Retrieval-Augmented Generation für den Mittelstand
KI-Innovationswettbewerb – Generative KI für den Mittelstand

Projekt ID: 01MK250104 **Projektstart:** 01.02.2025 **Laufzeit:** 36 Monate

Ergebnis 1.1: Anforderungen an die Basisversion

Publikationslevel	Öffentlich
Zieldatum	M3, 30.04.2025
Abschlussdatum	M3, 30.04.2025
Arbeitspaket	AP1– Anforderungsanalyse
Ergebnis	E1.1
Typ	Report
Status	Final
Version	1.0

Kurzzusammenfassung:
Die Anforderungen an die Basisversion werden zusammengefasst in Tabellenform beschrieben. Darüber hinaus wird ein erster Ausblick auf die Anforderungen für den späteren Plattformprototypen gegeben.

Gefördert durch:

Bundesministerium für Wirtschaft und Klimaschutz

aufgrund eines Beschlusses des Deutschen Bundestages

Dieses Projekt wurde finanziert vom Bundesministerium für Wirtschaft und Klimaschutz (01MK250104).

History

Version	Datum	Änderung	Author
0.1	24.04.2025	Erster Entwurf	Fraunhofer IEM
0.2	25.04.2025	Ergänzung um den Input zu Anforderungen von der USU	Fraunhofer IEM
0.3	28.04.2025	Review des Entwurfs	UPB
0.4	30.04.2025	Review der Prioritäten	Fraunhofer IEM
1.0	30.04.2025	Finale Version abgeschlossen und veröffentlicht	UPB

Zusammenfassung

Die Anforderungen an die Basisversion wurden über Interviews mit Industrie-Experten und projektinterne Workshops erhoben. Sie werden zusammengefasst in Tabellenform beschrieben.
Darüber hinaus wird ein erster Ausblick auf die Anforderungen für den späteren Plattformprototypen gegeben, die unabhängig vom Anwendungsfall bereits früh im Anschluss an die Entwicklung der Basisversion mitberücksichtigt werden sollten.

Abkürzungen

RAG	Retrieval-Augmented Generation
UPB	Paderborn University
SLA	Service Level Agreement

List of Tables

Table of Contents

1 Einleitung

Die Anforderungen an die Basisversion wurden im Rahmen von Interviews mit Industrie-Experten erhoben. Es wurden insgesamt drei Interviews geführt. Darüber hinaus wurden zwei projektinterne Workshops durchgeführt, um ergänzend zu den Interviews weitere Anforderungen zu erheben. Daneben wurden einige Anforderungen aus der Gesamtvorhabenbeschreibung (GVB) des Learn2RAG Projekts extrahiert. Die Anforderungen wurden anschließend im Projekt mit der MoSCoW-Priorisierung eingeordnet. Die 4 verschiedenen Priorisierungen sind in Tabelle 1 aufgeführt. Die Anforderungen werden zusammengefasst in Tabellenform in Kapitel 2 beschrieben.

Tabelle 1: Prioritäten der MoSCoW-Priorisierung.

Priorisierung	Erläuterung
MUST	Unbedingt erforderlich.
SHOULD	Sollte umgesetzt werden, wenn alle MUST-Anforderungen trotzdem erfüllt werden können.
COULD	Kann umgesetzt werden, wenn die Erfüllung von höherwertigen Anforderungen nicht beeinträchtigt wird.
WON'T	Wird diesmal nicht umgesetzt, ist aber für die Zukunft vorgemerkt.

Anforderungen, die im Rahmen der Workshops und Interviews genannt wurden und über den Umfang der Basisversion hinausgehen, wurden in einer separaten Tabelle zusammengefasst (Kapitel 3). Die Anforderungen dienen als Ausblick für die Weiterentwicklung der Basisversion zum Plattformprototypen und sollen – sofern möglich – an geeigneten Stellen mitberücksichtigt werden, um später in die Entwicklung des Plattformprototypen im Anschluss an die Basisversion einzufließen. Im weiteren Projektfortschritt werden die Anforderungen an den Plattformprototypen weiter ergänzt und auf feinerer Granularitätsstufe angegeben, sowie in Muss- und Wunsch-Anforderungen geteilt.

2 Anforderungen an die Basisversion

Tabelle 2: Anforderungen an die Basisversion.

Geschäftliche und strategische Anforderungen		
Anforderung	**Erhebungsart**	**Priorität**
Vendor-Lock-in vermeiden -offene Standards und Self-Hosting	Workshops	MUST
Wachstumsfähigkeit und skalierbares Datenhandling -skalierbar von Pilot-Use-Case (1 Tsd. Dokumente) bis Unternehmensbreite (>10 Mio Dok.) ohne Neuarchitektur	Workshops	COULD
Mehrsprachiger Support und SLA - Deutsch und Englisch - Reaktionszeit < 8 Stunden bei kritischen Incidents	Workshops	WON'T
Recht, Compliance und Governance		
Anforderung	**Erhebungsart**	**Priorität**
EU-AI-Act Konformität	Workshops	MUST
GDPR/DSGVO -z.B. personenbezogene Daten sind erkennbar, Pseudonymisierung und Lösch-Workflows	Workshops/ Interviews	MUST
Datenschutz der Anwender -Die Eingaben der Nutzer sollen geschützt und nicht einsehbar sein	Interviews	MUST
Security und Datenschutz		
Anforderung	**Erhebungsart**	**Priorität**
Data-in-Transit und -at-Rest Verschlüsselung -TLS 1.3, AES-256 oder ähnlich	Workshops	SHOULD
Zugriffsrechtemanagement und Berechtigungskonzept -Der Zugriff auf Dokumente im RAG-System soll konsistent mit den Zugriffsrechten auf Informationen einstellbar sein -Metadaten für Kontext- und Berechtigungsfilter	Interviews	COULD
Betrieb und IT-Service-Management		
Anforderung	**Erhebungsart**	**Priorität**
One-Click-on-Deploy on-prem und Cloud -Docker-Compose/K8s Helm Chart inkl. Default Konfiguration	Workshops	COULD
Self-Service Konnektorverwaltung -Fachanwender können neue Datenquellen konfigurieren ohne Dev-Eingriff	Workshops	COULD
Speicherortverwaltung -verschiedene Speicherorte sollen technisch zugänglich und verwaltbar sein	Interviews	COULD
Funktionale Anforderungen		
Anforderung	**Erhebungsart**	**Priorität**
RAG-Pipeline Die Basisversion muss eine RAG-Pipeline-basierte Lösung implementieren	GVB	MUST
Nachverfolgbarkeit Die Basisversion könnte bereits eine Nachverfolgbarkeit unterstützen (Herkunft von Daten bzw. Datenschnipseln als Teil der Antwort)	Workshops	COULD
Modell-Flexibilität -Umschaltbar zwischen lokalen Open-Weight-Modellen und externen API-Modellen	Workshops	COULD
Mehrsprachige Suche und Antwortgenerierung -mind. Deutsch und Englisch	Workshops	SHOULD
Kontrolle über Quellen -System soll Kontrolle darüber ermöglichen, welche Informationsquellen einbezogen werden (z.B. Ausschluss einzelner Dokumente oder Websuche deaktivieren)	Interviews	MUST
Kontextspezifisches Retrieval -Filtern möglich	Workshops	COULD
Usability und Change Management		
Anforderung	**Erhebungsart**	**Priorität**

Grafisch unterstützter Setup-Assistent - möglichst unter acht Eingaben bis zur lauffähigen Pipeline	Workshops	MUST
Mehrsprachiges User Interface - Deutsch und Englisch - konsistente Terminologie - Übersetzungsfiles extern editierbar	Workshops	COULD
Qualität, Wartbarkeit und Erweitbarkeit		
Anforderung	**Erhebungsart**	**Priorität**
Modularer Aufbau -klare standardisierte API-Schnittstellen pro Schicht (bspw. Ingestion, Embedding, Retrieval, Generation)	Workshops	COULD
Konfiguration-als-Code -gesamte Pipeline in YAML/JSON, versionierbar in Git	Workshops	COULD
Dokumentation in geringen Umfängen (für Basisversion) -Entwicklungs-, Admin- und Benutzerhandbuch, stets synchron mit dem Release	Workshops	MUST

3 Ausblick: Anforderungen an den Plattformprototypen

Tabelle 3: Ausblick Anforderungen an den Plattformprototypen.

Recht, Compliance und Governance	
Anforderung	**Erhebungsart**
Compliance -Prozessseitige Abbildung der Compliance-Anforderungen des Unternehmens	Interviews
Security und Datenschutz	
Anforderung	**Erhebungsart**
Individualisierbares Zugriffsrechtemanagement -Das Zugriffsrechtemanagement des Systems soll individualisierbar sein	Interviews
Server aus der EU/Deutschland -das System soll auf Cloud-Infrastrukturen nur dann zugreifen, wenn sie sich in Deutschland oder der EU befinden	Interviews
Betrieb und IT-Service-Management	
Anforderung	**Erhebungsart**
Schnittstellen -Identifikation und Umsetzung geeigneter Schnittstellen zur Anbindung an Unternehmensinfrastruktur	Interviews
Universaler Aufbau für Anschlussfähigkeit an breite Infrastruktur	Interviews
Wartung und Pflege ohne hohen Aufwand -die Wartung und Pflege soll ohne hohen Aufwand und ohne tiefgehende Expertise möglich sein, z.B. mithilfe einer kurzen Schulung	Interviews
Updates -Aktualität gewährleisten und automatische Updates ingesten	Workshops
Funktionale Anforderungen	
Anforderung	**Erhebungsart**
Datenquellen-Abdeckung -Dateisystem, SharePoint, ERP, E-Mail-Postfächer, SQL/NoSQL, CAD-Programme, Aufgabenspezifische Software z.B. Qualitätsmanagement-Software -Integration und Kombination verschiedener Datenquellen möglich -Unterstützung strukturierter, semi-strukturierter und unstrukturierter Datenquellen	Interviews/ Workshops
Verarbeitung unterschiedlicher Datenformate möglich -PDF, TXT, Daten aus CAD-Programmen, Qualitätsmanagementsoftware etc., gängige Formate wie Office-Excel, Word	Interviews
Feedbackmöglichkeiten -Feedback-Loop zur Bewertung der Antwortqualität und Nützlichkeit	Interviews
Unterstützung verschiedener Interaktionsmöglichkeiten -bspw. Spracheingabe - Dialogfähigkeit / Rückfragen möglich	Interviews/ Workshops
Nachträgliche Anpassung und Eingabe von Daten möglich	Interviews
Überprüfbarkeit -System soll Überprüfung der Antworten durch Anzeige und Öffnung der genutzten Quellen ermöglichen	Interviews
Parameteroptimierung -automatisierte Parameteroptimierung für neue Anwendungsfälle und Datenquellen	Workshops
Kontextspezifisches Retrieval -Query Expansion	Workshops
Usability und Change Management	
Anforderung	**Erhebungsart**
Intuitive Bedienung -Bedienung intuitiv möglich und wenig komplex	Interviews
Qualität, Wartbarkeit und Erweitbarkeit	
Anforderung	**Erhebungsart**
Ausführliche Dokumentation -Entwicklungs-, Admin- und Benutzerhandbuch; stehts synchron mit dem Release in höherem Umfang als bei der Basisversion	Workshops
Antwortqualität -Antworten möglichst richtig und begründet, Belastbarkeit der Antworten nahezu 100%	Interviews