Webinar

IHK Spezial

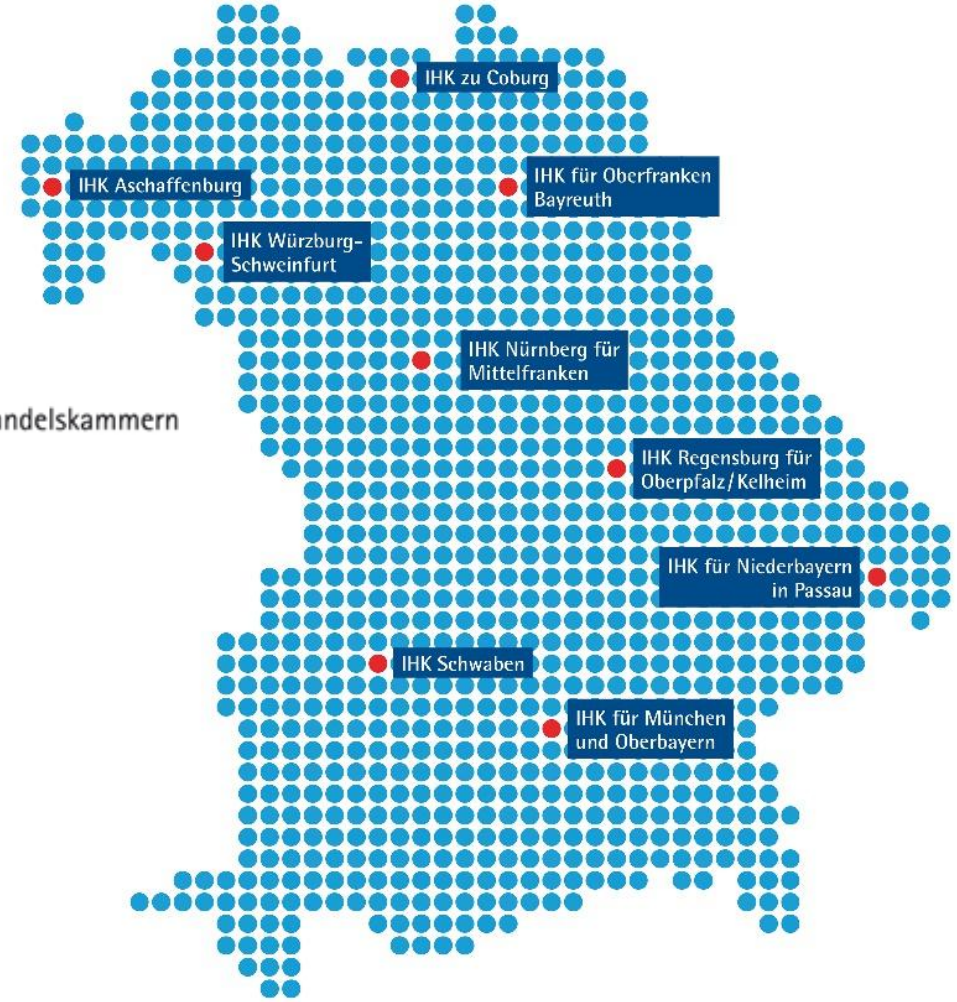
Mehr Nutzen, weniger Risiko: Wie Retrieval-Augmented Generation (RAG) KI im Mittelstand verlässlich macht

IHK Schwaben

BIHK-Reihe „Künstliche Intelligenz verstehen und nutzen“

IHK Schwaben

IHK Industrie- und Handelskammern in Bayern



https://www.bihk.de/themen/digitalisierung/webinarreihe-ki

Kooperation

IHK Industrie- und Handelskammern in Bayern

Bayerisches Staatsministerium für Digitales

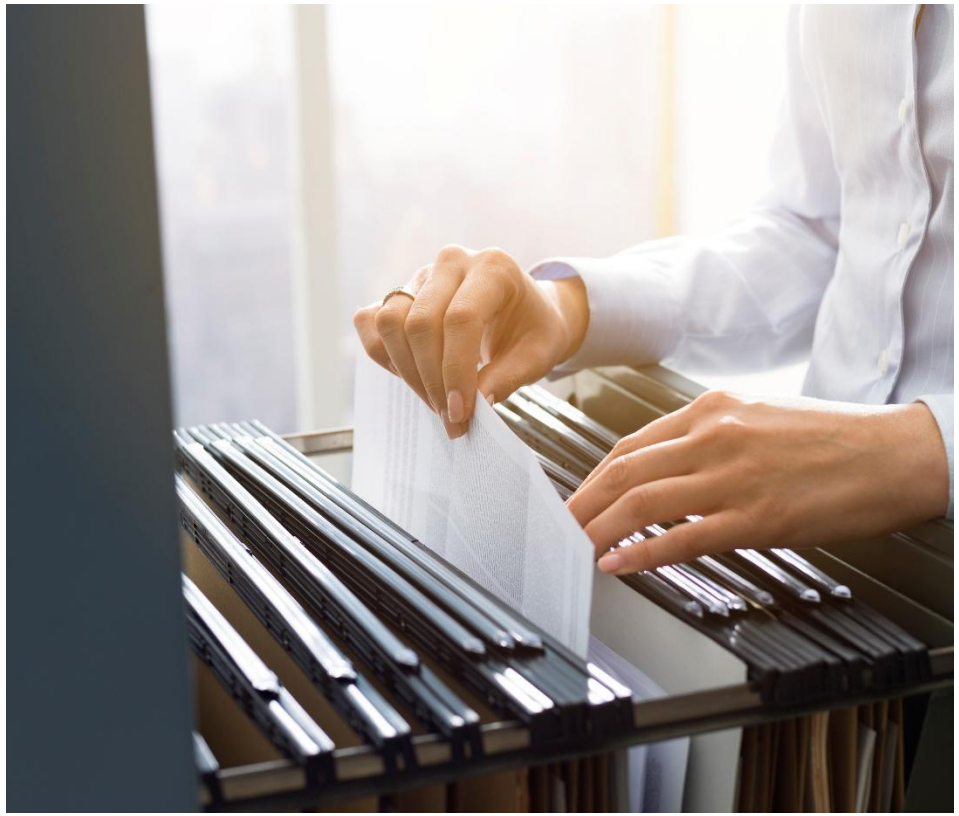
- Digitalimpulse im Rahmen des bayerischen Pakts für berufliche Weiterbildung
- https://www.kommweiter.bayern.de/mit-unterstuetzung/pakt-fuer-berufliche-weiterbildung/
- https://www.bihk.de/themen/digitalisierung/webinarreihe-ki

Aktuelle Zahlen

Digital-Report 2026:

- 66 % der Unternehmen nutzen generative KI
- 76 % der KI-nutzenden Unternehmen nutzen kommerzielle Standardsoftware
- Primäre Nutzung: Marketing & Vertrieb

Mehr Nutzen – Weniger Risiko



- RAG-Grundlagen
- Anwendungsfelder
- Strategische Vorteile

Nikolas Heinloth
Ehrenmüller GmbH
Experte für Effizienzsteigerung durch intelligente Softwarelösungen

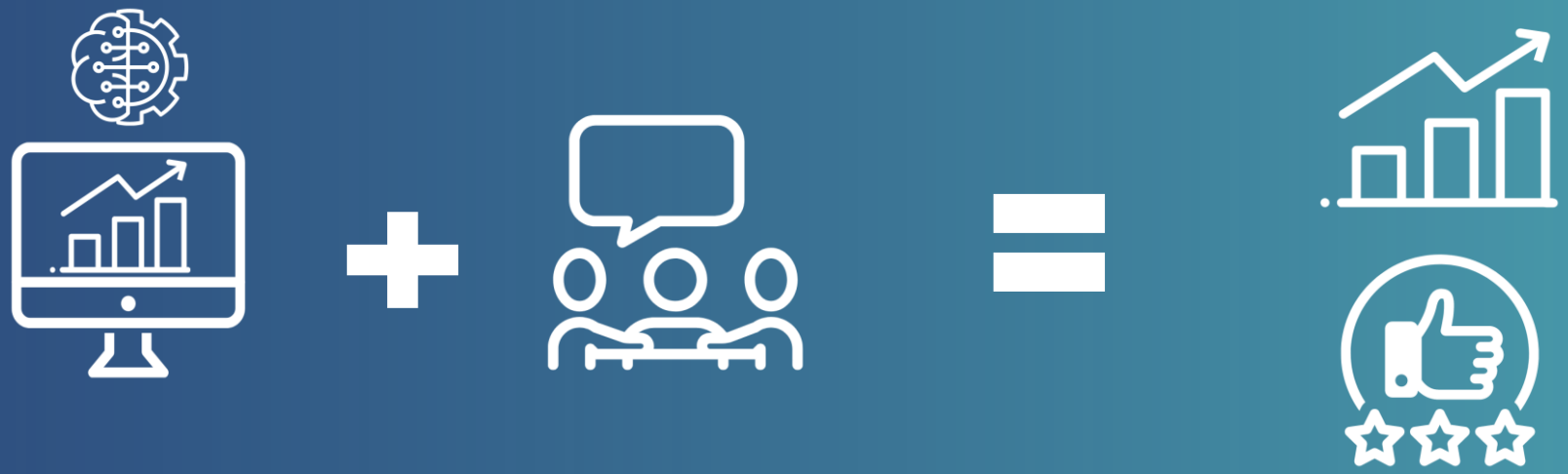
Data Science & Künstliche Intelligenz

Individuelle KI-Entwicklung für den innovativen Mittelstand

Sitz in Kempten (Allgäu) | Gründung 2019

EHRENMUELLER.AI

EHRENMUELLER.AI



EHRENMUELLER.AI

Firmeninterne Policy

Wie lauten die Erstattungsrichtlinien für die Einrichtung eines Homeoffice?

Unternehmen bieten in der Regel zwischen 500 und 1000 für die Einrichtung eines Homeoffice an.

EHRENMUELLER.AI

Firmeninterne Policy

Firmenpolicy: Erstattung von Home-Office Setup

Pro Mitarbeitendem steht ein Budget von bis zu 350 € pro Kalenderjahr zur Verfügung.
Erstattungen müssen vorab von der jeweiligen Führungskraft genehmigt werden.

Wie lauten die Erstattungsrichtlinien für die Einrichtung eines Homeoffice?

Unternehmen bieten in der Regel zwischen 500 und 1000 für die Einrichtung eines Homeoffice an.

Retrieval Augmented Generation

EHRENMUELLER.AI

Wie lauten die Erstattungsrichtlinien für die Einrichtung eines Homeoffice?

Hier ist unsere firmeninterne Policy:

Firmenpolicy: Erstattung von Home-Office Setup

Pro Mitarbeitendem steht ein Budget von bis zu 350 € pro Kalenderjahr zur Verfügung.
Erstattungen müssen vorab von der jeweiligen Führungskraft genehmigt werden.

In Ihrem Unternehmen steht ein Budget von 350 Euro pro Jahr für die Einrichtung eines Homeoffice an.

Retrieval

Firmeninterne Policy

Firmenpolicy: Erstattung von Home-Office Setup

Pro Mitarbeitendem steht ein Budget von bis zu 350 € pro Kalenderjahr zur Verfügung.
Erstattungen müssen vorab von der jeweiligen Führungskraft genehmigt werden.

Augmented

Wie lauten die Erstattungsrichtlinien für die Einrichtung eines Homeoffice?

Hier ist unsere firmeninterne Policy:

Firmenpolicy: Erstattung von Home-Office Setup

Pro Mitarbeitendem steht ein Budget von bis zu 350 € pro Kalenderjahr zur Verfügung.
Erstattungen müssen vorab von der jeweiligen Führungskraft genehmigt werden.

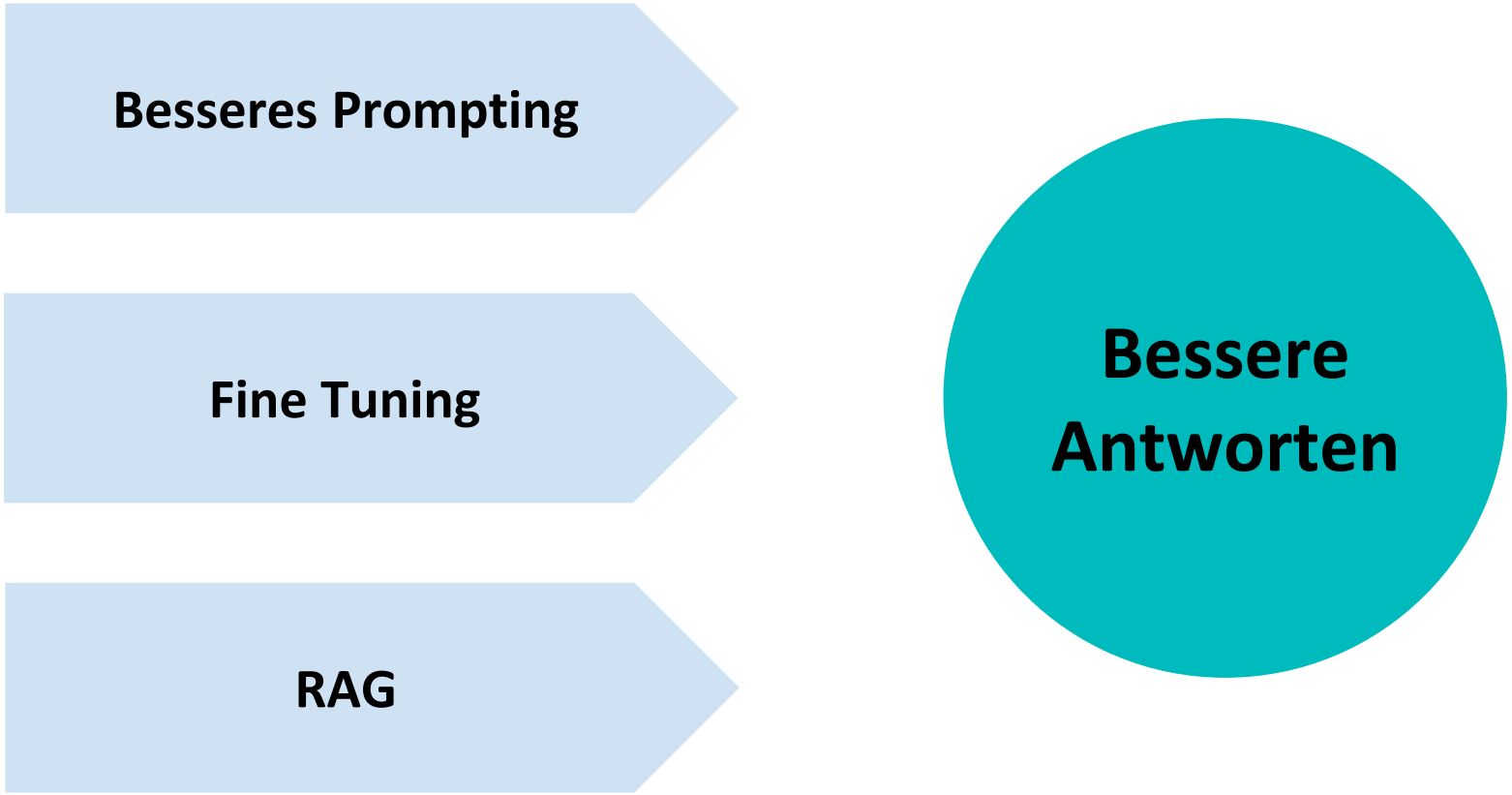
Generation

In Ihrem Unternehmen steht ein Budget von 350 Euro pro Jahr für die Einrichtung eines Homeoffice an.

EHRENMUELLER.AI

- ❑ Wie können wir basierend auf einem Prompt relevante Informationen finden?
- ❑ Wie können wir unseren Prompt mit dieser Information bestmöglich verbessern?
- ❑ Wie können wir eine passende Antwort-Generierung sicherstellen?

Also ab jetzt immer RAG?



EHRENMUELLER.AI

Besseres Prompting: Vorteile

Vorteile und Eigenschaften:

- Einfach umzusetzen
- Moderne Sprachmodelle haben relativ große Kontextfenster
- Eignet sich gut um dem LLM direkte Anweisungen mitzugeben (z.B. Verbot unerwünschter Inhalte)

Besseres Prompting: Nachteile

Herausforderungen bei großen Kontexten:

- Informationen gehen bei großen Kontextlängen verloren
- Skaliert nicht unbeschränkt (insb. bei Multiturn-Conversations)
- Erhöhte Laufzeit & Kosten

Finetuning: Vorteile

Vorteile und Eigenschaften:

- Informationen werden in Modellgewichten “verinnerlicht”
- → Informationen müssen nicht mehr im Prompt stehen
- Dadurch kann das Modell neue Konzepte erlernen
- Bspw. Jargon, Sprachstil oder (Programmier)sprache

Finetuning: Nachteile

Herausforderungen und Limitationen:

- Trainings- & Hosting-aufwand des angepassten Modells
- Änderungen an Use Case/Informationen erfordert neues Modell-Training
- Trainingsdaten erforderlich
- Modell kann “alte” Fähigkeiten vergessen

EHRENMUELLER.AI

Vorteile von RAG

RAG vs. Finetuning

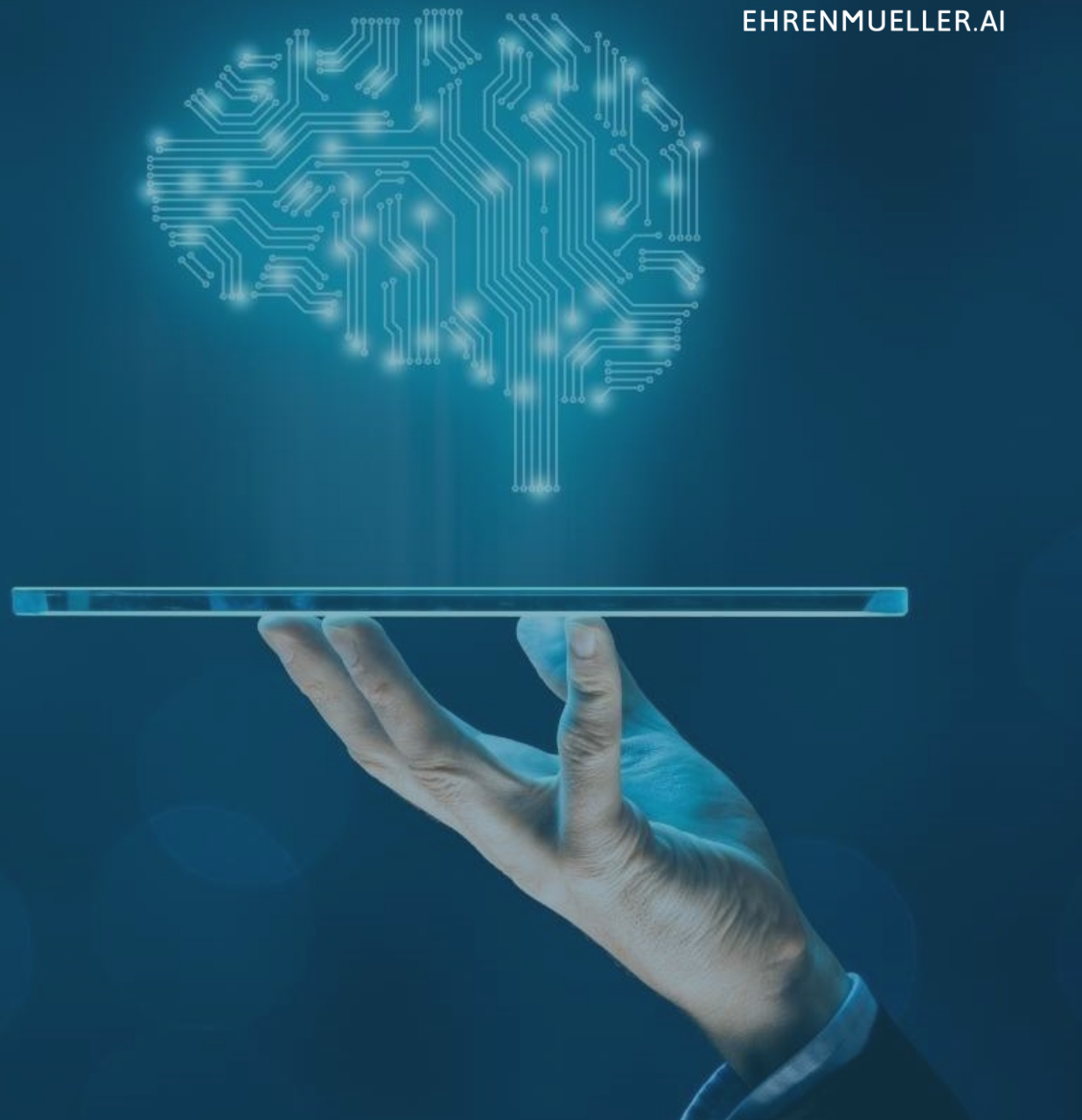
- Einfach anzupassen & steuerbar
- Flexibel: Ein Modell für viele Use-Cases via Prompts
- Kostengünstig:
  - Kein Training
  - Kein separates Hosting

RAG vs. Kontext

- Präzision: Nur relevante Infos passend zur Frage
- Effizienz: Kleinerer Fokus führt zu besseren Antworten

EHRENMUELLER.AI

EHRENMUELLER.AI



Wie funktioniert RAG?

Retrieval

Wie lauten die **Erstattungsrichtlinien** für die Einrichtung eines **Homeoffice**?

EHRENMUELLER.AI

Interne Policy Dokumente

Select * from internal_policy where content like “%Homeoffice%”

Lösungsansatz: Semantic Search

EHRENMUELLER.AI

Anstatt nach Matches an Wörtern zu suchen, suchen wir nach ähnlichen Bedeutungen in der Nutzerfrage und den uns zur Verfügung stehenden Dokumenten.

Mitarbeiter dürfen bis zu drei Tage pro Woche im Home-Office arbeiten.	[0.05, 0.90, -0.15, 0.60, 0.11, …]
Die Arbeitszeiten im Home-Office entsprechen den regulären Bürozeiten.	[-0.12, 0.70, -0.52, 0.25, 0.80, …]
Alle Home-Office-Tage müssen vorher mit der Führungskraft abgestimmt werden.	[-0.40, 0.55, -0.10, -0.05, 0.48, …]

Word Embeddings

EHRENMUELLER.AI

> *“You shall know a word by the company it keeps”*
>
> — J.R. Firth (Linguist)

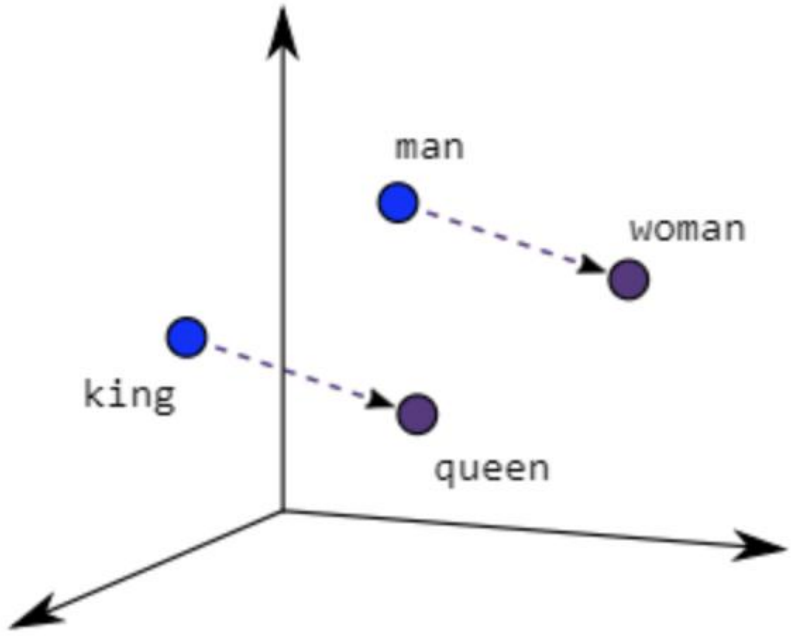
Ansatz:

- repräsentiere Wörter als Vektoren
- **Wörter die häufig in ähnlichen Kontexten verwendet werden** (wie zum Beispiel Hotel und Motel) **sollten ähnliche Vektoren haben**

Lösungsansatz Semantic Search

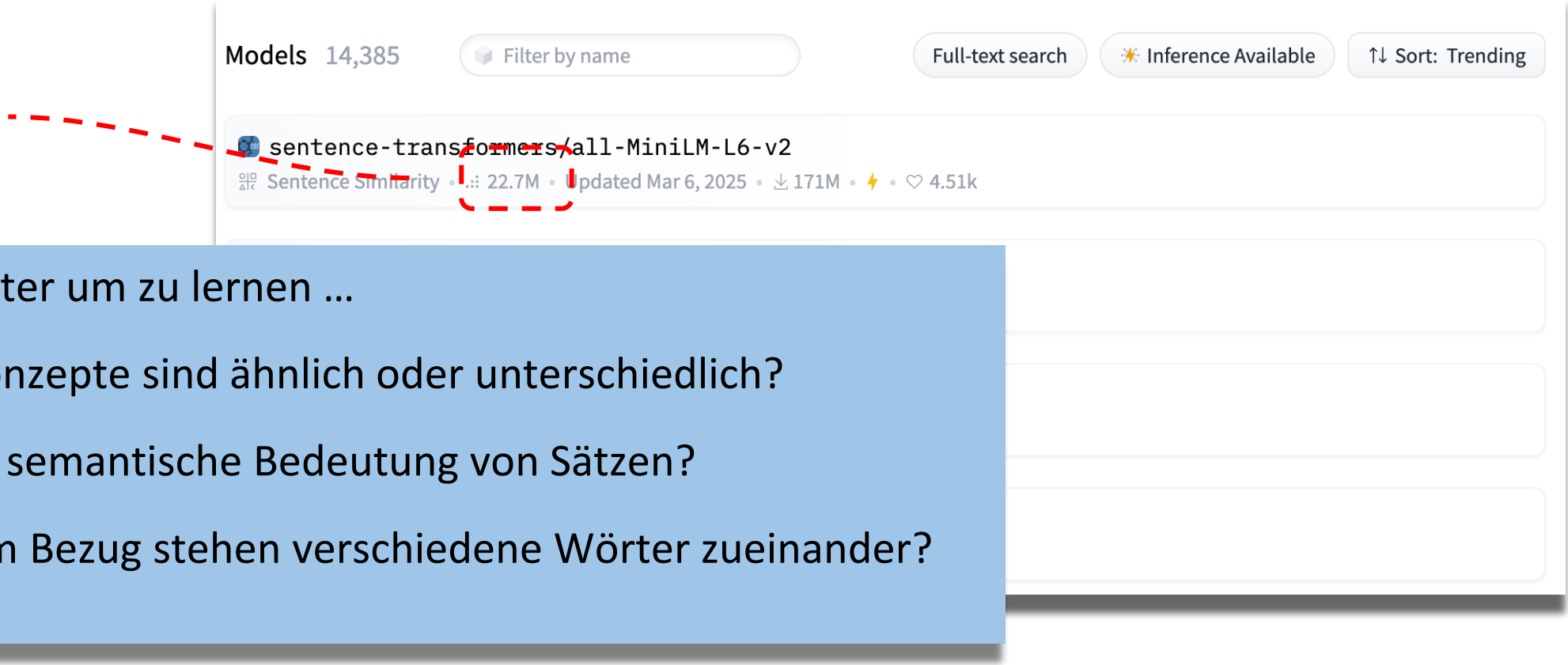
EHRENMUELLER.AI

Wenn wir nun Dokumente auf Ähnlichkeit untersuchen suchen wir nun nach den räumlich am nächsten liegenden Inhalten, nicht mehr den Dokumenten mit dem höchsten Word-count



Wie kommen wir zu dieser Darstellung?
Embedding Modelle

Anzahl an Parametern



22M Parameter um zu lernen …

… Welche Konzepte sind ähnlich oder unterschiedlich?

…Was ist die semantische Bedeutung von Sätzen?

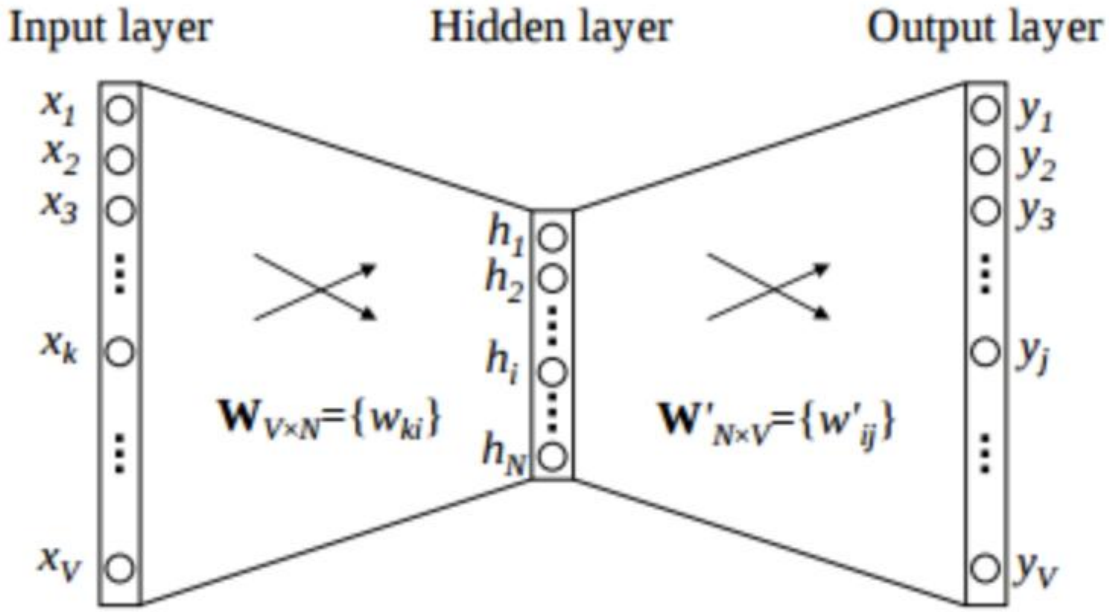
… In welchem Bezug stehen verschiedene Wörter zueinander?

Wie kommen wir zu dieser Darstellung?
Embedding Modelle

EHRENMUELLER.AI

Spezialisierte **Deep Learning Modelle**, die auf einer großen Menge an Daten vortrainiert wurden.

Wir verwenden diese bestehenden Modelle um unsere Dokumente/Texte in eine entsprechende mathematische Darstellung überführen zu können.



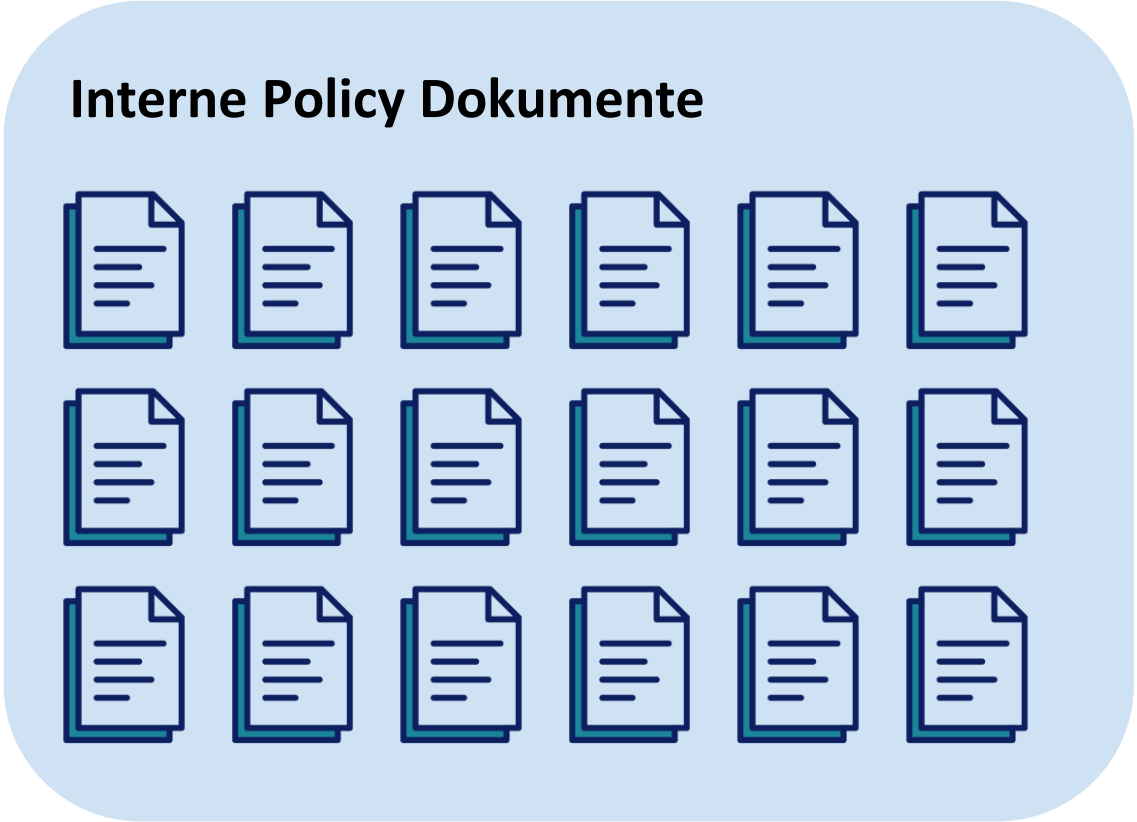
word2vec model architecture

Vektor Datenbanken

- spezialisierter Wissensspeicher
- Nutzen fortschrittliche Indexierungsalgorithmen (wie Approximate Nearest Neighbor), um in riesigen Datensätzen in Millisekunden die relevantesten Informationen zu finden

→ Ohne sie wäre das "Retrieval" (Abrufen) in RAG bei großen Datensätzen praktisch nicht performant umsetzbar

EHRENMUELLER.AI



100 Dokumente x Dimension 384 = 384000 Vergleiche für jede Query

Chunking

EHRENMUELLER.AI

Definition
Aufteilen eines Dokuments in kleinere, fokussierte Abschnitte

Gängige Strategien

- **Fixed-size chunking**
  Feste Zeichen- oder Wort- oder Tokenanzahl
- **Sentence based chunking**
  Trennung nach Satzende
- **Paragraph based chunking**
  Erhalt der Absatzstruktur

Beispiel-Dokument (Home-Office Policy)

Chunk 1
Mitarbeiter dürfen bis zu drei Tage pro Woche im Home-Office arbeiten.

Chunk 2
Die Arbeitszeiten im Home-Office entsprechen den regulären Bürozeiten.

Chunk 3
Alle Home-Office-Tage müssen vorher mit der Führungskraft abgestimmt werden.

→ Experimentieren Sie mit der passenden Strategie für Ihren spezifischen Use-Case.

Ingestion Process (RAG Pipeline)

EHRENMUELLER.AI

LOAD

Dokumenten-sammlung

Beispielsweise aus internem Laufwerk, Wiki, ERP-System

SPLIT

Aufteilen der Dokumente

In kleinere Abschnitte (Chunks) zerlegen. Kontext soll bestehen bleiben.

EMBED

Text zu Vektor

Embeddings mittels Sprachmodell erzeugen.

```
[1, 0.4, 1.8, 1.1, …]
[0.2, 3, 1.7, 1.4, …]
```

STORE

Datenbank

Vektoren persistent in einer passenden Vektordatenbank abspeichern.

Retrieval Process (RAG Pipeline)

EHRENMUELLER.AI

STEP 1

Nutzeranfrage zu Vektor

Embedding mittels Sprachmodell erzeugen.

`[1, 0.4, 1.8, 1.1, …]`

STEP 2

Vektorsuche

Ähnliche Vektoren in Datenbank finden.

`[0.9, 0.45, 1.7, 1.0, …]`

STEP 3

GENERATION

Antwortgenerierung

Durch LLM aus Nutzeranfrage UND informativen Textchunks.

RAG – VERBESSERUNGSMÖGLICHKEITEN

EHRENMUELLER.AI

Hybride Suche: Retrieval mit “klassischer” Suche kombinieren, Kombination mehrerer Rankings

Multimodaler RAG: Erweitert herkömmliche textbasierte RAG-Systeme, um andere Datentypen wie Bildern, Tabellen, Audio …

Query Rewriting/Expansion: Nutzeranfrage zu mehreren, besseren Suchanfragen umformulieren lassen

Reranking: Gefundene Ergebnisse mit Reranking Modell neu sortieren

RAG als Agent-Tool: Agent kann selbst Retriever per Query ansprechen, iteratives Vorgehen möglich

Context compression: Gefundene Ergebnisse zusammenfassen, deduplizieren, filtern und restrukturieren

EHRENMUELLER.AI

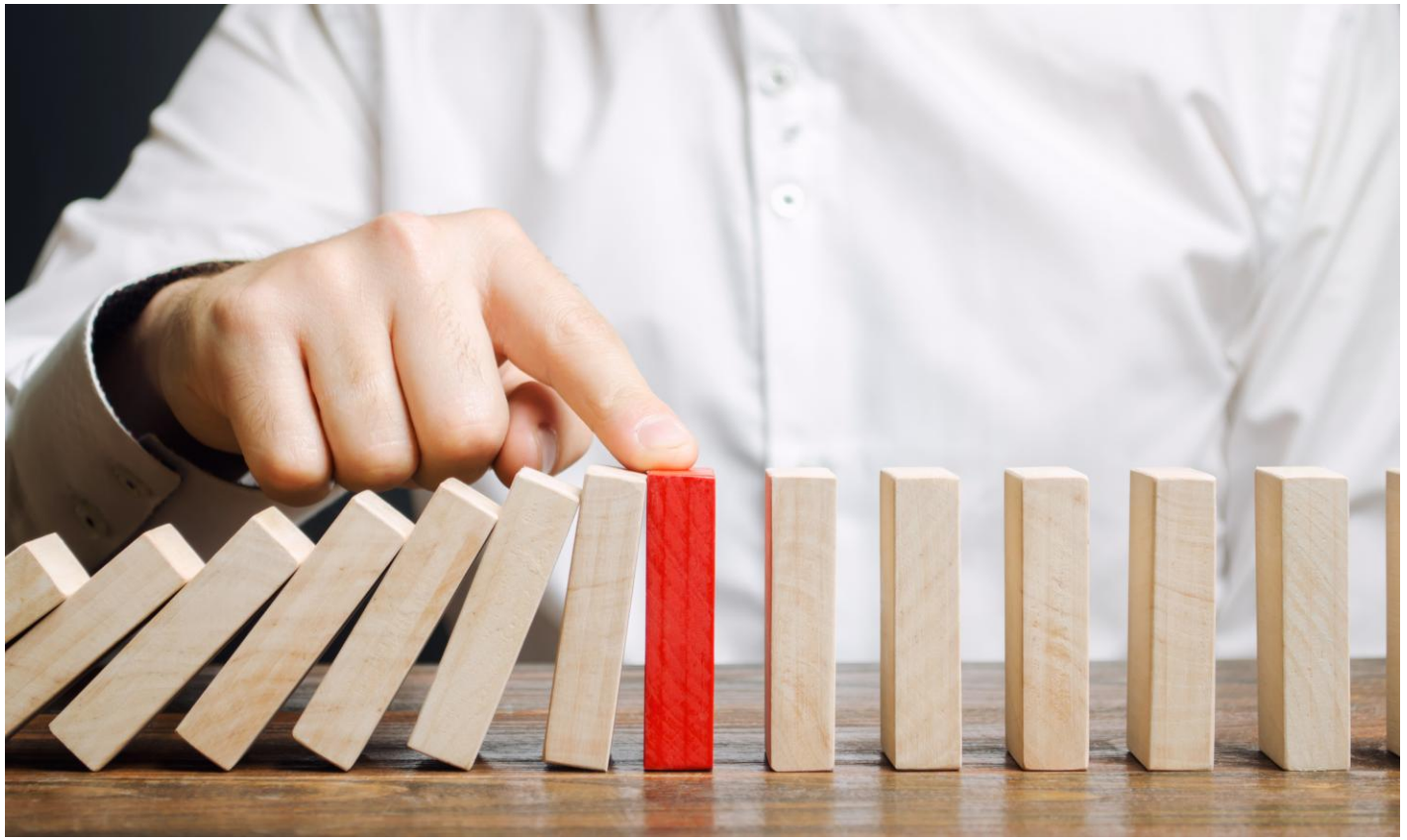
www.ehrenmueller.ai

Nikolas Heinloth
nikolas.heinloth@ehrenmueller.ai

Veranstaltungstipp

IHK Schwaben

Business Continuity Management | 1.-2. Juli 2026



Jetzt anmelden!

KI-Kompetenz-Check

IHK Schwaben

Erfüllt Ihr Unternehmen die Anforderungen des AI Acts?

kostenfrei anmelden:

kostenloses Handbuch:

Veranstaltungstipp

IHK Schwaben

Unsere nächsten Webinare im Themenfeld Digitalisierung und KI

GEO verstehen: Wie KI Ihr Unternehmen empfiehlt

Donnerstag, 07.05.2026 | 15:00 – 16:00 Uhr

- Steigerung der (regionalen) Sichtbarkeit in KI-gestützten Suchsystemen
- Neue Regeln für die Content-Architektur
- SEO vs. GEO

Benjamin Knecht

Geschäftsführer | MXP GmbH

Täuschend echt – IT-Sicherheitsprävention zwischen Phishing, Spoofing und KI

Mittwoch, 10.06.2026 | 10:00 - 11:00 Uhr

- Cyber Security Awareness
- Passwortsicherheit
- Deepfakes &Phishing
- aktuelle Betrugsmaschen

David Wilpert | Tim Buchwald

Kriminalpolizei Augsburg, Fachdezernat Cybercrime

Kommen Sie bei Fragen gerne auf mich zu!

Luzia Ghoggal

Digitale Innovation und Künstliche Intelligenz

luzia.ghoggal@schwaben.ihk.de

0821/3162-230

IHK Schwaben

Vielen Dank für Ihre Aufmerksamkeit!

Weitere Informationen unter
ihk.de/schwaben/ihkspezial