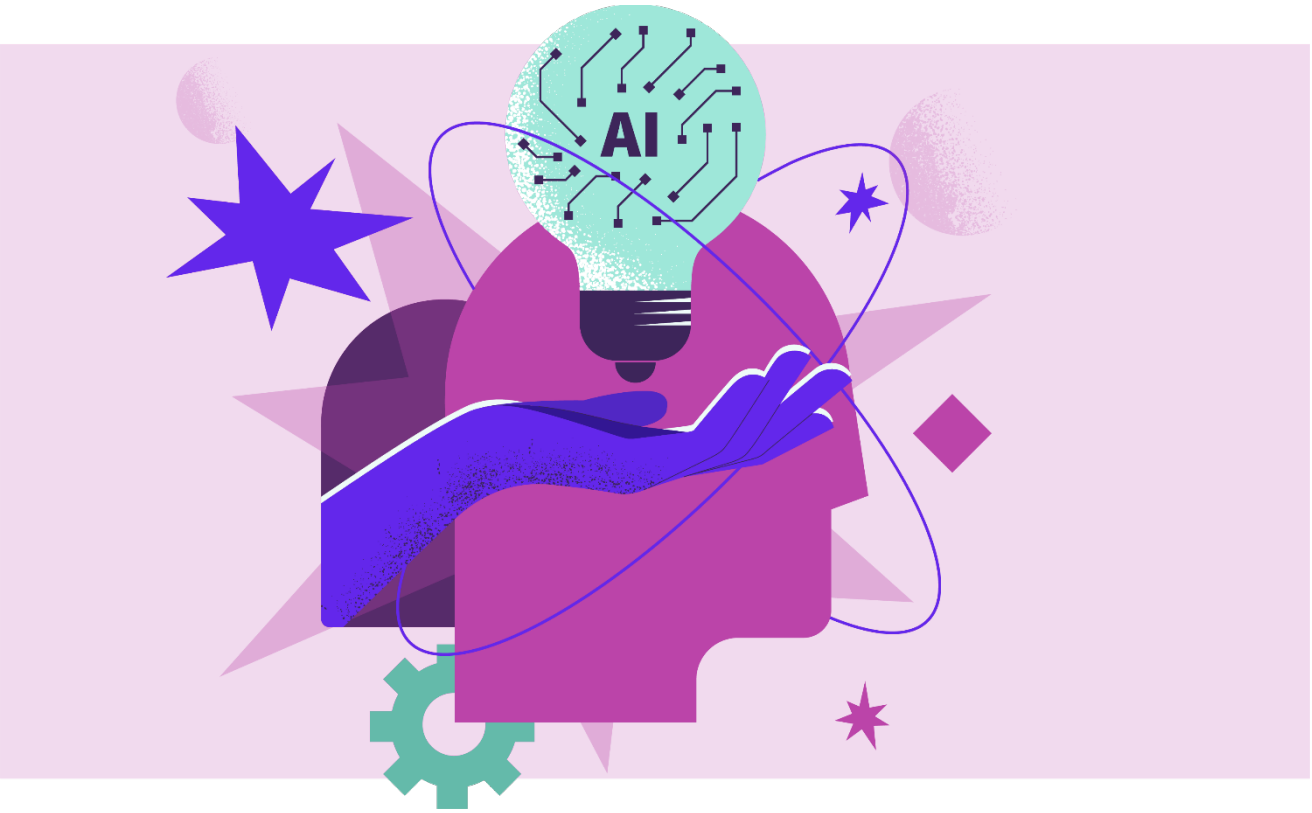
EBCONT proconsult GmbH
Millennium Tower, Handelskai 94-96, A-1200 Wien
T: +43 1 2700 202 • F: +43 2772 512 69 9
office@ebcont.com • http://www.ebcont.com
UID ATU68126645 • HG St. Pölten FN 399976a

ebcont
proconsult



KI-Gutachtenerstellung: Innovation mit Graph- und Vektordatenbanken

Inhalt

Kontakt:
Mag. Gregor Sieber
Prokurist/Executive Vice President
EBCONT proconsult GmbH

Die hier beschriebenen Inhalte, Konzepte und Ansätze sind ausschließlich für die Nutzung des Auftraggebers und die Präsentation auf der IÖB Innovationsplattform gedacht. Eine Verwendung durch Dritte oder zu anderen Zwecken darf nur mit Genehmigung des Erstellers erfolgen.

EBCONT proconsult GmbH
Millennium Tower, Handelskai 94-96, A-1200 Wien
T: +43 1 2700 202 • F: +43 2772 512 69 9
office@ebcont.com • http://www.ebcont.com
UID ATU68126645 • HG St. Pölten FN 399976a

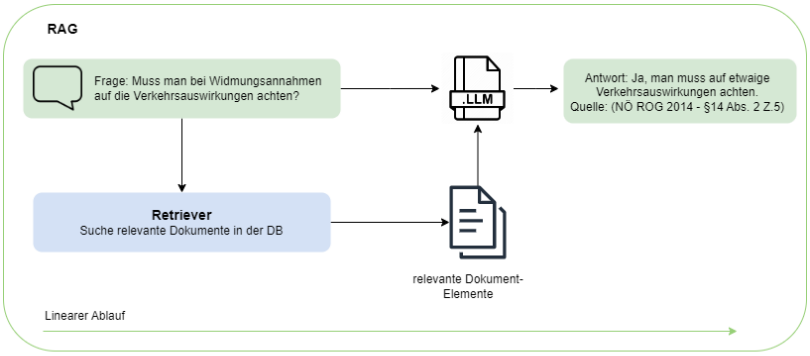
1 BESCHREIBUNG DES LÖSUNGSANSATZES

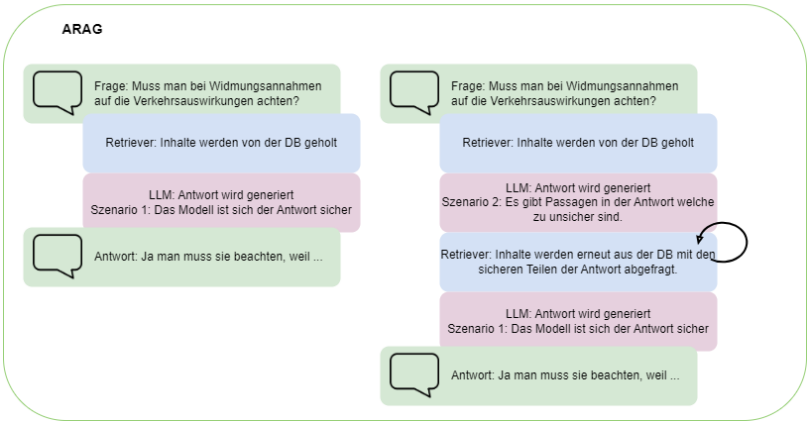
Unser integrierter Ansatz für die KI-basierte Erstellung raumordnungsfachlicher Gutachten kombiniert die generative Logik des Active Retrieval Generation (ARAG) mit semantischer Suche in Vektor- und Graphendatenbanken.

Zunächst werden die vorhandenen Daten von einem Dateiserver oder aus einer Datenbank mit einer OCR-Pipeline geladen. Im Anschluss erfolgt die Aufbereitung in kleine semantische Bausteine mittels Chunking mit vordefinierten Logiken, die auf Prüfinhalten, Widmungsänderungstypen und anderen charakteristischen Dokumenteneigenschaften basieren. Zusätzlich werden Metadaten und Beziehungen extrahiert. Dadurch wird unter Bewahrung des Zusammenhangs von Dateien & Dokumenten eine strukturierte Aufteilung der Daten erreicht.

Für die Beziehungsdarstellung zwischen den Datenpunkten nutzen wir ein Graphennetzwerk. Auch Dokumente ohne offensichtliche Beziehungen werden durch semantische Ähnlichkeit berücksichtigt. Dadurch werden relevante Zusammenhänge erfasst und verschiedene Versionen können erfolgreich abgebildet werden. Die dynamische Datenhaltung kann durch eine Schnittstellenanbindung und regelmäßige Delta Loads oder direkte Datenabfrage integriert werden.

Nach der strukturierten Datenaufbereitung erfolgt die Integration des Large Language Models (LLM) und der Retriever-Logik. Eine Ähnlichkeits- und Metadatensuche ermöglicht das Filtern früherer Einreichunterlagen, die der aktuellen Anfrage ähneln, und lassen bereits erstellte Inhalte einfließen. Durch die Verwendung dieser Datenpunkte in Verbindung mit einer vom LLM generierten Graph Query werden dem LLM nur für den Kontext relevante Inhalte zugespielt, um neue Gutachtentextbausteine zu generieren. Active Retrieval Generation (ARAG) und Few-Shot Prompting ermöglichen die effektive Nutzung des Sprachmodells ohne aufwendigen Feinabstimmungsprozess, insbesondere in einem Proof-of-Concept (PoC). Flexible Hosting-Optionen für das LLM umfassen den Azure OpenAI Service oder Open-Source LLMs wie Mistral oder LLama, die auch on premise gehostet werden können.





In einem Ausbauschritt kann auf einen multimodalen Ansatz erweitert werden, um zusätzlich Bild- und Objekterkennung einzuführen und die Daten weiter anzureichern. Die Umsetzung der Logik erfolgt in einer Test-Webanwendung, die einen Upload-Bereich für Unterlagen zur Verfügung stellt und ein Chat-Interface für die Generierung gewünschter Textelemente bietet. In Zusammenarbeit mit unserem UI/UX-Team kann für ein optimales Benutzererlebnis im Folgeprojekt ein UI/UX-Konzept erarbeitet werden.

Für das geplante Vorhaben können wir Erfahrungen aus erfolgreichen KI Projekten im juristischen Bereich (BMJ, Projekt “Entscheidungsanonymisierung”), in der UX bei der Bearbeitung von Inhalten die sich auf Gesetzestexte beziehen (MANZ Verlag, Bundesanzeiger), im Bereich fachlicher Textgenerierung mittels KI und Nutzung von LLMs auf Unternehmensdaten, sowie Expertise als Implementierungspartner der vorgeschlagenen Technologien einbringen.

2 MEHRWERT

Unsere Methode für die KI-gestützte Gutachtenerstellung bringt deutlichen Mehrwert im Vergleich zu herkömmlichen LLM-/RAG-Ansätzen. Der Einsatz von ARAG in Kombination mit Vektor- und Graphdatenbank eliminiert die Notwendigkeit für aufwändiges Modelltraining. Dabei werden Abhängigkeiten zwischen den Inhalten transparent dargestellt. Die generierten Antworten gehen über die bloße Momentaufnahme der aktuellen Useranfrage hinaus, denn sie beinhalten mögliche Abhängigkeiten zu Folgethemen. Dadurch können dier Ergebnisse z.B. in Bezug zur jeweils gültigen Rechtsfassung gesetzt und Gesetzesnovellen berücksichtigt werden.

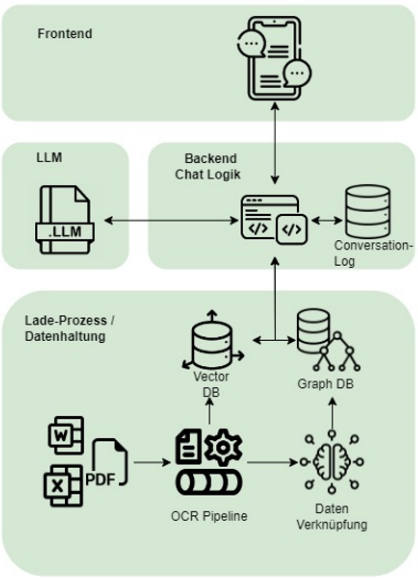
Die Graphdatenbank ermöglicht eine umfassende Darstellung von Beziehungen zwischen verschiedenen Dokumententypen wie Gutachten, Prüfprotokollen und Gesetzestexten. Abhängigkeiten sowie Metadaten (z.B. Prüfinhalte, Prüfer und Widmungsänderungstypen) können erfasst werden. Diese Struktur erlaubt eine effektive Berücksichtigung der zeitlichen Komponente gesetzlicher Entwicklungen und bietet tiefere Einblicke in Datenabhängigkeiten im Vergleich zu herkömmlichen Modellen (explainable AI). Die Vektordatenbank optimiert die Informationssuche und erleichtert den Zugriff auf relevante Einreichunterlagen.

Die Kombination beider Datenmodelle zu einem ARAG-Ansatz ermöglicht die Entwicklung eines zukunftsweisenden KI-Systems, das nicht nur die aktuelle User-Frage isoliert betrachtet, sondern auch bereits eingefügte Inhalte und deren Abhängigkeiten berücksichtigt.

3 ARCHITEKTUR UND VORGEHENSWEISE

3.1 Tools & Architektur

Die Architektur der Gutachter-KI nutzt zwei verschiedene Datenbanktypen. Zum einen wird auf einen Suchindex zurückgegriffen, welcher eine effiziente Erstellung, Speicherung und Suche von Vektoreinbettungen im großen Maßstab ermöglicht. Zum anderen werden die Zusammenhänge zwischen den Daten in einer Graphdatenbank, etwa mit ArangoDB, abgebildet. ArangoDB wird aufgrund seiner umfassenden Funktionalität für Graphdatenbanken empfohlen und bietet skalierbare Lösungen für diverse Anwendungsfälle.



Für den Aufbau der OCR-Pipeline verwenden wir z.B. ABBYY, einen OCR-Reader mit zusätzlichen AI-Features zur effizienten Verarbeitung der Daten. Alle diese Komponenten können je nach den Präferenzen des Auftraggebers in der Cloud oder On Premise bereitgestellt werden.

3.2 Projektplan

Datengetriebene Projekte bei EBCONT folgen einem agilen Ansatz. Nach einer initialen Kickoff- und Analysephase erfolgt das Setup der Infrastruktur. In der Umsetzung wird der Auftraggeber frühzeitig eingebunden, um durch etwaige Systemevaluierung optimale Ergebnisse zu erzielen. Zusätzlich werden diese Phasen durch regelmäßige Abstimmungen – etwa 2-3 pro Woche – begleitet.

Abbildung: EBCONT Delivery Model – Proof of Value

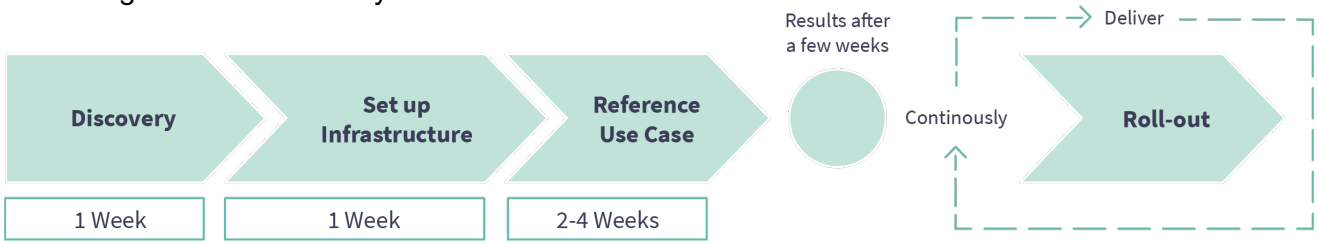


Abbildung: Roadmap-Vorschlag

Roadmap 2024 Epic	Woche 1	Woche 2	Woche 3	Woche 4	Woche 5	Woche 6
Administration, Doku, Tests						
Infrastruktur Setup						
Daten laden und verarbeiten						
LLM Hosting						
Retriever Logik						
WebApp Erstellung & Deployment						