BUSINESS MEETS AI

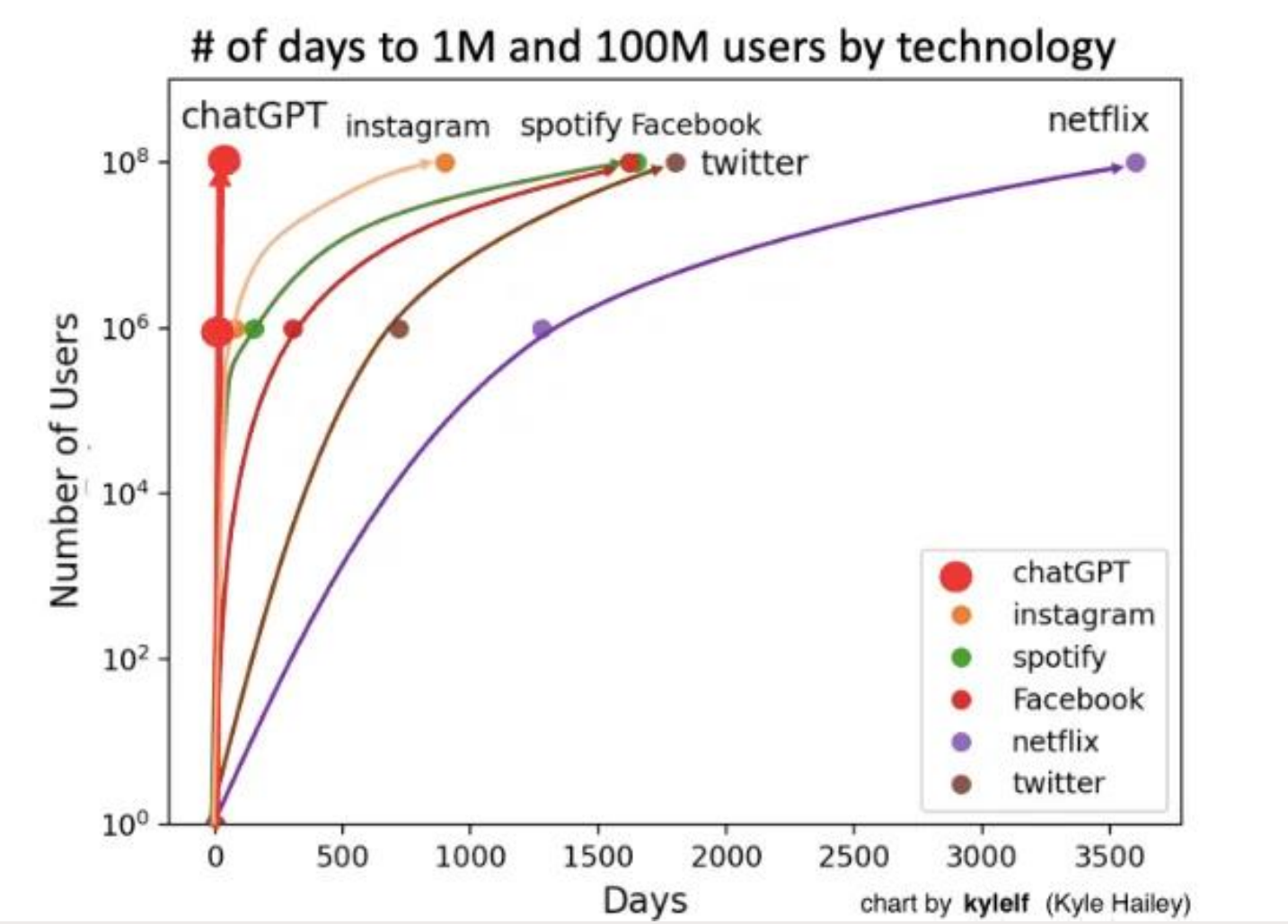
11 | 04 | 2024



RAG

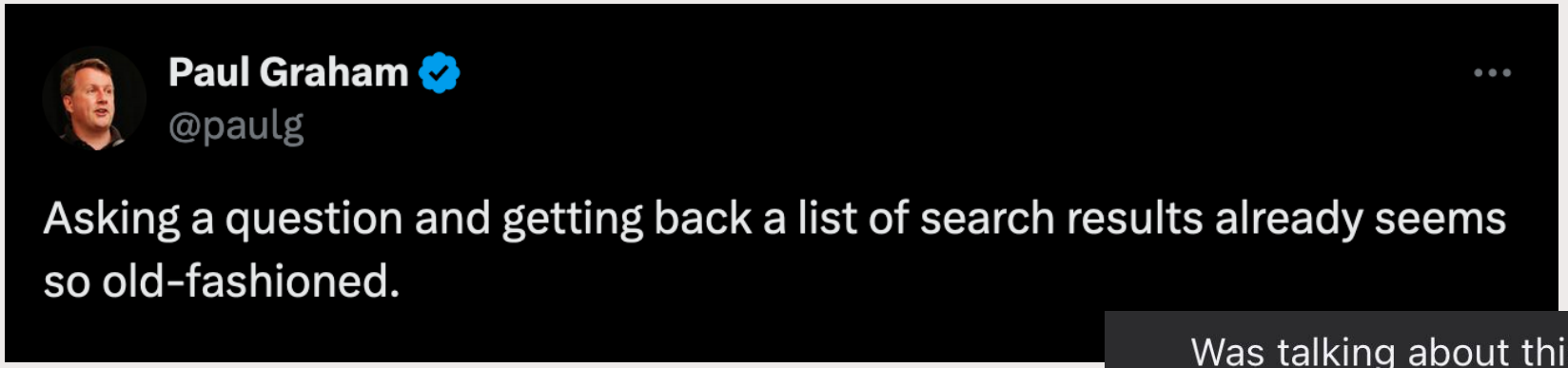
Eigene Daten mit AI durchsuchbar machen – Semantische Suche über Embeddings in der Private Cloud

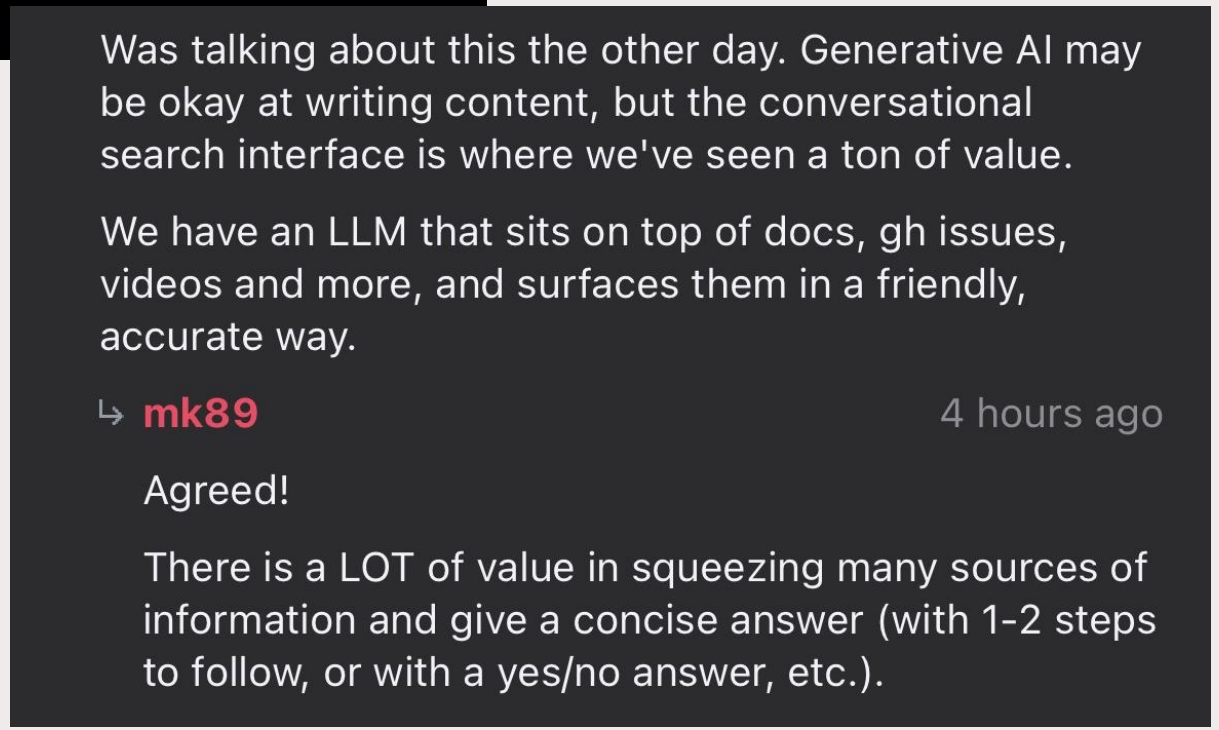
BENUTZERWACHSTUM CHATGPT



www.argcom.de

STÄRKEN VON CHATGPT/LLMs





www.aracom.de

DIE LIMITATIONEN VON CHATGPT/LLMs

Veraltetes öffentliches Wissen

what is your knowledge cuttoff date?

...

My knowledge is up-to-date until March 2023. If you have any questions or need information on topics prior to that, feel free to ask!

Kein internes Wissen

Was ist die Regelung für Urlaubstage?

...

Die Regelung für Urlaubstage kann je nach Land, Region und manchmal auch je nach Unternehmen oder Arbeitsvertrag variieren. In Deutschland zum Beispiel ist der gesetzliche Mindesturlaub im Bundesurlaubsgesetz (BUrlG) geregelt. Demnach hat jeder Arbeitnehmer einen Anspruch auf mindestens 24 Werktage Urlaub pro Jahr bei einer Sechs-Tage-Woche. Bei einer Fünf-Tage-Woche sind es entsprechend 20 Arbeitstage. Das bedeutet, dass Wochenenden nicht als Urlaubstage gezählt werden, da sie keine Werktage sind.

www.aracom.de

EIGENES WISSEN INTEGRIEREN

Prompt Engineering

Lernen über den Kontext

What do you want to train a custom model to do?*

What datasets (or types of data) do you expect to train the model on?*

☐ I acknowledge that it may take several months to train custom models, and that pricing starts at $2–3 million

Submit

RAG

Neue Fakten temporär kontextabhängig lernen

Retrieval Augmented Generation

Darum geht es heute vorrangig

www.qrqcom.de

RAG

Demo

Chat-Oberfläche / Anzeige von Quellen

Mehrsprachigkeit

Unmögliche Fragen

Upload neuer Daten

System Prompt

www.ragcom.de

WARUM IN AZURE?

DATENSCHUTZ

Unterstützung für VNET und Private Endpoints

Keine Verwendung der Daten für Training

Keine Weitergabe der Daten an OpenAI

Auswahlmöglichkeit der Region

INTEGRATION

Integration mit Azure KI Search

Integration mit anderen Azure Diensten wie Application Insights, EntraID etc.

www.aragcom.de

VERSCHIEDENE ANSÄTZE IN AZURE



Kein Code

Copilot Studio

Schnelles Ausprobieren

Beschränkte Möglichkeiten für eigene Daten

Wenig Code

Azure OpenAI Studio

Automatische Indizierung von Inhalten

Keine ACLs

Viel Code

Eigene Oberfläche mit Python, .NET, …

Volle Anpassbarkeit der Oberfläche

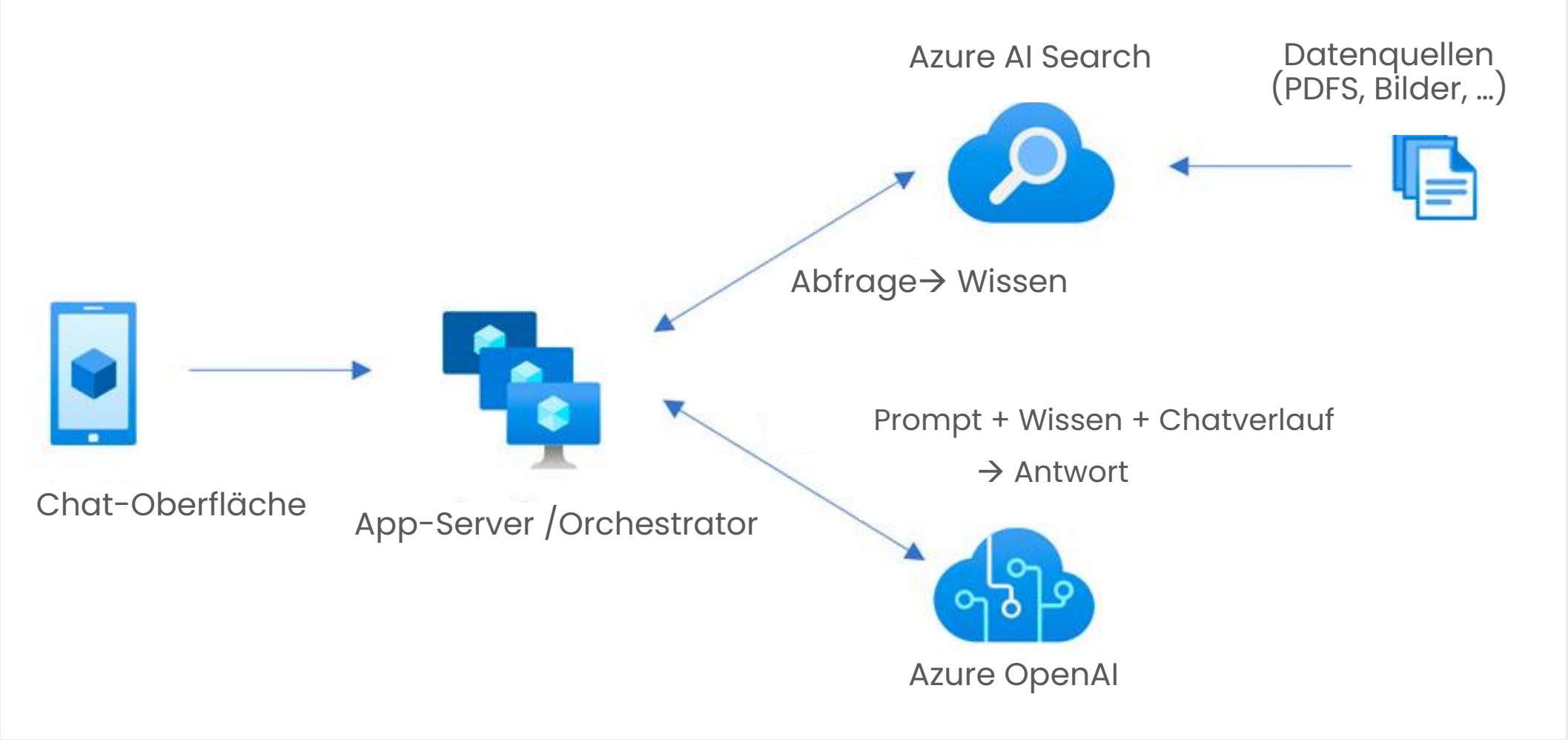
Kontrolle über Suchprozess

Integration Metadaten

Integration ACLs

www.aragcom.de

ARCHITEKTUR DER GROSSEN VARIANTE



www.aragcom.de

DER „WISSENS-ABFRAGE“-SCHRITT

Garbage in, garbage out

Enthalten die Suchergebnisse keine gute Antwort, kann das LLM die Anfrage nicht/nicht richtig beantworten

Zu viel Input

Erhält das LLM zu viele Information, kann es möglicherweise die richtige Antwort im Rauschen finden

Was kann durchsucht werden?

Strukturierte Daten

Indizierte Daten mit Vektoren in Azure AI Search, PostgreSQL (mit pgvector) uvm.

Unstrukturierte Daten

Unstrukturierte Daten (PDFs, Worddokumente, HTML, Bilder, ...) müssen mit einem Extraktionsschritt vorbereitet und in einem Index gespeichert werden

www.aracom.de

DER „WISSENS-ABFRAGE“-SCHRITT - SUCHSTRATEGIEN

Vektor-Suche

Findet die inhaltlich verwandten Dokumente. Verwendet sog. Embeddings.

Schlüsselwort-Suche

Findet Dokumente mit exakter wörtlicher Übereinstimmung

Hybride Suche

Kombiniert Vektor-Suche und Schlüsselwort-Suche, kombiniert die Ergebnisse und sortiert sie mit einem ML-Model.

www.arqcom.de

UPLOAD-PROZESS



Azure Blob Storage

Speicherort für die Quelldateien (zur späteren Anzeige)



Azure Document Intelligence

Extraktion von Text aus den Dateien. Unterstützt PDF, HTML, DOCX, PPTX, ...,



Python

Aufteilung des Texts in Segmente entlang von Token-Längen



Azure OpenAI

Berechnung von Vektoren/

Embeddings



Azure AI Search

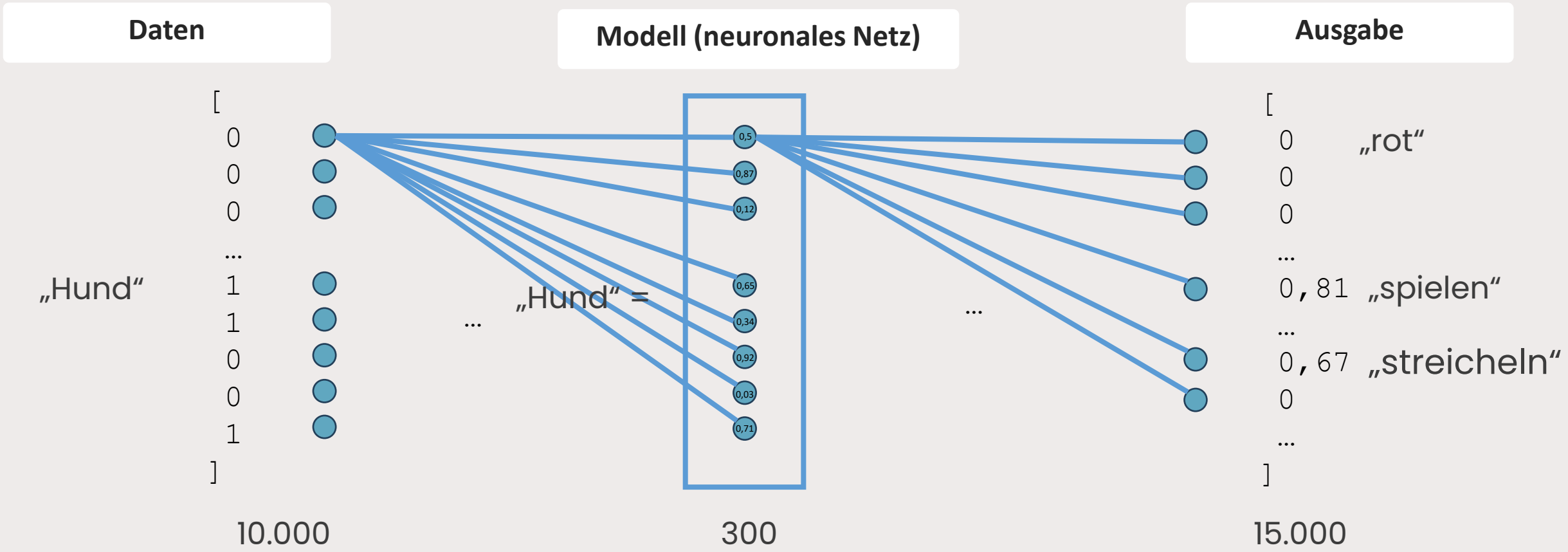
Speichern der Vektoren, Metadaten in einem Suchindex

www.aragcom.de

EMBEDDINGS

Text in Zahlen verwandeln

Idee: Ähnliche Wörter/Bedeutung = Verwendung in ähnlichem Kontext



www.argcom.de

LIMITATIONEN

Kontextlänge

Max. Kontextlänge von Embedding-Model und LLM
→ Tokens

Content-Filter

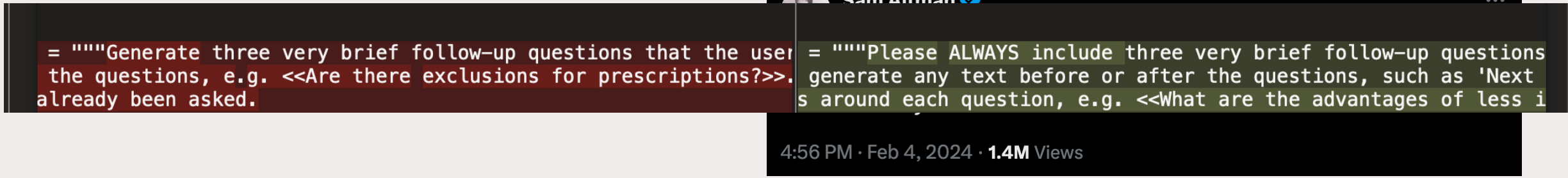
Kuriose Fälle von blockierten Inhalten

Performance

Geschwindigkeit von aktuellen LLMs degradiert mit zunehmender Kontextlänge

Faule LLMs

ChatGPT 4 wurde Anfang des Jahres arbeitsscheu



www.aracom.de

KOSTEN

Azure App Service

ca. 50 - 100 EUR / Monat abhängig von der Skalierung

Azure AI Document Intelligence

€ 1.386 für 1000 Seiten

Azure OpenAI

Abrechnung per Tokens:
€ 0.010 für 1000 Tokens Eingabe
€ 0.028 für 1000 Tokens Ausgabe

Erfahrungswerte:
1500 Eingabe-Tokens und 150 Ausabe-Tokens für eine Chat-Ausgabe → 15ct pro Chat-Schritt mit GPT4.

Azure AI Search

€68.09 / Monat für 2 GB Index-Speicher
€226.52 /Monat für 25 GB Index-Speicher

Azure Blob Storage

€ 0.180 pro GB

www.aracom.de

AUSBLICK

Kontextlänge

Neue Modelle wie Claude 2.1 haben eine Kontextlänge von 200k+ Tokens (ca. 400 Seiten Text) → mehr Wissen für RAG

Model-Performance

Claude 2.1 kann aus 400 Seiten Text zuverlässig Fakten an zufälliger Stelle abrufen („Needle in a haystack“ challenge)

Bessere Integration für Low-Code RAG

Kontinuierliche Verbesserungen in Azure

ChatGPT5

Chat GPT5 soll den gleichen Sprung von v4 machen wie v4 von v3

LSTMs

Neue Architektur (statt “Transformer“) für LLMs mit vorteilhaften Eigenschaften

www.aracom.de

DANKE

FRAGEN?

