Konferenz der unabhängigen

Datenschutzaufsichtsbehörden des Bundes und der Länder

Orientierungshilfe zu datenschutzrechtlichen Besonderheiten generativer KI-Systeme mit RAG-Methode

Version 1.0

Stand:
Oktober 2025

Inhalt

Abstract[1]

In diesem Papier wird die Umsetzbarkeit der Grundsätze der DSGVO in KI-Systemen mit Retrieval Augmented Generation (RAG), sog. RAG-Systemen, untersucht. Aufbauend auf einem typischen Anwendungsszenario werden die technischen Komponenten eines RAG-Systems dargestellt. Dabei zeigt sich, dass die RAG-Methode beispielsweise hinsichtlich der Sicherstellung kontextbezogener und überprüfbarer Inhalte positive Effekte auf die Richtigkeit und Nachvollziehbarkeit der Ausgaben eines KI-Systems entfalten kann und in einem RAG-System die Vertraulichkeit und Integrität von zusätzlich eingebundenen personenbezogenen Daten verbessert werden können. Gleichzeitig entstehen auch neue datenschutzrechtliche Herausforderungen. Ferner ermöglicht die Methode in vielen Fällen, ein LLM ggf. on-premise einzusetzen, das weniger umfangreiche Trainingsdaten benötigt, wodurch ggf. weniger personenbezogene Daten aus dem KI-Modell extrahierbar sind. Zwar bleibt insbesondere die datenschutzrechtliche Beurteilung des Trainings des verwendeten LLM als solches unberührt, doch kann die Implementierung der RAG-Methode in einem KI-System im Ergebnis zu einer differenzierten datenschutzrechtlichen Bewertung führen.

1. Einleitung

In diesem Papier werden spezifische Fragestellungen zu RAG-Systemen behandelt, wobei der Fokus auf Systemen mit Embeddings und Vektordatenbanken gelegt wird. Solche Systeme haben sich als technische Grundlage für viele RAG-Systeme etabliert, da sie eine an der Semantik orientierte Suche in großen Wissensbeständen ermöglichen. Andere technische Umsetzungen von RAG-Systemen, etwa auf Basis von API-Abfragen oder klassischen relationalen Datenbankzugriffen, werden in diesem Papier nicht behandelt. Ein RAG-System ändert insbesondere nichts daran, wie ein generatives KI-Modell trainiert wurde. Ein rechtswidrig trainiertes KI-Modell bleibt auch in einem RAG-System ein rechtswidrig trainiertes KI-Modell, an dessen rechtskonformen Einsatz hohe Anforderungen gestellt werden (siehe EDSA-Stellungnahme 28/2024, insbesondere Szenario 2, Rn. 124 ff.).

Zugleich ist zu beachten, dass sowohl LLM als auch RAG zahlreiche Anwendungsfälle abdecken können und die Aussagen in diesem Papier daher allgemein gefasst werden müssen. Die datenschutzrechtliche Bewertung des konkreten Einsatzes muss einzelfallabhängig erfolgen.

Der Fokus liegt im Folgenden auf den speziellen Auswirkungen des Einsatzes des RAG-Systems auf die datenschutzrechtliche Bewertung eines generativen KI-Systems auf Basis eines LLMs. Die datenschutzrechtliche Bewertung des RAG-Subsystems ist hierfür die entscheidende Grundlage. Es wird – wenn nicht anders beschrieben – ein Einsatzszenario zugrundegelegt, in dem das gesamte RAG-System on-premise bei einem Verantwortlichen betrieben wird. Aus der Verteilung von Systemteilen auf verschiedene Betreiber in Form einer Auftragsverarbeitung oder gemeinsamer Verantwortlichkeit können sich weitere Aspekte ergeben.

[1] Ein Glossar mit den wichtigsten Begriffen befindet sich am Ende des Dokuments.

Bislang haben vor allem KI-Systeme besondere Aufmerksamkeit erregt, die als KI-Modell ein großes Sprachmodell (LLM – Large Language Model) wie z. B. generative vortrainierte Transformer (GPT) nutzen. Ihre Popularität verdankt diese Technologie u. a. der Tatsache, dass sie es ermöglicht, natürlichsprachliche Texte zu analysieren und syntaktisch korrekte Texte zu erzeugen. Die Nutzung von LLMs birgt jedoch zahlreiche Risiken für den Datenschutz. Diese ergeben sich unter anderem aus dem Training mit Daten, zu denen typischerweise auch personenbezogene Daten und solche besonderer Kategorien gehören, sowie aus dem Phänomen des Halluzinierens, was die Richtigkeit der (ggf. personenbezogenen) Ausgabe beeinträchtigt. Darüber hinaus werden LLMs in der Regel mit Daten bis zu einem bestimmten Datum trainiert. Eingabeprompts, die z. B. Fragen zu Ereignissen enthalten, die nach diesem Datum stattgefunden haben, führen daher regelmäßig zu falschen Antworten. Auch sind Angriffe auf ein Sprachmodel zu berücksichtigen,[2] z. B. Membership Inference Attacks oder die Extraktion von Trainingsdaten aus dem KI-Modell (Model Inversion Attacks). Zudem müssen Verantwortliche u. a. gewährleisten, dass die Rechte der betroffenen Personen nach Kapitel 3 der DSGVO gewahrt werden, z. B. das Recht auf Berichtigung nach Art. 16 DSGVO und das Recht auf Löschung nach Art. 17 DSGVO.[3] In diesem Papier wird beschrieben, wie sich ein RAG-System auf den datenschutzkonformen Einsatz eines generativen KI-Systems auswirken kann.

Um ein LLM für einen bestimmten Anwendungskontext zu optimieren, d. h. Wissen einer bisher unbekannten oder unvollständigen Wissensdomäne zu nutzen, sind mehrere Ansätze möglich.[4] Einmal kann Nachtraining oder Finetuning des Sprachmodells genutzt werden – ein anderer Weg ist die Anwendung der RAG-Methode. Bei der RAG-Methode wird – vereinfacht gesagt – ein LLM mit einer Datenbasis kombiniert, um Informationen, die (nur) dem Verantwortlichen zur Verfügung stehen und nicht Teil der Trainingsdaten des LLMs waren, über ein KI-System zugänglich und nutzbar zu machen.

Es ist zu überprüfen, in welchem Ausmaß ein RAG-System die datenschutzrechtliche Bewertung eines generativen KI-Systems auf Basis eines LLMs beeinflusst, gerade im Hinblick auf die in der EDSA-Stellungnahme 28/2024 angesprochenen Fragestellungen. Diese betreffen insbesondere die Rechtsgrundlagen für das Training mit personenbezogenen Daten aus dem Webscraping, die Nachnutzung bei unrechtmäßigem Training sowie die Wahrnehmung der zuvor genannten Betroffenenrechte.

Dieses Papier analysiert den Einfluss der Anwendung der RAG-Methode auf die datenschutzrechtliche Bewertung von generativen KI-Systemen und identifiziert zusätzliche Herausforderungen, aber auch Erleichterungen, die mit der Erweiterung des Sprachmodells um ein RAG-

[2] Vgl. Rigaki/Garcia, A Survey of Privacy Attacks in Machine Learning, 2020.

[3] Siehe DSK, OH Künstliche Intelligenz und Datenschutz, Version 1.0, Mai 2024, insbesondere Rn. 26 ff.

[4] Lewis et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, arXiv:2005.11401v4, 12. April 2021; Daniel Jurafsky and James H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition, 12. Januar 2025, insbesondere Kapitel 14, https://web.stanford.edu/~jurafsky/slp3; Yunfan Gao et al., Retrieval-Augmented Generation for Large Language Models: A Survey, insbesondere Kapitel II.D, https://arxiv.org/pdf/2312.10997.

Subsystem einhergehen. Die RAG-Methode kann in unterschiedlichen Kontexten wie z. B. für natürliche Sprache, multimodale Daten-Formate oder in Kombination mit nicht-generativer KI (z. B. in Klassifikations- oder Empfehlungs-Systemen) eingesetzt werden. Betrachtet werden hier exemplarisch nur RAG-Systeme mit generativen Sprachmodellen.

2. Definition RAG-System[5]

Im Folgenden soll anhand eines ausgewählten Szenarios die grundlegende Funktionsweise eines RAG-Systems vorgestellt werden. In diesem Szenario enthalten die Referenzdokumente auch personenbezogene Daten, welche über das RAG-System abrufbar sein sollen. Auf unterschiedliche, mögliche Varianten des Szenarios wird nur an den Stellen eingegangen, an denen sie für das grundlegende Verständnis hilfreich sind.

In RAG-Systemen wird die Eingabe der Nutzenden (Eingabeprompt) durch das RAG-Subsystem um Texte aus Referenzdokumenten ergänzt, bevor sie an ein LLM gesandt wird. Damit soll erreicht werden, dass das generative KI-System wie z. B. ein LLM diese relevanten Informationen einbezieht und in den Fokus der zu generierenden Antwort stellt. Im gewählten Szenario sollte das faktische Wissen einer Antwort vollständig aus den Referenzdokumenten stammen und das LLM lediglich der Spracherzeugung dienen.[6]

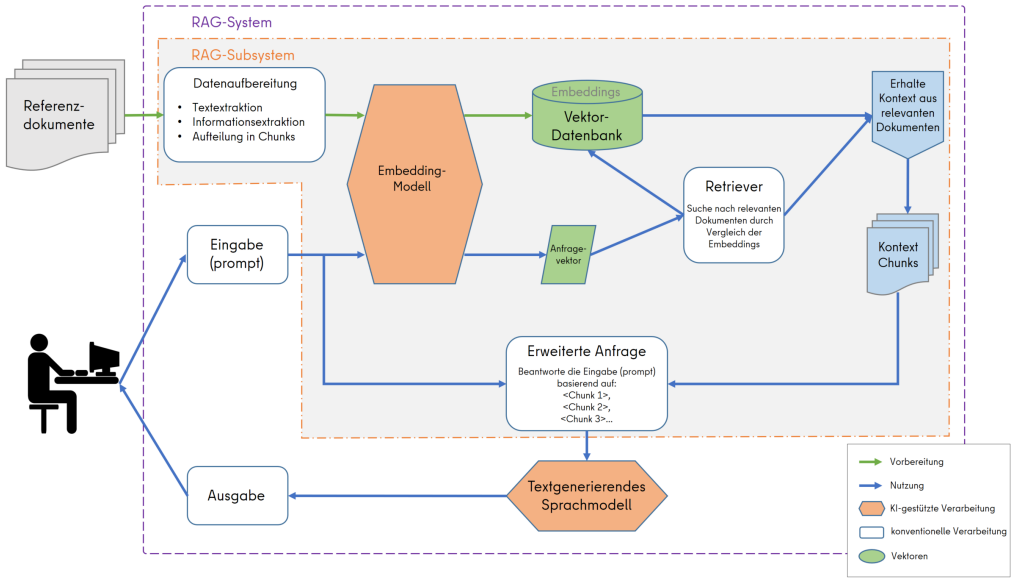

Abb. 1: Beantwortung von Fragen durch ein RAG-System

Wie in **Abb. 1** dargestellt, ergänzt die RAG-Methode das LLM um eine Suchfunktion (Retriever) und eine Vektordatenbank. Im hier zugrunde gelegten Szenario wurden vor der Nutzung des

[5] Siehe dazu auch: Lewis et al. 2020, https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf; Gao et al. 2023, https://simg.baai.ac.cn/paperfile/25a43194-c74c-4cd3-b60f-0a1f27f8b8af.pdf; Chen et al. 2024, https://ojs.aaai.org/index.php/AAAI/article/view/29728.

[6] Derzeit wird dieser Idealfall technisch nicht immer erreicht. Auch faktisches Wissen des LLMs kann ungewollt Teil des Outputs werden, was ggf. zusätzliche Maßnahmen erforderlich macht.

RAG-Systems die Referenzdokumente aufbereitet. Dabei umfasst das Verfahren neben weiteren Schritten auch die Aufteilung der Referenzdokumente in kürzere Absätze (Chunks). Diese Chunks wurden durch das Embedding-Modell in Vektoren überführt, welche den jeweiligen Chunk repräsentieren. Diese Vektoren wurden zusammen mit den Chunks in der Vektordatenbank gespeichert.

Bei der Nutzung des RAG-Systems wird auch die Eingabe mit Hilfe desselben Embedding-Models in einen Anfragevektor überführt. Danach wird der Anfragevektor an den Retriever weitergegeben. Dieser durchsucht die Vektordatenbank nach Vektoren der Chunks, welche eine möglichst geringe Distanz zum Anfragevektor haben, um so für die Anfrage relevante Chunks zu finden. Eine geringe Distanz wird hier mit einer großen semantischen Nähe assoziiert.[7]

Im nächsten Schritt wird die Eingabe um das Suchergebnis des Retrievers ergänzt. Die generative KI erzeugt dann aus der ergänzten Eingabe die Ausgabe. Je nach Ausgestaltung des KI-Systems werden neben der Ausgabe ergänzende Informationen, z. B. zu den Referenzdokumenten, angezeigt. Ein RAG-System im Sinne dieser Ausarbeitung ist damit ein System, welches im Kern aus Retriever, Embedding-Modell, Vektordatenbank und generativem Sprachmodell besteht.

Der Retriever (Suchfunktion) und das oben genannte Embedding-Modell, sind entscheidend für das typische Verhalten eines RAG-Systems. Der Retriever und das Embedding-Modell stellen eine Suchfunktion zur Verfügung, die insbesondere KI-Algorithmen wie Transformer und vektorbasierte Suchmechanismen enthalten kann. Die Suche nach relevanten Chunks ist möglich, da das Embedding-Modell ähnliche Texte in Vektoren überführt, die eine geringere Distanz zueinander aufweisen als die Vektoren von Texten, die weniger ähnlich sind. Dies wird durch konservative Verfahren oder auch durch KI-Modelle ermöglicht, welche wie das LLM auch, auf sogenannten Transformer-Modellen beruhen und deshalb genau wie LLM, dieselben datenschutzrechtlichen Probleme beim Training und in der Anwendung aufweisen können.

Wichtig ist, dass in vielen RAG-Systemen keine rein syntaktische Suche zur Anwendung kommt, sondern in der Regel die Suche nach der semantischen Nähe gemeint ist. D. h. es werden inhaltlich relevante Textabschnitte in der Suche berücksichtigt und nicht wortähnliche Treffer. Das RAG-System erzeugt so eine Ausgabe auf eine Eingabe eines Benutzers, basierend auf zusätzlichen relevanten Informationen, die aus einer dedizierten Datenquelle stammen. Der Retriever dient im Wesentlichen der Erkennung der relevanten Daten in den aufbereiteten Daten durch einen Vektorvergleich.

2.1. Beschreibung eines RAG-Systems im Einzelnen

Im Folgenden werden ausgewählte Komponenten und Verfahren, welche im vorherigen Abschnitt kurz angesprochen wurden, eingehender beschrieben.

[7] KI-Modelle bilden die Semantik natürlicher Sprache näherungsweise nach, indem sehr große Datenmengen analysiert und daraus indirekt Text-Bedeutungen abgeleitet werden. Ein echtes Textverständnis der Textinhalte liegt nicht vor.

Datenaufbereitung

Zuerst erfolgt die Datenaufbereitung der Datenquelle, v. a. von Referenzdokumenten. Dies geschieht in unserem Szenario durch die Aufteilung von Referenzdokumenten in einzelne Textabschnitte (sog. Chunks). Dabei handelt es sich um einen wichtigen Schritt, der die Qualität der späteren Antworten beeinflusst. Es können unterschiedliche Vorgehensweisen gewählt werden, von denen zwei im Folgenden beispielhaft skizziert werden.

Im einfachsten Fall werden gleichgroße Abschnitte z. B. mit einer festen Zeichenanzahl erzeugt. Dieses Vorgehen kann semantische Zusammenhänge in den Referenztexten zerstören und damit zu qualitativen Nachteilen beim darauffolgenden Embedding führen. Ein anderes Beispiel ist das Chunking mit LLM. Durch den Einsatz von zum Teil auch kleineren, generativen KI-Modellen, können die Referenzdokumente anhand der inhaltlichen, semantischen Bedeutung aufgeteilt werden. Dieses Vorgehen erhält den Kontext der Inhalte innerhalb eines Chunks besser und sorgt damit auch für eine bessere Qualität bei der Erstellung der Vektoren. Da sich diese Art der Aufteilung positiv auf die spätere Suche (Retrieval) auswirkt, wird auch die Qualität der Ausgabe des RAG-Systems verbessert. Wie bereits in der Einleitung erwähnt, muss letzteres Beispiel bezüglich des Trainings und der Anwendung aus Sicht des Datenschutzes wie andere LLM bewertet werden.

Embeddings

Das Embedding-Modell wandelt die jeweiligen Textabschnitte (und ebenso die spätere Eingabe) in Tokens bzw. Tokensequenzen (z. B. Worte, Zeichenfolgen, Zeichen) und anschließend in eine Vektordarstellung um. Entscheidend ist, dass die vom Embedding-Modell erzeugten Vektoren die Bedeutung (basierend auf den häufig in gemeinsamen Kontexten vorkommenden Wörtern aus den Trainingsdaten) der jeweiligen Informationen repräsentieren. Diese Vektoren können zusammen mit einer Verknüpfung zu den Textabschnitten oder den Textabschnitten selbst in einer Datenbank gespeichert werden. Auch hier gibt es eine Reihe unterschiedlicher Vorgehensweisen, deren Einsatz vom Konzept und der Zielsetzung des RAG-Systems abhängt.

Erkennen relevanter Daten (Retriever, RAG-Subsystem)

Die Aufgabe des Retrievers besteht darin, die Eingabe mit inhaltlich möglichst nahen Textabschnitten in Verbindung zu bringen. Dazu erzeugt das Embedding-Modell aus der Eingabe ebenfalls einen Vektor. Diesen vergleicht der Retriever mit den Vektoren in der Datenbank. Die Textabschnitte mit den ähnlichsten Vektoren (größte Nähe) werden zur Anreicherung der Eingabe verwendet. Bei diesem Verfahren wird die Nähe der Vektoren zueinander als Maß für die Relevanz des Chunks verstanden.

Dabei ist die erweiterte Anfrage des Retrievers nicht mit dem Training oder Nachtraining[8] eines LLM zu vergleichen. Die erweiterte Anfrage wird jedes Mal neu aus der Datenquelle erstellt. Im Gegensatz zu den Trainingsdaten verändert die erweiterte Anfrage das LLM nicht. Die Informationen in der Vektordatenbank des vorliegenden Szenarios können die Ausgabe des LLM wesentlich beeinflussen. Im Gegensatz zu den im LLM erlernten Informationen lassen sie sich jedoch einfach aktualisieren, löschen und beauskunften, was der Umsetzung der datenschutzrechtlichen Anforderungen, bezüglich der personenbezogenen Daten aus den Referenzdokumenten, entgegenkommt.

2.2. Grenzen der RAG-Methode

Die Einbettung von Chunks und die Anreicherung der LLM-Anfrage um semantisch benachbarte Textabschnitte kann nur den Eingabeprompt für ein LLM direkt beeinflussen. Antrainiertes Wissen des LLM kann dadurch nicht verändert werden (anders als bei einem Fine-Tuning oder Nachtraining). Allerdings wird durch die Gestaltung und den Inhalt der erweiterten Anfrage Einfluss auf die Ausgabe des LLM genommen. Gerade dies ist der Zweck von RAG-Systemen: Das LLM soll im gewählten Szenario nur die Sprachfähigkeit beisteuern.

Die Idee der automatisierten Anreicherung des Eingabeprompts unterliegt Beschränkungen. Die Eingabe in ein LLM darf nicht beliebig lang sein. Somit sind die Textpassagen, um die der Eingabeprompt ergänzt wird, in ihrer Länge begrenzt. Die Grenze hängt vom verwendeten Sprachmodell ab. Allerdings spielt die tatsächliche Länge der Dokumente oft keine Rolle mehr, wenn diese sinnvoll in Chunks aufgeteilt werden können, wodurch selbst sehr umfangreiche Dokumente handhabbar werden. Weiterhin findet die Anreicherung nur im semantisch benachbarten Einbettungsraum statt. Komplexe Wenn-Dann-Gedankenketten über lange Textpassagen hinweg liegen evtl. nicht mehr semantisch benachbart, sind unter Umständen nicht in demselben Chunk enthalten und werden damit unvollständig der erweiterten Anfrage hinzugefügt. Daraus ergeben sich systembedingte Schwächen in der Erkennung von relevanten Informationen, die zwar logisch zusammengehören, aber keine ausreichende semantische Nähe aufweisen.[9]

[8] Inwieweit die Qualität der Ergebnisse durch ein Nachtraining (Fine-Tuning) des KI-Modells weiter verbessert werden kann, lässt sich nicht pauschal sagen. Nach Ansicht des Fraunhofer-Instituts IESE „sollte Fine-Tuning aber nicht dafür eingesetzt werden, einem LLM Wissen anzutrainieren“, es kann aber helfen einen bestimmten Antwortstil zu berücksichtigen, auf bestimmte Details zu achten oder mit speziellen Suchergebnisformaten besser umgehen zu können (Fraunhofer IESE, 13. Mai 2024, https://www.iese.fraunhofer.de/blog/retrieval-augmented-generation-rag/). In einer anderen Studie wird beobachtet, dass RAG und Fine-Tuning unabhängig voneinander zu einer Verbesserung der Ergebnisse führen können (Balaguer et al., RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture, https://doi.org/10.48550/ arXiv.2401.08406).

[9] Es wird versucht, dieses Problem durch KI-Agenten-Systeme zu lösen, welche ggf. iterativ mehrere Anfragen an den Retriever schicken.

3. Generative Sprachmodelle mit und ohne RAG und Auswirkungen auf die datenschutzrechtliche Bewertung

Entlang der Grundsätze für die Verarbeitung personenbezogener Daten nach Art. 5 DSGVO und der Rechte der betroffenen Personen nach Kapitel 3 der DSGVO werden hier generative KI-Systeme mit einem Sprachmodell betrachtet und wie sich der Einsatz der RAG-Methode auf die datenschutzrechtliche Bewertung des KI-Systems auswirkt. Es sei darauf hingewiesen, dass ein RAG-System an den Eigenschaften seiner LLM-Komponente nichts ändert, sondern diese lediglich durch zusätzlichen Kontext beeinflusst und ergänzt.

3.1. Richtigkeit

KI-Systeme mit einem LLM müssen den datenschutzrechtlichen Grundsatz der Richtigkeit personenbezogener Daten erfüllen. Generative KI-Modelle stoßen insbesondere an Grenzen, wenn es darum geht, selten vorhandenes Wissen in den Trainingsdaten (spezifisches Wissen eines Verantwortlichen oder Long-tail Knowledge[10]) zu nutzen oder aktuelle Daten zu verwenden, die in den Trainingsdaten nicht vorhanden waren. Dies kann zur Ausgabe unrichtiger personenbezogener Daten führen. In einem RAG-System wird der Kontext der LLM-Eingabe spezifisch angepasst. Werden bestimmte Voraussetzungen eingehalten, kann sich dies positiv auf die Richtigkeit der generierten Ausgabe auswirken. Es kann aber nicht vollständig ausgeschlossen werden, dass durch das LLM trotzdem noch unrichtige personenbezogene Daten ausgegeben werden können. Im Unterschied zu den in einem LLM enthaltenen Informationen besteht ein großer Vorteil darin, dass die Daten in den Referenzdokumenten, wenn sie fehlerhaft oder veraltet sind, gelöscht oder aktualisiert werden können. Der Retriever, als Komponente eines RAG-Systems, kann im Vergleich zu einem allein betriebenen LLM, die Erzeugung unrichtiger Daten bereits durch die Anreicherung der Informationen im Eingabeprompt reduzieren. RAG-Systeme bieten daher eine Maßnahme zur Vermeidung unrichtiger personenbezogener Daten in der generierten Ausgabe, wobei dafür entsprechende Informationen in den Referenzdokumenten vorhanden sein müssen.

Man spricht von KI-Halluzinationen, wenn KI-Modelle Informationen erzeugen, die zwar plausibel klingen, aber nicht durch ihre Trainingsdaten gestützt sind.

In unserem Szenario basieren die vom RAG-System generierten Antworten idealerweise auf den Informationen aus den Referenzdokumenten. Daher können mit der RAG-Methode Halluzinationen reduziert und die Ausgabe unrichtiger personenbezogener Daten im Sinne von Art. 5 Abs. 1 Buchst. d DSGVO verringert werden. Dabei bleiben Herausforderungen bestehen, insbesondere:

- Qualität der Referenzdokumente: Die Zuverlässigkeit von RAG-Systemen hängt stark von der Qualität, Aktualität und Vollständigkeit der verwendeten Referenzdokumente

[10] Kandpal et al., Large Language Models Struggle to Learn Long-Tail Knowledge, https://doi.org/10.48550 /arXiv.2211.08411.

ab. Unvollständige oder veraltete Daten führen zu unrichtigen Ausgaben. Die verantwortliche Stelle muss daher regelmäßig überprüfen, ob die verwendeten Referenzdokumente diesen Anforderungen entsprechen.

- Qualität der Datenaufbereitung: Bei der Umsetzung sollte hinsichtlich der Parametrisierung und Datenaufbereitung berücksichtigt werden, dass vor der Wandlung eines Dokumentes in Textabschnitte oder Chunks störende Daten wie Kopf- und Fußzeilen oder Seitennummern bereinigt werden sollten und das Dokument in einen Fließtext überführt wird. Die Überlappung und die Größe der Textabschnitte beeinflussen ebenfalls die Qualität der Daten, welche zur Anreicherung genutzt werden können. Es sollte darauf geachtet werden, dass einzelne Chunks möglichst abschließende Sinnabschnitte enthalten. Soweit personenbezogene Daten nicht erforderlich sind, sind personenbezogene Daten zu entfernen oder zu anonymisieren.
- Qualität des Embeddings: Bei der Auswahl eines Embedding-Modells sollten die spezifischen Anforderungen berücksichtigt werden, z. B. dass dieses mit deutschsprachigen Texten trainiert wurde, wenn deutschsprachige Referenzdokumente verwendet werden. Anderenfalls können beispielsweise dem Eingabeprompt weniger relevante Chunks zugeordnet werden.
- Kontexttreue des verwendeten LLMs: Abhängig vom Design des RAG-Systems und Training des LLM können die Berechnungen im LLM dazu tendieren, insbesondere bei widersprüchlichen Inhalten das Wissen aus den Trainingsdaten wiederzugeben und die Inhalte aus den Referenzdokumenten zu verwerfen.

  Um dem zu begegnen, ist es abhängig von der gewählten Konzeption denkbar, dass weitere technische Maßnahmen ergriffen werden müssen. In Betracht kommt vor allem ein Systemprompt, der das RAG-System anweist, ausschließlich mithilfe der referenzierten Quellen zu antworten.

Ein RAG-System kann grundsätzlich auch mit externen Datenquellen betrieben werden oder diese zusätzlich nutzen. In diesem Fall wird z. B. eine temporäre Vektordatenbank erstellt und mit Chunks und Embeddings befüllt. Bei der Einbindung externer Datenquellen, z. B. Dokumentendatenbanken Dritter, muss die Rechtmäßigkeit der Verwendung, die Eignung und Richtigkeit der Daten und der damit erzielten Ergebnisse ausreichend geprüft und sichergestellt werden. Diese Anforderungen werden regelmäßig bei Einbindung einer Websuche nicht erfüllt. Die Einbindung von externen Daten kann die generierten Texte zwar aktuell oder spezifisch erscheinen lassen, kann sich aber auf die Richtigkeit auswirken und ggf. eine Priorisierung der internen und externen Datenquellen im RAG-Subsystem erforderlich machen.

3.2. Transparenz

Soweit eine Dokumentation der für die Antwort genutzten Quellen (Referenzen auf Chunks oder Dokumente) stattfindet, ermöglicht diese die Nachvollziehbarkeit und Transparenz des Inputs des LLMs, welcher aufgrund der Inhalte der Referenzdokumente erzeugt wurde. Eine

Erhöhung der datenschutzrechtlichen Transparenz in Bezug auf das eingesetzte LLM kann hingegen nicht erreicht werden.

Gleiches gilt für die in der Vektordatenbank gespeicherten Embeddings. Es kann weder nachvollzogen werden, warum den jeweiligen Chunks die entsprechenden Embeddings zugeordnet werden, noch welche genaue Bedeutung sie haben.

Folglich ist die Transparenz in einem RAG-System darauf beschränkt, Aussagen über die erweiterte Anfrage an die verwendete LLM-Komponente zu treffen. Aussagen darüber, wie die Ausgaben der KI-Modelle und somit auch des RAG-Systems entstehen, sind sehr schwer zu treffen und somit intransparent.

Bei der Auswahl der Referenzdokumente ist zu beachten, dass Informationspflichten zur Gewährleistung des Transparenzgrundsatzes umgesetzt werden. Dokumente aus unbekannten oder nicht vertrauenswürdigen Quellen sollten daher nicht verwendet werden.

3.3. Integrität und Vertraulichkeit

Mit Blick auf die Vertraulichkeit kann im Rahmen des RAG-Subsystems datenschutzrechtlichen Anforderungen an die Datenbank mit etablierten Maßnahmen begegnet werden. So können in einem RAG-Subsystem bewährte technische und organisatorische Maßnahmen, wie z. B. die Mandantentrennung/funktionale Trennung[11] und das Rechte- und Rollenkonzept[12], angewendet werden. Das Rechte- und Rollenkonzept bezieht sich hier auf Zugangsbeschränkungen zu Bereichen der Vektordatenbank und den Referenzdokumenten. Im Vergleich dazu kann im Sprachmodell (LLM) selber nicht gesteuert werden, auf welche Informationen bestimmte Nutzer Zugriff haben dürfen und auf welche Informationen kein Zugriff bestehen darf. Bei Erfüllung der datenschutzrechtlichen Anforderungen können daher auch personenbezogene Daten mit höherem Schutzbedarf, z. B. Daten nach Art. 9 und 10 DSGVO, verarbeitet werden, da diese nicht dauerhaft im LLM verbleiben, sofern kein gezieltes Training oder Nachtraining des LLM mit diesen Daten stattfindet, sondern in den Referenzdokumenten und der Vektordatenbank separat gespeichert sind.[13]

Da beim Einsatz eines RAG-Systems typischerweise die Informationen aus dem RAG-Subsystem für die Generierung einer Antwort genutzt werden sollen, kann ein LLM verwendet werden, das weniger faktische Informationen enthält. Damit kommen für den Einsatz eine größere Auswahl an LLMs in Betracht und unter Umständen kann das RAG-System on-premise betrieben werden. Dadurch kann vermieden werden, dass personenbezogene Daten an Online-Betreiber großer Sprachmodelle übertragen werden.

---

[11] WP 203, mehrere Passagen. Beispielsweise auf Abteilungsebenen einer Organisation, da auf unterschiedlichen Akten(-gruppen) gearbeitet wird.

[12] Beispielsweise auf Basis von Individuen, die verschiedenen Organisationseinheiten zugeordnet werden.

[13] Die Daten werden immer im KI-Modell verarbeitet. Siehe dazu DSK, OH Künstliche Intelligenz und Datenschutz, Version 1.0, Mai 2024, insbesondere Rn. 15 ff. und Rn. 32 ff.

Aus denselben Gründen kann auch die Integrität der Daten in einem RAG-Subsystem in der Regel besser gewahrt werden als in einem KI-System ohne RAG-Methode. Allerdings müssen auch hier Angriffsvektoren wie z. B. Membership Inference Attacks oder Data Poisoning Attacks auf die Daten im RAG-Subsystem berücksichtigt werden.[14] Dabei muss erwähnt werden, dass die RAG-Methode auch als Maßnahme gegen Data Poisoning (bezogen auf die Trainingsdaten der LLM-Komponente) eingesetzt werden kann.[15]

3.4. Zweckbindung

Eine Bereitstellung bestimmter Dokumente für das LLM kann zielgerichtete Abfragen von personenbezogenen Daten ermöglichen, die strikt auf den definierten Verarbeitungszweck beschränkt sind. Hierfür müssen den Mitarbeitenden, die das RAG-System für verschiedene Zwecke benutzen, verschiedene Rollen zugewiesen werden. Es muss sichergestellt werden, dass die Mitarbeitenden vor der Abfrage die für den Verarbeitungszweck korrekte Rolle erhalten. Die Zweckbindung der personenbezogenen Daten in der Vektordatenbank lässt sich durch eine Mandantentrennung oder funktionale Trennung der Daten technisch umsetzen. Bereits bei der Aufbereitung der Referenzdokumente sollte überprüft werden, ob eine erforderliche Mandantentrennung umgesetzt werden kann, beispielsweise durch die Aufteilung der Dokumente.

Dabei ist u. a. zu prüfen, ob die Einbettung von Tokensequenzen noch dem ursprünglichen Zweck dient oder schon einem anderen Zweck zugeordnet ist. Dies ist relevant, da die Zielrichtung hier deutlich auf die Verarbeitung durch ein Sprachmodell abzielt und auch eine semantische Suche ermöglicht wird.

Gleichzeitig stellt die Verwendung eines RAG-Systems eine Bedrohung für die Zweckbindung dar. Es ist davon auszugehen, dass Verantwortliche regelmäßig nicht wissen, welche personenbezogenen Daten in dem verwendeten KI-Modell des LLM enthalten sind. Mit der Übergabe von personenbezogenen Daten aus der Vektordatenbank an das LLM kann es zu einer Verkettung von personenbezogenen Daten aus der Vektordatenbank mit den personenbezogenen Daten im KI-Modell des LLM und somit zu einer Verletzung der Zweckbindung kommen. Problematisch ist dies insbesondere, da eine auf diese Weise durchgeführte Verkettung ggf. nicht in der Ausgabe des RAG-Systems erkennbar ist. Es muss bereits im Rahmen der Konzeption geprüft werden, welche wirksamen Maßnahmen ergriffen werden können.

3.5. Datenminimierung und Speicherbegrenzung

Die Datenminimierung kann im RAG-Subsystem dadurch unterstützt werden, dass bestimmt wird, welche Dokumente in der Vektordatenbank gespeichert werden. Es müssen nicht unnötig viele personenbezogene Daten hinzugefügt werden. Der Umfang von personenbezogenen

[14] Data Poisoning Attacks, z. B. Zou et al, PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models, https://doi.org/10.48550/arXiv.2402.07867.
[15] OWASP, https://genai.owasp.org/llmrisk/llm042025-data-and-model-poisoning/.

Daten im LLM eines KI-Systems mit oder ohne Einsatz der RAG-Methode verringert sich durch den Einsatz der RAG-Methode nicht.

Da an die LLM-Komponente weniger hohe Anforderungen in Bezug auf das Memorieren von Fakten gestellt werden müssen (siehe 3.3. Integrität und Vertraulichkeit), kann der Einsatz eines kleinen Sprachmodells (SLM) einen Vorteil bieten. Für ein kleines Sprachmodell bestehen die gleichen datenschutzrechtlichen Herausforderungen wie für ein großes Sprachmodell, insbesondere hinsichtlich der Umsetzung der Rechte der betroffenen Personen. Sie memorieren jedoch u. U. weniger personenbezogene Daten. Die Eignung des Sprachmodells sollte immer überprüft und dokumentiert werden.

Auch können Referenzdokumente und Einträge in der Vektordatenbank gelöscht werden. Die Löschung erfolgt technisch durch Löschung von Einträgen in einer Datenbank und das Entfernen aus den Referenzdokumenten. Diese Einträge sind direkt adressierbar und damit auch personenbezogene Daten gezielt löschbar. Klassische Mechanismen zur Löschung nach Ablauf von Aufbewahrungsfristen können damit sowohl für die Vektordatenbank als auch die Referenzdokumente genutzt werden.

Ungeachtet der Vorteile eines RAG-Systems bleiben die Probleme bei der Datenlöschung im Sprachmodell bestehen.

3.6. Rechtmäßigkeit

Bei der Bewertung des rechtskonformen Einsatzes eines KI-Systems müssen die Komponenten und ihr Zusammenwirken betrachtet werden. Bei RAG-Systemen ist zu beachten, dass bei einem ggf. rechtswidrig trainierten LLM unabhängig von der Integration im RAG-System das Training weiterhin rechtswidrig bleiben würde. Allerdings können einige der damit einhergehenden Risiken für die Rechte und Freiheiten von betroffenen Personen durch die Verwendung der RAG-Methode gemindert werden. Die konkrete Auswirkung eines RAG-Systems lässt sich jedoch nur im Einzelfall überprüfen.

Für den Einsatz eines RAG-Systems bedarf es einer Rechtsgrundlage, so werden unter anderem personenbezogene Daten aus den Referenzdokumenten durch das Embedding-Modell verarbeitet und in einer Vektordatenbank gespeichert. Kommt der Einsatz eines RAG-Systems ohnehin in Betracht, kann dies auch Vorteile aus datenschutzrechtlicher Sicht mit sich bringen. Sollten für den Einsatz eines KI-Systems ohne RAG-Methode die Voraussetzungen für eine datenschutzrechtliche Rechtsgrundlage nicht vorliegen, weil z. B. die Interessen der betroffenen Personen die Interessen des Verantwortlichen überwiegen, kann geprüft werden, ob durch den Einsatz der RAG-Methode risikomindernde Maßnahmen ergriffen werden können, die über die Maßnahmen hinausgehen, zu denen der Verantwortliche ohnehin gesetzlich ver-

pflichtet ist.[16] Dies ist im Einzelfall zu prüfen. Es ist immer zu beachten, dass bei einer Verarbeitung personenbezogener Daten, insbesondere bei der Verarbeitung der Referenzdokumente und in der Vektordatenbank, eine datenschutzrechtliche Rechtsgrundlage vorhanden sein muss.

3.7. Umgang mit Betroffenenrechten

Da RAG-Systeme Sprachmodelle nutzen, sind die Fragen von Betroffenenrechten im Zusammenhang mit Sprachmodellen hier ebenfalls relevant. Insgesamt ist festzustellen, dass die Umsetzung von Betroffenenrechten in KI-Modellen, insbesondere LLM, weitgehend ungelöst ist. Dies wirkt sich auch auf RAG-Systeme aus.

Unabhängig von der Frage, ob das verwendete LLM personenbezogen oder anonym ist, stehen Betroffenen insbesondere in Bezug auf den Eingabeprompt und die Ausgabe, die Referenzdokumente und die Vektordatenbank die Rechte aus Art. 15 ff. DSGVO zu, soweit hierbei personenbezogene Daten verarbeitet werden. Wie in den vorherigen Abschnitten dargelegt, können Transparenz, Auskunft, Berichtigung und Löschung für diese in der Regel umgesetzt werden.[17]

4. Fazit

Mit den dargelegten Auswirkungen auf die datenschutzrechtlichen Grundsätze wird deutlich, dass ein KI-System mit Verwendung der RAG-Methode einige Schwächen etwa bezüglich Halluzination und unrichtigen Ausgaben von KI-Systemen ohne RAG-Methode in gewissem Umfang reduzieren kann. Herausforderungen, wie fehlende Transparenz, Zweckbindung und Umsetzung der Betroffenenrechte im gesamten RAG-System, und Erleichterungen bei der Verwendung der RAG-Methode sind von der verantwortlichen Stelle im Einzelfall zu prüfen. Dies gilt zusätzlich zu den allgemeinen datenschutzrechtlichen Herausforderungen. Je nach Gestaltung, kann die Nutzung der RAG-Methode somit gegebenenfalls als eine von verschiedenen mitigierenden Maßnahmen im Sinne der EDSA-Stellungnahme 28/2024 erachtet werden.

[16] Siehe EDSA-Stellungnahme 28/2024, Rn. 96 ff.; EDSA-Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, Version 1.0, 8 October 2024, Rn. 57.

[17] Siehe EDSA-Guidelines 01/2022, https://www.edpb.europa.eu/system/files/2023-04/edpb_guidelines_202201_data_subject_rights_access_v2_en.pdf.

5. Glossar der wichtigsten Begrifflichkeiten

An dieser Stelle werden Begriffsdefinitionen und Fachbegriffe näher erläutert, welche im Text zur Beschreibung der speziellen IT-Komponenten genutzt werden. Wo möglich, wird auf die Begriffsdefinitionen zurückgegriffen, welche in der KI-Verordnung[18] definiert sind, oder in der „Orientierungshilfe zu empfohlenen technischen und organisatorischen Maßnahmen bei der Entwicklung und beim Betrieb von KI-Systemen“[19] genutzt worden sind.

Anfragevektor	Ergebnis der Anwendung des Embedding-Modells auf die Eingabe
Ausgabe	Die Ausgabe ist als Ergebnis des textgenerierenden Sprachmodells und damit des RAG-Systems zu verstehen.
Chunk (Textabschnitt)	Dies ist ein Textabschnitt mit vordefinierter Maximallänge. Zusatzdokumente, welche einem RAG-System als Zusatzinformationen zur Verfügung gestellt werden, werden in diese Textabschnitte unterteilt.
Data-Poisoning	Eine Angriffsmethode für datengetriebene Lernalgorithmen (also auch KI-Algorithmen), bei welcher versucht wird, die Trainingsdaten mit fehlerhaften Datensätzen anzureichern, um die Qualität der Lernergebnisse zu verschlechtern und im besten Fall ein unbrauchbares Modell zu verursachen.
Datenaufbereitung	Die Datenaufbereitung dient z. B. dem entfernen irrelevanter Inhalte und der Aufteilung der Referenzdokumente in Textabschnitte (Chunks), die im Rahmen des Embeddings weiterverarbeitet werden.
Einbettungsraum	Ist ein hochdimensionaler Vektorraum, welcher die inhaltliche Bedeutung von Textsequenzen kodieren kann. So kodierte Texte dienen dem Sprachmodell als Eingabe sowie dem Retriever zur Suche relevanter Zusatzinformationen.
Eingabe(prompt)	Es handelt sich um die Eingabe durch einen Benutzer oder ein vorgeschaltetes IT-System in das RAG-System.
Embedding	Beschreibt die mathematische Abbildungsvorschrift, mit der eine Textsequenz in einen Vektorraum abgebildet werden kann.
Embedding-Modell	Ist in der Regel ein KI-Modell (meist ein Sprachmodell), welches aus Textabschnitten (Chunks) die Abbildung in einen Vektorraum

[18] https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32024R1689.

[19] https://www.datenschutzkonferenz-online.de/media/oh/DSK-OH_KI-Systeme.pd.

	durchführen kann. Auch dieses Modell ist ein Bestandteil bzw. eine Komponente des RAG-Systems. Es kann sich dabei um ein Sprachmodell oder ein anderes, konservatives Verfahren zur inhaltlichen Kodierung von Text (ggf. ohne ausgeprägte generative Fähigkeiten) handeln.
Erweiterte Anfrage / Erweiterter Prompt	Der Erweiterte Prompt besteht aus dem Eingabeprompt, den relevanten Chunks und ggf. weiteren Textabschnitten, die als Kontext (z. B. „Ausgabe in deutscher Sprache“) hinzugefügt werden.
KI-Modell	Siehe OH-KI[20] sowie Art. 3 Nr. 63 KI-VO für KI-Modelle mit allgemeinem Verwendungszweck. Ein LLM ist eine Art von KI-Modell.
KI-System	Ist ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können, vgl. Art. 3 Nr. 1 KI-VO.
LLM / Textgenerierendes Sprachmodell	Die Abkürzung für Large Language Model stellt in aller Regel ein KI-Modell mit allgemeinem Verwendungszweck dar, da das Modell in der Lage ist, Textaufgaben zu lösen. Es nimmt dazu die Anfrage sowie den im Dialogverlauf erzeugten Text entgegen und generiert dabei das nächst wahrscheinlichste Wort. Die Generation der Worte erfolgt, bis ein Abbruchkriterium erfüllt wird. Es gibt LLMs, welche über weniger Parameter verfügen, aber dennoch generative Sprachfähigkeiten aufweisen. Sie gehören ebenfalls zur Gruppe der Sprachmodelle und werden auch als SLM (Small Language Model) bezeichnet.
LLM-Komponente	Ist der Teil des RAG-Systems, welcher das LLM implementiert.
Nachtraining / Finetuning	Im Kontext von Sprachmodellen wird der Begriff genutzt, um mitzuteilen, dass das ursprüngliche Sprachmodell verändert wird. Dabei werden für einen Anwendungsbereich spezifische Dokumente (z. B. mit einem bestimmten Sprachstil) genutzt, um die Modellparameter anzupassen. Damit ändert sich die Vorhersage

[20] DSK, Orientierungshilfe zu empfohlenen technischen und organisatorischen Maßnahmen bei der Entwicklung und beim Betrieb von KI-Systemen, Juni 2025, Version 1.0, https://www.datenschutzkonferenz-online.de/media/oh/DSK-OH_KI-Systeme.pdf.

	des nächst wahrscheinlichsten Wortes (siehe LLM) und gewissermaßen der Sprachstil. Aber auch die Unterdrückung bestimmter Ausgaben kann damit nachträglich verändert oder herbeigeführt werden.
Prompt	Die (Text-)Eingabe an das RAG-System, welche vom Nutzer gestellt wird.
RAG-System	Stellt die Gesamtheit eines funktionsfähigen Systems dar, welches RAG durchführen kann. D.h. in einem RAG-System ist eine Komponente vorhanden, welche das Überführen der Zusatzdokumente in den Einbettungsraum durchführen kann (Embedding), eine Komponente zum Speichern der vorverarbeiteten Zusatzdokumente (Vektordatenbank), eine Komponente zur Abfrage des Speichersystems (Retriever) sowie eine sprachgenerierende Komponente (als eigentliches Sprachmodell). Abb. 1 stellt ein RAG-System vollständig dar. Ein RAG-System ist ein KI-System nach Art. 3 Nr. 1 KI-VO.
RAG-Subsystem	Aus dem Eingabeprompt wird ein Anfragevektor erzeugt und damit eine Ähnlichkeitssuche in der Vektordatenbank durchgeführt, um die relevantesten Chunks zu finden. Diese werden zusammen mit dem Eingabeprompt als erweiterte Anfrage an das textgenerierende Sprachmodell übergeben.
RAG-Methode	Die Abkürzung für Retrieval Augmented Generation beschreibt eine Methode, welche zu Anfragen an ein Sprachmodell diese zusätzlich um passende Informationen aus Zusatzquellen anreichert, um Anfrage und Zusatzinformationen als veränderte und angereicherte Anfrage an das Sprachmodell zu stellen.
Referenzdokumente	Referenzdokumente können sowohl innerhalb als auch außerhalb des RAG-Systems liegen und stellen die Grundlage für die durch das RAG-System zu verarbeitenden Daten dar.
Retriever	Ist die Komponente des RAG-Systems, welche für die Abfrage von Zusatzinformationen zuständig ist, welche zur Eingabe an das Sprachmodell passen. Die Zusatzinformationen können aus einer Vektordatenbank stammen.
SLM	Abkürzung für Small Language Model. Vgl. LLM.
Sprachmodell	Ist der Oberbegriff für KI-Modelle, welche mit Sprache trainiert

	wurden. Zu den Sprachmodellen gehören z. B. LLMs und Embedding-Modelle
Token(sequenz)	Als Vorstufe zum Embedding werden die Textsequenzen der Anfrage bzw. der Zusatzinformationen in Wörter/Wortteile zerlegt, welche dann „Token“ darstellen. Tokensequenzen sind damit nur eine komprimierte Darstellung der Textsequenzen und eine Vorstufe des Embeddings. Welche Wörter/Wortteile welchen Token ergeben, wurde ebenfalls im Vorfeld gelernt bzw. programmiert. Die Wandlung von Text in Tokensequenzen stellt daher ebenfalls ein KI-Modell dar, ist aber eher mit der Funktion eines Wörterbuches vergleichbar.
Transformer	Eine spezifische Architektur eines Sprachmodells zur Vorhersage des nächstwahrscheinlichsten Wortes.
Vektorbasierter Suchmechanismus	Sucht in der Vektordatenbank Einträge mit ähnlichem Einbettungs-Vektor. Aus dem Prompt wird der Embedding-Vektor des Prompts berechnet. Einträge in der Nähe dieses Vektors werden als relevante Zusatzinformationen als Suchergebnisse zurückgegeben. Diese Art der Suche wird vom Retriever genutzt.
Vektordatenbank	Speichert z. B. die Tripel, die sich aus den Elementen Chunk, Tokensequenz des Chunks (Textabschnitt) und Vektor (Embedding des Chunks) zusammensetzen, in einer Datenbank.