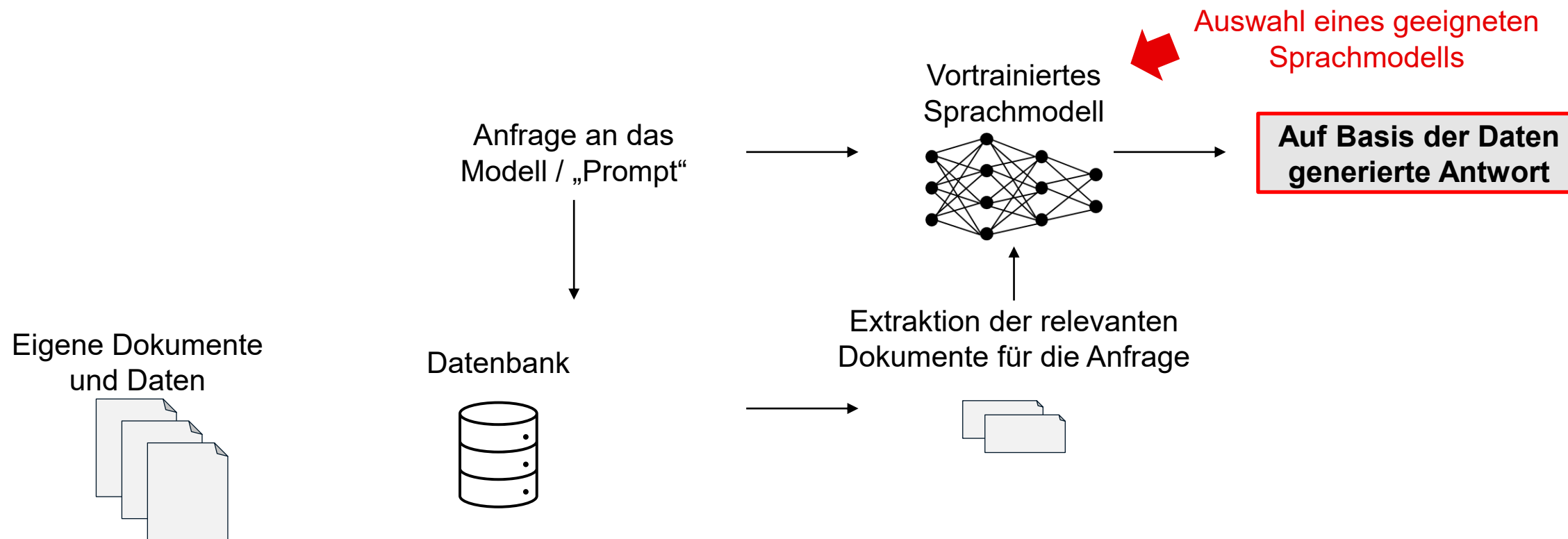
Vortrainiertes
Sprachmodell



**Auf Basis der Daten
generierte Antwort**

Extraktion der relevanten
Dokumente für die Anfrage





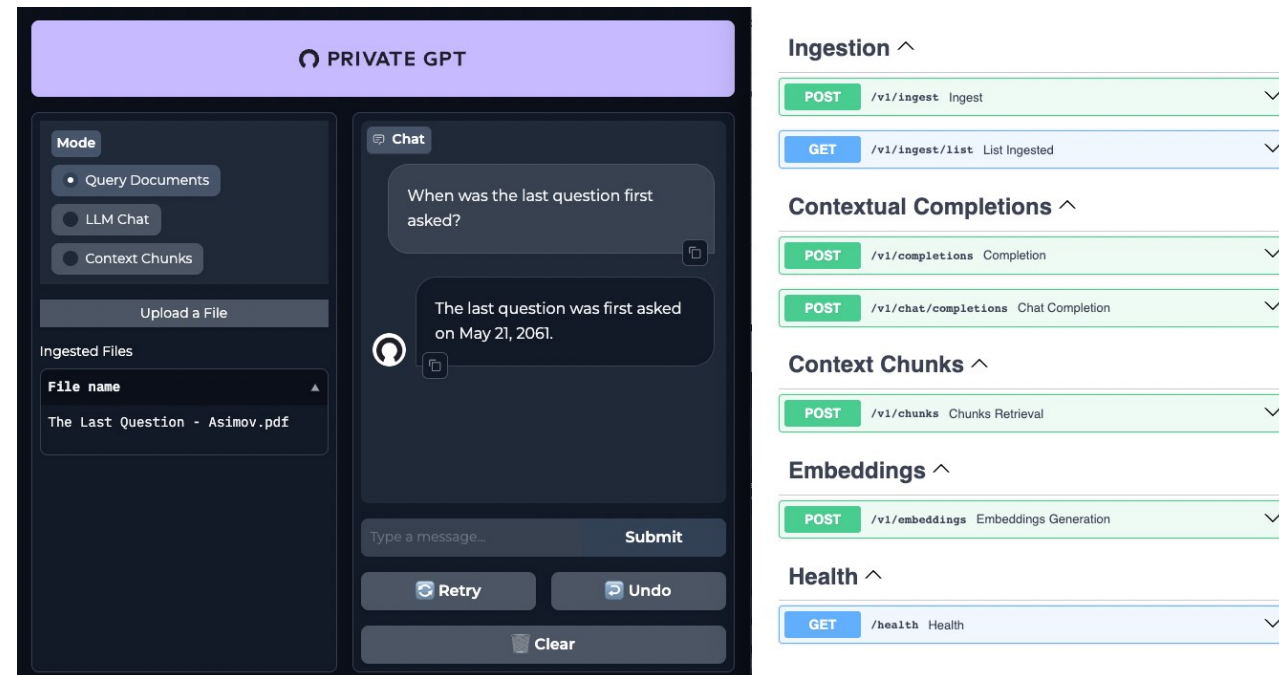
CustomGPT

vs.

RAG-Ansatz

— Programmvorschlge:

- LM-Studio
- PrivateGPT (Wechsel auf die kommerzielle Software Zylon.ai mglich)
- pdfGPT
- GPT4All
- Axolotl
- AnythingLLM



Berücksichtigung der DSGVO besonders einfach, wenn RAG-Anbieter deutsch ist.

Deutsche Anbieter für RAG-Systeme:

<https://www.it-p.de/leistungen/wisbee/>

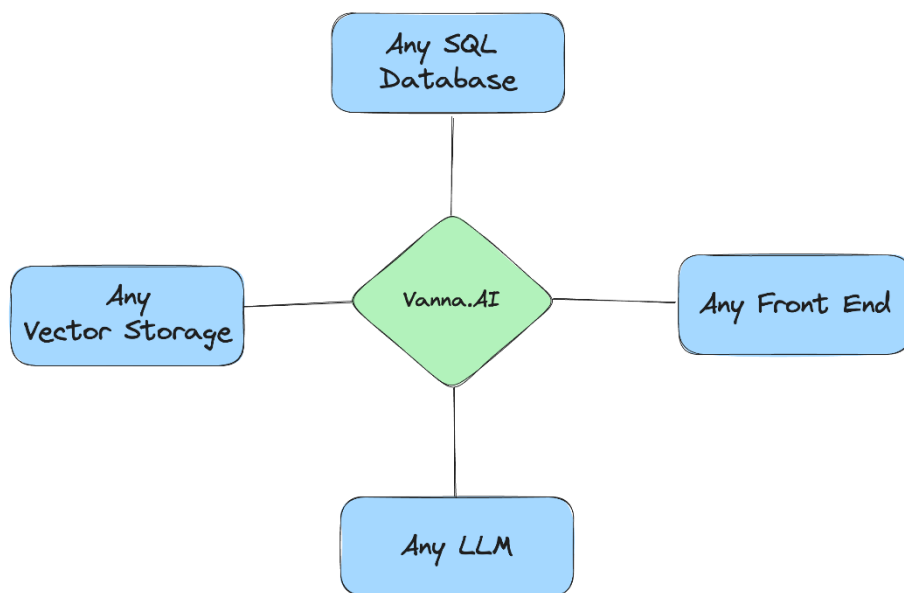
<https://www.kamium.de>

<https://www.kencube.com/de/loesungen/produktion.htm>

→ Es gibt noch viele weitere!

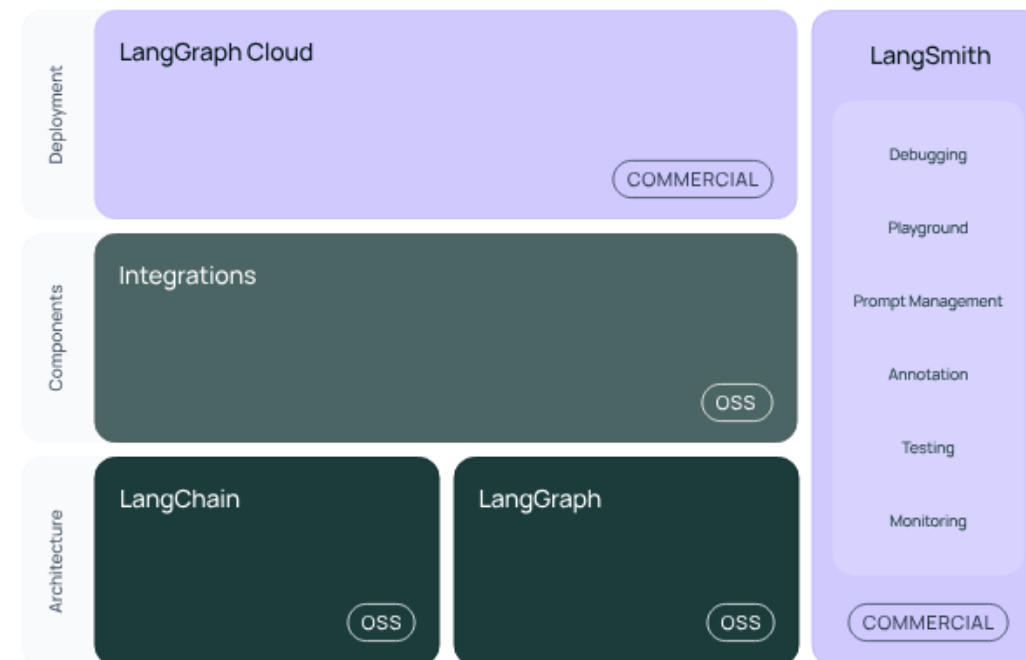


Use AI to Interact With Your Database



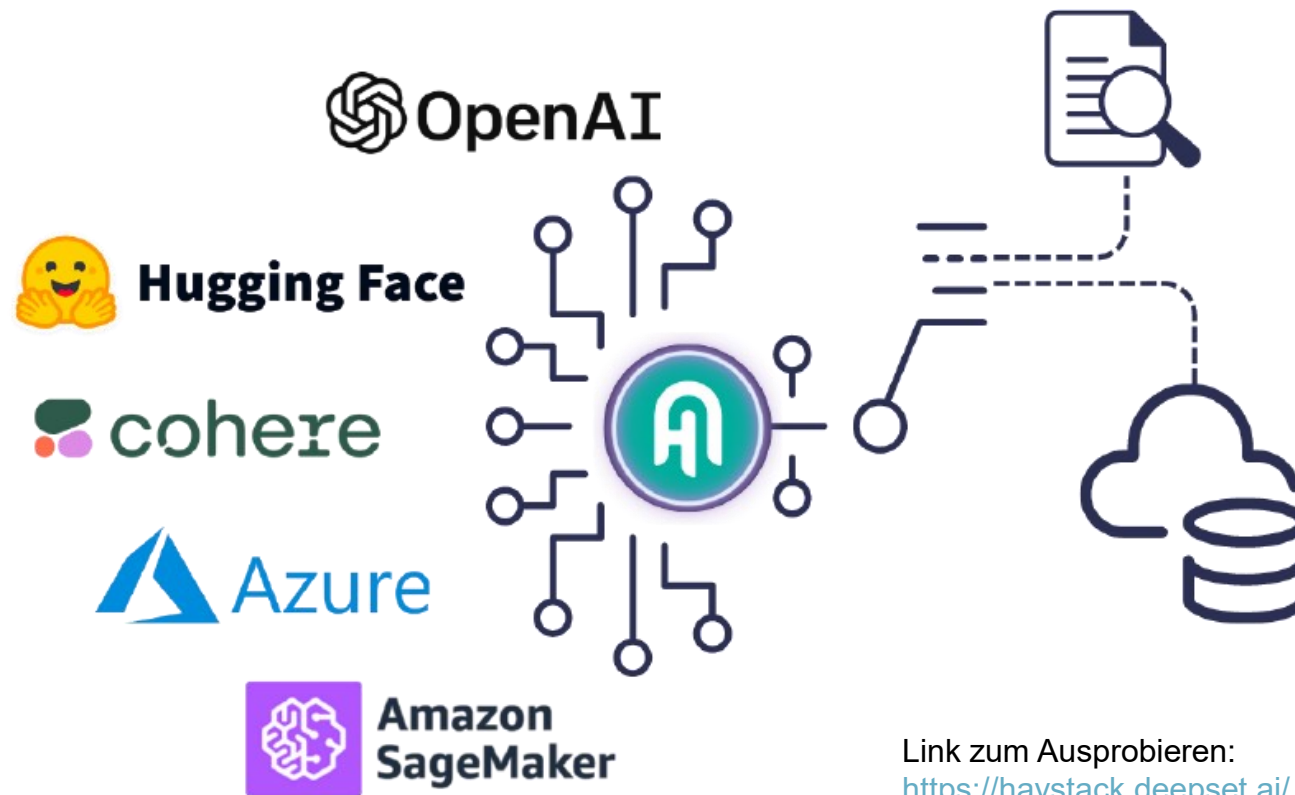
Link zum Ausprobieren:

<https://github.com/vanna-ai/vanna>



Link zum Ausprobieren:

<https://js.langchain.com/docs/introduction/>



Link zum Ausprobieren:
<https://haystack.deepset.ai/>

Kriterium	Open Source	Proprietär
Leistung	Gut	Sehr gut
Kostenmodell	Einmalige Infrastrukturkosten	Variable API-Kosten
Anpassbarkeit	Vollständig (bei Lizenz erlaubt)	Eingeschränkt, meist API-basiert
Datenschutz/Kontrolle	Sehr hoch (on-premise möglich)	Geringer, abhängig vom Anbieter
Support	Community-getrieben, eigene Verantwortung	Professioneller Support, SLAs etc.
Lizenz	Offen (teilweise mit Einschränkungen)	Nutzungsbedingungen strikt

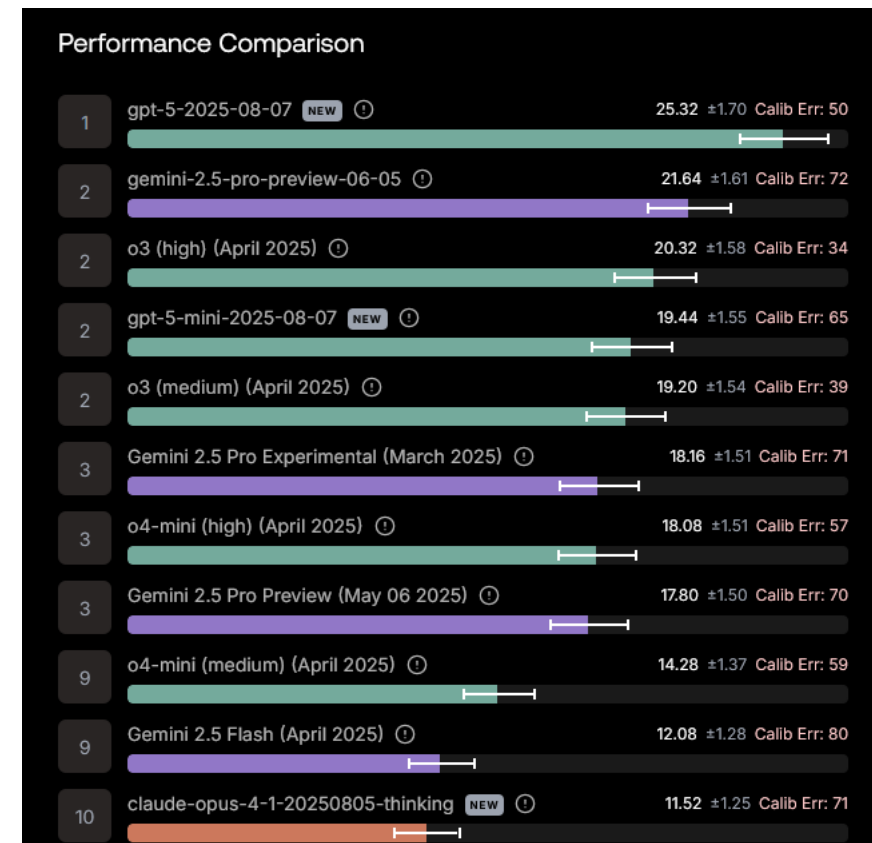
Verschiedene Benchmarks (Auszug):

- Scale.com – SEAL Leaderboard
- LLM-Stats.com
- Vellum AI – LLM Leaderboard
- LM Arena Leaderboard

- Bestenlisten abhängig von Art der Tests
- Viele Änderungen bei Anpassung der Modelle
- Austauschbarkeit kann von Vorteil sein!

Hinweis:

Unter <https://lmarena.ai/> können Sie verschiedene Modelle gegeneinander antreten lassen und anschließend selber bewerten!



Quelle: https://scale.com/leaderboard/humanitys_last_exam Stand 07.08.2025



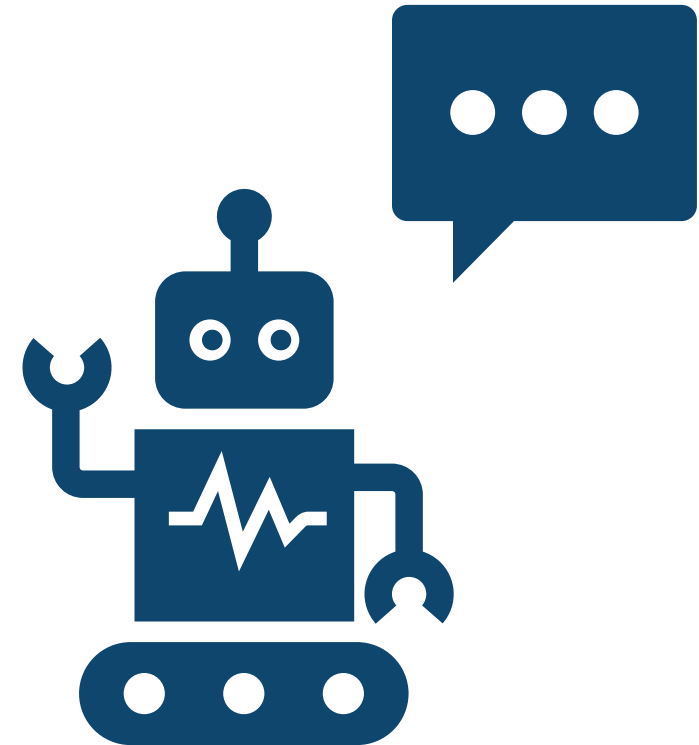
Der Hamburgische Beauftragte für
Datenschutz und Informationsfreiheit

13.11.2023

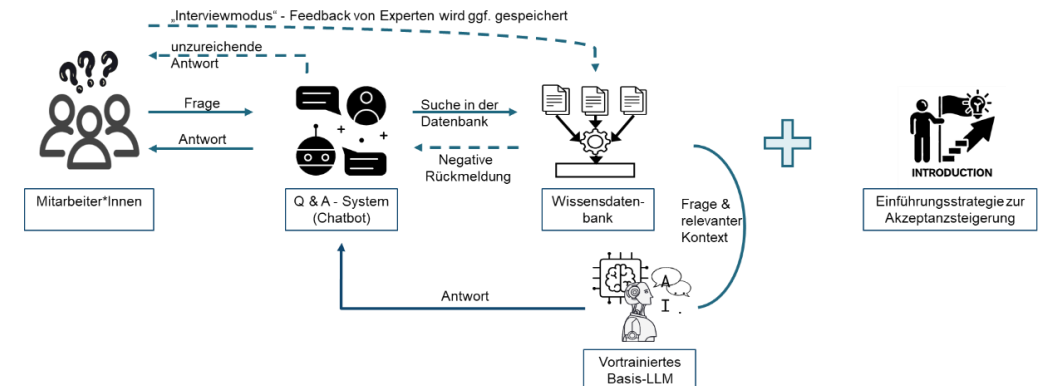
Checkliste zum Einsatz LLM-basierter Chatbots

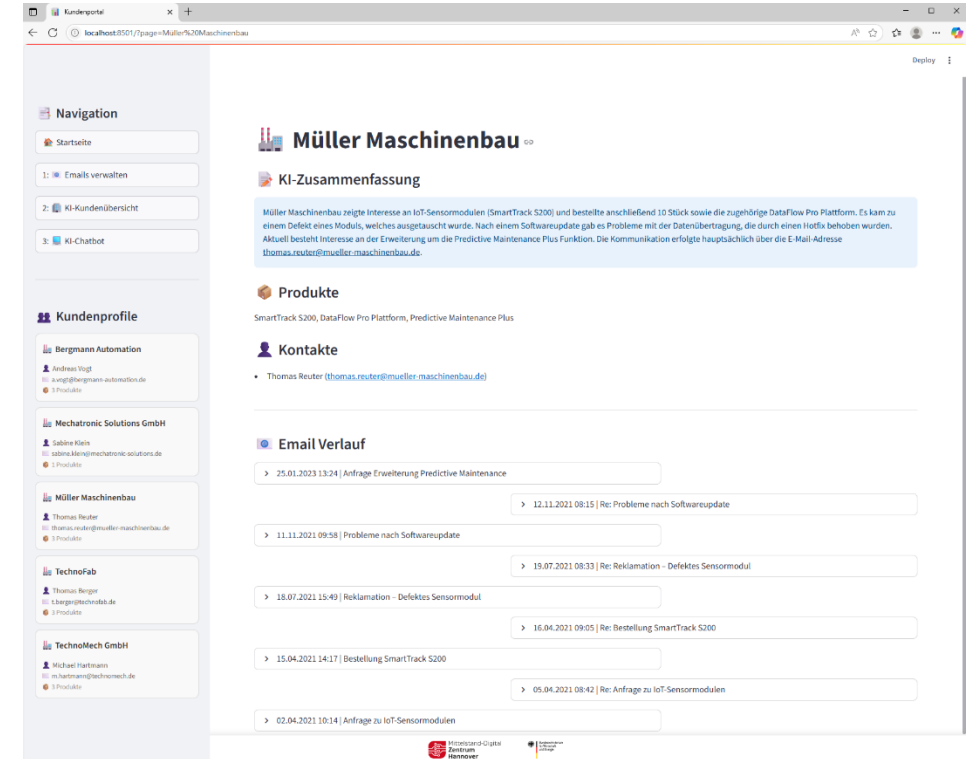
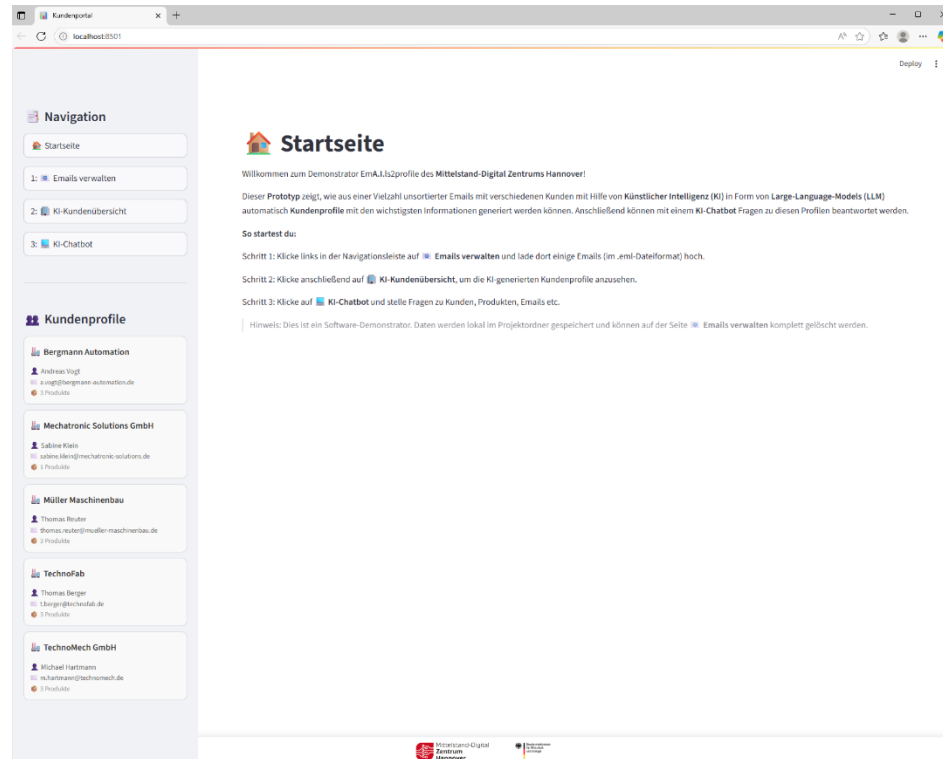
Generative KI in Form von Chatbots bietet die Möglichkeit, schnell und unkompliziert Inhalte zu erstellen. Bekannte Large Language Models (LLM) sind ChatGPT, Luminous oder Bard. In vielen Einrichtungen sind die Tools mittlerweile Teil des Arbeitsalltags geworden, oft jedoch ohne verbindliche Vorgaben zur Nutzung. Dass die Sprachmodelle üblicherweise in einer Cloud betrieben werden, birgt verschiedene Datenschutzrisiken. Zum einen ist der Schutz vertraulicher Daten ge-

- Für Anwendungen, in denen die Qualität der Ausgabe sehr wichtig ist, können KI-basierte Agenten genutzt werden.
- Sog. Agentic RAG haben trainierte KI-Agenten, die die Aussagen auf Plausibilität und Richtigkeit prüfen.
- Anwendungsfälle wären z. B., wenn Kunden im Kontakt mit der KI stehen und die Ausgabe mit hoher Sicherheit richtig sein muss oder wenn es um sicherheitsrelevante Themen geht.



- Inanspruchnahme von innovativen Dienstleistungen rund um CustomGTPs und RAG-Modelle.
 - Mögliche Förderprogramme sind z. B. Innovationsgutschein, Digitalbonus, usw.
- Partizipation an Forschungsprojekten, die sich mit Sprachmodellen und der Wissensspeicherung beschäftigen, z. B.:
 - Entwicklung eines KI-basierten, unternehmensindividuellen Chatbots zur Aufnahme und Nutzung von explizitem und implizitem Expertenwissen im Einkauf (**EnKi-Bot**)
 - Befähigung von produzierenden KMUs zum selbstständigen und sicheren Umgang mit dem AI Act und weiteren KI-Regularien (**AI Act-Ready**)
 - Befähigung von KMU zur Implementierung nachhaltiger Automatisierungslösungen in der Produktion und Entwicklung einer Einführungsstrategie (**AutoSus**)
 - Und viele mehr: <https://www.iph-hannover.de/de/forschung/forschungsprojekte/>







M. Sc. Alexander Poschke

IPH – Institut für Integrierte Produktion Hannover gGmbH
Hollerithallee 6
30419 Hannover



+49 (0)511 27976-229



poschke@iph-hannover.de



www.iph-hannover.de



M. Sc. Paulina Merkel

IPH – Institut für Integrierte Produktion Hannover gGmbH
Hollerithallee 6
30419 Hannover



+49 (0)511 27976-331



merkel@iph-hannover.de



www.iph-hannover.de