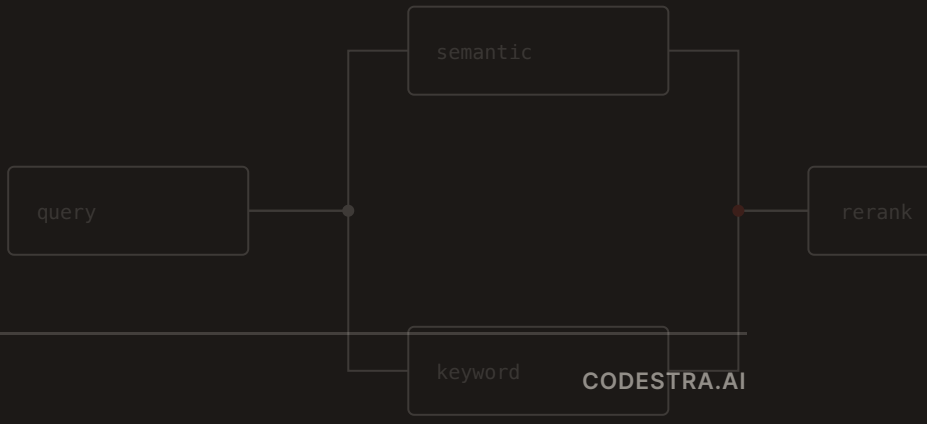
codestra

WHITEPAPER · KI-GESTÜTZTE SUCHE

Hybrid-Suche für technische Dokumentation im Maschinen- und Anlagenbau

Warum klassische Suchlösungen an ihre Grenzen stoßen - und wie moderne KI-gestützte Suche technisches Wissen zuverlässig nutzbar macht.



KURZFASSUNG

Technisches Wissen ist vorhanden - nur selten dort auffindbar, wo es gebraucht wird.

Technische Dokumentation im Maschinen- und Anlagenbau ist umfangreich, heterogen und geschäftskritisch. Betriebsanleitungen, Wartungspläne, Ersatzteillisten, Prüfprotokolle, Schaltpläne und Konformitätserklärungen enthalten Wissen, das Servicetechniker, Vertrieb und Konstruktion täglich benötigen. In der Praxis ist dieses Wissen oft schwer auffindbar: Es liegt verteilt über mehrere Systeme, in unterschiedlichen Versionen und Sprachen.

Untersuchungen zur Wissensarbeit zeigen, dass ein erheblicher Teil der Arbeitszeit auf das Suchen und Zusammentragen von Informationen entfällt[1]. Im Service schlägt sich das unmittelbar in einer zentralen Kennzahl nieder: der Erstlösungsquote (First-Time Fix Rate). Ob ein Techniker einen Fehler beim ersten Einsatz behebt, hängt nicht nur von Ersatzteilen und Qualifikation ab, sondern auch vom Zugriff auf die richtige Information zum richtigen Zeitpunkt[7].

Klassische Suchlösungen lösen dieses Problem nur teilweise. Eine reine Stichwortsuche findet exakte Begriffe, scheitert aber an Synonymen, Fachterminologie und Fragen in natürlicher Sprache. Eine reine KI-Vektorsuche versteht Bedeutung, schwächelt jedoch bei exakten technischen Werten, Artikelnummern und Normbezeichnungen[3,4]. Forschungsergebnisse belegen, dass kein einzelnes Suchverfahren über alle Anwendungsfälle hinweg konsistent überlegen ist[3].

Die Antwort darauf ist die hybride Suche, häufig als Bestandteil von Retrieval-Augmented Generation (RAG) eingesetzt: eine Kombination aus semantischer und schlüsselwortbasierter Suche, ergänzt um eine Bewertungsstufe und eine sprachlich aufbereitete, quellenbasierte Antwort. Dieser Ansatz spielt seine Stärke gerade dort aus, wo technische Dokumentation hohe terminologische Präzision mit natürlicher Sprache verbindet.

Dieses Whitepaper erläutert, warum klassische Suche im Industrieumfeld an Grenzen stößt, wie hybride Suche funktioniert, wo ihre Grenzen liegen und welchen wirtschaftlichen Nutzen Unternehmen im Maschinen- und Anlagenbau erwarten können.

01 Ausgangslage: Technisches Wissen liegt oft brach

1.1 Die Dokumentenrealität

Ein Unternehmen im Maschinen- und Anlagenbau erzeugt über den Lebenszyklus seiner Produkte hinweg Tausende technischer Dokumente: Betriebsanleitungen in mehreren Sprachen, Wartungspläne, Ersatzteillisten, Änderungshistorien, Prüfzertifikate, Schulungsunterlagen und projektspezifische Anpassungsdokumentationen.

Diese Dokumente liegen typischerweise verteilt über ERP-Systeme, PDM- und PLM-Systeme, Dokumentenmanagementsysteme sowie lokale Netzlaufwerke und E-Mail-Postfächer. Für die einzelne Mitarbeiterin oder den einzelnen Mitarbeiter bedeutet das: Wer eine konkrete Information benötigt, muss im besten Fall wissen, in welchem System sie liegt, und im schlechtesten Fall mehrere Systeme nacheinander durchsuchen oder einen Kollegen fragen.

1.2 Die Kosten schlechter Auffindbarkeit

Die Kosten unzureichender Auffindbarkeit werden selten direkt gemessen, sind aber relevant. Das McKinsey Global Institute berichtet, dass Wissensarbeiter einen erheblichen Anteil ihrer Arbeitswoche mit dem Suchen nach internen Informationen und dem Kontaktieren von Kollegen verbringen, und dass durchsuchbare, gut strukturierte Wissensspeicher diesen Aufwand spürbar senken können[1].

Im Service ist der Effekt besonders konkret messbar. Die Erstlösungsquote ist eine etablierte Kennzahl des technischen Außendienstes. Sie misst den Anteil der Einsätze, die beim ersten Besuch abgeschlossen werden, ohne dass zusätzliche Teile, Informationen oder ein zweiter Termin nötig werden[7]. In der Fertigung und im Anlagengeschäft liegt diese Quote je nach Komplexität in einem breiten Bereich; leistungsstarke Serviceorganisationen erreichen Werte um die 88 Prozent[7]. Eine niedrige Erstlösungsquote wirkt sich nachweislich negativ auf Anlagenverfügbarkeit, Kundenzufriedenheit und Vertragstreue aus[7]. Der Zugriff auf die richtige Information ist dabei einer der Hebel, der die Erstlösungsquote direkt beeinflusst.

Eine belastbare Aussage über die Kosten im eigenen Unternehmen lässt sich nur aus den eigenen Daten ableiten. Eine einfache Modellrechnung verdeutlicht jedoch die Größenordnung. Sie ist als Beispiel zu verstehen, nicht als gemessenes Ergebnis:

MODELLRECHNUNG · ILLUSTRATIVES BEISPIEL

Angenommen, 100 technische Mitarbeitende verbringen pro Arbeitstag durchschnittlich 30 Minuten mit der Suche nach Dokumenten und Informationen. Das ist eine konservative Annahme, die deutlich unter der von McKinsey berichteten Größenordnung liegt[1]. Bei rund

220 Arbeitstagen pro Jahr ergeben sich daraus etwa 11.000 Arbeitsstunden, die nicht in die eigentliche Wertschöpfung fließen.

30 Min	220	≈ 11.000
SUCHZEIT PRO TAG UND PERSON	ARBEITSTAGE PRO JAHR	STUNDEN P.A. OHNE WERTSCHÖPFUNG (BEI 100 MITARBEITENDEN)

Der monetäre Wert dieser Stunden hängt vom jeweiligen Vollkostensatz ab. Entscheidend ist nicht die exakte Zahl, sondern der Mechanismus: Jede eingesparte Minute Suchzeit, jeder vermiedene zweite Serviceeinsatz und jede schnellere Angebotserstellung wirkt unmittelbar auf Produktivität und Servicequalität.

02 Warum klassische Suche an Grenzen stößt

2.1 Stichwortsuche: präzise, aber bedeutungsblind

Die in vielen Dokumentenmanagementsystemen integrierte Volltext- oder Stichwortsuche, technisch meist auf Verfahren wie BM25 aufbauend, findet Dokumente anhand exakter oder ähnlicher Begriffe. Sie ist stark, wenn die Suchanfrage seltene, eindeutige Begriffe enthält: Artikelnummern, Fehlercodes, Normbezeichnungen[4,6]. Sie scheitert jedoch, sobald:

- Synonyme oder unterschiedliche Fachbegriffe verwendet werden. Eine Suche nach „Hydraulikzylinder“ findet ein Dokument nicht, das von einem „doppeltwirkenden Linearaktuator“ spricht.
- Fragen in natürlicher Sprache gestellt werden, etwa „Welches Anzugsmoment gilt für die Flanschverbindung der Hochdruckpumpe?“.
- Inhalte umschrieben statt wortgleich benannt sind.

2.2 KI-Vektorsuche: bedeutungsstark, aber ungenau im Detail

Moderne KI-Vektorsuche bildet Anfrage und Dokument in einem semantischen Raum ab und findet inhaltlich verwandte Passagen auch ohne exakte Wortübereinstimmung. Sie löst das Synonymproblem und versteht Kontext, auch sprachübergreifend.

Ihre Schwäche liegt jedoch genau dort, wo technische Dokumentation besonders anspruchsvoll ist: bei exakter Präzision. Studien zeigen, dass dichte Vektormodelle seltene Begriffe, Eigennamen und Fachterminologie weniger zuverlässig treffen als klassische Stichwortverfahren und in eng spezialisierten Fachdomänen schlechter generalisieren[3,4]. Eine rein semantische Suche nach einem spezifischen Drehmomentwert für eine bestimmte Schraubengröße kann eine thematisch passende, aber im Detail falsche Passage zurückliefern. Im industriellen Kontext ist das nicht nur unbefriedigend, sondern potenziell sicherheitsrelevant.

Stichwortsuche `BM25 · sparse`

+ Exakte Treffer bei Artikelnummern, Fehlercodes, Normen
– Blind für Synonyme & natürliche Sprache

KI-Vektorsuche `dense · semantic`

+ Versteht Bedeutung & Kontext, sprachübergreifend
– Ungenau bei exakten Werten & Fachterminologie

2.3 Die zentrale Erkenntnis

Der umfassende BEIR-Benchmark, der zehn moderne Suchverfahren über achtzehn Datensätze hinweg verglichen hat, kommt zu einem klaren Ergebnis: Über alle Aufgaben und Domänen hinweg konsistent gut abzuschneiden, ist für ein einzelnes Verfahren schwierig. Die klassische Stichwortsuche bleibt eine robuste Referenz, während dichte Modelle je nach Aufgabe deutlich darunter liegen können[3]. Die Stärken und Schwächen beider Verfahren sind komplementär. Genau hier setzt die hybride Suche an.

03 Hybride Suche: das Prinzip

3.1 Was hybride Suche kombiniert

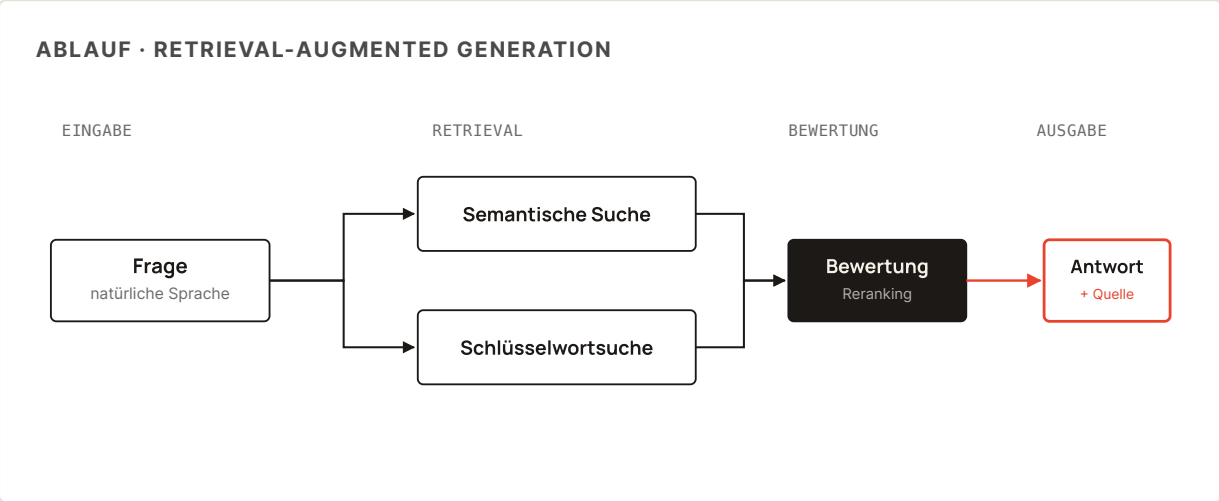
Hybride Suche verbindet zwei Suchverfahren, die sich gegenseitig ergänzen:

Semantische Suche versteht die Bedeutung einer Frage. Sie findet relevante Inhalte auch bei abweichender Wortwahl, erkennt Zusammenhänge und arbeitet sprachübergreifend.

Schlüsselwortbasierte Suche findet exakte Treffer. Sie ist präzise bei Artikelnummern, Maßangaben, Fehlercodes und Normbezeichnungen.

Eine Bewertungsstufe (Reranking) gewichtet und kombiniert die Ergebnisse beider Suchwege und stellt die relevantesten Passagen nach oben.

Im Rahmen von Retrieval-Augmented Generation wird das Ergebnis anschließend von einem Sprachmodell zu einer verständlichen Antwort zusammengefasst, mit direktem Verweis auf die zugrunde liegenden Originaldokumente. Das Verfahren der Retrieval-Augmented Generation wurde 2020 wissenschaftlich eingeführt und verbindet ein Sprachmodell mit einer externen, durchsuchbaren Wissensquelle, sodass Antworten auf belegbaren Dokumenten beruhen statt allein auf dem im Modell gespeicherten Wissen[2].



3.2 Warum dieser Ansatz für den Maschinen- und Anlagenbau passt

Technische Dokumentation in diesem Sektor hat Eigenschaften, die hybride Suche besonders wertvoll machen:

Hohe terminologische Dichte. Normen, interne Artikelnummern und maschinenspezifische Bezeichnungen erfordern eine Präzision, die die schlüsselwortbasierte Komponente liefert. Studien zeigen, dass genau hier dichte Vektormodelle allein an Grenzen stoßen[4].

Mehrsprachigkeit. Viele Unternehmen pflegen Dokumentation auf Deutsch und Englisch, teils in weiteren Sprachen. Die semantische Komponente ermöglicht sprachübergreifende Suche.

Heterogene Formate. Vom strukturierten PDF über Office-Dokumente bis zur eingescannten Altdokumentation: Die richtige Aufbereitung dieser Quellen ist Voraussetzung für brauchbare Ergebnisse.

Hohe Anforderung an Korrektheit. Im Gegensatz zu vielen Anwendungen im Privatkundenbereich ist im industriellen Umfeld entscheidend, dass eine Antwort belegbar ist. Das Prinzip der Quellenbindung, bei dem die Antwort auf konkrete Dokumente verweist, ist hier ein wichtiger Mechanismus zur Nachvollziehbarkeit.

04 Anwendungsfälle

Die folgenden Szenarien illustrieren typische Einsatzfelder. Sie beschreiben den Mechanismus, nicht garantierte Ergebnisse; der konkrete Nutzen hängt von Datenlage und Umsetzung ab.

01 Servicetechniker im Einsatz

Ein Techniker steht vor einer Anlage. Ein Fehlercode wird angezeigt. Mit klassischer Suche findet er im System mehrere Dokumente zu diesem Code und muss entscheiden, welches für diese Anlagengeneration und diesen Revisionsstand gilt, häufig mit einer Rückfrage im Büro.

Mit hybrider Suche stellt er die Frage in natürlicher Sprache, etwa: „Was bedeutet dieser Fehlercode bei dieser Baureihe und wie behebe ich ihn?“ Er erhält eine Antwort mit konkreter Handlungsanweisung und Verweis auf das passende Kapitel der gültigen Dokumentation. Da der Zugriff auf die richtige Information ein direkter Hebel auf die Erstlösungsquote ist[7], wirkt dieser Anwendungsfall unmittelbar auf Servicequalität und Anlagenverfügbarkeit.

02 Angebotserstellung im Vertrieb

Ein Vertriebsingenieur erstellt ein Angebot für eine kundenspezifische Anlage und benötigt technische Spezifikationen, vergleichbare frühere Projekte und Angaben zur Normkonformität. Statt mehrere Systeme manuell zu durchsuchen, stellt er eine gebündelte Frage und erhält eine konsolidierte Antwort aus Datenblättern, Projektberichten und Zertifikaten, jeweils mit Quellenverweis.

03 Einarbeitung neuer Mitarbeitender

Erfahrene Mitarbeitende verfügen über implizites Wissen, das selten vollständig dokumentiert ist. Neue Kolleginnen und Kollegen finden relevante Unterlagen oft nicht und fragen daher häufig nach. Mit hybrider Suche können sie Fragen direkt an die vorhandene Dokumentation stellen, etwa zu Unterschieden zwischen zwei Baureihen, und werden schneller eigenständig handlungsfähig. Das entlastet zugleich die erfahrenen Mitarbeitenden.

05 Grenzen und Voraussetzungen für belastbare Ergebnisse

Eine seriöse Darstellung benennt auch die Grenzen der Technologie. Drei Punkte sind besonders wichtig.

Quellenbindung reduziert Fehler, schließt sie aber nicht aus.
Retrieval-Augmented Generation senkt das Risiko erfundener Inhalte deutlich, weil Antworten auf abgerufenen Dokumenten beruhen. Die Forschung zeigt jedoch, dass auch bei korrekter Dokumentengrundlage Antworten entstehen können, die über die Quelle hinausgehen oder ihr widersprechen[5,6]. Quellenangaben in der Antwort bedeuten daher nicht automatisch, dass jede Aussage korrekt ist. Mechanismen zur Prüfung und Bewertung der Antwortqualität gehören zu einer verantwortungsvollen Umsetzung.

Die Qualität der Suche bestimmt die Qualität der Antwort.
Wird die falsche Passage abgerufen, kann auch das beste Sprachmodell keine korrekte Antwort erzeugen[5]. Gerade deshalb ist die hybride Kombination aus semantischer und schlüsselwortbasierter Suche und eine saubere Bewertungsstufe entscheidend, nicht ein einzelnes Verfahren.

Die Qualität der Dokumente bestimmt die Qualität des Systems.
Veraltete, widersprüchliche oder schlecht strukturierte Dokumente führen zu schlechten Ergebnissen. Eine initiale Sichtung und Bereinigung des relevanten Dokumentenbestands ist Teil jeder ernsthaften Einführung.

06 Datensicherheit und Souveränität

Für Unternehmen im Maschinen- und Anlagenbau ist der Schutz technischen Know-hows existenziell. Konstruktionsdaten, Fertigungsparameter und kundenspezifische Lösungen sind Kern des Wettbewerbsvorteils. Eine tragfähige Lösung für diesen Sektor sollte daher:

- in einer dedizierten Cloud-Umgebung in der EU oder On-Premises betreibbar sein,
- rollenbasierte Zugriffssteuerung unterstützen, sodass nicht jede Person jedes Dokument sieht,
- DSGVO-konform dokumentiert und auditierbar sein,
- für besonders sensible Umgebungen ohne ständige externe Anbindung funktionieren.

Ein häufig übersehener Punkt verdient Klarstellung: Vollständige Datensouveränität setzt voraus, dass auch das Sprachmodell lokal oder in der kontrollierten Umgebung betrieben wird. Wird für die Antworterzeugung eine externe Modell-Schnittstelle genutzt, verlassen Inhalte diese Umgebung. Wer maximale Souveränität benötigt, kombiniert daher einen lokalen Suchindex mit einem lokal betriebenen Sprachmodell. Dieser Abwägung zwischen Souveränität, Aufwand und Antwortqualität sollte bewusst und dokumentiert getroffen werden.

07 Der Weg zur Einführung

Eine Einführung verläuft sinnvollerweise in aufeinander aufbauenden Schritten, die sich nach Inhalt und Ergebnis gliedern, nicht nach festen Zeitvorgaben.

01

Pilot

Auswahl und Aufbereitung eines klar abgegrenzten Korpus, etwa der Wartungsdokumentation einer Produktlinie. Aufbau des Suchindex und erster Zugang für ausgewählte Nutzerinnen und Nutzer. Ziel ist der Nachweis des konkreten Nutzens an einem realen Ausschnitt.

02

Integration

Anbindung an vorhandene Systeme, Erweiterung des Dokumentenbestands, Einweisung der Nutzer und Feinjustierung der Suchergebnisse auf Basis des Feedbacks aus dem Pilotbetrieb.

03

Betrieb & Pflege

Breiterer Zugriff, laufende Überwachung von Nutzung und Antwortqualität sowie kontinuierliche Aktualisierung bei neuen oder geänderten Dokumenten.

Voraussetzungen auf Unternehmensseite sind ein benannter Verantwortlicher für das Dokumentenmanagement, ein klares Konzept für Zugriffsrechte, die Bereitschaft zur initialen Dokumentenbereinigung und die Freigabe der erforderlichen Infrastruktur.

08 Wirtschaftlicher Nutzen

Der Nutzen hybrider Suche lässt sich auf vier konkrete Hebel zurückführen:

Schnellere Serviceprozesse
Besserer Informationszugriff wirkt direkt auf die Erstlösungsquote - eine etablierte Kennzahl mit dokumentiertem Einfluss auf Anlagenverfügbarkeit und Kundenzufriedenheit[7].

Kürzere Vertriebszyklen
Schnellerer Zugriff auf technische Informationen und vergleichbare Projekte beschleunigt die Angebotserstellung und senkt das Risiko fehlerhafter Spezifikationen.

Sicherung von Wissen
Implizites Wissen wird über die durchsuchbare Dokumentation zugänglich, bevor es durch Fluktuation oder altersbedingtes Ausscheiden verloren geht.

Schnellere Einarbeitung
Neue Mitarbeitende werden früher eigenständig produktiv, erfahrene Kolleginnen und Kollegen werden entlastet.

Die belastbare Quantifizierung dieser Effekte erfolgt am besten in einem Pilotprojekt anhand der eigenen Kennzahlen. Damit wird aus einer plausiblen Erwartung ein messbarer Geschäftsfall.

09 Fazit

Der Maschinen- und Anlagenbau steht unter anhaltendem Innovations- und Effizienzdruck, während erfahrene Fachkräfte das Unternehmen verlassen und implizites Wissen mitnehmen. Technisches Wissen verlässlich nutzbar zu machen, ist damit ein realer Wettbewerbsfaktor.

Klassische Suche allein leistet das nicht: Stichwortsuche ist präzise, aber bedeutungsblind, KI-Vektorsuche bedeutungsstark, aber ungenau im Detail. Die Forschung zeigt klar, dass beide Verfahren komplementäre Stärken haben und kein einzelnes konsistent überlegen ist[3,4]. Die hybride Kombination, eingebettet in eine quellenbasierte Antwortgenerierung, ist der Ansatz, der den spezifischen Anforderungen technischer Dokumentation gerecht wird, sofern Suchqualität, Dokumentenqualität und Antwortprüfung mitgedacht werden.

DER NÄCHSTE SCHRITT

Der sinnvolle erste Schritt ist klein und überprüfbar: Ein Pilot mit einem klar abgegrenzten Dokumentenbestand zeigt am eigenen Material, welchen Nutzen der Ansatz für Ihr Unternehmen liefert.

---

codestra

Über codestra

codestra ist ein auf KI-Architekturen spezialisiertes Beratungs- und Implementierungsunternehmen für den Mittelstand und größere Industrieunternehmen im deutschsprachigen Raum. Wir begleiten die Einführung KI-gestützter Wissens- und Suchsysteme von der Konzeption über die Evaluierung bis zum produktiven Betrieb, mit besonderem Fokus auf Datenqualität, Bewertbarkeit und Datensouveränität.

hello@codestra.ai · codestra.ai

Dieses Whitepaper wurde von codestra erstellt. Die genannte Modellrechnung in Abschnitt 1.2 ist ein illustratives Beispiel und kein gemessenes Ergebnis. Konkrete Effekte sollten anhand der eigenen Kennzahlen in einem Pilotprojekt ermittelt werden.

Quellen

1 **McKinsey Global Institute (2012):** The Social Economy: Unlocking value and productivity through social technologies. Bericht zur Wissensarbeit; interaction worker verbringen knapp 20 Prozent der Arbeitswoche mit der Suche nach internen Informationen, durchsuchbare Wissensspeicher können diesen Aufwand um bis zu 35 Prozent senken. mckinsey.com

2 **Lewis, P., Perez, E., Piktus, A., et al. (2020):** Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems 33 (NeurIPS 2020). arXiv:2005.11401

3 **Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I. (2021):** BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. NeurIPS 2021 Datasets and Benchmarks Track. arXiv:2104.08663. Zentrales Ergebnis: Kein einzelnes Verfahren ist über alle Datensätze konsistent stark; die klassische Stichwortsuche bleibt eine robuste Referenz.

4 **Chen, X., Lakhotia, K., Oguz, B., et al. (2022):** Salient Phrase Aware Dense Retrieval: Can a Dense Retriever Imitate a Sparse One? Findings of the ACL: EMNLP 2022, S. 250–262 (arXiv:2110.06918). Befund: Dichte Vektormodelle treffen seltene Begriffe und Eigennamen weniger zuverlässig als klassische Stichwortverfahren und generalisieren schwächer auf fachfremde Daten.

5 **Ni, B., Liu, Z., Wang, L., et al. (2025):** Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey. arXiv:2502.06872 (eingereicht 8. Februar 2025). Übersichtsarbeit; Retrieval-Augmented Generation reduziert Halluzinationen, führt aber neue Risiken ein, die für vertrauenswürdige Systeme adressiert werden müssen.

6 **Xiong, G., He, Z., Liu, B., Sinha, S., Zhang, A. (2025):** Toward Faithful Retrieval-Augmented Generation with Sparse Autoencoders. arXiv:2512.08892. Befund: Die Bindung an abgerufene Dokumente beseitigt unfaithful generierte Inhalte nicht vollständig; Antworten können der Quelle widersprechen oder über sie hinausgehen.

7 **PTC: What is First-Time Fix Rate** (ptc.com) sowie **IBM: What is first-time fix rate** (ibm.com). Definition und Branchenwerte der Erstlösungsquote; leistungsstarke Serviceorganisationen lösen rund 88 Prozent der Fälle beim ersten Einsatz. Eine von der Aberdeen Group berichtete Erstlösungsquote unter 70 Prozent wirkt sich negativ auf Anlagenverfügbarkeit, Kundenzufriedenheit, Vertragstreue und SLA-Einhaltung aus.