German Q&A-Gen

Erstellung von Datensätzen für das Fine-Tuning von Information Retrieval Modellen

Persona-basierte Generierung extraktiver Frage-Antwort-Paare zur Domänenadaption von Retrieval-Modellen für deutschsprachige Verwaltungstexte

Ein Projekt finanziert im Rahmen des Programms **KIRAS/K-PASS**

Wissenschaftlicher Bericht

Helmut Ablinger[1] · Antonio Girlasu-Lauhard[2]

[1] JAAS GmbH, Wien [2] Language Services Semantica e.U.

Oktober 2026

Das Projekt „German Q&A-Gen – Erstellung von Datensätzen für das Fine-Tuning von Information Retrieval Modellen“ wird gefördert bzw. finanziert im Rahmen des Cybersicherheitsforschungsförderprogramms K-PASS durch das Bundesministerium für Finanzen (BMF) und von der Österreichischen Forschungsförderungsgesellschaft (FFG) abgewickelt.

Kurzfassung

Die Antwortqualität von Retrieval-Augmented-Generation-Systemen (RAG) hängt maßgeblich vom Retrieval ab, für dessen Anpassung an deutschsprachige Verwaltungsvorschriften Trainings- und Testdaten fehlen. Der Bericht stellt einen Ansatz vor, der solche Daten on-premise mit einem Open-Weight-Sprachmodell erzeugt: Zu jeder Quellpassage entstehen Anfragen aus Sicht frei definierbarer Personas, die Sprachstil, Rolle und Nutzungskontext abbilden; die Antwort ist stets ein wörtliches Zitat der Quelle. Im Vergleich mit 1.000 manuell erstellten Gold-Standard-Paaren war die maschinelle Erstellung formal fehlerfrei (manuell: 18,2 Prozent Formatfehler), in einer Stichprobe ebenso ohne falsche Antworten und um ein Vielfaches schneller, erreichte das menschliche Niveau aber nur mit ausführlichen Personas und übertrieb bei sehr informellen Anfragen. Daraus folgt eine Arbeitsteilung: Fachleute definieren Personas und übernehmen heikle informelle Anfragen, die Maschine den Großteil der Datenerstellung. Die erzeugte Ground Truth dient zudem als Testwerkzeug für RAG-Systeme.

Schlüsselwörter: Retrieval-Augmented Generation; Information Retrieval; synthetische Trainingsdaten; Personas; Ground Truth; Verwaltungssprache

Abstract

The answer quality of retrieval-augmented generation (RAG) systems depends largely on retrieval, yet training and test data for adapting it to German-language administrative regulations are lacking. This report presents an approach that generates such data on-premises with an open-weight language model: for each source passage, queries are generated from the perspective of freely definable personas capturing language style, role and context of use; the answer is always a verbatim quotation from the source. Compared with 1,000 manually created gold-standard pairs, machine generation was formally error-free (manual: 18.2 percent format errors), equally free of incorrect answers in a sample and many times faster, but reached human quality only with detailed personas and exaggerated very informal queries. This suggests a division of labour: experts define personas and handle sensitive informal queries, while the machine produces the bulk of the data. The resulting ground truth also serves as a testing tool for RAG systems.

Keywords: retrieval-augmented generation; information retrieval; synthetic training data; personas; ground truth; administrative language

Inhaltsverzeichnis

Kurzfassung **2**
Abstract **2**
1 Einleitung **6**
1.1 Ausgangslage und Motivation 6
1.2 Zielsetzung, Forschungsfragen und Beiträge 7
2 Grundlagen: Retrieval-Augmented Generation **8**
2.1 Architektur von RAG-Systemen 8
2.2 Der Retrieval-Schritt als kritischer Hebel 8
2.3 Retrieval-Verfahren 9
2.4 Evaluation von Retrieval- und RAG-Systemen 10
3 Problemstellung **11**
3.1 Verwaltungsvorschriften als eigenes Textgenre 11
3.2 Sprach- und Domänenlücke vortrainierter Retrieval-Modelle 12
3.3 Heterogenität der Nutzeranfragen 12
3.4 Fehlende Trainings- und Testdaten 13
4 Verwandte Arbeiten **14**
4.1 Synthetische Trainingsdaten für Retrieval-Modelle 14
4.2 Persona-basierte Datengenerierung 15
4.3 Ground Truth, Gold-Standards und menschliche Evaluation 15
5 Methodik **16**
5.1 Forschungsdesign 16
5.2 Korpusanalyse und Festlegung zulässiger Anfragen 16
5.3 Personas als Steuerungsinstrument 17
5.4 Datenmodell: n Anfragen, eine extraktive Antwort 18
5.5 Maschinelle Generierung und Validierung 19
5.6 Manuelle Gold-Standard-Erstellung 20
5.7 Vergleichsdesign und Bewertungskriterien 20
6 Softwarearchitektur und Einsatzszenarien **22**
6.1 Modularer Aufbau 22
6.2 Modellunabhängigkeit und On-Premise-Betrieb 22
6.3 Einsatz als Testwerkzeug für RAG-Systeme 23
7 Ergebnisse **24**
7.1 Formale Korrektheit 24
7.2 Inhaltliche Richtigkeit und Quellentreue 24
7.3 Effizienz der Erstellung 25
7.4 Einfluss von Persona-Definitionen und Domänenwissen 25
7.5 Natürlichkeit der Anfragen 26
7.6 Zusammenfassung der Ergebnisse 26
8 Diskussion **27**
8.1 Interpretation der Ergebnisse 27
8.2 Implikationen für Training und Evaluation 27
8.3 Limitationen 28

8.4 Übertragbarkeit 28
9 Ausblick und weiterführende Forschung 29
10 Fazit 30
Literaturverzeichnis 31

Abbildungsverzeichnis

Abbildung 1: Schematischer Ablauf eines RAG-Systems 8
Abbildung 2: Heterogene Anfragen zu derselben Quellpassage 12
Abbildung 3: Forschungsdesign mit maschinellem und manuellem Erstellungspfad 16
Abbildung 4: Datenmodell eines Q&A-Paares mit n persona-spezifischen Anfragen und einer extraktiven Antwort 19
Abbildung 5: Prozess der manuellen Gold-Standard-Erstellung mit dreistufigem Korrektorat 20
Abbildung 6: Modulare Architektur der Software (vereinfacht) 22
Abbildung 7: Einsatz der generierten Ground Truth als Regressionstest beim Komponententausch in RAG-Systemen 24
Abbildung 8: Kumulierter Aufwand manueller und maschineller Erstellung (schematisch) 25

Tabellenverzeichnis

Tabelle 1: Vergleich gängiger Retrieval-Verfahren 10
Tabelle 2: Einordnung verwandter Ansätze zur synthetischen Erzeugung von Retrieval-Trainingsdaten 14
Tabelle 3: Dimensionen einer Persona-Definition mit illustrativen Beispielen 18
Tabelle 4: Bewertungskriterien und Operationalisierung 21
Tabelle 5: Zusammenfassender Vergleich manuell und maschinell erstellter Q&A-Paare 26
Tabelle 6: Agenda für weiterführende Forschung 29

1 Einleitung

1.1 Ausgangslage und Motivation

Große Sprachmodelle (Large Language Models, LLMs) werden zunehmend auch in der öffentlichen Verwaltung eingesetzt, etwa um Bedienstete bei der Suche in umfangreichen internen Regelwerken zu unterstützen. Ein Sprachmodell allein ist dafür ungeeignet: Es kennt interne Dokumente nicht, sein Wissen ist auf den Stand der Trainingsdaten eingefroren, und es kann plausibel klingende, aber unbelegte Aussagen erzeugen. Retrieval-Augmented Generation (RAG) begegnet diesen Grenzen, indem zu jeder Anfrage zunächst passende Textpassagen aus einem Dokumentenbestand gesucht und dem Sprachmodell als Grundlage für die Antwort übergeben werden (Lewis et al. 2020; Gao et al. 2023). Die Qualität der Antwort hängt damit maßgeblich davon ab, ob dieser Retrieval-Schritt die richtigen Quellen findet (Abschnitt 2.2).

Für deutschsprachige Verwaltungstexte ist das keineswegs selbstverständlich. Für nicht-englische Frage-Antwort-Systeme gilt der Mangel an annotierten Datensätzen als zentrale Herausforderung (Möller et al. 2021), Verwaltungsvorschriften folgen einer eigenen Fachsprache, und Nutzerinnen und Nutzer formulieren denselben Informationsbedarf sehr unterschiedlich (Kapitel 3). Ein naheliegender Weg ist die Domänenadaption eines Retrieval-Modells durch Fine-Tuning. Dafür werden Paare aus Anfrage und relevanter Passage benötigt, die es für interne Dokumente nicht gibt, weil diese nicht öffentlich zugänglich sind. Eine rein manuelle Erstellung skaliert nicht. Arbeiten zur synthetischen Datengenerierung zeigen, dass Sprachmodelle geeignete Anfragen zu vorhandenen Passagen erzeugen können (Bonifacio et al. 2022). Offen ist, wie sich dabei die Vielfalt realer Anfragen gezielt abbilden lässt und wie maschinell erzeugte Daten im Vergleich zu sorgfältig manuell erstellten abschneiden.

Hier setzt das Projekt German Q&A-Gen an, das im Rahmen des Programms KIRAS/K-PASS (Innovation AKUT) gemeinsam mit einer österreichischen Bundesbehörde als Bedarfsträgerin durchgeführt wurde. Inhalte, Umfang und Struktur des zugrunde liegenden Dokumentenbestands werden aus Gründen der Vertraulichkeit nicht wiedergegeben; alle Beispiele in diesem Bericht sind frei erfunden.

1.2 Zielsetzung, Forschungsfragen und Beiträge

Ziel war die Entwicklung und Erprobung einer Methode und einer Software, die aus deutschsprachigen Verwaltungsdokumenten automatisiert Frage-Antwort-Paare (Q&A-Paare) erzeugt. Die Paare sollen die sprachliche und situative Vielfalt realer Anfragen abbilden, jede Antwort wörtlich aus der Quelle übernehmen und maschinenlesbar vorliegen, damit sie sowohl als Trainingsdaten für Retrieval-Modelle als auch als Ground Truth für die Prüfung von RAG-Systemen dienen können. Die Software sollte vollständig auf der Infrastruktur der Bedarfsträgerin mit Open-Weight-Sprachmodellen betrieben werden können. Daraus ergeben sich drei Forschungsfragen:

– **FF1:** Wie lassen sich aus deutschsprachigen Verwaltungsvorschriften automatisiert Q&A-Paare erzeugen, die die Heterogenität realer Nutzeranfragen abbilden und zugleich quellentreu bleiben?
– **FF2:** Wie unterscheiden sich maschinell erzeugte Q&A-Paare von manuell durch linguistische Expertise erstellten Gold-Standard-Paaren hinsichtlich formaler Korrektheit, inhaltlicher Richtigkeit, Effizienz und Natürlichkeit?
– **FF3:** Welchen Beitrag leisten explizit definierte Personas, um Sprachstil, Relevanz und Nutzungskontext der generierten Anfragen zu steuern?

Der Bericht leistet dazu vier Beiträge: einen Ansatz, bei dem zu einer Quellpassage beliebig viele Anfragen aus der Perspektive frei definierbarer Personas mit genau einer wörtlichen Antwort erzeugt werden; eine modulare, modellunabhängige Software für den On-Premise-Betrieb; eine Vergleichsstudie gegenüber einem manuell erstellten Gold-Standard; sowie die Nutzung der erzeugten Ground Truth als anpassbares und kostengünstiges Testwerkzeug, mit dem sich der Austausch einzelner Komponenten eines RAG-Systems überprüfen lässt. Das Training eines Retrieval-Modells mit diesen Daten ist nicht Gegenstand des Berichts.

Kapitel 2 und 3 behandeln Grundlagen und Problemstellung, Kapitel 4 ordnet die Arbeit in die Literatur ein. Kapitel 5 und 6 beschreiben Methodik und Software, Kapitel 7 und 8 präsentieren und diskutieren die Ergebnisse, Kapitel 9 und 10 schließen mit Ausblick und Fazit.

2 Grundlagen: Retrieval-Augmented Generation

2.1 Architektur von RAG-Systemen

Gao et al. (2023) beschreiben die Grundform eines RAG-Systems („Naive RAG“) als Abfolge von Indexierung, Retrieval und Generierung. Bei der Indexierung wird der Dokumentenbestand bereinigt und in kleinere Textabschnitte (Chunks) zerlegt. Ein Embedding-Modell überführt jeden Abschnitt in eine Vektorrepräsentation; Abschnitte und Vektoren werden gemeinsam in einem Index abgelegt. Beim Retrieval werden die k Abschnitte ausgewählt, die der Anfrage am ähnlichsten sind. Bei der Generierung werden Anfrage und ausgewählte Abschnitte zu einem Prompt zusammengeführt, auf dessen Grundlage das Sprachmodell antwortet (Gao et al. 2023). Lewis et al. (2020) sprechen von einem nicht-parametrischen Gedächtnis in Form eines dichten Vektorindex, auf den über einen neuronalen Retriever zugegriffen wird. Abbildung 1 fasst den Ablauf zusammen.

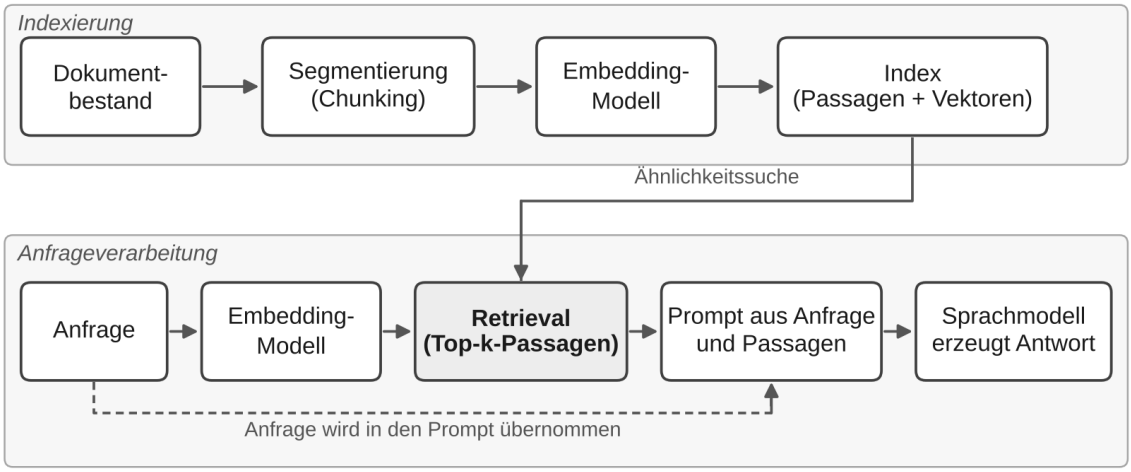

Abbildung 1: Schematischer Ablauf eines RAG-Systems
Quelle: eigene Darstellung in Anlehnung an Gao et al. (2023) und Lewis et al. (2020).

2.2 Der Retrieval-Schritt als kritischer Hebel

Aus dieser Architektur folgt eine einfache, aber folgenreiche Abhängigkeit: Das Sprachmodell kann nur mit den Abschnitten arbeiten, die ihm das Retrieval übergibt. Fehlt die maßgebliche Passage unter den k Treffern, ist eine korrekte und belegte Antwort nicht mehr möglich, unabhängig davon, wie leistungsfähig das Sprachmodell ist. Die Trefferquote des Retrievals bildet damit eine Obergrenze für den Anteil der Anfragen, die das Gesamtsystem quellengestützt beantworten kann.

Barnett et al. (2024) beschreiben auf Basis von drei Fallstudien sieben typische Fehlerquellen von RAG-Systemen. Drei davon betreffen unmittelbar die Bereitstellung der Quellen: Die Frage lässt sich mit den vorhandenen Dokumenten nicht beantworten; die Antwort steht in einem Dokument, das aber nicht hoch genug gerankt wird; oder die passenden Dokumente werden gefunden, gelangen aber nicht in den Kontext, aus dem die Antwort erzeugt wird. Auch die Reihenfolge der Treffer spielt eine Rolle: Sprachmodelle nutzen relevante Information in der Mitte langer Kontexte schlechter als an deren Anfang oder Ende (Liu et al. 2024).

Daraus folgt zweierlei. Das Retrieval ist der Hebel, an dem Verbesserungen eines RAG-Systems zuerst ansetzen sollten. Und jede gezielte Verbesserung setzt voraus, dass sich die Retrieval-Qualität messen lässt, wofür Anfragen benötigt werden, zu denen bekannt ist, welche Passage gefunden werden muss. Genau solche Daten erzeugt der hier beschriebene Ansatz.

2.3 Retrieval-Verfahren

Lexikalische Verfahren vergleichen Anfrage und Dokument auf Ebene der verwendeten Wörter. Über einen invertierten Index werden alle Dokumente gefunden, die die Terme der Anfrage enthalten; die Rangfolge ergibt sich aus Termgewichten. Das bekannteste Gewichtungsschema ist BM25, nach Robertson und Zaragoza (2009) eines der erfolgreichsten Verfahren des Textretrievals. Lexikalische Verfahren erweisen sich auch im Vergleich mit neuronalen Modellen als robuste Referenz (Thakur et al. 2021). Ihre Stärke ist der exakte Abgleich von Termen, ihre zentrale Schwäche das Vokabularproblem: Anfragen verwenden andere Begriffe als die relevanten Dokumente (Nogueira et al. 2019). Im Deutschen verschärft die Komposition dieses Problem, da frei zusammengesetzte Wörter das Vokabular stark vergrößern (Koehn und Knight 2003). Ein eigenes Beispiel: Eine Anfrage nach der „Abrechnung der Reisekosten“ teilt mit einem Dokument über die „Reisekostenabrechnung“ ohne zusätzliche Vorverarbeitung keinen einzigen Term.

Dichte Verfahren repräsentieren Anfragen und Passagen als Vektoren in einem gemeinsamen semantischen Raum. Karpukhin et al. (2020) zeigen mit Dense Passage Retrieval (DPR), dass sich solche Repräsentationen mit einem einfachen Dual-Encoder aus einer vergleichsweise kleinen Zahl von Fragen und Passagen lernen lassen; die Ähnlichkeit wird als Skalarprodukt der Vektoren berechnet. In ihren Experimenten übertrifft DPR ein starkes BM25-System um 9 bis 19 Prozentpunkte bei der Treffergenauigkeit unter den ersten 20 Passagen. Das Training erfolgt kontrastiv: Das Modell lernt, die passende

Passage näher an die Frage heranzurücken als andere Passagen, etwa jene der übrigen Fragen im selben Trainingsbatch (In-Batch Negatives; Karpukhin et al. 2020). Da Anfrage und Passage getrennt kodiert werden, lassen sich die Passagenvektoren vorab berechnen und in Indizes für die approximative Nächste-Nachbarn-Suche ablegen. Für das Deutsche stehen mehrsprachige Modelle zur Verfügung, etwa M3-Embedding mit Unterstützung für mehr als 100 Sprachen (Chen et al. 2024). Allgemein gültige Empfehlungen gibt es jedoch nicht: Keine Embedding-Methode dominiert über alle Aufgaben hinweg (Muennighoff et al. 2023). Die Eignung eines Modells für eine konkrete Domäne muss daher mit domänenspezifischen Daten geprüft und gegebenenfalls durch Training hergestellt werden.

Hybride Suche führt lexikalisches und dichtes Retrieval parallel aus und vereinigt die Ergebnislisten, verbreitet etwa über Reciprocal Rank Fusion, ein einfaches Verfahren zur Kombination der Ranglisten mehrerer Systeme (Cormack et al. 2009). Da nur Rangplätze eingehen, müssen die unterschiedlich skalierten Scores nicht normalisiert werden. Beim *Re-Ranking* wird zunächst eine größere Kandidatenmenge ermittelt und anschließend von einem aufwendigeren Modell neu bewertet, das Anfrage und Passage gemeinsam verarbeitet (Cross-Encoder); Nogueira und Cho (2019) steigerten damit auf MS MARCO den MRR@10 gegenüber dem damaligen Stand der Technik von 28,1 auf 35,8, also relativ um rund 27 Prozent. Tabelle 1 fasst die Verfahren zusammen.

Tabelle 1: Vergleich gängiger Retrieval-Verfahren

Verfahren	Stärken	Schwächen	Bedeutung für Verwaltungstexte
Lexikalisch (BM25)	Exakter Abgleich von Fachbegriffen und Kennungen; ohne Training einsetzbar	Vokabularproblem; Komposita und Flexion	Stark bei feststehender Terminologie, schwach bei freien Formulierungen
Dicht (Bi-Encoder)	Erfasst inhaltliche Ähnlichkeit trotz abweichender Wortwahl; effizient durch vorab berechnete Vektoren	Abhängig von Trainingssprache und -domäne	Profitiert von Fine-Tuning auf Domänendaten
Hybrid (z. B. RRF)	Kombiniert komplementäre Stärken; keine Score-Normalisierung nötig	Mehrere Indizes und Verfahren zu betreiben	Ausgleich zwischen Fachsprache und Alltagssprache
Re-Ranking (Cross-Encoder)	Hohe Genauigkeit bei der Reihung der Kandidaten	Rechenaufwendig; nur auf Kandidatenmenge anwendbar	Verbessert die Reihenfolge der an das LLM übergebenen Passagen

Quelle: eigene Darstellung auf Basis von Robertson und Zaragoza (2009), Karpukhin et al. (2020), Cormack et al. (2009) sowie Nogueira und Cho (2019).

2.4 Evaluation von Retrieval- und RAG-Systemen

Die Retrieval-Qualität wird mit rangbasierten Kennzahlen gemessen. Recall@k gibt an, bei welchem Anteil der Anfragen die relevante Passage unter den ersten k Treffern ist, und ist

für RAG-Systeme besonders aussagekräftig, weil nur diese Treffer an das Sprachmodell gehen. Der Mean Reciprocal Rank (MRR) belohnt eine möglichst weit vorne liegende Platzierung des ersten relevanten Treffers, nDCG@k berücksichtigt zusätzlich abgestufte Relevanzurteile und dient etwa im BEIR-Benchmark als Hauptkennzahl (Thakur et al. 2021). Benchmarks wie BEIR oder MTEB (Muennighoff et al. 2023) erlauben allgemeine Modellvergleiche, bilden aber naturgemäß keine internen Dokumentenbestände ab.

Für ganze RAG-Pipelines schlagen Es et al. (2024) mit RAGAs eine referenzfreie Bewertung vor, die die Qualität des Retrievals, die getreue Nutzung der Passagen durch das Sprachmodell und die Qualität der Antwort einschätzt. Übersichtsarbeiten zur Evaluation von Sprachmodellen betonen zudem, dass uneinheitliche Evaluationsaufbauten zu inkonsistenten Befunden führen (Laskar et al. 2024). Rangbasierte Kennzahlen setzen jedoch Relevanzurteile voraus: Für jede Anfrage muss bekannt sein, welche Passage als richtig gilt. Referenzfreie Verfahren ergänzen diese Sicht, ersetzen eine solche Ground Truth aber nicht, wenn die Retrieval-Qualität gezielt gemessen und über Systemversionen hinweg verglichen werden soll. Für interne Dokumentenbestände existiert sie in der Regel nicht.

3 Problemstellung

3.1 Verwaltungsvorschriften als eigenes Textgenre

Dienst- und Verwaltungsvorschriften regeln, wie eine Behörde intern vorzugehen hat. Sprachlich sind sie von präziser Fachterminologie, Nominalstil, langen Komposita und standardisierten Formulierungen geprägt; formal folgen sie einer stark gegliederten Struktur mit nummerierten Abschnitten, Aufzählungen und Querverweisen auf andere Regelungen. Für das Retrieval hat das zwei Folgen. Erstens unterscheidet sich die Sprache der Dokumente deutlich von der Sprache, in der Bedienstete fragen (Abschnitt 3.3). Zweitens ist die Struktur bei der Segmentierung zu berücksichtigen, weil etwa eine Aufzählung ohne ihren Einleitungssatz kaum verständlich ist.

Die Sichtung des Dokumentenbestands im Projekt bestätigte weitere genretypische Merkmale. Anfang und Ende der Dokumente enthalten meist organisatorische Angaben mit geringer inhaltlicher Relevanz, während der verwertbare Inhalt im Hauptteil liegt. Zudem kommen inhaltlich gleichartige Regelungen in mehreren Varianten vor, etwa für unterschiedliche Partner oder Geltungsbereiche. Für das Retrieval konkurrieren solche

Varianten um dieselben Anfragen; für die Erstellung von Trainingsdaten bergen sie die Gefahr, dieselbe Frage mehrfach zu stellen und den Datensatz einseitig zu gewichten.

3.2 Sprach- und Domänenlücke vortrainierter Retrieval-Modelle

Interne Verwaltungsvorschriften sind in der Regel nicht öffentlich zugänglich und können daher weder in den Trainingsdaten verbreiteter Embedding-Modelle noch in öffentlichen Benchmarks enthalten sein. Hinzu kommt die Sprachlücke: Möller et al. (2021) nennen den Mangel an annotierten Datensätzen als zentrale Herausforderung für nicht-englische Frage-Antwort-Systeme und zeigen, dass maschinell übersetzte Trainingsdaten handannotierte Daten in der Zielsprache nicht vollständig ersetzen können. Da die Leistung eines Modells in seiner Trainingsdomäne kein verlässlicher Indikator für seine Leistung in anderen Domänen ist (Thakur et al. 2021), lässt sich die Eignung eines Modells für solche Bestände nur mit domänenspezifischen Daten beurteilen.

3.3 Heterogenität der Nutzeranfragen

Zur Fachsprache der Dokumente kommt die Vielfalt auf Seite der Anfragen. Penha et al. (2022) halten fest, dass Anfragen, die verschiedene Personen zu ein und demselben Informationsbedarf stellen, äußerst variabel sind, und zeigen, dass Retrieval-Pipelines gegenüber solchen Variationen nicht robust sind; in ihren Experimenten auf zwei Datensätzen sinkt die Effektivität im Mittel um rund 20 Prozent. Im Verwaltungskontext unterscheiden sich Anfragen mindestens in folgenden Dimensionen, die Abbildung 2 an frei erfundenen Beispielen veranschaulicht:

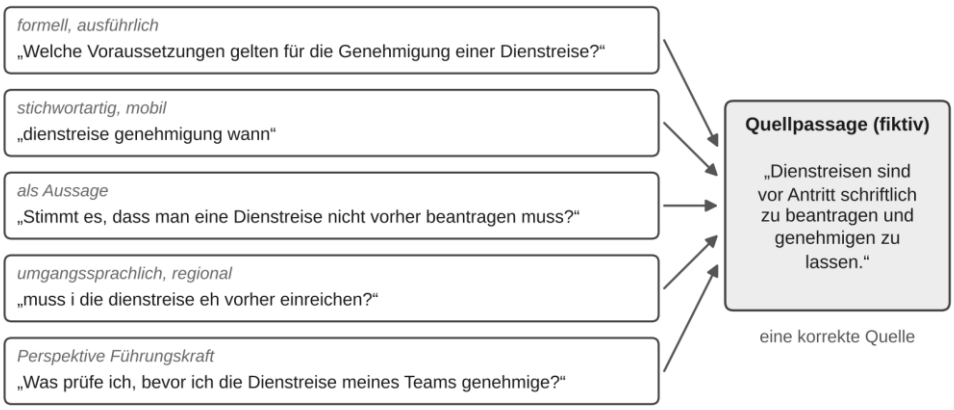

Abbildung 2: Heterogene Anfragen zu derselben Quellpassage
Quelle: eigene Darstellung; Anfragen und Passage frei erfunden.

- **Sprachstil und Register:** formell oder umgangssprachlich, gegebenenfalls regional oder dialektal gefärbt.
- **Länge und Form:** ausführliche Frage, stichwortartige Eingabe oder Aussage, die bestätigt oder widerlegt werden soll („Stimmt es, dass …“). Dass Nutzerinnen und Nutzer im Chatbot häufig Aussagen statt Fragen formulieren, wurde im Projekt ausdrücklich berücksichtigt.
- **Rolle und Relevanz:** Eine Person mit Führungsverantwortung interessiert sich für andere Aspekte derselben Regelung als eine Person ohne diese Verantwortung; manche Inhalte sind für eine Rolle zentral, für eine andere selbsterklärend.
- **Situation und Endgerät:** Wer am Schreibtisch mit mehreren Bildschirmen recherchiert, fragt anders als jemand, der unterwegs unter Zeitdruck und Belastung auf einem Mobilgerät rasch eine Antwort sucht.

Alle fünf Anfragen in Abbildung 2 zielen auf dieselbe Passage. Ein Retrieval-Modell muss alle diese Eingaben auf dieselbe Quelle abbilden. Daraus folgt die zentrale Anforderung an Trainingsdaten: Sie müssen die Heterogenität der Anfragen abbilden, während die zugehörige Quelle für alle Varianten identisch bleibt.

3.4 Fehlende Trainings- und Testdaten

Für die Domänenadaption eines Retrieval-Modells wie für die Prüfung eines RAG-Systems werden somit Daten benötigt, die zu jeder Quellpassage viele unterschiedliche, realistische Anfragen enthalten. Für interne Dokumentenbestände liegen solche Daten nicht vor. Eine manuelle Erstellung durch Fachleute ist aufwendig, skaliert nicht mit dem Umfang des Bestands und spiegelt die Formulierungsgewohnheiten weniger Personen wider. Ein geeignetes Verfahren zur automatisierten Erzeugung muss daher vier Anforderungen erfüllen:

- **Heterogenität:** Die Anfragen bilden Stil, Form, Rolle und Situation der späteren Nutzerinnen und Nutzer ab.
- **Quellentreue:** Die Antwort ist eine wörtliche Passage aus dem Originaldokument und damit eindeutig überprüfbar.
- **Maschinenlesbarkeit:** Die Paare liegen in einem formal korrekten, strukturierten Format vor, das Trainings- und Evaluationsprozesse direkt verarbeiten können.

– **Datensouveränität:** Die Erzeugung erfolgt vollständig auf eigener Infrastruktur, sodass die Dokumente die Organisation nicht verlassen.

4 Verwandte Arbeiten

4.1 Synthetische Trainingsdaten für Retrieval-Modelle

Die Idee, zu vorhandenen Passagen passende Anfragen maschinell zu erzeugen, hat sich im Information Retrieval schrittweise entwickelt. Doc2query sagt vorher, welche Anfragen zu einem Dokument gestellt werden könnten, und erweitert das Dokument um diese (Nogueira et al. 2019). InPars nutzt die Few-Shot-Fähigkeiten großer Sprachmodelle als Generatoren synthetischer Trainingsdaten; Modelle, die ausschließlich auf diesen Daten feinjustiert werden, übertreffen starke Baselines wie BM25 (Bonifacio et al. 2022). GPL geht von der Beobachtung aus, dass dichte Retriever große Mengen an Trainingsdaten benötigen, die für die meisten Domänen nicht verfügbar sind, und dass ihre Leistung bei einem Domänenwechsel stark nachlässt. Das Verfahren kombiniert einen Anfragegenerator mit Pseudo-Labels eines Cross-Encoders und übertrifft einen ohne Anpassung eingesetzten State-of-the-Art-Ansatz für dichtes Retrieval um bis zu 9,3 Punkte nDCG@10 (Wang et al. 2022). Wang et al. (2024) schließlich erzeugen mit proprietären Sprachmodellen vielfältige synthetische Daten für Hunderttausende Embedding-Aufgaben in 93 Sprachen und trainieren darauf Embedding-Modelle mit einer kontrastiven Verlustfunktion.

Der vorliegende Ansatz knüpft an diese Arbeiten an, setzt aber drei andere Schwerpunkte (Tabelle 2): Die Vielfalt der Anfragen wird über explizit definierte Personas gesteuert; zu jeder Passage entstehen mehrere Anfragen mit genau einer wörtlich übernommenen Antwort; und die Erzeugung erfolgt mit Open-Weight-Modellen vollständig auf der Infrastruktur der Organisation, deren Dokumente verarbeitet werden.

Tabelle 2: Einordnung verwandter Ansätze zur synthetischen Erzeugung von Retrieval-Trainingsdaten

Ansatz	Grundidee	Verwendung der erzeugten Daten
doc2query (Nogueira et al. 2019)	Vorhersage möglicher Anfragen zu einem Dokument	Erweiterung der Dokumente im Index
InPars (Bonifacio et al. 2022)	Sprachmodell erzeugt Few-Shot synthetische Trainingsdaten	Fine-Tuning von Retrieval-Modellen
GPL (Wang et al. 2022)	Anfragegenerator kombiniert mit Pseudo-Labels eines Cross-Encoders	Domänenadaption dichter Retriever

Ansatz	Grundidee	Verwendung der erzeugten Daten
Wang et al. (2024)	Proprietäre Sprachmodelle erzeugen Daten für viele Embedding-Aufgaben und Sprachen	Kontrastives Training von Embedding-Modellen
Dieser Bericht	Mehrere persona-gesteuerte Anfragen je Passage, eine wörtliche Antwort; Open-Weight-Modell on-premise	Trainingsdaten für Retrieval-Modelle und Ground Truth für Tests von RAG-Systemen

Quelle: eigene Darstellung auf Basis der genannten Arbeiten.

4.2 Persona-basierte Datengenerierung

Ge et al. (2024) schlagen eine persona-gesteuerte Datensynthese vor, die die vielfältigen Perspektiven innerhalb eines Sprachmodells nutzt, und stellen mit Persona Hub eine Sammlung von einer Milliarde aus Webdaten kuratierter Personas bereit. Personas fungieren dabei als verteilte Träger von Weltwissen. Yu et al. (2023) zeigen für die Erzeugung von Trainingsdaten, dass einfache klassenbedingte Prompts die Vielfalt der erzeugten Daten begrenzen und systematische Verzerrungen des Sprachmodells übernehmen können, während Prompts mit gezielt variierten Attributen in ihren Experimenten zur Textklassifikation zu besseren Modellen führen. Der vorliegende Ansatz überträgt diese Idee auf das Retrieval in einer geschlossenen Domäne: Statt einer großen Zahl allgemeiner Personas wird eine kleine Zahl fachlich definierter Personas verwendet, die Rolle, Verantwortung, Arbeitskontext, Eingabesituation und Sprachstil der tatsächlichen Nutzergruppen beschreiben.

4.3 Ground Truth, Gold-Standards und menschliche Evaluation

Auch synthetisch erzeugte Daten benötigen einen menschlichen Bezugspunkt. In einem Praxisleitfaden zur Erstellung von Ground Truth für Frage-Antwort-Systeme halten Stuart et al. (2025) fest, dass von Sprachmodellen erzeugte Ground Truth die Einbindung von Fachexpertinnen und -experten nicht ersetzt. Sie empfehlen, mit einem kleinen, von Menschen kuratierten Datensatz zu beginnen, den Prüfaufwand am akzeptierten Risiko auszurichten und Sprachmodelle als Prüfinstanz nur zur Verringerung, nicht zur Beseitigung menschlicher Kontrolle einzusetzen. Für die Verlässlichkeit manuell erstellter Daten ist schließlich die Übereinstimmung zwischen mehreren Annotierenden maßgeblich: Daten gelten als verlässlich, wenn gezeigt werden kann, dass die Annotierenden sich einig sind (Artstein und Poesio 2008).

Der vorliegende Bericht folgt diesen Empfehlungen, indem den maschinell erzeugten Paaren ein von einem Linguisten mit Expertise in Verwaltungssprache manuell erstellter Gold-Standard gegenübergestellt wird (Abschnitt 5.6). Die Grenzen dieses Vorgehens,

insbesondere hinsichtlich der Übereinstimmung mehrerer Annotierender, werden in Abschnitt 8.3 diskutiert.

5 Methodik

5.1 Forschungsdesign

Das Forschungsdesign verbindet zwei parallele Erstellungspfade mit einem systematischen Vergleich (Abbildung 3). Ausgangspunkt beider Pfade ist eine gemeinsame Korpusanalyse mit verbindlichen Festlegungen, die mit der Bedarfsträgerin abgestimmt wurden (Abschnitt 5.2). Im manuellen Pfad erstellte ein Linguist mit Expertise in Verwaltungssprache einen Gold-Standard (Abschnitt 5.6). Im maschinellen Pfad erzeugte die entwickelte Software Q&A-Paare mithilfe von Personas und einem Open-Weight-Sprachmodell (Abschnitte 5.3 bis 5.5). Erkenntnisse aus der manuellen Erstellung flossen laufend in die Definition der Personas und die Gestaltung der Prompts ein. Abschließend wurden beide Datensätze anhand von fünf Kriterien verglichen (Abschnitt 5.7).

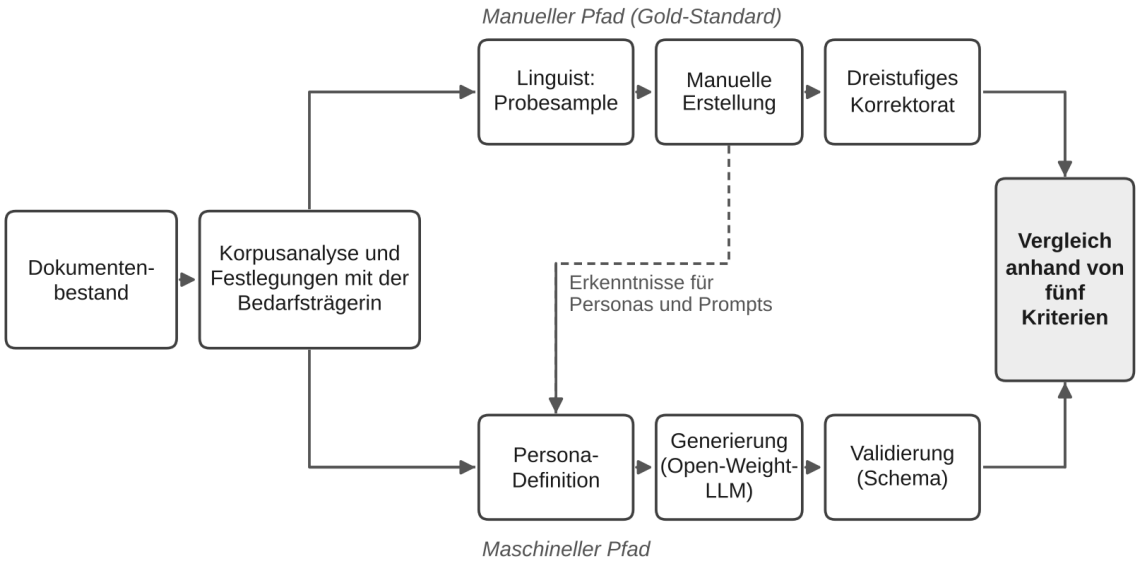

Abbildung 3: Forschungsdesign mit maschinellem und manuellem Erstellungspfad
Quelle: eigene Darstellung.

5.2 Korpusanalyse und Festlegung zulässiger Anfragen

Zu Beginn wurden gemeinsam mit der Bedarfsträgerin der konkrete Anwendungsfall, die Dokumentenbasis, die relevanten Dokumentteile, Art, Länge und Formalitätsgrad der

Anfragen sowie Lieferformat und Gesamtumfang festgelegt. Als Dokumentenbasis dienten die Hauptdokumente ohne Anhänge. Um eine repräsentative Abdeckung des Bestands zu erreichen, wurde die Zahl der Q&A-Paare je Dokument aus Kennzahlen zum Umfang der Dokumente abgeleitet und eine Obergrenze je Dokument festgelegt.

Als zulässige Anfrageformen wurden neben Fragen auch Aussagen festgelegt, die bestätigt oder widerlegt werden sollen. Diese Erweiterung erfolgte nach Abnahme eines ersten Probesamples, weil Nutzerinnen und Nutzer in Chatbots häufig Aussagen statt Fragen formulieren. Darüber hinaus wurden drei Regeln für die Auswahl der Passagen abgeleitet: Organisatorische Rahmenteile am Anfang und Ende der Dokumente werden nachrangig behandelt; bei gleichartigen Regelungen in mehreren Varianten werden unterschiedliche Fragen gestellt, statt dieselbe Frage je Variante zu wiederholen; und die Anfragen werden ausgewogen auf die Themen verteilt, die für die verschiedenen Nutzergruppen relevant sind.

5.3 Personas als Steuerungsinstrument

Die in Abschnitt 3.3 beschriebene Heterogenität der Anfragen lässt sich nicht allein über eine Vorgabe wie „formell“ oder „umgangssprachlich“ erzeugen. Der Ansatz verwendet daher Personas: textuelle Beschreibungen typischer Nutzerinnen und Nutzer, die dem Sprachmodell bei der Generierung mitgegeben werden. Eine Persona legt nicht nur den Sprachstil fest, sondern auch Rolle und Funktion, Verantwortung, Arbeitskontext, Eingabesituation und typische Informationsinteressen (Tabelle 3). Damit kann das Sprachmodell nicht nur unterschiedlich formulieren, sondern auch einschätzen, welche Aspekte einer Regelung für die jeweilige Person relevant sind. Eine Person mit Führungsverantwortung fragt etwa nach den Voraussetzungen einer Genehmigung, eine Person im Außendienst nach der konkreten Handlungsanweisung.

Die Zahl der Personas ist frei wählbar, von einer bis zu beliebig vielen; jede Persona erzeugt zu einer Passage eine eigene Anfrage. Der Ansatz greift damit die persona-gesteuerte Datensynthese von Ge et al. (2024) und den Befund von Yu et al. (2023) auf, dass gezielt variierte Attribute die Vielfalt synthetischer Daten erhöhen. Anders als dort werden die Personas jedoch nicht aus allgemeinen Webdaten gewonnen, sondern von Fachleuten für die konkreten Nutzergruppen definiert. Voraussetzung ist, dass sie vor der Generierung klar und ausführlich beschrieben werden; wie stark die Qualität davon abhängt, zeigt Abschnitt 7.4.

Tabelle 3: Dimensionen einer Persona-Definition mit illustrativen Beispielen

Dimension	Persona A: Sachbearbeitung	Persona B: Außendienst	Persona C: Führungskraft
Rolle und Funktion	Sachbearbeiterin in einer Fachabteilung	Bediensteter im operativen Außendienst	Leiterin einer Organisationseinheit
Verantwortung	Bearbeitung von Einzelfällen	Umsetzung von Vorgaben vor Ort	Personalverantwortung, Genehmigungen
Arbeitskontext	Büro, Recherche ohne Zeitdruck	Unterwegs, unter Zeitdruck und Belastung	Zwischen Terminen, Überblick gefragt
Eingabesituation	Arbeitsplatz mit mehreren Bildschirmen	Smartphone, kurze Eingaben	Laptop oder Smartphone
Sprachstil	Formell, fachsprachlich, vollständige Sätze	Umgangssprachlich, knapp, teils regional gefärbt	Sachlich, ergebnisorientiert
Typische Interessen	Verfahrensschritte, Fristen, Zuständigkeiten	Konkrete Handlungsanweisung	Pflichten als Vorgesetzte, Voraussetzungen von Genehmigungen

Quelle: eigene Darstellung. Frei erfundene Beispiele; nicht die im Projekt verwendeten Personas.

5.4 Datenmodell: n Anfragen, eine extraktive Antwort

Ein Q&A-Paar besteht aus n Anfragen, einer pro Persona, und genau einer Antwort (Abbildung 4). Die Anfragen können als Frage oder als Aussage formuliert sein. Die Antwort ist kein vom Sprachmodell formulierter Text, sondern eine wörtlich aus dem Originaldokument übernommene Passage. Ergänzt wird das Paar um Metadaten zum Quelldokument und zur Lage der Passage im Dokument. Jedes Paar wird als JSON-Objekt ausgegeben, damit nachgelagerte Prozesse es ohne manuelle Nacharbeit einlesen können.

Dieses Datenmodell erfüllt mehrere Zwecke zugleich. Die wörtliche Antwort macht jedes Paar nachvollziehbar und bindet es eindeutig an eine Quelle. Für das Training eines Retrieval-Modells liefert ein Paar n positive Anfrage-Passage-Paare zur selben Passage, die sich gerade in Formulierung und Perspektive unterscheiden. Für die Prüfung eines RAG-Systems ist zu jeder Anfrage bekannt, welche Passage gefunden werden muss. Die Gold-Standard-Paare entsprechen dem Sonderfall n = 2 (Abschnitt 5.6).

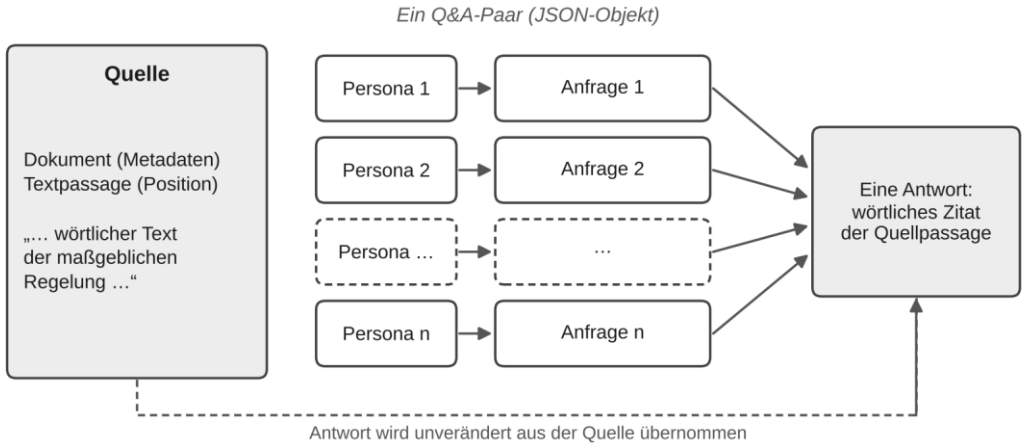

Abbildung 4: Datenmodell eines Q&A-Paares mit n persona-spezifischen Anfragen und einer extraktiven Antwort
Quelle: eigene Darstellung.

5.5 Maschinelle Generierung und Validierung

Die Software verarbeitet die hochgeladenen Dokumente nacheinander und extrahiert zunächst den Text. Die Erzeugung eines Paares folgt dann einem Vorgehen von der Antwort zur Frage: Das Sprachmodell wählt zuerst eine geeignete Textpassage des Dokuments als Antwort aus und formuliert erst anschließend, ausgehend von dieser Passage, die Anfragen der einzelnen Personas. Für die Auswahl enthält der Systemprompt Hinweise, etwa Deckblatt, Inhalts- und Abbildungsverzeichnisse sowie organisatorische Rahmenteile am Anfang und Ende der Dokumente zu meiden und nicht zu viele Paare je Seite zu erzeugen. Als letzter Schritt wird jede Ausgabe gegen ein vorgegebenes JSON-Schema validiert; nur schemakonforme Paare gelangen in den Datensatz.

Dieses Vorgehen entspricht der Grundidee synthetischer Retrieval-Daten, zu vorhandenen Passagen passende Anfragen zu erzeugen (Abschnitt 4.1), und ist für die Qualität entscheidend. Da die Antwort nicht generiert, sondern wörtlich übernommen wird, entfallen typische Fehlerquellen generierter Antworten wie Halluzinationen, Vereinfachungen, Verkürzungen oder ungenaue Umformulierungen. Das Sprachmodell muss nicht zu einer Frage eine Lösung suchen, sondern zu einer bekannten Lösung passende Fragen formulieren. Die verbleibende Fehlerquelle liegt damit in der Passung der Anfrage zur Antwort, nicht in der Antwort selbst (Abschnitte 7.2 und 8.3).

Für die Generierung wurde das Open-Weight-Modell Gemma 4 verwendet, das über einen lokal betriebenen Inferenzserver angebunden ist; es ist austauschbar, ohne dass die Software angepasst werden muss (Kapitel 6). Die in Abschnitt 5.2 beschriebenen

Auswahlregeln und die Erkenntnisse aus der manuellen Erstellung wurden in die Prompt-Vorlagen übernommen. Personas, Prompt-Vorlagen und Modellparameter lassen sich in der Benutzeroberfläche ohne Programmierkenntnisse anpassen. Insgesamt wurden im Projekt auf diese Weise 5.000 Q&A-Paare erzeugt.

5.6 Manuelle Gold-Standard-Erstellung

Den Gold-Standard erstellte ein Linguist mit Erfahrung in behördlicher Kommunikation vollständig manuell. Er umfasst 1.000 Q&A-Paare mit je zwei Anfragen aus Sicht zweier Personas, jeweils im Kontext der Bedarfsträgerin: einer formell formulierenden Person mit akademischem Hintergrund und Führungsverantwortung sowie einer umgangssprachlich formulierenden Mitarbeiterin bzw. eines Mitarbeiters ohne Führungsfunktion. Die Anfragen sind als Frage oder Aussage formuliert; die Antwort ist wörtlich aus dem Dokument übernommen. Entsprechend der Empfehlung, mit einem kleinen, von Menschen kuratierten Datensatz zu beginnen (Stuart et al. 2025), wurde zunächst ein Probesample erstellt und von der Bedarfsträgerin freigegeben, bevor die übrigen Paare entstanden.

Die Qualitätssicherung erfolgte in einem dreistufigen Korrektorat (Abbildung 5): einem ersten menschlichen Korrekturdurchgang, einer automatisierten Prüfung mit maschineller Bereinigung der gefundenen Fehler und einem abschließenden menschlichen Durchgang. Da im Projekt nur ein Linguist tätig war, wurde das Vier-Augen-Prinzip durch zeitlich versetzte Durchgänge mit jeweils mehreren Wochen Abstand angenähert.

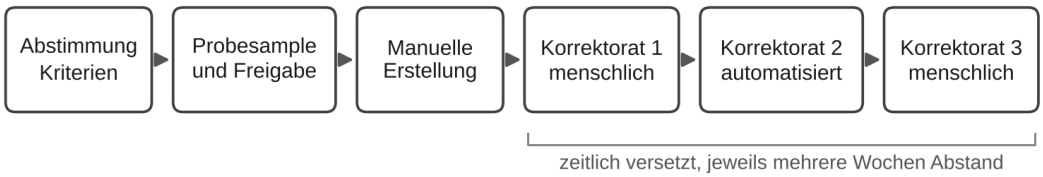

Abbildung 5: Prozess der manuellen Gold-Standard-Erstellung mit dreistufigem Korrektorat
Quelle: eigene Darstellung.

5.7 Vergleichsdesign und Bewertungskriterien

Für den Vergleich wurden 300 Paare zufällig aus den 1.000 Gold-Standard-Paaren gezogen. Zu denselben Quellpassagen erzeugte die Software maschinelle Paare. Um vergleichbare Bedingungen herzustellen, erhielt sie dieselben zwei Personas, die auch der manuellen Erstellung zugrunde lagen, jeweils mit einer rund zweiseitigen Beschreibung. So

liegen zu jeder der 300 Passagen eine manuell und eine maschinell erstellte Fassung vor. Die fünf Kriterien und ihre Operationalisierung fasst Tabelle 4 zusammen.

Formale Korrektheit bezeichnet die Maschinenlesbarkeit: Ein Paar gilt als formal korrekt, wenn es ein gültiges JSON-Objekt gemäß dem vorgegebenen Schema ist, also etwa keine fehlenden Trennzeichen oder nicht geschlossenen Klammern enthält. Geprüft wurde mit einem Parser und einer Schema-Validierung sowie zusätzlich durch ein Sprachmodell, und zwar für alle 1.000 manuell erstellten und alle 5.000 maschinell erzeugten Paare. Die inhaltliche Richtigkeit und die Natürlichkeit der Anfragen beurteilten in der Stichprobe zwei Personen: der Linguist sowie ein Projektmanager mit Erfahrung in KI-Projekten, der weder Linguist noch Experte für behördliche Fachsprache ist. Die Beurteilung erfolgte nicht verblindet; beiden war bekannt, ob ein Paar manuell oder maschinell erstellt worden war. Da die Natürlichkeit keine Ja-Nein-Entscheidung ist, sondern graduelle und teils subjektive Urteile erfordert, wurde sie qualitativ ausgewertet. Für die Effizienz wurde der Zeitaufwand beider Pfade erfasst, ergänzt um eine qualitative Betrachtung der Fixaufwände. Den Einfluss der Personas untersuchten mehrere Testläufe ohne Personas, mit kurz gefassten und mit ausführlich beschriebenen Personas, deren Ergebnisse jeweils mit dem Gold-Standard verglichen wurden.

Tabelle 4: Bewertungskriterien und Operationalisierung

Kriterium	Operationalisierung	Datenbasis	Auswertung
Formale Korrektheit	Gültiges JSON gemäß Schema; Prüfung per Parser und Schema-Validierung, zusätzlich per Sprachmodell	Alle 1.000 manuellen und alle 5.000 maschinellen Paare	Quantitativ
Inhaltliche Richtigkeit	Manuelle Prüfung, ob die wörtliche Antwort die Anfragen beantwortet	300 zufällig gezogene Paare und maschinelle Paare zu denselben Passagen	Quantitativ und qualitativ
Effizienz	Arbeits- und Durchlaufzeit im Vergleich zur Generierungszeit; Fixaufwände für Software und Personas	Erstellung von 1.000 Paaren	Quantitativ und qualitativ
Personas und Vorwissen	Testläufe ohne Personas, mit kurz und mit ausführlich beschriebenen Personas	Mehrere Testläufe, verglichen mit dem Gold-Standard	Qualitativ
Natürlichkeit	Beurteilung, ob Stil und Inhalt der Anfragen zur jeweiligen Persona passen	Stichprobe wie oben	Qualitativ

Quelle: eigene Darstellung.

6 Softwarearchitektur und Einsatzszenarien

6.1 Modularer Aufbau

Die Software ist modular aus vier Komponenten aufgebaut (Abbildung 6). Über eine Web-Oberfläche werden die Dokumente hochgeladen, Personas, Prompt-Vorlagen und Modellparameter konfiguriert und die erzeugten Q&A-Paare exportiert. Ein Backend mit Datenbank verwaltet Konfigurationen und Ergebnisse. Die KI-Komponente extrahiert den Text aus den Dokumenten, steuert die Generierung und prüft die Ausgabe gegen das JSON-Schema. Das Sprachmodell selbst läuft auf einem lokalen Inferenzserver. Durch diese Trennung lassen sich einzelne Komponenten weiterentwickeln oder ersetzen, ohne die übrigen anzupassen.

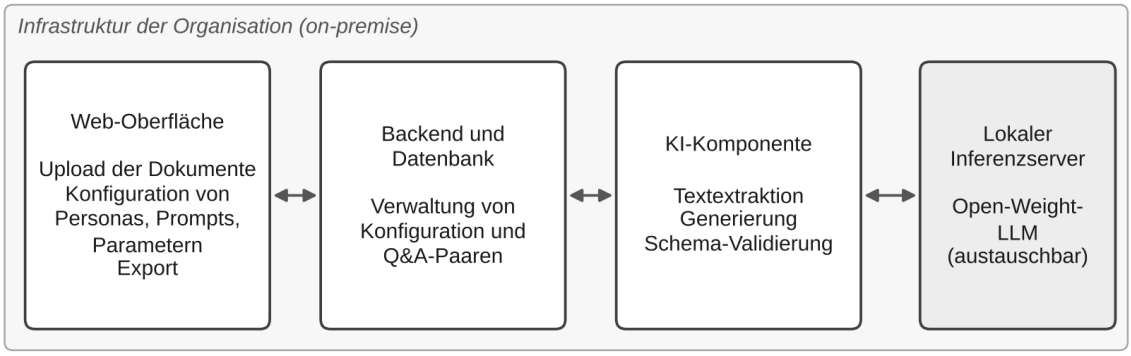

Abbildung 6: Modulare Architektur der Software (vereinfacht)
Quelle: eigene Darstellung.

6.2 Modellunabhängigkeit und On-Premise-Betrieb

Die KI-Komponente spricht das Sprachmodell über eine standardisierte Schnittstelle des Inferenzservers an. Dadurch ist die Software modellunabhängig: Ein leistungsfähigeres Open-Weight-Modell kann eingesetzt werden, ohne die Software zu verändern. Das ist auch deshalb relevant, weil sich die in Abschnitt 7.5 beschriebenen Schwächen bei der Natürlichkeit informeller Anfragen voraussichtlich mit künftigen Modellgenerationen verringern lassen.

Alle Komponenten werden on-premise auf der Infrastruktur der Organisation betrieben, sodass weder Dokumente noch erzeugte Daten die Organisation verlassen. Gerade bei internen Verwaltungsvorschriften ist das eine Grundvoraussetzung. Zugleich vermeidet dieser Betrieb die Abhängigkeit von externen Anbietern. Wie schwierig die Anforderungen

der Datenschutz-Grundverordnung beim Einsatz von Sprachmodellen allgemein zu erfüllen sind, zeigen Feretzakis et al. (2025): Zentrale Bestimmungen wie das Recht auf Löschung, das Auskunftsrecht, das Recht auf Berichtigung und die Beschränkungen automatisierter Entscheidungen werden durch die intransparente und verteilte Natur von Sprachmodellen herausgefordert.

6.3 Einsatz als Testwerkzeug für RAG-Systeme

Über die Erzeugung von Trainingsdaten hinaus ist die Software ein Testwerkzeug für RAG-Systeme. Da zu jeder Anfrage bekannt ist, welche Passage gefunden werden muss, bildet die erzeugte Ground Truth eine Testbasis für das Retrieval und für das Gesamtsystem. Damit lässt sich vor einer Änderung prüfen, wie sich der Austausch einer einzelnen Komponente auswirkt, etwa des Embedding-Modells, der Segmentierung, des Re-Rankers, des Sprachmodells oder eines Prompts (Abbildung 7). Die bisherige und die geänderte Variante werden mit denselben Anfragen geprüft und anhand von Kennzahlen wie Recall@k und MRR (Abschnitt 2.4) sowie einer Bewertung der Antworten verglichen. Referenzfreie Verfahren wie RAGAs (Es et al. 2024) lassen sich ergänzend einsetzen.

Barnett et al. (2024) folgern aus ihren Fallstudien, dass die Validierung eines RAG-Systems erst im Betrieb möglich ist. Eine domänenspezifische Ground Truth ersetzt die Beobachtung im Betrieb nicht, erlaubt aber, zumindest das Retrieval bereits vorab systematisch und reproduzierbar zu prüfen. Weil die Personas auf die tatsächlichen Nutzergruppen zugeschnitten werden und die Erzeugung nach der Einrichtung mit geringem Aufwand wiederholt werden kann, entstehen anpassbare Testdaten hoher Qualität zu vergleichsweise geringen Kosten, auch für neue oder geänderte Dokumente.

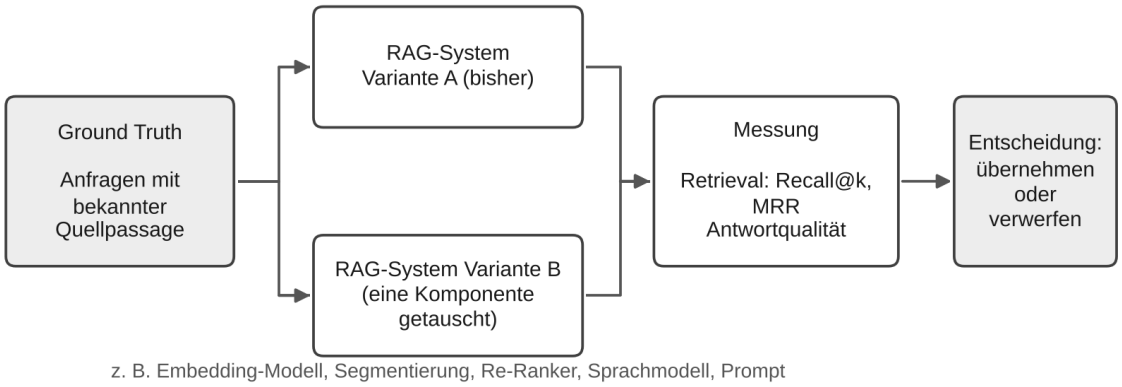

Abbildung 7: Einsatz der generierten Ground Truth als Regressionstest beim Komponententausch in RAG-Systemen
Quelle: eigene Darstellung.

7 Ergebnisse

7.1 Formale Korrektheit

Von den 1.000 manuell erstellten Paaren erfüllten 18,2 Prozent die formalen Vorgaben nicht. Die Fehler waren durchweg syntaktischer Art, etwa fehlende Beistriche oder geöffnete, aber nicht geschlossene Klammern, und machten die betroffenen Paare für eine automatische Weiterverarbeitung unbrauchbar, bis sie im automatisierten Korrekturschritt bereinigt wurden. Bei den 5.000 maschinell erzeugten Paaren lag der Anteil bei 0 Prozent. Das ist auch eine Folge des Verfahrens: Da die Schema-Validierung den letzten Schritt der Erstellung bildet, gelangen formal fehlerhafte Ausgaben gar nicht erst in den Datensatz. Die manuelle Erstellung von Daten in einem streng maschinenlesbaren Format erweist sich damit als fehleranfällig, während die automatisierte Erstellung dieses Kriterium vollständig erfüllt.

7.2 Inhaltliche Richtigkeit und Quellentreue

In der Stichprobe von 300 Paaren wurde weder bei den manuell noch bei den maschinell erstellten Paaren eine vollständig falsche Antwort gefunden. Dass die Antwort stets eine wörtlich übernommene Passage ist und die Maschine die Anfragen ausgehend von dieser Antwort formuliert (Abschnitt 5.5), schließt grobe inhaltliche Abweichungen weitgehend aus. Bei den maschinell erzeugten Paaren passte jedoch vereinzelt die umgangssprachliche Anfrage weniger gut zur Antwort als die formelle. Diese Abweichungen betreffen die

Formulierung der Anfrage, nicht die Antwort, und werden in Abschnitt 7.5 behandelt. Eine feinere Abstufung der inhaltlichen Passung wurde nicht erhoben.

7.3 Effizienz der Erstellung

Für 1.000 Paare mit je zwei Anfragen und einer Antwort benötigte der Linguist einige hundert Arbeitsstunden. Da sich eine solche Tätigkeit nicht ohne Unterbrechung mit gleichbleibender Konzentration ausführen lässt und die Erstellung in zwei Durchgängen erfolgte, betrug die gesamte Durchlaufzeit fast drei Monate. Die Software erzeugte dieselbe Menge in rund 90 Minuten.

Dieser Vergleich der reinen Erstellungszeit greift allerdings zu kurz. Auf maschineller Seite fallen Fixaufwände an, die unabhängig von der Zahl der erzeugten Paare entstehen: die Entwicklung der Software, ihr Betrieb auf eigenen Servern, Wartung und Updates sowie der fachliche Aufwand, Personas zu erstellen, zu pflegen und zu erweitern. Für einen einzelnen Durchlauf verringert sich der Vorteil dadurch deutlich. Da der variable Aufwand je Paar jedoch sehr gering ist, wächst der Vorteil mit jedem weiteren Durchlauf, etwa für neue Dokumente oder neue Nutzergruppen (Abbildung 8).

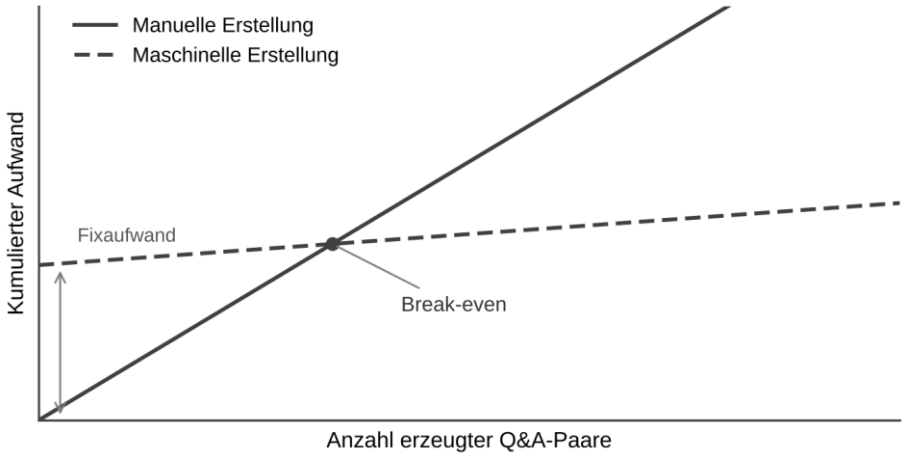

Abbildung 8: Kumulierter Aufwand manueller und maschineller Erstellung (schematisch)
Quelle: eigene Darstellung; schematisch, ohne Maßstab.

7.4 Einfluss von Persona-Definitionen und Domänenwissen

Der Linguist verfügte durch seine Erfahrung, auch mit behördlicher Kommunikation, über einen deutlichen Vorsprung. Ein Sprachmodell ohne weitere Anpassung kann mit der Sprache der Verwaltungsvorschriften dagegen nur wenig anfangen. Testläufe zeigten das deutlich: Ohne Personas oder mit nur kurz beschriebenen Personas, etwa der bloßen

Vorgabe einer formellen oder informellen Formulierung, blieb die Qualität klar hinter dem Gold-Standard zurück. Erst mit ausführlich formulierten Personas erreichte die maschinelle Erstellung das Qualitätsniveau des Menschen, mit Ausnahme sehr informeller Anfragen (Abschnitt 7.5). Das Domänenwissen der Fachleute ersetzt sich also nicht von selbst, sondern muss in die Persona-Definitionen übertragen werden.

7.5 Natürlichkeit der Anfragen

Bei der Natürlichkeit der Anfragen zeigte sich der deutlichste Unterschied. Formelle Anfragen und die übrigen Formulierungen passten in der Stichprobe gut zur jeweiligen Persona. Bei den vollständig informellen Anfragen wurde dagegen wiederholt beobachtet, dass das Sprachmodell übertreibt: ein überzeichneter Dialekt, zu knappe Formulierungen oder eine stilistisch unangemessen saloppe Ausdrucksweise, die im behördlichen Umfeld kaum vorkommen dürfte. Reale Bedienstete hätten dieselbe Frage meist etwas sorgfältiger formuliert. Da die Beurteilung graduell und teils eine Frage des Sprachgefühls ist, wurden diese Beobachtungen nicht quantifiziert.

7.6 Zusammenfassung der Ergebnisse

Tabelle 5 fasst die Ergebnisse zusammen. Die maschinelle Erstellung ist formal überlegen, in der Stichprobe ebenso frei von falschen Antworten wie die manuelle und ab einer gewissen Menge deutlich effizienter. Ihre Qualität hängt jedoch wesentlich davon ab, wie sorgfältig die Personas beschrieben sind, und bei sehr informellen Anfragen bleibt sie hinter dem Menschen zurück. Aus der manuellen Erstellung ergaben sich zudem die in Abschnitt 5.2 beschriebenen Auswahlregeln sowie eine organisatorische Erkenntnis: Der erste menschliche Korrekturdurchgang ließe sich an eine weitere Person übertragen, was Kosten und Durchlaufzeit verbessern würde.

Tabelle 5: Zusammenfassender Vergleich manuell und maschinell erstellter Q&A-Paare

Kriterium	Manuell (Gold-Standard)	Maschinell	Einordnung
Formale Korrektheit	18,2 % der 1.000 Paare mit Formatfehlern	0 % von 5.000 Paaren (Schema-Validierung als letzter Schritt)	Vorteil Maschine
Inhaltliche Richtigkeit	Keine vollständig falsche Antwort in der Stichprobe	Keine vollständig falsche Antwort; informelle Anfrage vereinzelt weniger passend	Kein Unterschied bei falschen Antworten; gesichert durch wörtliche Antwort
Effizienz	Einige hundert Arbeitsstunden, fast drei Monate Durchlaufzeit	Rund 90 Minuten für 1.000 Paare, zuzüglich Fixaufwände	Vorteil Maschine, wächst mit der Menge

Kriterium	Manuell (Gold-Standard)	Maschinell	Einordnung
Personas und Vorwissen	Erfahrung des Linguisten als Vorsprung	Erst mit ausführlichen Personas auf menschlichem Niveau	Expertenwissen muss in Personas einfließen
Natürlichkeit	Referenzniveau	Informelle Anfragen wiederholt übertrieben	Vorteil Mensch

Quelle: eigene Darstellung.

8 Diskussion

8.1 Interpretation der Ergebnisse

Die Ergebnisse beantworten die drei Forschungsfragen wie folgt. Zu FF1: Persona-gesteuerte Generierung mit wörtlich übernommenen Antworten erzeugt aus deutschsprachigen Verwaltungsvorschriften Q&A-Paare, die unterschiedliche Formulierungen abbilden, eindeutig an ihre Quelle gebunden und durchgängig maschinenlesbar sind. Zu FF2: Die maschinelle Erstellung ist der manuellen formal überlegen, in der Stichprobe ebenso frei von falschen Antworten und ab einer gewissen Menge deutlich effizienter; bei sehr informellen Anfragen bleibt sie hinter dem Menschen zurück. Zu FF3: Die Personas sind der entscheidende Hebel für die Qualität. Sie steuern nicht nur den Sprachstil, sondern auch, welche Aspekte einer Regelung für eine Nutzergruppe relevant sind. Ihre Wirkung entfalten sie allerdings erst, wenn sie ausführlich beschrieben sind.

Der Mensch wird durch die Maschine damit nicht ersetzt, sondern verlagert seine Rolle. Sein Wissen über Sprache und Nutzergruppen wird nicht mehr in jedes einzelne Paar eingebracht, sondern einmalig in die Persona-Definitionen, und dort wirkt es auf alle erzeugten Paare.

8.2 Implikationen für Training und Evaluation

Für das Training eines Retrieval-Modells liefert jedes Paar mehrere positive Anfrage-Passage-Paare, die sich in Formulierung und Perspektive unterscheiden. Das passt zum kontrastiven Training dichter Retriever (Karpukhin et al. 2020) und zur Domänenadaption mit synthetischen Anfragen (Wang et al. 2022). Die in Abschnitt 3.1 beschriebenen Varianten gleichartiger Regelungen bieten sich zudem als schwierige Negativbeispiele an: Xiong et al. (2020) zeigen in ihren Experimenten zum Dense Retrieval, dass Negativbeispiele, die nicht repräsentativ für die irrelevanten Dokumente im Einsatz sind, das Training dichter Retriever beeinträchtigen. Die übertriebenen informellen Anfragen sind

dabei kein bloßes Schönheitsproblem: Gehen sie ungefiltert in das Training ein, lernt das Modell eine Verteilung von Anfragen, die so in der Praxis kaum vorkommt. Sie sollten daher vor dem Training geprüft oder gezielt nachbearbeitet werden.

Für die Evaluation erlaubt die Ground Truth, Retrieval und Gesamtsystem reproduzierbar zu prüfen (Abschnitt 6.3). Da jede Anfrage einer Persona zugeordnet ist, lassen sich die Ergebnisse zudem nach Nutzergruppen auswerten. So wird sichtbar, ob ein System etwa formelle Anfragen gut, knappe mobile Eingaben aber schlecht bedient.

8.3 Limitationen

Die Ergebnisse unterliegen mehreren Einschränkungen. Erstens wurden inhaltliche Richtigkeit und Natürlichkeit qualitativ und nicht verblindet von zwei Personen beurteilt, von denen eine zugleich den Gold-Standard erstellt hatte. Eine formale Messung der Übereinstimmung zwischen mehreren unabhängigen Annotierenden, wie sie für die Verlässlichkeit annotierter Daten maßgeblich ist (Artstein und Poesio 2008), fand nicht statt. Zweitens wurde der Gold-Standard von einer einzigen Person erstellt; das Vier-Augen-Prinzip wurde nur durch zeitlich versetzte Durchgänge angenähert. Drittens begünstigt das Kriterium der formalen Korrektheit strukturell die Maschine, da ein Mensch, der JSON von Hand schreibt, naturgemäß fehleranfälliger ist und die Maschine fehlerhafte Ausgaben bereits während der Erstellung aussortiert. Der Befund bleibt dennoch praktisch relevant, weil genau diese Fehler die Weiterverarbeitung verhindern. Viertens stellt die wörtliche Antwort sicher, dass sie aus der Quelle stammt, nicht aber, dass sie die Anfrage in jedem Fall vollständig beantwortet. Fünftens wurde im Projekt kein Retrieval-Modell mit den Daten trainiert, sodass ihre Wirkung auf die Retrieval-Qualität noch nicht gemessen ist. Schließlich beziehen sich die Ergebnisse auf einen Dokumentenbestand und den aktuellen Stand eines Sprachmodells.

8.4 Übertragbarkeit

Der Ansatz ist nicht auf Verwaltungsvorschriften beschränkt. Er eignet sich für jeden Dokumentenbestand, dessen Inhalte sich durch wörtlich zitierbare Passagen beantworten lassen und dessen Nutzergruppen sich beschreiben lassen, etwa für andere Verwaltungen, Rechtstexte, technische Regelwerke oder interne Wissensdatenbanken von Unternehmen. Voraussetzungen sind Fachleute, die die Personas definieren, und eine Infrastruktur, auf der ein Open-Weight-Modell betrieben werden kann.

9 Ausblick und weiterführende Forschung

Aus den Ergebnissen und Limitationen ergeben sich mehrere Ansatzpunkte für weiterführende Forschung (Tabelle 6). Vordringlich ist der Nachweis der Wirkung auf das Retrieval: Ein mit den Daten feinjustiertes Embedding-Modell sollte gegenüber dem Ausgangsmodell sowie gegenüber lexikalischer und hybrider Suche anhand von Recall@k, MRR und nDCG verglichen werden. Die Beurteilung der Natürlichkeit sollte mit mehreren unabhängigen Annotierenden und einer Messung ihrer Übereinstimmung wiederholt werden. Um die menschliche Prüfung zu skalieren, bietet sich der Einsatz von Sprachmodellen als Bewertungsinstanz an. Eine solche Prüfung kann den menschlichen Aufwand verringern, aber nicht beseitigen (Stuart et al. 2025), und müsste daher an menschlichen Urteilen kalibriert werden. Inzwischen stehen zudem spezialisierte Entscheidungsmodelle (Decision Models) zur Verfügung, die statt freien Texts unmittelbar Wahrscheinlichkeiten für Bewertungslabels ausgeben. Li et al. (2026) setzen ein solches Modell als Bewertungsinstanz ein, die sichere Urteile übernimmt und unsichere an ein schlussfolgerndes Sprachmodell weitergibt, und erreichen in ihren Tests eine mit diesem vergleichbare Genauigkeit bei deutlich geringeren Kosten, mit Unterschieden je nach Aufgabe. Zum Zeitpunkt der Projektdurchführung standen solche Modelle noch nicht frei zur Verfügung. Sie könnten die Bewertung vieler Passagen nach mehreren Kriterien erheblich beschleunigen; für interne Dokumente müssten sie allerdings ebenfalls on-premise betrieben werden können. Schließlich verdienen die Personas selbst systematische Untersuchung: Welcher Detaillierungsgrad ist nötig, und wie lassen sich Persona-Beschreibungen methodisch entwickeln und prüfen?

Tabelle 6: Agenda für weiterführende Forschung

Forschungsfrage	Möglicher Ansatz	Bezug
Verbessern die Daten das Retrieval messbar?	Fine-Tuning eines Embedding-Modells; Vergleich mit Ausgangsmodell, BM25 und hybrider Suche anhand von Recall@k, MRR und nDCG	Abschnitte 2.4, 8.3
Wie verlässlich ist die Beurteilung der Natürlichkeit?	Mehrere unabhängige Annotierende, Messung der Übereinstimmung	Abschnitte 7.5, 8.3
Lässt sich die Prüfung automatisieren?	Sprachmodell oder Entscheidungsmodell als Bewertungsinstanz, kalibriert an menschlichen Urteilen	Abschnitte 4.3, 9
Helfen Regelungsvarianten als Negativbeispiele?	Varianten gleichartiger Regelungen als schwierige Negativbeispiele im kontrastiven Training	Abschnitte 3.1, 8.2
Wie detailliert müssen Personas sein?	Systematische Variation des Detaillierungsgrads; Leitfaden für Persona-Beschreibungen	Abschnitt 7.4
Verbessern neuere Modelle informelle Anfragen?	Wiederholung der Evaluation mit neueren Open-Weight-Modellen	Abschnitte 6.2, 7.5

Quelle: eigene Darstellung.

10 Fazit

Das Retrieval entscheidet maßgeblich darüber, ob ein RAG-System korrekt und belegbar antwortet. Für deutschsprachige Verwaltungsvorschriften fehlen jedoch die Daten, um Retrieval-Modelle anzupassen und zu prüfen. Der vorliegende Bericht zeigt, dass sich solche Daten mit einem Open-Weight-Sprachmodell vollständig auf eigener Infrastruktur erzeugen lassen: Zu jeder Quellpassage entstehen Anfragen aus der Perspektive frei definierbarer Personas, die alle auf dieselbe, wörtlich übernommene Antwort verweisen. Im Vergleich mit einem manuell erstellten Gold-Standard ist die maschinelle Erstellung formal fehlerfrei, in der Stichprobe ebenso ohne falsche Antworten und um ein Vielfaches schneller, bleibt aber bei sehr informellen Anfragen hinter dem Menschen zurück.

Die wichtigste Lehre ist daher eine Arbeitsteilung zwischen Mensch und Maschine:

– **Der Mensch** beschreibt die Personas, verfeinert sie anhand der Ergebnisse und übernimmt die bekannten Schwachstellen, insbesondere die Prüfung und gegebenenfalls Formulierung heikler informeller Anfragen.
– **Die Maschine** übernimmt den Großteil der Datenerstellung, garantiert das korrekte Format und skaliert auf große Dokumentenbestände und wiederholte Durchläufe.

Über das Training hinaus entsteht so ein anpassbares und kostengünstiges Testwerkzeug, mit dem sich jede Änderung an einem RAG-System anhand einer domänenspezifischen Ground Truth überprüfen lässt, bevor sie im Betrieb wirksam wird.

Literaturverzeichnis

Artstein, Ron, und Massimo Poesio. 2008. „Inter-Coder Agreement for Computational Linguistics." *Computational Linguistics* 34 (4): 555–596. https://doi.org/10.1162/coli.07-034-R2. Frei zugänglich: https://aclanthology.org/J08-4004/.

Barnett, Scott, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly und Mohamed Abdelrazek. 2024. „Seven Failure Points When Engineering a Retrieval Augmented Generation System." In *3rd International Conference on AI Engineering – Software Engineering for AI*. Lissabon. https://arxiv.org/abs/2401.05856.

Bonifacio, Luiz, Hugo Abonizio, Marzieh Fadaee und Rodrigo Nogueira. 2022. „InPars: Data Augmentation for Information Retrieval using Large Language Models." arXiv:2202.05144. https://arxiv.org/abs/2202.05144.

Chen, Jianlv, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian und Zheng Liu. 2024. „M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation." arXiv:2402.03216. https://arxiv.org/abs/2402.03216.

Cormack, Gordon V., Charles L. A. Clarke und Stefan Büttcher. 2009. „Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods." In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 758–759. https://doi.org/10.1145/1571941.1572114. Frei zugänglich: https://cormack.uwaterloo.ca/cormacksigir09-rrf.pdf.

Es, Shahul, Jithin James, Luis Espinosa Anke und Steven Schockaert. 2024. „RAGAs: Automated Evaluation of Retrieval Augmented Generation." In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16.

Feretzakis, Georgios, Evangelia Vagena, Konstantinos Kalodanis, Paraskevi Peristera, Dimitris Kalles und Athanasios Anastasiou. 2025. „GDPR and Large Language Models: Technical and Legal Obstacles." *Future Internet* 17 (4): 151. https://doi.org/10.3390/fi17040151.

Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang und Haofen Wang. 2023. „Retrieval-Augmented Generation for Large Language Models: A Survey." arXiv:2312.10997. https://arxiv.org/abs/2312.10997.

Ge, Tao, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi und Dong Yu. 2024. „Scaling Synthetic Data Creation with 1,000,000,000 Personas." arXiv:2406.20094. https://arxiv.org/abs/2406.20094.

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen und Wen-tau Yih. 2020. „Dense Passage Retrieval for Open-Domain Question Answering." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550.

Koehn, Philipp, und Kevin Knight. 2003. „Empirical Methods for Compound Splitting." In *10th Conference of the European Chapter of the Association for Computational Linguistics*, 187–194. Budapest. https://aclanthology.org/E03-1076/.

Laskar, Md Tahmid Rahman, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan et al. 2024. „A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13785–13816. https://doi.org/10.18653/v1/2024.emnlp-main.764.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. 2020. „Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." In *Advances in Neural Information Processing Systems 33*, 9459–9474.

https://papers.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Li, Yubo, Yidi Miao, Ramayya Krishnan und Rema Padman. 2026. „JEV-as-a-Judge: Accept When Confident, Escalate When Unsure.“ arXiv:2609.26550. https://arxiv.org/abs/2609.26550.

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni und Percy Liang. 2024. „Lost in the Middle: How Language Models Use Long Contexts.“ *Transactions of the Association for Computational Linguistics* 12: 157–173. https://doi.org/10.1162/tacl_a_00638.

Möller, Timo, Julian Risch und Malte Pietsch. 2021. „GermanQuAD and GermanDPR: Improving Non-English Question Answering and Passage Retrieval.“ In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, 42–50. https://doi.org/10.18653/v1/2021.mrqa-1.4.

Muennighoff, Niklas, Nouamane Tazi, Loic Magne und Nils Reimers. 2023. „MTEB: Massive Text Embedding Benchmark.“ In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2014–2037. https://doi.org/10.18653/v1/2023.eacl-main.148.

Nogueira, Rodrigo, und Kyunghyun Cho. 2019. „Passage Re-ranking with BERT.“ arXiv:1901.04085. https://arxiv.org/abs/1901.04085.

Nogueira, Rodrigo, Wei Yang, Jimmy Lin und Kyunghyun Cho. 2019. „Document Expansion by Query Prediction.“ arXiv:1904.08375. https://arxiv.org/abs/1904.08375.

Penha, Gustavo, Arthur Câmara und Claudia Hauff. 2022. „Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators.“ In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022*, Lecture Notes in Computer Science 13185, 397–412. https://doi.org/10.1007/978-3-030-99736-6_27. Frei zugänglich: https://arxiv.org/abs/2111.13057.

Robertson, Stephen, und Hugo Zaragoza. 2009. „The Probabilistic Relevance Framework: BM25 and Beyond.“ *Foundations and Trends in Information Retrieval* 3 (4): 333–389. https://doi.org/10.1561/1500000019. Frei zugänglich: https://staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf.

Stuart, Samantha, Ivan Cui, Philippe Duplessis-Guindon und Rahul Jani. 2025. „Ground Truth Generation and Review Best Practices for Evaluating Generative AI Question-Answering with FMEval.“ *AWS Machine Learning Blog*, 5. März 2025. https://aws.amazon.com/blogs/machine-learning/ground-truth-generation-and-review-best-practices-for-evaluating-generative-ai-question-answering-with-fmeval/.

Thakur, Nandan, Nils Reimers, Andreas Rücklé, Abhishek Srivastava und Iryna Gurevych. 2021. „BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.“ arXiv:2104.08663. https://arxiv.org/abs/2104.08663.

Wang, Kexin, Nandan Thakur, Nils Reimers und Iryna Gurevych. 2022. „GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval.“ In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2345–2360. https://doi.org/10.18653/v1/2022.naacl-main.168.

Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder und Furu Wei. 2024. „Improving Text Embeddings with Large Language Models.“ In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11897–11916. https://doi.org/10.18653/v1/2024.acl-long.642.

Xiong, Lee, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed und Arnold Overwijk. 2020. „Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval.“ arXiv:2007.00808. https://arxiv.org/abs/2007.00808.

Yu, Yue, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen und Chao Zhang. 2023. „Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias.“ In *37th Conference on Neural Information Processing Systems (NeurIPS 2023), Track on Datasets and Benchmarks*. https://arxiv.org/abs/2306.15895.