Enterprise RAG – Checkliste & Strategieleitfaden für AI Leads im Mittelstand

Für CIOs & CTOs, AI Leads und Tech-Teams

Zielgruppen des Dokuments:

- AI-Teams die **RAG selbst entwickeln,** also interne Prozesse mit RAG optimieren oder eigene Produkte mit KI anreichern möchten
- AI Leads, die **interne Implementierungen challengen**, Probleme & Risiken identifizieren wollen
- AI Leads, die Kriterien für den **Einkauf eines RAG-Systems** verstehen möchten

Autor:
Marco Bürckel, Gründer der axio concept GmbH

Über axio concept:
Robuste KI für Industrie und regulierte Branchen seit 2017. Anbieter Enterprise KI-Plattform, individuelle Lösungen und Beratung für KI.
20 Mitarbeiter.

Beratung/Entwicklung RAG für:
Pharma/Chemie, Med-Tech, Kanzleien, Fachverlage, Normung, Zertifizierer, Software-Hersteller

Beispiele aus der Praxis (umgesetzt von uns):

- **Sales/Service:** RAG für Datenblätter, Specs etc. im Anlagen-/Maschinenbau
- **Einkauf:** Prüfung von Compliance-Nachweisen, Abgleich von ABs, RFQs etc.
- **Wartung/Instandhaltung:** Unterstützung bei Troubleshooting an Anlagen
- **Kanzleien/Beratung:** Recherche in Beratungsthemen
- **Regulatory Compliance:** Automatische Gap-Analyse gegen Normen, Impact-Analyse bei Normenänderungen
- **Fachverlage:** Durchsuchen großer Content-Bestände, Ausleitung neuer Formate, Prüfung auf interne Leitfäden/Richtlinien

---

Warum RAG in der Praxis scheitert

- **Unterkomplexe RAG-Pipeline:** Ingestion oder Retrieval zu primitiv für die Komplexität der Dokumente – welche i.d. meisten Unternehmen HOCH ist. Unternehmen arbeiten mit sehr komplexen/verschiedenartigen Dokumenten, während die meisten RAG-Pipelines Spielzeug sind.
  Symptome: Halluzinationen, fehlende Antworten, Themen werden durcheinandergeworfen.

- **RAG-System nicht an Unternehmen angepasst:** RAG muss ein Stück weit an Prozesswelt / Fragestellungen / fachliche Domäne / Fachbegriffe / Dokumentarten / etc. angepasst werden, sonst werden Ergebnisse immer Mittelmaß sein. Merkregel: “KI von der Stange” = Illusion.
  Symptome: Ebenfalls Halluzinationen, schlechte Antworten, …

- **Fehlende Fallback-Strategien:** Abhängigkeit von der Verfügbarkeit einzelner Cloud-Modelle und RAG-Services
  Symptome: Provider-Ausfälle schlagen direkt auf produktive Prozesse durch. Updates von KI-Modellen und RAG-Services sorgen plötzlich für vermehrt Fehler.

- **Kostenexplosion:** Die eigene RAG-Pipeline schluckt deutlich mehr Token als erwartet – Kosten werden unkontrollierbar. Oder angemietete RAG-Systeme in der Cloud verschlingen hohe, monatliche Gebühren
  Symptome: KI-Lösungen rechnen sich kaum noch

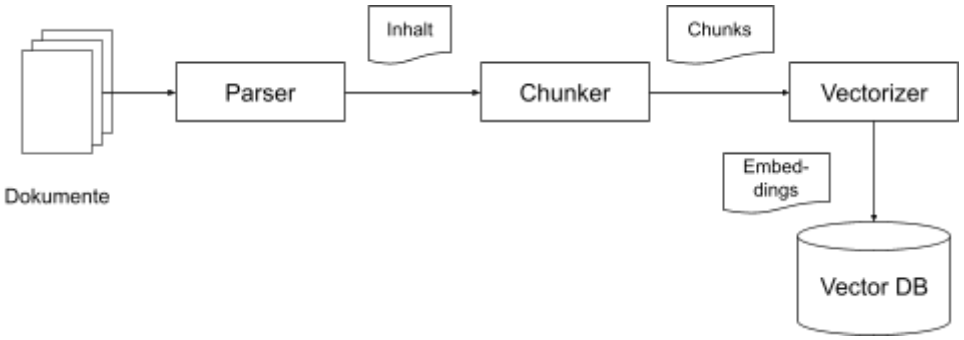
- **Fehlende Sicherheitsarchitektur:** Zugriffsrechte werden nicht konsequent abgebildet, hohe Zahl an Angriffsvektoren im Betrieb (z.B. Wartungszugänge) oder andere vermeidbare Fehler im Bereich IT-Sicherheit.
  Symptome: Personen haben unberechtigt Zugriff (Klassiker: Kein Zugriff auf SharePoint, aber Zugriff auf Daten in SharePoint über KI-Chatbot)

- **Fehlende Quality-Gates:** Keine Benchmarks, Modell-Cards, versionierte Systeme.
  Symptome: Keine Möglichkeit, Modell-Drift zu erkennen oder Optimierungen vorzunehmen, weil man im Blindflug ist.

Beobachtung: KI-Projekte scheitern häufig **nicht** am Modell, sondern am Engineering der RAG-Pipeline selbst (sei es bei eingekauften oder selbst geschaffenen Lösungen)

Naives RAG – Wie die meisten RAG-Implementierungen starten und enden

Die typische RAG-Implementierung: Primitive Ingestion + Retrieval auf Basis von Embeddings.



Inhalte werden auf Basis semantischer Ähnlichkeit in einem hochdimensionalen Vektorraum (Latent Space) abgebildet. Ähnliche Inhalte liegen nahe beieinander. Richtungen kodieren (implizit) Bedeutungen.
Suche nach Kontext = Umkreissuche

Ziele der Vektorsuche:

- Relevante Inhalte für Anfrage erkennen, unabhängig der Formulierung
- Schneller Suchalgorithmus

Naive RAG-Systeme holen Rohtext aus Dokumenten und führen eine Embedding-Suche nach obigem Schema aus. **Zum Scheitern verurteilt.**

Typische Schwächen:

- **Inhaltlich ähnlich != relevant**
  Beispiel: Ähnliche Wartungsprotokolle für unterschiedliche Anlagen – gleiche Schritte, unterschiedliche Grenzwerte. Vermischung gefährlich. Und vorprogrammiert.

- **Seltene Fachbegriffe, Abkürzungen etc. haben keinen eindeutigen Bedeutungsvektor**
  Suche nach diesen Begriffen funktioniert nicht (gut).

- **Keine Adaption an verschiedene Fragestellungen**
  “Wie ging der Rechtsstreit Meier ./. Müller aus” vs.
  “Welche Rechtsfälle haben wir für Herrn Meier bearbeitet?”

- **Kein temporales Verständnis oder Verständnis von Gültigkeit/Versionierung**
  “Was lief im letzten Meeting” liefert Punkte aus einem oder mehreren älteren Meetings
  “Wie ist die aktuelle Richtlinie zu …” liefert Stand aus alter Richtlinie

- **Schwache Ingestion: Alles wird Rohtext**
  Unternehmensdokumente sind selten Prosatext. Mehrspaltige Dokumente, Entscheidungstabellen, Formeln, Spezifikationen – alles wird als Rohtext herausgezogen, Kontext geht verloren. In Layout kodierte Informationen gehen verloren. Relevante Metadaten werden nicht extrahiert.

- **Keine Berücksichtigung von Rollen und Rechten**
  Jeder sieht alles. Entfernte Rechte im Primärsystem (ERP, DMS, …) werden nicht gespiegelt.

- **Keine Verknüpfungsfähigkeit**
  Fragen erfordern oft mehrere Retrieval-Schnitte
  (“Welche Kunden sind von der Preiserhöhung betroffen und haben zugleich Sonderkonditionen im Rahmenvertrag?”)

Checkliste: Ist Ihr RAG-System enterprise-ready?

Die folgenden Kriterien leiten sich direkt aus den typischen Schwächen naiver RAG-Systeme ab. Sie helfen bei der Bewertung einer internen Implementierung oder eines externen Anbieters.

Ingestion & Dokumentenverarbeitung

☐ Werden Tabellen, Formeln und Layouts strukturiert verarbeitet – oder wird alles als Fließtext extrahiert?

☐ Gibt es Strategien, um grafisch kodierte Informationen (Flussdiagramme, aufwändige Präsentationen etc.) gezielt zu ingesten?

☐ Werden Metadaten (Datum, Version, Dokumenttyp, Quelle) mit indexiert?

☐ Gibt es eine Strategie für verschiedene Dokumentarten (PDFs, CAD, E-Mails, ERP-Exporte etc.)? Sind diese flexibel erweiterbar?

☐ Werden Ihre branchenspezifischen Formate und Daten durch geeignete Strategien berücksichtigt (Bspw. S1000D in Defense)?

☐ Werden Aktualisierungen in Quellsystemen automatisch erkannt und nachindexiert?

☐ Werden RBAC-relevante Aktualisierungen (Rechte, Rollen, Gruppenzuordnungen) in Quellsystemen automatisch erkannt und nachindexiert?

Retrieval & Suchqualität

☐ Wird ausschließlich auf Embedding-Suche gesetzt – oder gibt es ergänzende Strategien (z.B. Keyword-basierte Suche, Filter, Reranking)?

☐ Werden Fachbegriffe und Abkürzungen zuverlässig erkannt?

☐ Kann das System zwischen verschiedenen Fragetypen unterscheiden? (Faktenabfrage vs. Vergleich vs. Zusammenfassung vs. Recherche)

☐ Gibt es ein temporales Verständnis? Werden aktuelle Versionen gegenüber veralteten priorisiert?

☐ Sind mehrstufige Abfragen möglich? (z.B. "Welche Kunden haben Rahmenvertrag UND sind von Preiserhöhung betroffen?"), Stichwort Agentic RAG / Multi-Hop-QA

Sicherheit & Berechtigungen

☐ Werden Zugriffsrechte aus Primärsystemen (ERP, DMS, SharePoint etc.) konsequent durchgereicht?

☐ Sind Angriffsvektoren im Betrieb identifiziert und abgesichert? (Wartungszugänge, API-Endpunkte, Prompt Injection etc.)

☐ Falls externes RAG-System: Hat der Anbieter dauerhaften Zugriff über einen Wartungs-Account? **Gefährlich!! – Immer nur temporäre Zugriffe.**

☐ Falls externes RAG-System: Wie geht der Anbieter mit Testdaten um, die er von Ihnen benötigt? Landen die direkt auf den Rechnern der Entwickler? **Gefährlich!! Gerade bei Daten im regulierten / KRITIS-Umfeld.**

☐ Liegen Unternehmensdaten ausschließlich in kontrollierter Infrastruktur – oder werden sie an Drittanbieter-APIs geschickt?

☐ Falls On-Prem: Gibt es irgendwelche zentralen Komponenten, die einen Datenabfluss verursachen (z.B. zentralisiertes Logging beim Anbieter)?

Betrieb & Qualitätssicherung

☐ Gibt es Benchmarks und automatisierte Tests für die Antwortqualität?

☐ Sind System und Modelle versioniert – kann man zu einem früheren Stand zurückrollen?

☐ Wird Modell-Drift erkannt? (z.B. nach Provider-Updates, Modellwechsel)

☐ Gibt es Fallback-Strategien bei Ausfall einzelner Cloud-Modelle oder Services?

Kosten & Skalierbarkeit

☐ Sind die Token-Kosten pro Anfrage bekannt und kalkulierbar?

☐ Gibt es eine Strategie zur Kostenoptimierung bei steigendem Volumen? (z.B. Caching, kleinere Modelle für Standardfragen)

☐ Besteht ein Lock-in zu einem einzelnen Cloud-Anbieter – oder ist das System modell-agnostisch?

☐ Falls On-Prem: Wie skaliert die Lösung im Parallelbetrieb, durch gleichzeitige Anfrage mehrerer Benutzer? Bedeutung für Hardware-Aufwand?

☐ Falls On-Prem: Kann der Anbieter ein kostengünstiges Hardware-Setup anbieten, das sich an den konkreten Erfordernissen orientiert?

Scoring:

- **20+ Punkte erfüllt:** Solide Basis. Feintuning und Skalierung stehen im Fokus.
- **13–19 Punkte:** Relevante Lücken. Gezielte Optimierung nötig, bevor das System in Produktion geht.
- **Unter 13 Punkte:** Hohes Risiko. Die Architektur sollte grundlegend überprüft werden.

Nächste Schritte

RAG für Industrieunternehmen ist kein Standardprojekt. Die Komplexität liegt nicht in der Technik allein, sondern in der Verbindung von **Technik, Fachprozessen und Unternehmensrealität.**

Wir unterstützen auf zwei Wegen:

→ **Sie haben ein Dev-Team und wollen RAG selbst bauen?**

RAG-Workshop & Architekturberatung Wir challengen Ihre Architektur, identifizieren Schwachstellen und geben konkrete Empfehlungen für Ingestion, Retrieval und Betrieb.

Ergebnis: Klare Roadmap, technische Entscheidungsgrundlage, vermiedene Sackgassen.

→ **Sie wollen eine fertige, erprobte Lösung?**

KogniLink Enterprise KI-Plattform Unsere Plattform löst die in diesem Dokument beschriebenen Probleme out of the box: Robuste Ingestion, intelligentes Retrieval, Rechtemanagement, Quality-Gates und Fallback-Strategien – **auch integrierbar in Softwaresysteme**

Ergebnis: Produktionsreifes RAG-System, angepasst an Ihre Prozesse und Dokumentenwelt.

Kontakt:
https://www.axioconcept.com
Direkt Termin buchen: https://calendly.com/axioconcept/30min