WHITE PAPER

Ist RAG der Schlüssel zur datenschutz-konformen KI?

Was die Datenschutzkonferenz zur RAG-Methode sagt - und warum wir mit Wissensgraphen einen Schritt weitergehen

findic

Inhalt

Über dieses White Paper

Große Sprachmodelle werden datenschutzrechtlich weiterhin kritisch betrachtet: Trainingsdaten lassen sich aus fertigen Modellen kaum gezielt entfernen, Auskunft und Löschung stoßen an strukturelle Grenzen. Mit der Orientierungshilfe zur RAG-Methode hat die Datenschutzkonferenz im Oktober 2025 erstmals eine Architektur beschrieben, die diesen Konflikt entschärfen kann, allerdings nur unter klaren Voraussetzungen. Dieses White Paper ordnet die Orientierungshilfe aus der Betriebspraxis ein: welche Aussagen im Betrieb tragfähig sind, wo die Methode endet und warum wir bei findic bereits einen weiteren Schritt gehen. Unsere Plattform organisiert Wissen als Graphen, in dem jeder Inhalt adressierbar, zweckgebunden und nachweisbar löschbar ist; als Datenschutzarchitektur, nicht nur als Datenbank.

1 · Datenschutz beginnt bei der Architektur

Die Grundprobleme großer Sprachmodelle mit dem Datenschutz liegen weniger im einzelnen Verhalten eines Systems als in seiner Architektur. Werden personenbezogene Daten für das Training verwendet, gehen sie in eine Vielzahl von Parametern ein und sind später kaum noch isoliert erreichbar. Dass Trainingsinhalte aus fertigen Modellen extrahiert werden können, hat die Sicherheitsforschung früh und deutlich gezeigt.[1] Der Europäische Datenschutzausschuss zieht daraus eine Konsequenz: Ein mit personenbezogenen Daten trainiertes Modell ist nicht automatisch anonym; Anonymität muss im Einzelfall belegt werden.[2] Die Datenschutzkonferenz formuliert entsprechend, dass die Umsetzung von Betroffenenrechten in Sprachmodellen weitgehend ungelöst ist.[3]

Wie Kontrolle über lernende Systeme zurückgewonnen werden kann, beschäftigt die deutsche Rechtswissenschaft seit Jahren.[4,5,6] Die derzeit naheliegende Antwort ist pragmatisch: Daten werden nicht in das Modell verlagert, sondern außerhalb des Modells gehalten. Für ein Institut, das KI auf eigene Bestände anwenden will, etwa Kreditakten, Beratungsprotokolle oder Beschwerdemanagement, ist daher weniger entscheidend, welches Modell eingesetzt wird. Entscheidend ist, wo die Daten verarbeitet und gespeichert werden. Befinden sie sich im Modell, übernimmt das Unternehmen auch dessen datenschutzrechtliche Probleme. Bleiben sie in einer kontrollierten Schicht außerhalb des Modells, greifen dort bekannte Regeln: Rechte und Rollen, Fristen, Nachweise. An dieser Stelle setzt die RAG-Methode an.[7] Deshalb hat ihr die Aufsicht eine eigene Orientierungshilfe gewidmet, auf die wir hier eingehen.

---

[1] Carlini, N. et al. (2021): Extracting Training Data from Large Language Models. USENIX Security 2021

[2] EDSA (2024): Stellungnahme 28/2024 zu gewissen Datenschutzaspekten im Zusammenhang mit der Verarbeitung personenbezogener Daten im Kontext von KI-Modellen, 17. Dezember 2024

[3] DSK (2025): Orientierungshilfe zu datenschutzrechtlichen Besonderheiten generativer KI-Systeme mit RAG-Methode. Version 1.0, Stand Oktober 2025, Kap. 1 und 3.7.

[4] Martini, M. (2019): Blackbox Algorithmus – Grundfragen einer Regulierung künstlicher Intelligenz, Springer

[5] Wischmeyer, T. (2018): Regulierung intelligenter Systeme. Archiv des öffentlichen Rechts 143, 1–66

[6] Vogel, P. (2022): Künstliche Intelligenz und Datenschutz. Nomos

[7] 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

2 · Die Aufsicht hat geantwortet: die Orientierungshilfe im Kern

Am 17. Oktober 2025 hat die Datenschutzkonferenz ihre Orientierungshilfe zu datenschutzrechtlichen Besonderheiten generativer KI-Systeme mit RAG-Methode veröffentlicht.[8] Es ist die dritte KI-Veröffentlichung der Aufsicht seit 2024 und die erste, die eine konkrete Systemarchitektur näher betrachtet. Für Dienstleister ist die Veröffentlichung deshalb in zweierlei Hinsicht relevant: als rechtliche Einordnung und als Hinweis darauf, welche technischen und organisatorischen Anforderungen im Betrieb erwartet werden können.

Der Kern: Die DSK bewertet ein RAG-System entlang des Grundsatzkatalogs des Art. 5 DSGVO und der Betroffenenrechte – Richtigkeit, Transparenz, Integrität und Vertraulichkeit, Zweckbindung, Datenminimierung, Rechtmäßigkeit. Ihr Befund ist differenziert. Die Methode kann positive Effekte auf Richtigkeit und Nachvollziehbarkeit der Ausgaben entfalten; die Vertraulichkeit und Integrität eingebundener personenbezogener Daten lassen sich verbessern; die Inhalte der Referenzschicht, anders als antrainiertes Wissen, können aktualisiert, beauskunftet und gelöscht werden.[9] Damit qualifiziert RAG als risikomindernde Maßnahme, die bei der Rechtmäßigkeitsprüfung zugunsten des Verantwortlichen ins Gewicht fallen kann.[10]

Zugleich ist die Veröffentlichung ausdrücklich kein Freibrief. Drei Klarstellungen gehören zu jedem Projektstart dazu. Erstens: Ein rechtswidrig trainiertes Sprachmodell bleibt rechtswidrig – die RAG-Methode heilt das Training nicht. Zweitens: Jede Verarbeitung ist im Einzelfall zu bewerten; eine Pauschalfreigabe gibt es nicht. Weiter entstehen mit der Referenzschicht neue Verarbeitungen, die selbst eine Rechtsgrundlage brauchen. Embeddings personenbezogener Daten in einer Vektordatenbank sind Verarbeitung personenbezogener Daten, kein technischer Zwischenschritt ohne Rechtsfolgen.

Für Institute kommt ein zweiter Rahmen hinzu, den die Orientierungshilfe nur streift: Die DSK betrachtet das RAG-Gesamtsystem – Retriever, Wissensspeicher und Sprachmodell – als KI-System im Sinne des Art. 3 Nr. 1 KI-VO. RAG als Methode ist dabei keine eigenständige Risikoklasse. Die Einstufung folgt dem Einsatzzweck. Bei Kreditwürdigkeitsprüfung und ähnlichen Anwendungsfällen nach Anhang III Nr. 5 KI-VO liegt die Hochrisikoeinstufung nahe. Mit dem Digital Omnibus wurden die Anwendungszeitpunkte verschoben: Die Hochrisiko-Pflichten für eigenständige Anhang-III-Systeme gelten nunmehr ab dem 2. Dezember 2027,[11] für Hochrisikosysteme als Sicherheitsbauteile von Produkten (Anhang I) ab dem 2. August 2028. Unverändert gelten bereits die Transparenzpflichten des Art. 50 KI-VO seit dem 2. August 2026 und die KI-Kompetenzpflicht des Art. 4 seit dem 2. Februar 2025. Die Verschiebung verschafft Vorlauf. Sie ändert aber nichts an den Pflichten selbst. Die datenschutzrechtliche Architektur, die dieses White Paper beschreibt, ersetzt also keine KI-VO-Compliance. Sie bietet eine Lösung für deren technisches Fundament, weil Risikomanagement, Datenqualität und Protokollierung auf denselben Kontrollpunkten aufsetzen.

[8] DSK (2025): Orientierungshilfe RAG-Methode, a.a.O.
[9] DSK (2025): Orientierungshilfe RAG-Methode, Kap. 3.1 und 3.7
[10] DSK (2025): Orientierungshilfe RAG-Methode, a.a.O., Kap. 3.6
[11] Verordnung (EU) 2024/1689 in der Fassung des „Digital Omnibus on AI“, ABl. 24. Juli 2026

Der Ton der Veröffentlichung ist für die Praxis relevant. Die Aufsicht erkennt an, dass RAG-Systeme eigenständig entwickelt, betrieben und kontrolliert werden können und dadurch sich verbesserter Datenschutz durch Technikgestaltung ergeben kann. Zugleich ordnet sie die Methode strategisch ein: Werden kleinere Modelle lokal betrieben, kann die Übermittlung personenbezogener Daten an Hyperscaler vermieden und die digitale Souveränität gestärkt werden. Für Anbieter und Anwender ergibt sich daraus kein Automatismus, aber ein klarer Arbeitsauftrag: Die Architektur muss so gebaut und dokumentiert sein, dass sie prüfbar bleibt.

RAG kann damit ein wichtiger Baustein für datenschutzkonforme KI sein. Die Orientierungshilfe macht aber ebenso deutlich, dass die Methode nur innerhalb bestimmter Grenzen trägt. Der weitere Text beschreibt, wie weit man innerhalb dieser Grenzen kommt, wenn die Architektur konsequent auf Kontrolle, Nachweisbarkeit und Löschbarkeit ausgelegt wird.

3 · Was die Methode trägt und wo sie endet

Hebel, die im Betrieb wirken

Richtigkeit statt bloßer Wahrscheinlichkeit. Die Antwort eines RAG-Systems basiert auf einer geprüften und kuratierten Dokumentenbasis. Halluzinationen lassen sich dadurch reduzieren; unrichtige oder veraltete Inhalte werden in den Referenzdokumenten korrigiert, nicht im Modell. Der Grundsatz der Richtigkeit nach Art. 5 Abs. 1 lit. d DSGVO wird damit technisch besser unterstützbar.[12]

Nachvollziehbarkeit durch Quellenbindung: Jede Antwort kann ihre Belege mitliefern; der Input der Generierung ist dokumentierbar. Das schafft Einblick, wenn auch begrenzt auf die erweiterte Anfrage. Das Innere des Modells bleibt auch in einem RAG-System dunkel, und die DSK sagt das deutlich.[13]

Vertraulichkeit mit bekannten Instrumenten: In der Datenbankschicht können Mandantentrennung sowie Rechte- und Rollenkonzepte umgesetzt werden. Das sind Instrumente, die im Sprachmodell selbst nur eingeschränkt zur Verfügung stehen. Dort lässt sich nicht verlässlich steuern, wer auf welche antrainierte Information Zugriff hat. Die Aufsicht geht davon aus, dass in der Referenzschicht auch personenbezogene Daten mit höherem Schutzbedarf verarbeitet werden können, weil diese Daten nicht dauerhaft im Sprachmodell verbleiben und die Zugriffskontrolle auf Datenbankebene erfolgt.[14]

Zweckbindung durch Rollen: Die Bereitstellung bestimmter Dokumente für bestimmte Rollen beschränkt Abfragen zielgerichtet auf den definierten Verarbeitungszweck, dabei technisch umgesetzt über funktionale Trennung der Bestände und saubere Rollenzuordnung vor der Abfrage.

Löschbarkeit im Regelbetrieb: Einträge in der Referenzschicht sind direkt adressierbar; klassische Fristenlogik kann darauf angewendet werden. Aus der Finanzwelt sind die entsprechenden Instrumente bekannt: Löschkonzepte, Aufbewahrungspflichten und revisionssichere Nachweise. Die KI-Anwendung wird damit stärker an die etablierten Kontrollmechanismen der übrigen IT angebunden.

[12] Kühling, J. & Buchner, B. (Hrsg.) (2024): Datenschutz-Grundverordnung, Bundesdatenschutzgesetz – DS-GVO/BDSG. 4. Auflage, C.H. Beck, Art. 5 DSGVO.
[13] DSK (2025): Orientierungshilfe RAG-Methode, Kap. 3.2
[14] DSK (2025): Orientierungshilfe RAG-Methode, Kap. 3.3

Grenzen, die man kennen muss

Erstens bleibt das Training des Sprachmodells, was es ist. Die datenschutzrechtliche Bewertung der LLM-Komponente ändert sich durch RAG nicht.[15] Zweitens entsteht ein Verkettungsrisiko, das man nicht unterschätzen darf: Gibt man personenbezogene Daten aus der Referenzschicht an ein Modell, dessen Trainingsinhalt man nicht kennt, muss man Vorkehrung treffen, dass sich beide Bestände nicht verbinden. Drittens sind Embeddings selbst intransparent: Warum ein Chunk einen bestimmten Vektor erhält, lässt sich nicht nachvollziehen; die Vektordatenbank ist damit ein eigener schutzwürdiger Speicher. Viertens kann die Referenzschicht eine neue Angriffsfläche schaffen. In kontrollierten

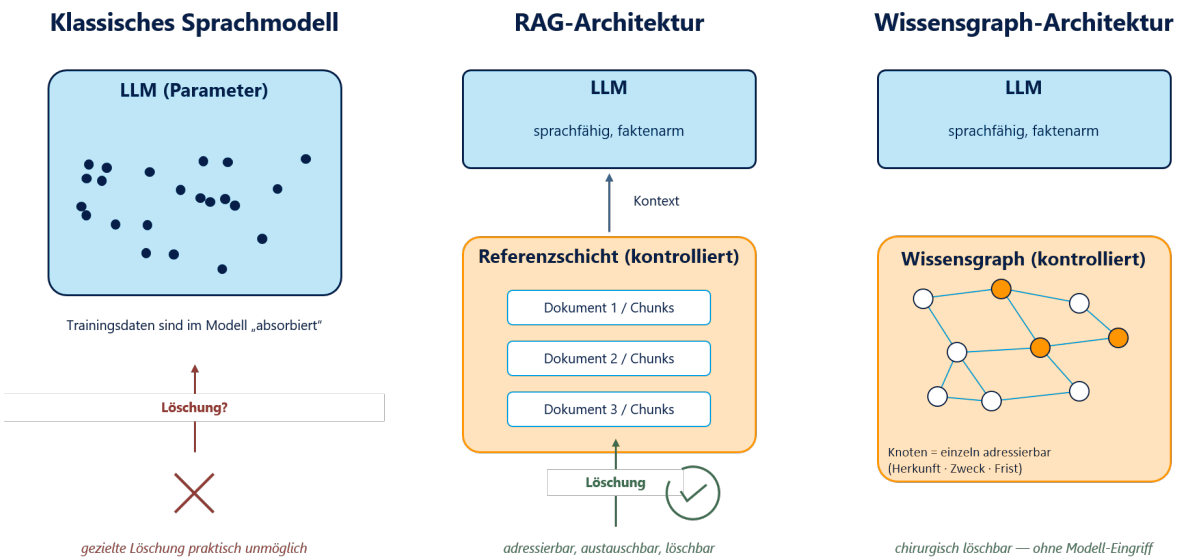

Abbildung 1 – Kontrollpunkte entlang des Grundsatzkatalogs: jede System-schicht erhält eigene, klassische Kontrollinstrumente

Laborversuchen zu Poisoning-Angriffen ließen sich RAG-Systeme bereits mit wenigen injizierten Schadtexten pro Zielfrage in neun von zehn Fällen manipulieren;[16] auch Jamming-Angriffe über gezielt platzierte Blocker-Dokumente sind nachgewiesen;[17] Membership-Inference- und Poisoning-Angriffe gehören deshalb in jede Bedrohungsanalyse.[18]

Wer diese Grenzen ernst nimmt, landet nicht bei Resignation, sondern bei einer Aufgabenliste. Die folgende Abbildung ordnet die Kontrollpunkte den Schichten des Systems zu.

[15] EDSA (2024): Stellungnahme 28/2024 zu gewissen Datenschutzaspekten im Zusammenhang mit der Verarbeitung personenbezogener Daten im Kontext von KI-Modellen, 17. Dezember 2024

[16] Zou, W., Geng, R., Wang, B. & Jia, J. (2025): PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. USENIX Security 2025, 3827–3844

[17] Shafran, A., Schuster, R. & Shmatikov, V. (2025): Machine Against the RAG: Jamming Retrieval-Augmented Generation with Blocker Documents. USENIX Security Symposium, 3787–3806

[18] Rigaki, M. & Garcia, S. (2020): A Survey of Privacy Attacks in Machine Learning. arXiv:2007.07646

4 · Aus dem Konzept in den Betrieb

Orientierungshilfen beschreiben den Rahmen; im Betrieb müssen daraus konkrete Kontrollpunkte werden. Aus unseren Aufträgen haben sich sechs Punkte bewährt, mit denen sich die Anforderungen der DSK in laufende Systeme übersetzen lassen.

Erstens: Folgenabschätzung als Auftakt, nicht als Anhang. Ein RAG-System umfasst Retriever, Wissensspeicher und Sprachmodell. Die Datenschutz-Folgenabschätzung nach Art. 35 DSGVO muss alle drei Komponenten und ihr Zusammenwirken erfassen, samt Dokumentation der erweiterten Anfragen und Datenflüsse.

Zweitens: Dokumentenhygiene vor jedem Embedding. Die Aufsicht nennt es selbst: Kopf- und Fußzeilen, Seitenzahlen und sonstige Störddaten sind zu bereinigen; Textabschnitte sollen abgeschlossene Sinnabschnitte bilden; nicht erforderliche personenbezogene Daten sind zu entfernen oder zu anonymisieren. Unsere eigene Ergänzung dieser Liste: Für deutschsprachige Bestände setzen wir ein Embedding-Modell ein, das mit deutschen Texten trainiert wurde. In der Praxis entscheidet diese unspektakuläre Arbeit über die halbe Antwortqualität.

Drittens: Quellenbindung per Systemprompt – und ihre Überwachung. Sprachmodelle tendieren bei Widersprüchen dazu, antrainiertes Wissen über den gelieferten Kontext zu stellen. Die DSK empfiehlt Systemprompts, die das Modell anweisen, ausschließlich aus den referenzierten Quellen zu antworten. Wir behandeln diese Anweisungen wie Konfiguration mit Versionspflicht: Jede Änderung wird dokumentiert und getestet.

Viertens: Modellwahl als Datenschutzentscheidung. Da das Faktenwissen in der Referenzschicht liegt, muss das Sprachmodell weniger memorieren. Kleinere Modelle, lokal betrieben, werden dadurch realistisch. Auch die Praxisforschung stützt diese Architektur: Für die Beisteuerung von Faktenwissen ist die Anreicherung im Abruf der flexiblere Weg; Fine-Tuning bleibt ein ergänzendes Instrument für Stil und Format, nicht der Ersatz für eine gepflegte Wissensbasis.[19,20]

Fünftens: Fristen- und Rechtemanagement der Referenzschicht. Löschfristen, Aufbewahrungspflichten und Rechte-Rollen-Konzepte laufen auf der Datenbankschicht mit, als wäre sie ein klassisches Fachverfahren. Genau das ist der Punkt: Sie ist eins.

Sechstens: Monitoring und Schulung. Ausgaben werden (stichprobenartig) gegen ihre Quellen geprüft, Protokolle lückenlos geführt, Anwenderinnen und Anwender geschult. Die schwierigste Stelle ist dabei selten die Technik. Es ist der Kalender. Die Frage, wer prüft, wann, mit welchem Ergebnis, muss festgeschrieben sein, sonst verkommt die risikomindernde Maßnahme zur Absichtserklärung.

[19]Balaguer, A. et al. (2024): RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. arXiv:2401.08406.

[20]Honroth, T., Siebert, J. & Kelbert, P. (2024): Retrieval Augmented Generation (RAG): Chatten mit den eigenen Daten. Fraunhofer IESE Blog, 13. Mai 2024

5 · Der findic-Schritt: der Wissensgraph als Datenschutzarchitektur

An dieser Stelle verlassen wir die Orientierungshilfe und beschreiben, was wir selbst gebaut haben und in Mandaten betreiben. Ausgangspunkt ist eine Beobachtung, die jede RAG-Implementierung früher oder später macht: Ein Chunk ist ein Textstück mit Vektor. Was er über eine Person enthält, wofür er verwendet werden darf, wann er zu löschen ist, all das muss das System mühsam separat organisieren. Die Datenschutzlogik hängt an der Datenbank wie ein Anhänger.

Wir haben die Plattform deshalb auf Wissensgraphen gebaut. Inhalte werden nicht zu anonymen Textstücken, sondern zu vernetzten Knoten, wobei jeder Knoten seine Rechtemetadaten bei sich trägt: Herkunft, Zweck, Löschfrist, Freigabeklasse. Die Struktur folgt demselben Grundgedanken, mit dem die GraphRAG-Forschung Dokumentenbestände in navigierbare Wissensräume überführt.[21] Wir haben ihn um die Dimension ergänzt, die im Finanzumfeld zählt: die rechtliche Adressierbarkeit jedes einzelnen Inhalts.

Was das verändert, zeigt sich am konkreten Fall. Eine Auskunft nach Art. 15 DSGVO wird zur Pfadabfrage: Welche Knoten enthalten Daten zu dieser Person, aus welcher Quelle, für welchen Zweck? Eine Berichtigung nach Art. 16 wird zum Knoten-Update, eine Löschung nach Art. 17 zum Entfernen eines Knotens samt seiner Kanten, protokolliert, in Sekunden, ohne dass ein Modell angefasst werden muss (Abbildung 2). Das meint unserer Ansicht nach die DSK, wenn sie die direkte Adressierbarkeit der Einträge als entscheidenden Vorteil beschreibt.[22] Wir haben das zur Grundlage der Architektur gemacht.

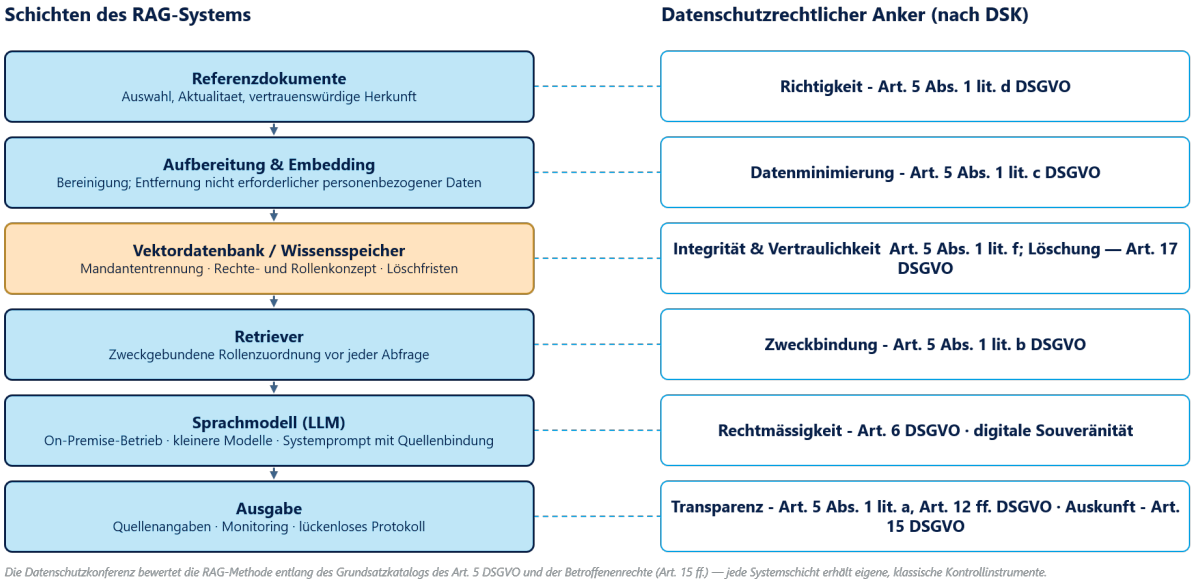

Abbildung 2 – Auskunft, Berichtigung und Löschung als Graph-Operation: jeder Knoten trägt Herkunft, Zweck und Frist bei sich

[21] Edge, D. et al. (2024): From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv:2404.16130.

[22] DSK (2025): Orientierungshilfe RAG-Methode, Kap. 3.7

Es gibt einen zweiten, fachlichen Grund für den Graphen, und er ist fachlich. Die klassische RAG-Suche hat eine strukturelle Schwäche, die in der Praxis schnell sichtbar wird. Semantische Nähe verfehlt logische Zusammenhänge. Komplexe Wenn-Dann-Ketten über lange Textpassagen hinweg landen nicht in einem Chunk und erreichen die Anfrage unvollständig. Finanzrecht besteht aus genau solchen Ketten: Ausnahmen, die auf Ausnahmen verweisen, Fristen, die von Definitionen abhängen. Ein Wissensgraph hält diese Beziehungen explizit fest, statt sie der Vektornähe zu überlassen. Unsere Branchenkonfiguration bildet die Verweisungslogik des Finanzrechts so ab, wie Juristinnen und Juristen sie denken, nicht, wie ein Embedding sie vermutet.

Aufgebaut wird der Graph weitgehend autonom aus den vorhandenen Unternehmensquellen. Das Domänenwissen steckt in der Strukturierung, nicht in manueller Nacharbeit. Automatisiert heißt dabei nicht ungeprüft: Neue Kanten werden nur mit Beleg in den Quelldaten übernommen, also dieselbe Nachweislogik, die unsere Graph-White Paper beschreiben. Konkret: Für eindeutige Entitäten bleibt die klassische Extraktion unverändert im Einsatz. Wo dieselbe Wortform mehrere Fachbegriffe trägt, werden Kanten dagegen nicht aus einer einmaligen Extraktion übernommen, sondern paarweise gegen die Quellstellen verifiziert. Manche Häuser prüfen derzeit noch erste RAG-Pilotprojekte. Bei uns werden Löschverlangen inzwischen mit einem Protokolleintrag beantwortet.

Auch ein Graph ist kein Zaubermittel. Die Metadatenpflege erfordert Disziplin, die Zwecktrennung gelebte Rollen. Was ein Graph nicht kann, kann er nicht, etwa die datenschutzrechtliche Bewertung des verwendeten Sprachmodells abnehmen. Und er löst das Löschproblem nur in seiner eigenen Schicht: Die erweiterten Anfragen ans Sprachmodell, deren Dokumentation die DSK verlangt, enthalten selbst personenbezogene Daten und liegen in Protokollen und Logs. Auch hier gelten Frist und Zweck; wir führen die Löschlogik des Graphen deshalb auf die Protokollschicht fort, statt sie dort enden zu lassen. Er macht aber den Teil der Architektur, den das Haus kontrolliert, vollständig kontrollierbar. Abbildung 3 fasst die drei Lagen noch einmal gegenüber.

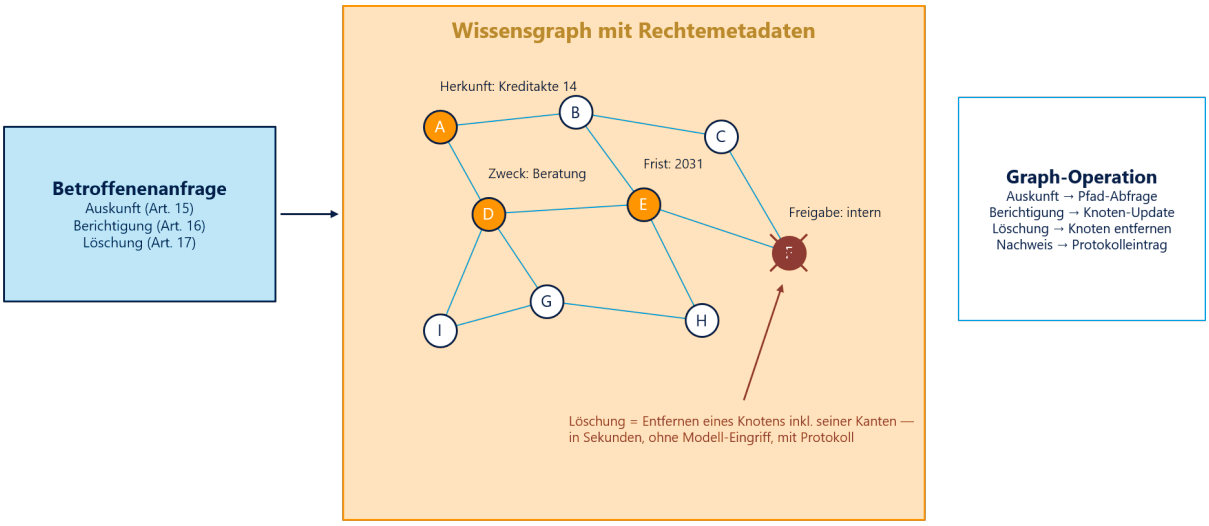

Abbildung 3 – Wo die personenbezogenen Daten verbleiben: im Modell, in der Referenzschicht oder im adressierbaren Wissensgraphen

6 · Erfahrungen aus dem Betrieb

Drei Episoden aus dem letzten Jahr, weil sie mehr erklären als jede Architekturfolie.

Der Löschfall. Während der Prüfungsvorbereitung eines Instituts fiel auf, dass ältere Schulungsunterlagen in der Wissensbasis Namen und Beurteilungsdetails von Mitarbeitenden enthielten, die dort nichts zu suchen hatten. In der früheren Vektorarchitektur hätte das bedeutet: Index analysieren, Chunks identifizieren, Neuaufbau gegenprüfen – Tage. Im Graph war es eine Abfrage, ein Löschlauf und ein Protokolleintrag. Die Datenschutzbeauftragte des Hauses hat den Vorgang anschließend als Muster für ihr Löschkonzept übernommen. Das eigentliche Kompliment war, dass keine Mail mehr nötig war.

Die Auskunft. Weiter verlangte eine Kundin Auskunft darüber, welche ihrer Daten dem KI-System zugänglich seien. Früher wäre das ein wochenlanges Suchprojekt durch Akten und Datenbanken gewesen. Über die Metadaten der Knoten, wie Herkunft, Zweck, Freigabeklasse, ließ sich die Antwort als Pfadabfrage formulieren und in einer Stunde dokumentieren.

Die Modellwahl. Unser erstes Deployment lief auf einem großen Modell, weil es verfügbar war und beeindruckte. Datenschutzfachlich war es die schwächste Option: mehr memorisiertes Wissen, mehr Abhängigkeit, mehr Erklärungsbedarf. Wir sind auf ein kleineres, lokal betriebenes Modell umgestiegen und haben die Kontexttreue über Graph und Systemprompt abgesichert und die Vorteile dieser Architektur genutzt. Die Antwortqualität blieb stabil. Zweimal haben wir außerdem Rollenkonzepte nachschärfen müssen.

Aus den bisherigen Aufträgen ergibt sich ein wiederkehrendes Muster: In Prüfungen steht weniger die Modellbezeichnung im Mittelpunkt als die Frage nach Kontrolle und Nachweisbarkeit. Wer zeigen kann, wer auf welche Inhalte zugreifen durfte, wann Daten gelöscht wurden und worauf diese Nachweise beruhen, reduziert den Abstimmungsaufwand erheblich.

7 · Fazit

Die Datenschutzkonferenz zeigt, wo die Grenze verläuft. RAG-Systeme können die datenschutzrechtliche Bewertung generativer KI verbessern, wenn Architektur und Betrieb die erforderlichen Kontrollen tatsächlich abbilden. Innerhalb dieser Grenze entscheidet die konkrete Bauweise. Eine Referenzschicht, deren Inhalte adressierbar, zweckgebunden und nachweisbar löschbar sind, entspricht dem Gedanken von Datenschutz durch Technikgestaltung. Mit dem Wissensgraphen wählen wir eine Bauweise, die diesen Anspruch in den Betrieb überführt. Eine Anfrage der Aufsicht sollte sich dann nicht nur mit einer Beschreibung beantworten lassen, sondern mit belastbaren Protokollen aus dem System.

8 · Quellen / weiterführende Literatur

Balaguer, A. et al. (2024). RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. arXiv:2401.08406. Vergleichsstudie, die RAG und Fine-Tuning als komplementäre Verfahren einordnet und die Beisteuerung von Wissen über die Anreicherung als Stärke der RAG-Methode bestätigt.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A. & Raffel, C. (2021). Extracting Training Data from Large Language Models. 30th USENIX Security Symposium, 2633–2650. Grundlegende Sicherheitsarbeit zur Extraktion von Trainingsdaten aus großen Sprachmodellen und zu den damit verbundenen Datenschutzrisiken.

Datenschutzkonferenz (2025). Orientierungshilfe zu datenschutzrechtlichen Besonderheiten generativer KI-Systeme mit RAG-Methode. Version 1.0, Stand Oktober 2025. Zentrale aufsichtsbehördliche Einordnung der RAG-Methode entlang der DSGVO-Grundsätze, der Betroffenenrechte und der technischen Systemgrenzen.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O. & Larson, J. (2024). From Local to Global: A Graph RAG Approach to Query-Focused Summarization. Microsoft Research / arXiv:2404.16130. Forschungsarbeit zur Erweiterung klassischer RAG-Ansätze durch Wissensgraphen und Community-Zusammenfassungen für komplexe Abfragen über große Dokumentbestände.

Europäisches Parlament und Rat der Europäischen Union (2026). Verordnung (EU) 2026/1744 vom 8. Juli 2026 zur Änderung der Verordnungen (EU) 2024/1689, (EU) 2018/1139 und (EU) 2023/1230 im Hinblick auf die Vereinfachung der Umsetzung harmonisierter Vorschriften für künstliche Intelligenz (Digital Omnibus on AI). Amtsblatt der Europäischen Union, 24. Juli 2026. Änderungsverordnung zur KI-VO; verschiebt die Hochrisiko-Pflichten für Anhang-III-Systeme auf den 2. Dezember 2027 und für Anhang-I-Systeme auf den 2. August 2028, belässt die Transparenzpflichten des Art. 50 KI-VO beim 2. August 2026.

European Data Protection Board (2024). Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. 17. Dezember 2024. Einordnung dazu, wann KI-Modelle als anonym gelten können, welche Nachweise Verantwortliche führen müssen und welche Folgen rechtswidrige Trainingsdaten für den späteren Einsatz haben können.

Honroth, T., Siebert, J. & Kelbert, P. (2024). Retrieval Augmented Generation (RAG): Chatten mit den eigenen Daten. Fraunhofer IESE Blog, 13. Mai 2024. Praxisorientierter Überblick, der empfiehlt, Wissen über die Anreicherung statt über Fine-Tuning beizusteuern und das Sprachmodell auf die Sprachfähigkeit zu beschränken.

Kühling, J. & Buchner, B. (Hrsg.) (2024). Datenschutz-Grundverordnung, Bundesdatenschutzgesetz: DS-GVO/BDSG. 4. Auflage, C.H. Beck. Standardkommentar zum Datenschutzrecht; hier herangezogen zum Grundsatzkatalog des Art. 5 DSGVO, entlang dessen die DSK RAG-Systeme bewertet.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33. Grundlagenpaper zur Verbindung parametrischer Sprachmodelle mit externer, abrufbarer Wissensbasis und damit Ausgangspunkt vieler heutiger RAG-Architekturen.

Martini, M. (2019). Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz. Springer Vieweg. Monographie zu den Grundfragen der KI-Regulierung und zur Frage, wie rechtliche Kontrolle über intransparente, lernende Systeme zurückgewonnen werden kann.

Rigaki, M. & Garcia, S. (2020). A Survey of Privacy Attacks in Machine Learning. arXiv:2007.07646. Systematischer Überblick über Angriffe auf die Vertraulichkeit lernender Modelle, darunter Membership-Inference- und Poisoning-Angriffe, die auch die DSK-Orientierungshilfe als relevante Bedrohungen benennt.

Shafran, A., Schuster, R. & Shmatikov, V. (2025). Machine Against the RAG: Jamming Retrieval-Augmented Generation with Blocker Documents. 34th USENIX Security Symposium, 3787–3806. Sicherheitsarbeit zu Angriffen auf RAG-Systeme über manipulierte oder blockierende Dokumente in der Wissensbasis.

Vogel, P. (2022). Künstliche Intelligenz und Datenschutz. Nomos, 262 S. Untersuchung zur Vereinbarkeit intransparenter KI-Systeme mit dem geltenden Datenschutzrecht und zu Regulierungsansätzen für eine transparente Entwicklung und Nutzung.

Wischmeyer, T. (2018). Regulierung intelligenter Systeme. Archiv des öffentlichen Rechts 143(1), 1–66. Grundlegende rechtswissenschaftliche Analyse der Regulierungsoptionen für lernende Systeme, von Transparenzanforderungen bis zu institutionellen Kontrollmodellen.

Zou, W., Geng, R., Wang, B. & Jia, J. (2025): PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. 34th USENIX Security Symposium, 3827–3844. Sicherheitsarbeit zu Poisoning-Angriffen auf die Wissensbasis von RAG-Systemen; zeigt Erfolgsquoten um 90 Prozent bereits bei wenigen injizierten Schadtexten pro Zielfrage.

9 · Kontakt

Udo Jacobasch
Partner
Udo.Jacobasch@findic.de
Office Berlin

Dr. Josef Dirnberger
Manager
Josef.Dirnberger@findic.de
Office München

Hamburg

Kurze Mühren 20,

D-20095 Hamburg

Phone +49 40/303740-0

E-Mail info@findic.de

Zürich

Gutenbergstrasse 1,

CH-8002 Zürich

Phone +41 44/56098-00

E-Mail info@findic.ch