ARCHITEKTUR & INFRASTRUKTUR — WHITEPAPER

# Vom Vektorraum zur multimodalen Suche: Warum Embedding LLMs das Fundament moderner KI-Architekturen sind

**Autor:** Michael Kettel (michael.kettel@rm-solutions.de)
Leiter IT, rms. Stuttgart | Veröffentlicht: Juni 2026

Wer über Künstliche Intelligenz im Unternehmenseinsatz spricht, denkt meist zuerst an generative Modelle: Große Sprachmodelle (LLMs), die präzise Antworten formulieren, Code schreiben oder komplexe Berichte zusammenfassen. Doch die mächtigste generative KI bleibt blind, wenn sie nicht mit den richtigen Informationen gefüttert wird. Das unscheinbare, aber technologisch entscheidende Bindeglied für präzise Wissensabfrage (Retrieval-Augmented Generation) sind sogenannte Embedding LLMs. Sie übersetzen unstrukturierte Daten in eine mathematische Landkarte.

## 1. Die Intuition: Eine Landkarte für Bedeutungen

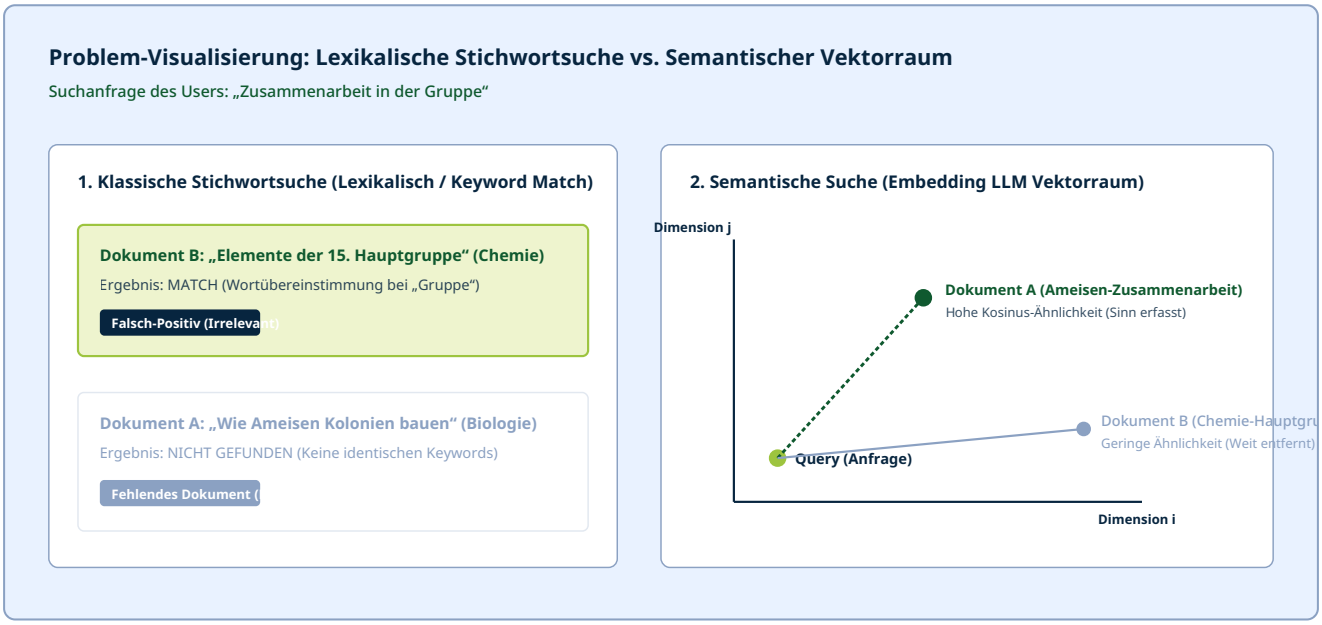
Um die Funktionsweise eines Embeddings zu verstehen, hilft eine Analogie aus dem analogen Alltag. Nehmen wir an, ein Bibliothekar soll in einem ungeordneten Archiv passende Dokumente zum Thema **„Zusammenarbeit in der Gruppe“** heraussuchen. Auf dem Tisch liegen zwei Optionen:

- **Dokument A:** *„Wie Ameisen gemeinsam komplexe Kolonien bauen.“*
- **Dokument B:** *„Die chemische Zusammensetzung von Elementen der 15. Hauptgruppe.“*

Eine klassische, algorithmische Keyword-Suche (wie die traditionelle Strg+F-Funktion) würde primär **Dokument B** auswählen. Der Grund: Hier findet die Zeichenkette **„Gruppe“** eine exakte, wortwörtliche Übereinstimmung – obwohl das Dokument von Chemie handelt und inhaltlich völlig am Kern vorbeigeht.

Jeder menschliche Leser erkennt jedoch sofort, dass **Dokument A** die gesuchte Semantik – den tieergehenden Sinn von Kooperation und kollektiver Arbeit – perfekt abbildet, obwohl das Wort „Gruppe“ oder „Zusammenarbeit“ im Titel gar nicht vorkommt.

Genau diese menschliche Fähigkeit zur semantischen Abstraktion überführen Embedding-Modelle in Mathematik. Sie analysieren Informationseinheiten und weisen ihnen feste Koordinaten in einem hochdimensionalen Raum zu. Inhalte mit ähnlicher Bedeutung landen in dieser mathematischen Landschaft in direkter Nachbarschaft, während semantisch irrelevante Inhalte weit voneinander entfernt positioniert werden.




## 2. Der technische Deep Dive: Vektoren und Distanzmaße

Aus technischer Perspektive handelt es sich bei Embedding-Modellen um spezialisierte Encoder-Architekturen oder modifizierte Teilschichten autoregressiver Transformatoren. Anstatt neue Token sequentiell zu generieren, extrahiert das Modell den internen Zustand (*Hidden State*) der finalen Neuronenschichten. Dies geschieht in der Regel über das sogenannte *Mean Pooling* über alle Token-Repräsentationen hinweg.

Das Resultat dieser Transformation ist ein dichter Vektor – ein Array aus Fließkommazahlen mit einer festen Dimensionalität (typischerweise $d = 768$, $1536$ oder $4096$).

### Mathematische Repräsentation

Ein Dokument oder Textabschnitt $T$ wird durch die Abbildung $f: T \rightarrow v$ in einen Vektor überführt:

$$v = [v1, v2, v3, ..., vd] \in R^d$$

Der semantische Vergleich zweier Entitäten erfolgt anschließend über geometrische Distanzmaße in diesem Vektorraum. Der Standard für das Retrieval ist die **Kosinus-Ähnlichkeit** (Cosine Similarity), welche den Kosinus des Winkels zwischen zwei Vektoren berechnet.

Ein wesentlicher Meilenstein moderner Embedding-Verfahren ist das **Matryoshka Representation Learning (MRL)**. Diese Trainingsmethodik zwingt das Modell dazu, die kritischsten semantischen Informationen in den vorderen Dimensionen des Vektors zu konzentrieren. Dadurch lässt sich ein Vektor

bei Bedarf von 4096 auf beispielsweise 256 Dimensionen kürzen. Die Einsparung an Speicherplatz und Rechenzeit innerhalb der nachgelagerten Vektordatenbank ist massiv, während der Verlust an Retrieval-Genauigkeit marginal bleibt.

## 3. Architektur-Klassen im Vergleich

In der Praxis stehen Systemarchitekten vor der Wahl zwischen verschiedenen Bereitstellungsmodellen. Die Wahl beeinflusst Latenz, Datenschutz und Betriebskosten fundamental.

| Modell-Klasse | Vorteile | Nachteile |
|---|---|---|
| **Proprietäre APIs** (z.B. OpenAI text-embedding-3, Cohere Embed v3) | Minimaler Integrationsaufwand, keine eigene GPU-Infrastruktur erforderlich. Standardmäßig hervorragende Performance und native MRL-Unterstützung. | Abfluss sensibler Unternehmensdaten (Data Outbound). Skalierungskosten bei massiven Datenmengen schwer kalkulierbar. Keine tiefe Gewichtsanpassung (Fine-Tuning) möglich. |
| **Open-Source Text-Embeddings** (z.B. BGE-M3, Jina Embeddings v3) | Vollständige Datenhoheit durch lokales Deployment. Exzellente Multilingualität und Unterstützung langer Kontexte (8k bis 32k Token). Zielgerichtetes Fine-Tuning via Triplet-Loss auf eigene Fachterminologie möglich. | Eigene Hosting- und Wartungsressourcen notwendig. Reine Text-Modelle versagen systematisch, sobald Dokumente komplexe Tabellen, Diagramme oder Scans enthalten. |

## 4. Die Grenze der Multimodalität: Qwen3-VL-Embedding-8B

Klassische RAG-Pipelines stoßen an eine harte Grenze, sobald die Datenbasis aus der realen Unternehmenspraxis stammt: PDF-Berichte mit verschachtelten Layouts, Bilanzen in Tabellenform, Infografiken oder Dashboards. Hier versagt die klassische Kombination aus optischer Zeichenerkennung (OCR) und reinem Text-Embedding oft, da die strukturelle Anordnung der Informationen verloren geht.

Mit der Einführung der **Qwen3-Familie** im Frühjahr 2026 hat Alibaba dieses Problem grundlegend adressiert. Das **Qwen3-VL-Embedding-8B** repräsentiert eine neue Generation nativer, multimodaler Embedding-Modelle (Vision-Language).

### Die technologischen Kernmerkmale des Modells:

- **Unified Semantic Space:** Im Gegensatz zu älteren Architekturen wie CLIP, die separate Encoder für Text und Bild über ein kontrastives Alignment abstimmen müssen, nutzt Qwen3-VL eine tief integrierte Architektur. Roher Text, komplexe Diagramme, Scans und UI-Screenshots werden direkt in denselben, deckungsgleichen Vektorraum projiziert.

- **32k-Token-Kontextfenster:** Für ein Vision-basiertes Modell ist dieses Kontextfenster außergewöhnlich groß. Es erlaubt die hochauflösende Verarbeitung mehrseitiger Dokumente oder sequentieller Video-Frames ohne zerstörerisches vorheriges Partitionieren (Chunking).
- **Natives Visual Document Retrieval (VDR):** Das Modell „sieht“ das visuelle Layout einer Bilanz. Es erfasst die räumliche Relation von Tabellenzellen direkt und macht fehleranfällige OCR-Zwischenschritte überflüssig.

### Infrastrukturelle Implikationen

Diese enorme Leistungsfähigkeit bringt veränderte Anforderungen an die IT-Infrastruktur mit sich. Mit 8 Milliarden Parametern ist das Modell um ein Vielfaches größer als klassische BERT-Abkömmlinge, die oft mit weniger als 512 Millionen Parametern operieren. Ein lokales Deployment erfordert dedizierten VRAM (z.B. NVIDIA A10G oder A100) sowie optimierte Inferenz-Frameworks wie vLLM oder SGLang mit aktivierter FlashAttention-2-Unterstützung. Zudem erfordert das Modell präzise Task-Instructions beim Abruf, um asymmetrische Suchanfragen (z.B. Text-to-Image) mit maximaler Präzision zu verarbeiten.

## 5. Strategisches Fazit für die Enterprise-Architektur

Die Entscheidung für das richtige Embedding-Modell orientiert sich strikt an der Beschaffenheit der Datenmatrix:

Liegt der Fokus auf rein textbasierten Repräsentationen – wie Source-Code-Repositories, bereinigten Markdown-Dokumenten oder strukturierten Datenbankexporten –, bleiben schlanke Text-Embeddings aufgrund minimaler Latenzen und geringer Compute-Kosten die wirtschaftlichste Wahl.

Sobald die Wissensdomäne jedoch durch **visuelle Dokumente, gescannte Verträge, komplexe Industrie-Diagramme oder gemischte Medien** definiert ist, führt kein Weg mehr an multimodalen Ansätzen vorbei. In diesem Segment definiert das **Qwen3-VL-Embedding-8B** den aktuellen State-of-the-Art für datenschutzkonforme, lokal betriebene Enterprise-Pipelines.

## Digitalagentur und KI-Spezialist in Stuttgart

**rms.** ist Ihr spezialisierter Partner für die nahtlose Integration von Künstlicher Intelligenz in Ihre digitalen Ökosysteme. Wir übersetzen die Komplexität moderner Large Language Models (LLM) und KI-Agents in klare, messbare Ergebnisse für Ihr Business.

### Vom hochperformanten Webportal bis zum KI-Chatbot

Als Digitalagentur konzentrieren uns darauf, Effizienz und User Experience in den Schlüsselbereichen zu maximieren:

> **Intelligente Chatbots & KI-Agents:**
> Von der Prozessautomatisierung bis zum hochperformanten Kundenservice entwickeln wir Conversational-AI-Lösungen, die das nächste Level an Interaktion bieten.

> **Content-Exzellenz:**
> Wir implementieren KI-Services zur Textgenerierung, Lokalisierung und Übersetzung direkt in Ihre Content Management Systeme (CMS). Dies beschleunigt Ihre globalen Content-Workflows und garantiert Konsistenz.

> **Smarte Suche mit LLM & RAG:**
> Wir ersetzen starre, veraltete Webseiten-Suchen durch zukunftsweisende Retrieval-Augmented Generation (RAG) Systeme. Ihre Nutzer erhalten dadurch präzise, kontextuelle Antworten auf komplexe Fragen – ein entscheidender Schritt zu besserer Usability.