Factbook

KI-Weltreise – Staffel 8

„RAG – Vom Daten-Sand zur Wissens-Pyramide“

Inhalt

IDEE, TEXT, KONZEPT & LERNAUFBEREITUNG: BIRGIT POHN & ROBERT HORTSCHITZ;
OPTIMIERT UND UNTERSTÜTZT MIT DEN KI SYSTEMEN CHATGPT,COPILOT, GEMINI,
MISTRAL, NOTEBOOKLM; EINE PRODUKTION DER MOGI BUSINESS CREATION COMANY
GMBH & STRO GMBH;

Dieses Factbook fasst die in Staffel 8 behandelten Fachbegriffe, Architekturelemente und regulatorischen Auswirkungen fundiert zusammen. Es richtet sich an Entscheidungsträger, IT-Architekten, Datenschutzbeauftragte und Compliance-Verantwortliche in Österreich und Deutschland.

Der Fokus liegt auf technischer Präzision, regulatorischer Einordnung und marktrelevanter Umsetzung.

1. Grundbegriffe und Buzzwords – verständlich und fachlich korrekt erklärt

Large Language Model (LLM)

Ein Large Language Model ist ein neuronales Netz, das auf großen Textmengen trainiert wurde und Wahrscheinlichkeiten für Wortfolgen berechnet.
Es generiert Text durch statistische Mustererkennung, nicht durch „Verstehen“ im menschlichen Sinne.

Eigenschaften:

- Pre-trained, statisch nach Trainingsphase
- Kontextabhängig im sogenannten Context Window
- Kein persistentes Unternehmenswissen

Regulatorische Relevanz:

- Nach EU AI Act meist General Purpose AI (GPAI)
- Anbieter unterliegen Transparenz- und Dokumentationspflichten
- Anwender müssen Einsatzkontext bewerten

Retrieval-Augmented Generation (RAG)

RAG ist eine Architektur, die ein LLM mit externen Datenquellen kombiniert.

Ablauf:

1. Nutzerfrage wird in einen Vektor umgewandelt
2. Semantische Suche identifiziert relevante Textabschnitte
3. Diese Abschnitte werden dem LLM als Kontext übergeben
4. Das LLM formuliert eine Antwort auf Basis dieser Inhalte

Ziel:

- Reduktion von Halluzinationen
- Nutzung interner Wissensbestände
- Nachvollziehbare Antworten

Regulatorische Einordnung:

- RAG selbst ist kein eigenständiges AI-System nach EU AI Act
- Die Anwendung bestimmt die Risikoklasse
- Dokumentierte Quellen erhöhen Auditierbarkeit

Embeddings

Embeddings sind numerische Vektorrepräsentationen von Text.

Eigenschaften:

- Mehrdimensionale Zahlenvektoren
- Abbildung semantischer Nähe im Vektorraum
- Grundlage für Semantic Search

Technische Relevanz:

- Modellwahl beeinflusst Qualität
- Sprach- und Domänenspezifika sind entscheidend

DSGVO-Aspekt:

- Embeddings können personenbezogene Daten enthalten
- Sie gelten als personenbezogen, wenn Rückschluss möglich ist
- Daher unterliegen sie denselben Schutzpflichten wie Originaldaten

Vektordatenbank

Eine spezialisierte Datenbank zur Speicherung und schnellen Abfrage hochdimensionaler Vektoren.

Typische Funktionen:

- Approximate Nearest Neighbor (ANN)
- Metadaten-Filterung
- Skalierbarkeit

Rechtlicher Aspekt:

- Technisch-organisatorische Maßnahmen nach Art. 32 DSGVO erforderlich
- Zugriffsschutz und Verschlüsselung verpflichtend

Semantic Search

Suche nach inhaltlicher Ähnlichkeit statt Wortgleichheit.

Technische Basis:

- Distanzberechnung im Vektorraum
- Kosinusähnlichkeit

Business-Relevanz:

- Reduktion von Suchzeit
- Erhöhung Trefferqualität

Chunking

Segmentierung großer Dokumente in kleinere Textabschnitte.

Ziel:

- Bessere Suchgranularität
- Effiziente Nutzung des Context Window

Risiko:

- Zu kleine Chunks verlieren Kontext
- Zu große Chunks verringern Präzision

Context Window

Maximale Textmenge, die ein Modell gleichzeitig verarbeiten kann.

Limitation:

- Begrenzte Tokenanzahl
- Beeinflusst Auswahl der Retrieval-Ergebnisse

Grounding

Verankerung der Antwort in konkreten Dokumentquellen.

Vorteile:

- Reduktion von Halluzinationen
- Nachvollziehbarkeit
- Auditfähigkeit

Halluzination

Erzeugung plausibler, aber faktisch falscher Inhalte durch ein LLM.

Ursache:

- Statistische Wahrscheinlichkeitsberechnung
- Fehlender Kontext

RAG reduziert, eliminiert jedoch nicht vollständig.

Prompt Injection

Manipulation durch bösartige Inhalte im Kontext.

Beispiel:
Ein Dokument enthält Anweisungen, Sicherheitsregeln zu ignorieren.

Schutzmaßnahmen:

- Trennung Systemprompt und Kontext

- Content-Validation
- Sicherheitsfilter

Role-Based Access Control (RBAC)

Zugriffssteuerung auf Basis von Rollen.

Relevanz:

- Pflicht bei sensiblen Daten
- Verhindert unautorisierte Informationsausgabe

Identity & Access Management (IAM)

System zur Verwaltung von:

- Benutzeridentitäten
- Rollen
- Berechtigungen

Unverzichtbar für produktive RAG-Systeme.

Data Governance

Gesamtheit der Richtlinien zur Verwaltung von Datenqualität, Zugriff und Lebenszyklus.

Bezug zu RAG:

- Dokumentenqualität beeinflusst Antwortqualität
- Versionierung entscheidend

Approximate Nearest Neighbor (ANN)

Effiziente Näherungssuche im Vektorraum.

Warum notwendig:

- Exakte Suche skaliert schlecht bei großen Datenmengen

Hybrid Search

Kombination aus:

- Keyword-basierter Suche
- Semantischer Suche

Erhöht Präzision bei strukturierten Begriffen.

2. Service- und Produktderivate im DACH-Markt

2.1 Enterprise Knowledge Assistant

Interne Wissensplattform auf Basis von RAG.

Funktionen:

- Dialogbasierter Zugriff auf Dokumente
- Quellenangaben
- Rollenbasierter Zugriff

Typische Einsatzbereiche:

- HR
- IT
- Produktmanagement
- Compliance

Regulatorik:

- Niedriges Risiko bei rein interner Nutzung
- Datenschutz-Folgenabschätzung empfohlen

2.2 Compliance-Assistant

RAG-System für:

- Vertragsanalyse
- Richtlinienabgleich
- Auditvorbereitung

Relevanz:

- Erhöhte Nachvollziehbarkeit
- Dokumentierte Entscheidungsgrundlage

AI Act:

- Kann als High-Risk gelten, wenn Entscheidungsautomatisierung erfolgt
- Menschliche Kontrolle erforderlich

2.3 Technischer Support-Assistent

Zugriff auf:

- Handbücher
- Wartungsprotokolle
- Tickets

Nutzen:

- Reduktion Supportzeit
- Standardisierung Antworten

DSGVO:

- Kundendaten müssen pseudonymisiert werden
- Logging erforderlich

2.4 Legal Research RAG

Juristische Datenbanken kombiniert mit internen Dokumenten.

Besonderheit:

- Erhöhte Haftungsanforderungen
- Dokumentierte Quellen zwingend

AI Act:

- Abhängig vom Einsatzzweck
- Kein High-Risk per se

3. Auswirkungen auf EU AI Act

Klassifizierung

RAG ist eine Architektur, kein eigenständiges AI-System.

Risikoklasse hängt ab von:

- Einsatzbereich
- Automatisierungsgrad
- Auswirkungen auf Rechte natürlicher Personen

Beispiele:

Anwendung	Risikoklasse
Interne Wissenssuche	Minimal
HR-Vorauswahl	High-Risk
Kreditbewertung	High-Risk
Medizinische Unterstützung	High-Risk

Transparenzpflichten

Erforderlich:

- Offenlegung bei Interaktion mit KI
- Dokumentation der Systemarchitektur
- Beschreibung Trainingsdatenquelle des LLM

Menschliche Aufsicht

Bei High-Risk-Systemen verpflichtend:

- Human-in-the-Loop
- Override-Möglichkeiten
- Dokumentierte Entscheidungskontrolle

4. Auswirkungen auf DSGVO (Österreich & Deutschland)

Rechtsgrundlage

Verarbeitung personenbezogener Daten erfordert:

- Art. 6 DSGVO Rechtsgrundlage
- Zweckbindung
- Datensparsamkeit

Embeddings als personenbezogene Daten

Wenn Rückschluss möglich:

- gelten als personenbezogen
- unterliegen Löschpflicht
- müssen korrigierbar sein

Datenübermittlung in Drittstaaten

Cloudbasierte LLMs außerhalb EU:

- Standardvertragsklauseln erforderlich
- Transfer Impact Assessment

Datenschutz-Folgenabschätzung (DPIA)

Empfohlen bei:

- großflächiger Verarbeitung sensibler Daten
- Profiling
- automatisierter Entscheidungsunterstützung

5. Marktüberblick Österreich & Deutschland

Relevante Branchen

- Industrie
- Energie
- Öffentlicher Sektor
- Banken
- Versicherungen
- Gesundheitswesen

Typische Anforderungen

- On-Premise oder EU-Hosting
- ISO 27001 Konformität
- Auditfähigkeit
- Mandantentrennung

6. Technologische Derivate

RAG + Agentic AI

Erweiterung um:

- Aufgabenautomatisierung
- Multi-Step Workflows

Erhöht regulatorisches Risiko.

RAG + Fine-Tuning

Kombination aus:

- Modellanpassung
- Retrieval-Architektur

Vorteil:

- Domänenspezifische Sprachqualität

Nachteil:

- Höhere regulatorische Komplexität

RAG + Structured Data Integration

Integration relationaler Datenbanken.

Ermöglicht:

- KPI-Auswertung
- Business-Analysen

7. Strategische Einordnung

RAG ist im DACH-Markt keine experimentelle Technologie mehr.
Sie wird zur Infrastrukturkomponente für Wissensmanagement.

Erfolgsfaktoren:

- Saubere Datenbasis
- Klare Zugriffskontrolle
- Dokumentierte Architektur
- Compliance-Einbindung

Schlussbetrachtung

Staffel 8 zeigt RAG nicht als Marketingbegriff, sondern als Systemarchitektur.

Die Technologie verbindet:

- Statistik
- Geometrie
- Dokumentenmanagement
- Zugriffskontrolle
- regulatorische Rahmenbedingungen

Im Markt Österreich und Deutschland entscheidet nicht allein technische Leistungsfähigkeit, sondern:

- Rechtskonformität
- Governance-Struktur
- Nachhaltige Integration

RAG ist damit weniger ein Produkt als eine Infrastrukturentscheidung.

Und wie jede Infrastruktur ist sie nur so stabil wie ihr Fundament.