

Leistungsbeschreibung

Beschaffung und Integration eines mandantenfähigen KI-Stacks

Stadt Gütersloh – Urban Stack



Gefördert durch das Bundesprogramm „Modellprojekte Smart Cities“ des Bundesministeriums für Wohnen, Stadtentwicklung und Bauwesen.

Inhalt

1	Ausgangslage	10
1.4	Einleitung.....	11
1.5	Zielsetzung	11
2	Gegenstand der Leistung.....	12
2.4	Abgrenzung der Leistung	12
	Bereits vorhandene Komponenten:	12
	Umfang der Erstimplementierung.....	13
	Nicht Bestandteil der Erstimplementierung	13
2.5	Beschreibung der Plattformbestandteile.....	14
2.6	Einordnung als Gesamtsystem.....	16
3	Lose-Struktur	17
3.4	Übersicht der Lose	17
3.5	Beschreibung der Lose.....	17
3.6	Übersicht der Lose	19
3.4	Beschreibung der Lose.....	19
3.5	Abgrenzung zwischen den Losen.....	23
3.6	Fachliche Konkretisierung Los 2 – Urban KI Hub	23
3.7	Stufenweise Zielarchitektur	24
4	Funktionale Anforderungen.....	25
4.1	KI-Dialogsystem	25
4.2	KI-Agenten-Plattform	33
4.3	Retrieval, RAG und Wissensmanagement	38
	Zielsetzung	38
	Architekturgrundsatz	38
4.4	Erweiterte Recherche (Deep Research).....	42
	Zielsetzung	42
	Architekturgrundsatz	42
4.5	Bildgenerierung	45
	Zielsetzung	45
	Architekturgrundsatz	45
4.6	Prompt Management (Prompt Registry)	46
	Zielsetzung	46
	Architekturgrundsatz	46
4.7	Zeitgesteuerte Aufgaben	49
	Zielsetzung	49

Architekturgrundsatz	49
4.8 Personalisierung.....	51
Zielsetzung	51
Architekturgrundsatz	51
4.9 ML-Services API und Umfang des Loses ML-Services	53
Zielsetzung	53
Architekturgrundsatz	53
4.10 Prompt Governance	56
Zielsetzung	56
Architekturgrundsatz	56
4.11 Tenant Context.....	58
Zielsetzung	58
Architekturgrundsatz	58
4.12 Citation Validator.....	60
Zielsetzung	60
Architekturgrundsatz	60
4.13 Entscheidungslogik & Human Approval	62
Zielsetzung	62
Architekturgrundsatz	62
4.14 Token- und Kostenmonitoring	64
Zielsetzung	64
Architekturgrundsatz	64
4.15 Performance-Monitoring	67
Zielsetzung	67
Architekturgrundsatz	67
4.16 Datenschutz und Administration.....	69
Zielsetzung	69
Architekturgrundsatz	69
5 Nicht-funktionale Anforderungen.....	73
Zielsetzung	73
5.1 Performance.....	73
Zielsetzung	73
Architekturgrundsatz	73
5.2 Skalierbarkeit.....	75
Zielsetzung	75
Architekturgrundsatz	75

5.3	Verfügbarkeit	77
	Zielsetzung	77
	Architekturgrundsatz	77
5.4	Sicherheit	79
	Zielsetzung	79
	Architekturgrundsatz	79
5.5	Datenschutz	82
	Zielsetzung	82
	Architekturgrundsatz	82
5.6	Erweiterbarkeit & Modularität	84
	Zielsetzung	84
	Architekturgrundsatz	84
6	Architektur- und Integrationsanforderungen	87
6.1	Trennung von Retrieval, Modell und Orchestrierung	88
6.2	OpenSearch als zentrale Suchkomponente innerhalb der Hybrid-Retrieval-Architektur	88
6.3	Flexible Modellanbindung und Modell-Routing	89
6.4	Modellunabhängigkeit	90
6.5	RAG als Standardmechanismus	91
6.6	Token- und Ressourcenoptimierung	91
6.7	Agent Runtime	92
6.8	Tool-Integration und ereignisgesteuerte Verarbeitung	93
6.9	Protokollierung und Nachvollziehbarkeit	93
6.10	Wissensgraph und Graph-RAG	94
6.11	Dokumentenverarbeitungspipeline	95
7	Betrieb & Infrastruktur	96
	Zielsetzung	96
	Architekturgrundsatz	96
7.1	Cloud-native Architektur	96
	Zielsetzung	96
	Architekturgrundsatz	96
7.2	Nutzung von Managed Services	98
	Zielsetzung	98
	Architekturgrundsatz	98
3.8	Verzicht auf eigene KI-Infrastruktur	99
	Zielsetzung	99

Architekturgrundsatz	99
3.9 Skalierung und Verfügbarkeit	101
Zielsetzung	101
Architekturgrundsatz	101
3.4 Monitoring Integration	102
7.6 Sicherheitsstandards	104
Zielsetzung	104
Architekturgrundsatz	104
8 Datenschutz & Informationssicherheit	106
Zielsetzung	106
Grundsätze	106
8.1 Einhaltung der DSGVO-Anforderungen	106
Zielsetzung	106
Datenschutzgrundsatz	106
8.2 PII-Erkennung und –Verarbeitung	108
Zielsetzung	108
Datenschutzgrundsatz	108
8.3 Datenschutz-Ampelsystem	110
Zielsetzung	110
Datenschutzgrundsatz	110
8.4 Trennung von Daten und Modellen	112
Zielsetzung	112
Datenschutzgrundsatz	112
8.5 KI-Inventar und EU AI Act	114
Zielsetzung	114
Compliance-Grundsatz	114
8.6 Informationssicherheit	116
Zielsetzung	116
Informationssicherheitsgrundsatz	116
8.7 Sicherheit von KI-Systemen (LLM Security)	117
Zielsetzung	117
Sicherheitsgrundsatz	117
8.8 Agentenverwaltung und Agentenkonfiguration	120
Zielsetzung	120
Governance-Grundsatz	120
9 Schnittstellen	122

Zielsetzung	122
Integrationsgrundsatz	122
9.1 API-Schnittstellen	122
Zielsetzung	122
Integrationsgrundsatz	123
3.4 Integration in den Urban Stack	125
Zielsetzung	125
Integrationsgrundsatz	125
9.3 Einbindung externer Systeme	127
Zielsetzung	127
Integrationsgrundsatz	127
9.4 Webhooks und event-basierte Integration	128
Zielsetzung	128
Integrationsgrundsatz	129
9.5 Interoperabilität und Erweiterbarkeit	131
Zielsetzung	131
Integrationsgrundsatz	131
9.6 Integration von MCP-Servern (Model Context Protocol)	132
Zielsetzung	132
Integrationsgrundsatz	132
10. Qualitätssicherung & Monitoring	135
Qualitätsgrundsatz	135
10.1 KI-Qualität	135
Zielsetzung	135
Qualitätsgrundsatz	135
10.1.1 Evaluation von Retrieval-, RAG- und Agentensystemen	139
Zielsetzung	139
Evaluierungsgrundsatz	139
10.2 Systemmonitoring	141
Zielsetzung	141
Monitoring-Grundsatz	141
10.3 KI- und Nutzungsmonitoring	143
Zielsetzung	143
Monitoring-Grundsatz	143
10.4 Protokollierung	145
Zielsetzung	145

Protokollierungsgrundsatz	145
10.5 KI-Governance & Vertrauenswürdigkeit	146
Zielsetzung	146
Governance-Grundsatz	146
10.5.1 Prompt Governance	148
Zielsetzung	148
Governance-Grundsatz	148
10.5.2 Halluzinationsvermeidung und Antwortqualität	150
Zielsetzung	150
Qualitätsgrundsatz	151
10.5.3 Human-in-the-Loop	153
Zielsetzung	153
Governance-Grundsatz	153
10.5.4 Auditierbarkeit und KI-Logging (EU AI Act)	155
Zielsetzung	155
Governance-Grundsatz	155
11. Schulung und Dokumentation	158
Zielsetzung	158
Dokumentations- und Schulungsgrundsatz	158
11.1 Administratorhandbuch	158
Zielsetzung	158
Dokumentationsgrundsatz	158
11.2 API-Dokumentation	160
Zielsetzung	160
Dokumentationsgrundsatz	160
11.3 Schulung und Wissenstransfer	161
Zielsetzung	161
Schulungsgrundsatz	161
11.4 Onboarding und Benutzerunterstützung	162
Zielsetzung	162
Onboarding-Grundsatz	162
11.5 Digitale Souveränität und Open Source	163
Zielsetzung	163
Grundsatz der digitalen Souveränität	163
12. Support & Service Level (SLA)	167
Zielsetzung	167

Support- und Servicegrundsatz	167
12.1 Supportorganisation	167
Zielsetzung	167
Supportgrundsatz	167
12.2 Supportzeiten und Priorisierung	169
Zielsetzung	169
Supportgrundsatz	169
12.4 Wartung und Betrieb	171
Zielsetzung	171
Betriebsgrundsatz	171
12.5 Monitoring und Reporting	172
Zielsetzung	172
Monitoring-Grundsatz	172
13 Abnahme	173
Zielsetzung	173
Abnahmegrundsatz	173
Abnahme von KI-Komponenten	174
13.1 Funktionale Tests	175
Zielsetzung	175
Testgrundsatz	175
13.2 Last- und Performancetests	175
Zielsetzung	175
Testgrundsatz	175
13.3 Sicherheitstests	176
Zielsetzung	176
Testgrundsatz	176
13.4 Dokumentations- und Betriebsabnahme	177
Zielsetzung	177
9.2 Abnahmegrundsatz	177
13.4 Abnahmekriterien und Dokumentation	178
Zielsetzung	178
Abnahmegrundsatz	178
13.5 Abnahmeentscheidung	179
Zielsetzung	179
Abnahmegrundsatz	179
13.6 Nachweis- und Abnahmeverfahren	180

Zielsetzung	180
Nachweisgrundsatz	180
14 Exit-Strategie	181
Zielsetzung	181
Grundsatz der Exit-Strategie	181
14.1 Datenexport.....	181
Zielsetzung	181
Exportgrundsatz	182
14.2 Offene Formate	183
Zielsetzung	183
Portabilitätsgrundsatz.....	183
14.3 Unterstützung beim Anbieterwechsel	184
Zielsetzung	184
Übergangsgrundsatz	184
14.4 Sicherstellung der Weiterverwendbarkeit	185
Zielsetzung	185
Architekturgrundsatz	185
15 Projektsteuerung und Projektmanagement.....	186
Zielsetzung	186
Projektmanagementgrundsatz.....	186
Zusammenarbeit zwischen den Auftragnehmern der einzelnen Lose	186
16 Zeitplan.....	188
Zielsetzung	188
Grundsatz der Terminplanung.....	188
Leistungsumfang	188
17 Schlussbestimmungen.....	189
Zielsetzung	189
Grundsatz.....	189

1 Ausgangslage

Die Stadt Gütersloh mit rund 100.000 Einwohnern ist ein wirtschaftlich starker und innovativer Standort in Ostwestfalen-Lippe. Im Rahmen des bundesweiten Förderprogramms „Modellprojekte Smart Cities“ gestaltet die Stadt die digitale Transformation aktiv und entwickelt nachhaltige, bürgerzentrierte Lösungen für urbane Herausforderungen.

Mit dem Smart-City-Projekt verfolgt die Stadt einen offenen und kooperativen Ansatz auf Basis von Open-Source-Prinzipien. Ziel ist es, eine betriebsbereite, leistungsfähige und auf andere Kommunen übertragbare Plattform bereitzustellen. Dabei stehen digitale Souveränität, offene Standards, Interoperabilität sowie DSGVO-konforme und möglichst europäische Technologien im Mittelpunkt. So soll eine nachhaltige, gemeinschaftlich weiterentwickelbare KI- und Dateninfrastruktur für Kommunen entstehen.

Hierfür wurde der Urban Stack entwickelt – eine modulare, mandantenfähige Plattformarchitektur für die Integration, Verarbeitung und Nutzung kommunaler Daten. Sie bündelt Daten aus Sensorik, Fachverfahren, Drittsystemen und IoT-Anwendungen und schafft die Grundlage für datenbasierte Anwendungen in Bereichen wie intelligente Quartiere, nachhaltige Mobilität, Energieversorgung und digitale Verwaltungsservices. Die Datenhoheit verbleibt dabei jederzeit bei den jeweiligen Mandanten.

Da marktverfügbare Standardlösungen die Anforderungen nicht vollständig erfüllten, wurde der Urban Stack als eigenständige kommunale Plattform entwickelt. Aufbauend darauf beabsichtigt die Stadt Gütersloh nun die Beschaffung und Integration eines KI-Stacks. Ziel ist es, die bestehende Daten- und Systemlandschaft um leistungsfähige KI-Komponenten zu erweitern, um Verwaltungsprozesse zu automatisieren, Mitarbeitende zu unterstützen und datenbasierte Entscheidungen zu verbessern.

1.4 Einleitung

Die Stadt Gütersloh beabsichtigt die Beschaffung eines KI-Stacks zur Unterstützung der internen Verwaltungsarbeit. Ziel ist es, Verwaltungsprozesse effizienter zu gestalten, Routineaufgaben zu automatisieren und Mitarbeitenden moderne KI-gestützte Werkzeuge bereitzustellen.

Der KI-Stack soll als integraler Bestandteil des Urban Stack – der modularen Plattformarchitektur der Stadt Gütersloh für datenbasierte Anwendungen und Dienste – in das bestehende digitale Ökosystem integriert werden. Insbesondere ist eine Anbindung an den Urban Data Hub sowie weitere Komponenten des Urban Stack vorgesehen, um eine durchgängige, interoperable und skalierbare Nutzung von Daten und KI-Diensten zu ermöglichen.

Im Sinne dieser Leistungsbeschreibung bezeichnet der Begriff KI-Stack (nachfolgend auch Plattform) das Gesamtsystem aus Infrastruktur, KI-Verarbeitung, Retrieval, Orchestrierung, Anwendungsschicht, Governance und Schnittstellen.

1.5 Zielsetzung

Effizienzsteigerung

Unterstützung der Mitarbeitenden bei Textarbeit, Recherche, Dokumentenverarbeitung und Wissensmanagement durch KI-gestützte Assistenten.

Wissensmanagement

Aufbau und Nutzung von Vektor-Wissensdatenbanken zur strukturierten Erschließung verwaltungsinterner Dokumente und Informationen.

Flexibilität

Ermöglichung der eigenständigen Erstellung und Konfiguration von spezialisierten KI-Agenten durch die Fachbereiche.

Datenschutz

Bereitstellung sowohl leistungsstarker API-basierter Sprachmodelle als auch datenschutzkonformer Open-Weight-Modelle mit transparenter Kennzeichnung des jeweiligen Datenschutzniveaus.

2 Gegenstand der Leistung

Gegenstand dieser Ausschreibung ist die Beschaffung, Implementierung und Integration eines mandantenfähigen KI-Stacks für die Stadt Gütersloh. Der KI-Stack soll als Erweiterung des bestehenden Urban Stack in das digitale Ökosystem der Stadt eingebunden werden und die vorhandene Dateninfrastruktur, insbesondere den Urban Data Hub, funktional ergänzen.

Ziel der Leistung ist die Bereitstellung einer leistungsfähigen, skalierbaren und interoperablen KI-Plattform zur Unterstützung verwaltungsinterner Prozesse sowie zur schrittweisen Erweiterung des Urban Stack um KI-gestützte Funktionen.

Die langfristige Zielarchitektur dient der Beschreibung der angestrebten Gesamtarchitektur. Verbindlicher Leistungsgegenstand dieser Ausschreibung ist ausschließlich der in Kapitel 2 definierte Umfang der produktiven Erstimplementierung sowie die nach Kategorien gekennzeichneten Anforderungen.

2.4 Abgrenzung der Leistung

Gegenstand der Leistung ist die Entwicklung, Integration, Konfiguration und Inbetriebnahme der in dieser Leistungsbeschreibung beschriebenen Komponenten des KI-Stacks. Schwerpunkt der Leistung ist die Integration neu entwickelter Komponenten in die bereits vorhandene Infrastruktur und Systemlandschaft der Auftraggeberin.

Die Architektur ist modular, komponentenbasiert und herstellernerneutral auszugestalten. Sämtliche Komponenten müssen über standardisierte Schnittstellen integriert werden und eine spätere funktionale Erweiterung ohne grundlegende Änderungen der Gesamtarchitektur ermöglichen.

Weiterentwicklungen, zusätzliche Fachverfahren, neue Anwendungsfälle sowie funktionale Erweiterungen nach der produktiven Erstimplementierung sind nicht Bestandteil dieser Herstellungsleistung und können im Rahmen gesonderter Beauftragungen umgesetzt werden.

Bereits vorhandene Komponenten:

Die Auftraggeberin stellt wesentliche Infrastruktur-, Plattform- und Basisdienste bereits bereit. Diese Komponenten sind grundsätzlich nicht neu bereitzustellen oder funktional gleichwertig zu ersetzen, sondern innerhalb der ausgeschriebenen Lösung zu integrieren.

Hierzu gehören insbesondere:

- Kubernetes-Cluster einschließlich Basisbetrieb
- S3-kompatibler Objektspeicher
- OpenSearch in Standardkonfiguration
- ClickHouse
- Authentifizierungs- und Autorisierungsinfrastruktur (Single Sign-On)
- Basisdienste für Monitoring, Logging und Netzwerk
- Urban Stack Basisplattform

Bestandteil der Leistung ist ausschließlich die fachliche und technische Integration, Konfiguration sowie Nutzung dieser Komponenten innerhalb des KI-Stacks.

Umfang der Erstimplementierung

Die Erstimplementierung umfasst die Bereitstellung einer produktiv nutzbaren Grundausbaustufe des KI-Stacks.

Diese umfasst insbesondere:

- Integration der bereitgestellten Infrastrukturkomponenten
- Aufbau des AI Harness einschließlich Agent Runtime und Context Engine
- Aufbau der Dokumentenverarbeitungspipeline
- Integration von OpenSearch als Such- und Retrieval-Komponente
- Bereitstellung des definierten Memory-Konzeptes U
- Umsetzung der Governance-, Sicherheits- und Guardrail-Funktionen
- -Bereitstellung der standardisierten APIs
- Bereitstellung des Urban KI Hub
- Integration der in dieser Leistungsbeschreibung definierten ML-Services

Die Erstimplementierung beschränkt sich zunächst auf folgende Referenzumfänge:

- Verarbeitung von Dokumenten aus einem bereitgestellten S3-kompatiblen Objektspeicher sowie über standardisierte REST-Schnittstellen
- Umsetzung von zwei Referenzagenten
- Ereignisgesteuerte Verarbeitung innerhalb der Dokumentenpipeline
- Nutzung der vorhandenen UI-Komponenten und des bestehenden Frontend-Stacks des Urban Stack
- Integration der bereitgestellten Infrastrukturkomponenten ohne deren Neuaufbau
- Sämtliche Komponenten sind so auszugestalten, dass weitere Dokumentenquellen, Agenten, Workflows, Ereignisquellen und Fachverfahren ohne grundlegende Änderungen der Architektur ergänzt werden können.

Nicht Bestandteil der Erstimplementierung

Nicht Bestandteil der Erstimplementierung sind insbesondere:

- Aufbau oder Betrieb der Infrastruktur
- Bereitstellung eines Kubernetes-Clusters
- Bereitstellung oder Betrieb von OpenSearch, ClickHouse oder des S3-Objektspeichers
- Bereitstellung eines Identity-Providers oder Single-Sign-On-Systems
- Bereitstellung eines Model Hubs oder eigener Foundation Models
- Training eigener Sprachmodelle oder Embedding-Modelle
- Entwicklung eigener OCR-, Speech- oder Vision-Modelle
- Integration weiterer Dokumentenquellen über S3 und REST-Schnittstellen hinaus
- Integration weiterer Fachverfahren, DMS-, SharePoint- oder Dateisysteme
- Entwicklung weiterer Agenten über die definierten Referenzagenten hinaus
- Entwicklung zusätzlicher Frontend-Anwendungen außerhalb des Urban KI Hub
- Aufbau einer Workflow-Automatisierungsplattform
- Migration bestehender Fachverfahren
- Erweiterungen nach der produktiven Erstimplementierung

Die in dieser Leistungsbeschreibung beschriebenen Anforderungen unterscheiden bewusst zwischen den im Rahmen der produktiven Erstimplementierung verbindlich umzusetzenden Leistungen, den architektonisch vorzubereitenden Erweiterungsmöglichkeiten sowie der langfristigen Zielarchitektur des KI-Stacks.

Verbindlicher Leistungsgegenstand dieser Ausschreibung sind ausschließlich die als Bestandteil der produktiven Erstimplementierung gekennzeichneten Anforderungen sowie die hierfür erforderlichen Integrations-, Konfigurations- und Dokumentationsleistungen. Anforderungen, die als Architekturvorbereitung, Soll-Anforderung oder Erweiterung der Zielarchitektur gekennzeichnet sind, begründen grundsätzlich keine Verpflichtung zur vollständigen Umsetzung im Rahmen dieser Ausschreibung. Sie dienen der Sicherstellung einer zukunftsfähigen, modularen und erweiterbaren Gesamtarchitektur und sind im Rahmen der Erstimplementierung lediglich soweit zu berücksichtigen, dass eine spätere Umsetzung ohne grundlegende Änderungen der Architektur möglich bleibt.

Maßgeblich für den Leistungsumfang sind die jeweiligen Kategorien der einzelnen Anforderungen sowie die in Kapitel 2 beschriebene Abgrenzung der Leistung.

2.5 Beschreibung der Plattformbestandteile

Der KI-Stack besteht aus mehreren logisch getrennten, jedoch interoperablen Komponenten: Der Urban KI Hub bildet die zentrale Benutzeroberfläche für Mitarbeitende der Stadtverwaltung. Über ihn werden KI-Dialoge, Agenten, Prompts, Wissensdatenbanken, Workflows und administrative Funktionen zugänglich gemacht.

Der Urban Gov Hub bildet eine vorgelagerte Daten- und Dokumentenquelle innerhalb der Gesamtarchitektur. Der KI-Stack muss in der Lage sein, Daten und Dokumente aus dem Urban Gov Hub, insbesondere über einen Datenmanager oder eine funktional gleichwertige Komponente, zu übernehmen und an die Dokumentenverarbeitungspipeline zu übergeben. Der Datenmanager im Urban Gov Hub stellt den primären Einstiegspunkt für Dokumente in die Dokumentenverarbeitungspipeline dar. Das KI-Dialogsystem ermöglicht Multi-Turn-Dialoge zur Unterstützung bei Recherche, Textarbeit, Dokumentenverarbeitung und verwaltungsinternem Wissenszugriff. Die KI-Agenten-Plattform ermöglicht die Erstellung, Konfiguration, Verwaltung und organisationsweite Bereitstellung spezialisierter KI-Agenten durch berechtigte Nutzerinnen und Nutzer bzw. Fachbereiche.

Die Retrieval- und Wissensdatenbank-Komponente stellt RAG-Funktionen bereit. Sie nutzt OpenSearch für semantische Suche, Volltextsuche, hybride Suche, Metadatenfilterung und berechtigungsabhängigen Zugriff auf verwaltungsinterne Informationen. Die Modellabstraktions- und Routing-Schicht ermöglicht die flexible Anbindung unterschiedlicher Sprachmodelle, sowohl API-basierter externer Modelle als auch datenschutzkonformer Open-Weight-Modelle. Die Auswahl der Modelle erfolgt anwendungsfallbezogen, administrierbar und für Nutzende transparent. Die Orchestrierungsschicht steuert Modellaufrufe, Retrieval-Prozesse, Tool-Integrationen, Kontextzusammenstellung, Guardrails, Berechtigungen und Protokollierung.

Der AI Harness bildet die zentrale Ausführungs- und Steuerungsschicht des KI-Stacks. Der AI Harness bündelt die Funktionen der Orchestrierung, Agent Runtime, Modellanbindung, Retrieval-Steuerung, Tool-Integration, Prompt-Verwaltung, Richtlinienumsetzung (Policies), Guardrails, Monitoring, Auditierung sowie Human-in-the-Loop-Prozesse in einer gemeinsamen Laufzeitumgebung.

Er stellt sämtliche KI-Funktionen über standardisierte Schnittstellen bereit und gewährleistet eine konsistente, nachvollziehbare und modellunabhängige Ausführung aller KI-Prozesse. Die Context Engine ist Bestandteil des AI Harness und übernimmt die strukturierte Zusammenstellung des für eine KI-Ausführung relevanten Gesamtkontexts.

Sie führt nutzer-, rollen-, mandanten-, agenten-, prompt-, tool-, memory-, policy- und wissensbezogene Kontextinformationen unmittelbar vor dem Modellaufruf kontrolliert zusammen. Ziel ist eine konsistente, nachvollziehbare und datenschutzkonforme Kontextbildung für KI-Dialoge, RAG-Prozesse, Agentenausführungen und Tool-Integrationen. Die Agent Runtime ist Bestandteil der Orchestrierungsschicht und übernimmt die operative Ausführung von KI-Agenten. Die Administrationsoberfläche umfasst Nutzer- und Rollenverwaltung, Modellverwaltung, Agentenverwaltung, Token- und Kostenmonitoring, Plattformkonfiguration, KI-Inventar sowie Monitoring- und Auditfunktionen.

Workflow-Automatisierungskomponenten, die außerhalb dieser Leistungsbeschreibung im Urban Stack bzw. über City Tools bereitgestellt werden, können als angebundene Drittsysteme genutzt werden. Der KI-Stack muss hierfür geeignete, dokumentierte Schnittstellen bereitstellen, damit KI-Funktionen, Agenten, ML-Services oder Retrieval-Funktionen durch externe Workflow-Automatisierungssysteme angesteuert werden können. Workflow-Automatisierungssysteme sind nicht Bestandteil der Leistung, stellen jedoch eine zentrale Integrationsoption innerhalb der Gesamtarchitektur dar.

2.6 Einordnung als Gesamtsystem

Der KI-Stack ist als modulares Gesamtsystem innerhalb des digitalen Ökosystems der Stadt Gütersloh zu verstehen. Er bildet keine isolierte Einzelanwendung, sondern eine querschnittliche Plattform für KI-gestützte Verwaltungsprozesse und Entscheidungsunterstützung.

Das Gesamtsystem verbindet Benutzerinteraktion, KI-Verarbeitung, Wissensmanagement, Modellanbindung, Tool-Integration, Datenschutz, Monitoring und Governance in einer gemeinsamen Architektur. Die einzelnen Lose übernehmen dabei klar abgegrenzte Aufgaben: Infrastruktur und Betrieb stellen die technische Grundlage bereit, der KI-Stack verantwortet Modell- und Retrieval-Funktionen, der Urban KI Hub bildet die zentrale Anwendungsschicht, ML-Services ergänzen spezialisierte Funktionen und die Workflow-Automatisierung verknüpft die Komponenten zu wiederverwendbaren Prozessen.

Die Plattform ist modellunabhängig, mandantenfähig, schnittstellenbasiert und erweiterbar auszugestalten. Sie muss so konzipiert werden, dass einzelne Komponenten unabhängig voneinander betrieben, ausgetauscht und erweitert werden können. Dadurch wird eine langfristig wartbare, skalierbare und anbieterunabhängige KI-Infrastruktur für die Stadtverwaltung Gütersloh geschaffen.

Sämtliche Anforderungen dieser Leistungsbeschreibung sind unter Berücksichtigung der technischen Umsetzbarkeit, Wirtschaftlichkeit und Verhältnismäßigkeit auszulegen. Gleichwertige technische Lösungen sowie angemessene alternative Umsetzungsansätze sind zulässig, sofern die funktionalen, qualitativen und architektonischen Zielsetzungen der Auftraggeberin in vergleichbarer Weise erfüllt werden. Die Auftraggeberin verfolgt ausdrücklich einen modularen, evolutiven und wirtschaftlich vertretbaren Ausbau der Plattformarchitektur.

Die Architektur ist im Sinne digitaler Souveränität so auszugestalten, dass offene Standards, transparente Schnittstellen, Open-Source-Komponenten sowie europäische und DSGVO-konforme Technologien bevorzugt unterstützt werden. Proprietäre Abhängigkeiten sind auf ein technisch und wirtschaftlich notwendiges Maß zu begrenzen. Soweit in dieser Leistungsbeschreibung konkrete Technologien, Standards, Protokolle oder Komponenten genannt werden, dienen diese primär der Beschreibung des angestrebten Funktions-, Qualitäts- und Architekturniveaus. Gleichwertige technische Lösungen sind zulässig, sofern sie die funktionalen und nichtfunktionalen Anforderungen in vergleichbarer Weise erfüllen und keine Einschränkungen hinsichtlich Interoperabilität, Erweiterbarkeit, Wartbarkeit oder Anbieterunabhängigkeit verursachen.

3 Lose-Struktur

3.4 Übersicht der Lose

Der zu beschaffende KI-Stack wird in mehrere fachlich und technisch abgegrenzte Lose unterteilt. Ziel dieser Aufteilung ist es, eine modulare, erweiterbare und anbieterunabhängige Systemarchitektur zu ermöglichen sowie die Vergabe an spezialisierte Auftragnehmer zu erleichtern.

Die Leistung gliedert sich in folgende Lose:

- **Los 1** – Backend / KI-Plattform
- **Los 2** – Frontend / Agentenplattform
- **Los 3** – ML-Hilfsdienste

Die Lose sind funktional miteinander verbunden, müssen jedoch so konzipiert sein, dass sie über standardisierte Schnittstellen interoperabel sind und unabhängig voneinander entwickelt, betrieben und erweitert werden können.

3.5 Beschreibung der Lose

Los 1 – Backend / KI-Plattform

Kurzbeschreibung

Los 1 umfasst die Bereitstellung des zentralen AI Harness als Backend des KI-Stacks. Der AI Harness bildet die zentrale Ausführungs-, Integrations- und Governance-Plattform der Gesamtlösung. Er umfasst insbesondere Modellanbindung, Modell-Routing, Orchestrierung, Agent Runtime, Retrieval-Steuerung, Tool-Integration, Prompt-Verwaltung, Policy Enforcement, Guardrails, Monitoring, Auditierung sowie die Bereitstellung standardisierter APIs.

Zielsetzung

Ziel ist die Bereitstellung einer modularen und modellunabhängigen KI-Architektur, die flexibel erweitert werden kann, unterschiedliche Modelle integriert und eine effiziente sowie nachvollziehbare KI-Verarbeitung ermöglicht.

Los 2 – Frontend / Agentenplattform

Kurzbeschreibung

Los 2 umfasst die Entwicklung und Bereitstellung der zentralen Benutzeroberfläche („Urban KI Hub“) als Einstiegspunkt für alle Nutzenden. Die Anwendung dient als Steuerungsschicht für KI-Funktionen, Agenten, Prompts und Workflows.

Zielsetzung

Ziel ist die Bereitstellung einer benutzerfreundlichen, transparenten und barrierefreien Oberfläche, die den Zugriff auf alle KI-Funktionen ermöglicht und eine intuitive Nutzung der Plattform durch Mitarbeitende der Verwaltung sicherstellt.

Los 3 – ML-Hilfsdienste**Kurzbeschreibung**

Los 3 umfasst die Bereitstellung spezialisierter Machine-Learning-Services als eigenständige, wiederverwendbare Komponenten. Dazu zählen insbesondere Dienste für Bildverarbeitung, Dokumentenanalyse, Sprachverarbeitung, Klassifikation sowie die Verarbeitung und Erkennung personenbezogener Daten.

Zielsetzung

Ziel ist die Schaffung modularer, skalierbarer und flexibel einsetzbarer ML-Services, die unabhängig vom zentralen KI-Stack betrieben und über standardisierte Schnittstellen in unterschiedliche Anwendungsfälle sowie in die Gesamtarchitektur des Urban Stack integriert werden können.

Technische Umsetzung der ML-Services:

Der Auftragnehmer stellt modulare ML-Services für folgende Anwendungsbereiche bereit:

- Dokumentenanalyse und -strukturierung
- Optical Character Recognition (OCR)
- Sprachtranskription (Speech-to-Text)
- Klassifikation und Informationsextraktion
- Bildanalyse
- Datenschutz- und PII-Erkennung sowie -Verarbeitung

Die Umsetzung soll vorrangig unter Verwendung etablierter Open-Source-Komponenten erfolgen, insbesondere unter Einsatz von Technologien wie Docling, PaddleOCR, Tesseract, faster-whisper, Presidio, scikit-learn, Hugging Face Transformers, vLLM, KServe oder funktional gleichwertigen Komponenten. Der Einsatz alternativer Technologien ist zulässig, sofern diese die funktionalen und nicht-funktionalen Anforderungen gleichwertig erfüllen.

Alle ML-Services sind als eigenständige, lose gekoppelte Dienste bereitzustellen und müssen über standardisierte, dokumentierte Schnittstellen (z. B. REST-APIs oder funktional gleichwertig) adressierbar sein. Die Schnittstellen müssen eine strukturierte, maschinenlesbare Kommunikation (insbesondere JSON oder vergleichbar) unterstützen und so gestaltet sein, dass eine Integration in den KI-Stack sowie durch externe Drittsysteme ohne zusätzliche Anpassungen möglich ist.

Die ML-Services müssen sich nahtlos in die bestehende Authentifizierungs- und Autorisierungsinfrastruktur des Urban Stack integrieren lassen. Hierzu ist die Unterstützung gängiger Verfahren (z. B. Single Sign-On, SAML 2.0 oder funktional gleichwertig) sowie eine rollen- und berechtigungsbasierte Zugriffskontrolle (RBAC) sicherzustellen.

Die Services sind mandantenfähig auszulegen. Es muss gewährleistet sein, dass mehrere Organisationseinheiten (Mandanten) parallel betrieben werden können, ohne dass es zu einer Vermischung von Daten, Konfigurationen oder Verarbeitungskontexten kommt. Die Mandantentrennung ist mindestens auf Ebene von Daten, Konfigurationen, Zugriffen und Protokollierung umzusetzen.

Die eingesetzten Komponenten sind modular aufzubauen und müssen austauschbar gestaltet sein. Ein Austausch einzelner Komponenten oder Technologien darf keine grundlegenden Änderungen an der Gesamtarchitektur oder den Schnittstellen erfordern. Die Integration in den KI-Stack hat ausschließlich über die definierten Schnittstellen zu erfolgen.

3.6 Übersicht der Lose

Der zu beschaffende KI-Stack wird in mehrere fachlich und technisch voneinander abgegrenzte Lose unterteilt. Ziel dieser Aufteilung ist die Umsetzung einer modularen, interoperablen und langfristig erweiterbaren Plattformarchitektur sowie die Möglichkeit, spezialisierte Auftragnehmer entsprechend ihrer fachlichen Schwerpunkte einzubinden.

Die Lose bilden gemeinsam die Zielarchitektur der Urban-KI-Plattform ab. Gegenstand dieser Ausschreibung ist jedoch ausschließlich die Umsetzung einer produktionsfähigen Erstimplementierung. Die Architektur ist dabei so auszugestalten, dass zukünftige Erweiterungen ohne grundlegende Änderungen der Gesamtarchitektur möglich sind.

Die Leistung gliedert sich in folgende Lose:

- **Los 1 – Backend / KI-Plattform**
- **Los 2 – Frontend / Urban KI Hub**
- **Los 3 – ML-Hilfsdienste**

Die Lose sind funktional miteinander verbunden und kommunizieren ausschließlich über standardisierte, dokumentierte und versionierte Schnittstellen. Eine enge technische Kopplung zwischen den Losen ist zu vermeiden.

3.4 Beschreibung der Lose

Los 1 – Backend / KI-Plattform

Kurzbeschreibung

Los 1 umfasst die Implementierung des zentralen AI Harness als Backend des KI-Stacks. Der AI Harness bildet die zentrale Ausführungs-, Integrations- und Governance-Plattform der Gesamtlösung und übernimmt insbesondere die Modellanbindung, Modellsteuerung, Agent Runtime, Orchestrierung, Retrieval-Steuerung, Tool-Integration, Prompt-Verwaltung, Context Engine, Policy Enforcement, Guardrails, Monitoring, Auditierung sowie die Bereitstellung standardisierter APIs.

Schwerpunkt dieses Loses ist die Integration, Konfiguration und Orchestrierung der bereits vorhandenen Infrastrukturkomponenten sowie der neu bereitzustellenden KI-Plattformdienste.

Zielsetzung

Ziel ist die Bereitstellung einer modularen, modellunabhängigen und erweiterbaren Backend-Plattform für den Betrieb der Urban-KI-Plattform.

Umfang der Erstimplementierung

Die Erstimplementierung umfasst insbesondere:

- Integration des bereitgestellten IONOS Model Hub
- Integration des bereitgestellten OpenSearch-Clusters als Such- und Retrieval-Komponente
- Integration des bereitgestellten S3-kompatiblen Objektspeichers als Dokumentenquelle
- Integration der vorhandenen Authentifizierungs- und Autorisierungsinfrastruktur
- Integration von ClickHouse zur Speicherung von Betriebs-, Monitoring-, Audit- und Nutzungsdaten
- Aufbau des AI Harness
- Bereitstellung der Context Engine
- Bereitstellung der Agent Runtime
- Bereitstellung der Dokumentenverarbeitungspipeline
- Bereitstellung der Retrieval- und RAG-Funktionen
- Bereitstellung der standardisierten APIs
- Umsetzung von zwei Referenzagenten
- Integration der ML-Hilfsdienste aus Los 3

Die Dokumentenverarbeitung ist in der Erstimplementierung auf Dokumente aus dem bereitgestellten S3-kompatiblen Objektspeicher sowie über standardisierte REST-Schnittstellen bereitgestellte Dokumente beschränkt.

Die Bereitstellung oder der Betrieb der zugrundeliegenden Infrastruktur ist nicht Bestandteil dieses Loses.

Los 2 – Frontend / Urban KI Hub

Kurzbeschreibung

Los 2 umfasst die Entwicklung und Bereitstellung des Urban KI Hub als zentrale Benutzeroberfläche der Plattform.

Der Urban KI Hub bildet die Anwendungsschicht des KI-Stacks und stellt sämtliche Funktionen des AI Harness für die Mitarbeitenden der Stadtverwaltung über eine einheitliche, barrierefreie Benutzeroberfläche bereit.

Die Umsetzung erfolgt auf Basis der bereits vorhandenen Frontend-Architektur, des bestehenden Designsystems sowie vorhandener UI-Komponenten des Urban Stack. Vorhandene Komponenten sind grundsätzlich wiederzuverwenden.

Zielsetzung

Ziel ist die Bereitstellung einer intuitiven, transparenten und barrierefreien Benutzeroberfläche für sämtliche KI-Funktionen der Plattform.

Umfang der Erstimplementierung

Die Erstimplementierung umfasst insbesondere:

- Chatoberfläche
- Dokumentenchat
- Auswahl und Nutzung der Referenzagenten
- Promptverwaltung
- Quellenanzeige
- Modellinformationen einschließlich Datenschutzniveau
- Administrationsoberfläche
- Feedbackfunktionen
- Anzeige von Verarbeitungsschritten
- Integration der durch Los 1 bereitgestellten Backend-Funktionen

Die Entwicklung eines eigenständigen Frontend-Frameworks, eines neuen Designsystems oder allgemeiner Basis-UI-Komponenten ist nicht Bestandteil dieses Loses.

Los 3 – ML-Hilfsdienste**Kurzbeschreibung**

Los 3 umfasst die Bereitstellung spezialisierter Machine-Learning-Hilfsdienste als eigenständige, lose gekoppelte und wiederverwendbare Komponenten.

Die Dienste ergänzen den zentralen KI-Stack um spezialisierte Funktionen und werden über standardisierte Schnittstellen durch den AI Harness sowie weitere Komponenten des Urban Stack genutzt.

Der Schwerpunkt dieses Loses liegt auf der Integration etablierter Open-Source-Komponenten sowie geeigneter Machine-Learning-Frameworks. Eigene Foundation Models oder eigene Machine-Learning-Modelle sind nicht Gegenstand dieses Loses.

Zielsetzung

Ziel ist die Bereitstellung einer zentralen ML-Service-Plattform für den Urban Stack. Die Plattform stellt wiederverwendbare Machine-Learning- und KI-Dienste als eigenständige Backend-Services bereit, die unabhängig vom AI Harness betrieben und flexibel erweitert werden können.

Die ML-Service-Plattform dient der Bereitstellung standardisierter KI- und Machine-Learning-Funktionen für unterschiedliche Anwendungsfälle innerhalb der Stadtverwaltung. Die Dienste sind über standardisierte, dokumentierte APIs bereitzustellen und müssen sowohl durch den AI Harness als auch durch weitere Anwendungen, Fachverfahren oder Komponenten des Urban Stack nutzbar sein.

Die technische Umsetzung kann auf geeigneten Open-Source-Frameworks, kommerziellen Produkten oder funktional gleichwertigen Technologien basieren. Die Auswahl der eingesetzten Frameworks bleibt dem Auftragnehmer überlassen, sofern die funktionalen Anforderungen dieser Leistungsbeschreibung erfüllt werden.

Umfang der Erstimplementierung

Die Erstimplementierung umfasst mindestens folgende Dienste:

ML-Service-Domäne	Beispiele der bereitgestellten Funktionen	API	Typ
Dokumentenverarbeitung	OCR, Layoutanalyse, Tabellenextraktion, Handschrifterkennung, Dokumentstruktur	☑	Muss
Dokumentenverständnis	Klassifikation, Zusammenfassung, Metadatenextraktion, Dokumentvergleich	☑	Muss
Informationsextraktion	Named Entity Recognition, Schlüsselwörter, Relation Extraction, Ereignisse, Termine	☑	Muss
Datenschutz & Compliance	PII-Erkennung, Redaction, Anonymisierung, Pseudonymisierung	☑	Muss
Sprachverarbeitung (NLP)	Language Detection, Übersetzung, Rechtschreibprüfung, Textnormalisierung	☑	Muss
Audioverarbeitung	Speech-to-Text, Text-to-Speech, Sprechertrennung	☑	Muss
Semantische Verarbeitung	Embeddings, Ähnlichkeitssuche, Clustering, semantischer Vergleich	☑	Muss
Bildverarbeitung	Bildklassifikation, Objekterkennung, Bildanalyse	☑	Soll
Knowledge Extraction	Graph-Extraktion, Wissensgraphen, Entitäten und Relationen	☑	Soll
Moderation & Qualität	Toxicity Detection, Spam-Erkennung, Qualitätsbewertung	☑	Soll
Prognose & Machine Learning	Klassische ML-Modelle, Vorhersagen, Regression, Klassifikation, Zeitreihen	☑	Soll
Geodaten & Karten	Luftbilder, Kartenanalyse, GeoML	☑	Kann

Die genannten Frameworks und Technologien dienen ausschließlich der Veranschaulichung möglicher Umsetzungen. Der Einsatz anderer funktional gleichwertiger Open Source ist zulässig, sofern die Anforderungen dieser Leistungsbeschreibung erfüllt werden.

Die Dokumentenverarbeitung erfolgt zunächst ausschließlich für Dokumente aus dem bereitgestellten S3-kompatiblen Objektspeicher sowie über standardisierte REST-Schnittstellen. Die Umsetzung erfolgt vorrangig durch Integration etablierter Open-Source-Komponenten und Machine-Learning-Frameworks. Die Entwicklung eigener Foundation Models oder eigener Machine-Learning-Modelle ist nicht Bestandteil dieses Loses.

Alle ML-Services sind als eigenständige Backend-Dienste bereitzustellen und ausschließlich über standardisierte, dokumentierte und abgesicherte APIs in den KI-Stack zu integrieren. Die Dienste müssen unabhängig voneinander betrieben, aktualisiert, erweitert und durch andere Anwendungen des Urban Stack wiederverwendet werden können.

3.5 Abgrenzung zwischen den Losen

Die Lose sind funktional eindeutig voneinander abgegrenzt.

Los 1 verantwortet sämtliche Backend-Funktionen des KI-Stacks, insbesondere AI Harness, Modellanbindung, Retrieval, RAG, Agent Runtime, Context Engine, Governance, Guardrails, Monitoring, Auditierung und die Bereitstellung standardisierter APIs.

Los 1 stellt keine eigenständige Benutzeroberfläche für Endnutzende bereit.

Los 2 verantwortet ausschließlich die Benutzerinteraktion und nutzt sämtliche KI-Funktionen über die von Los 1 bereitgestellten Schnittstellen. Eine eigene KI-Verarbeitungslogik ist nicht Bestandteil dieses Loses.

Los 3 stellt spezialisierte ML-Hilfsdienste bereit, die unabhängig betrieben und über standardisierte Schnittstellen in den AI Harness integriert werden.

Die Kommunikation zwischen den Losen erfolgt ausschließlich über dokumentierte, versionierte und standardisierte Schnittstellen.

3.6 Fachliche Konkretisierung Los 2 – Urban KI Hub

Der Urban KI Hub bildet die zentrale Anwendungsschicht für sämtliche Mitarbeitenden der Stadtverwaltung.

Die Erstimplementierung konzentriert sich auf die wesentlichen Funktionen für den produktiven KI-Einsatz.

Hierzu gehören insbesondere:

- dialogbasierte Interaktion mit KI-Systemen
- Dokumentenchat
- Auswahl und Nutzung der Referenzagenten
- Nutzung von Wissensbeständen
- Nutzung der bereitgestellten ML-Services
- Erstellung und Nutzung von Prompts
- Administrationsfunktionen
-

Die Benutzeroberfläche muss insbesondere die Anzeige der verwendeten Quellen, des eingesetzten Modells, des Datenschutzniveaus sowie der wesentlichen Verarbeitungsschritte unterstützen.

Die Umsetzung erfolgt auf Basis der vorhandenen UI-Komponenten des Urban Stack.

Die beigefügten Mockups dienen ausschließlich der Beschreibung der fachlichen Zielrichtung und stellen keine verbindlichen Designvorgaben dar.

3.7 Stufenweise Zielarchitektur

Die Leistungsbeschreibung beschreibt sowohl die langfristige Zielarchitektur der Urban-KI-Plattform als auch den Umfang der produktiven Erstimplementierung.

Die Erstimplementierung bildet eine wirtschaftlich umsetzbare Grundausbaustufe der Zielarchitektur.

Hierzu gehören insbesondere:

- AI Harness
- Context Engine
- Agent Runtime
- Urban KI Hub
- Dokumentenpipeline
- RAG
- Integration des IONOS Model Hub
- Integration von OpenSearch
- Integration von ClickHouse
- Integration des S3-kompatiblen Objektspeichers
- zwei Referenzagenten
- ML-Hilfsdienste gemäß Los 3

Die Architektur muss so ausgelegt sein, dass zusätzliche Dokumentenquellen, weitere Agenten, zusätzliche ML-Services, Fachverfahren, Workflow-Systeme, Eventquellen sowie weitere Integrationen ohne grundlegende Änderungen der Gesamtarchitektur ergänzt werden können.

Die Umsetzung dieser Erweiterungen ist nicht Bestandteil der Erstimplementierung, sofern sie in dieser Leistungsbeschreibung nicht ausdrücklich gefordert wird.

4 Funktionale Anforderungen

Die nachfolgenden funktionalen Anforderungen definieren die fachlichen Funktionen der Plattform.

4.1 KI-Dialogsystem

Das KI-Dialogsystem bildet die zentrale Interaktionskomponente der Urban-KI-Plattform. Es ermöglicht Mitarbeitenden der Stadtverwaltung die dialogbasierte Nutzung von Sprachmodellen, Retrieval-Augmented Generation, Wissensbeständen, Dokumentenfunktionen und KI-Agenten.

Die Erstimplementierung konzentriert sich auf einen produktiv nutzbaren, nachvollziehbaren und mandantenfähigen KI-Dialog. Erweiterte Komfortfunktionen und zusätzliche multimodale Interaktionsformen werden als Architekturvorbereitung oder spätere Ausbaustufe eingeordnet.

FA-001 – Multi-Turn-Dialogfähigkeit

Anforderungs- beschreibung	Die Plattform muss mehrstufige Dialoge (Multi-Turn-Dialoge) unterstützen und den Gesprächskontext über mehrere Interaktionen hinweg konsistent verwalten.
Kategorie	P1 - Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist im Rahmen der produktiven Erstimplementierung vollständig bereitzustellen.
Begründung	Die Multi-Turn-Dialogfähigkeit bildet die Grundlage sämtlicher KI-Interaktionen und ist Voraussetzung für einen produktiven Einsatz der Plattform.

FA-002 – Kontextverarbeitung

Anforderungs- beschreibung	Die Plattform muss den Konversationskontext bei der Antwortgenerierung berücksichtigen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist im Rahmen der produktiven Erstimplementierung vollständig bereitzustellen und über den AI Harness sowie die Context Engine umzusetzen.
Begründung	Die Berücksichtigung des Konversationskontexts ist erforderlich, damit Antworten fachlich anschlussfähig, nachvollziehbar und für mehrstufige Verwaltungsdialoge nutzbar sind.

FA-003 – RAG-Nutzung

Anforderungs- beschreibung	Bei wissensbasierten Anfragen muss die Retrieval-Schicht verpflichtend eingebunden werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist im Rahmen der produktiven Erstimplementierung vollständig bereitzustellen. Die Umsetzung erfolgt unter Nutzung der vorgesehenen Retrieval-Komponente, insbesondere OpenSearch.
Begründung	RAG ist der zentrale Mechanismus zur Nutzung verwaltungsinterner Wissensbestände und reduziert das Risiko unbelegter oder halluzinierter Antworten.

FA-004 – Quellenfähigkeit

Anforderungs- beschreibung	Antworten müssen auf zugrunde liegende Daten, Dokumente oder Wissensquellen referenzierbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist im Rahmen der produktiven Erstimplementierung vollständig bereitzustellen. Quellen müssen mindestens auf Ebene der verwendeten Dokumente oder Dokumentabschnitte nachvollziehbar sein.
Begründung	Quellenfähigkeit ist für Transparenz, Nachvollziehbarkeit und fachliche Prüfbarkeit von KI-generierten Antworten in der Verwaltung erforderlich.

FA-005 – Antwortqualität

Anforderungs- beschreibung	Die Plattform muss die Möglichkeit bieten, unbelegte Antworten zu kennzeichnen oder deren Ausgabe zu verweigern.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist im Rahmen der produktiven Erstimplementierung vollständig bereitzustellen. Sie ist mit den Mechanismen für RAG, Quellenprüfung, Guardrails und Antwortverweigerung zu verbinden.
Begründung	Die Kennzeichnung oder Verweigerung unbelegter Antworten ist eine zentrale Voraussetzung für vertrauenswürdige KI-Nutzung und reduziert fachliche Fehlinterpretationen.

FA-006 – Antwortverweigerung

Anforderungsbeschreibung	Bei unzureichender Datenlage muss die Plattform in der Lage sein, keine Antwort zu generieren oder transparent auf die unzureichende Informationsgrundlage hinzuweisen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist im Rahmen der produktiven Erstimplementierung vollständig bereitzustellen und in die Guardrail- und Governance-Mechanismen des AI Harness einzubinden.
Begründung	Eine kontrollierte Antwortverweigerung ist erforderlich, um spekulative oder fachlich nicht belastbare Antworten zu vermeiden.

FA-007 – Strukturierte Antworten

Anforderungsbeschreibung	Die Plattform soll Antworten strukturiert, insbesondere in Absätzen, Listen oder Tabellen, ausgeben können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist im Rahmen der produktiven Erstimplementierung bereitzustellen. Die konkrete Strukturierung kann über Prompt-Vorgaben, Agentenkonfigurationen oder Formatpräferenzen erfolgen.
Begründung	Strukturierte Antworten erhöhen die Verständlichkeit und Nutzbarkeit der Ergebnisse in verwaltungsinternen Arbeitsprozessen.

FA-008 – Mandantenfähigkeit

Anforderungsbeschreibung	Dialoge müssen mandantenfähig verwaltet und ausgeführt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist im Rahmen der produktiven Erstimplementierung vollständig bereitzustellen. Mandantenkontext, Berechtigungen und Datenisolation müssen über die gesamte Verarbeitungskette berücksichtigt werden.
Begründung	Mandantenfähigkeit ist eine grundlegende Architekturvoraussetzung des Urban Stack und muss bereits in der Erstimplementierung durchgängig umgesetzt werden.

FA-009 – Differenziertes Memory-Konzept

Anforderungs- beschreibung	Die Plattform muss ein differenziertes Memory-Konzept bereitstellen, das verschiedene Formen von Kontext- und Erinnerungsinformationen getrennt verwaltet und der KI ermöglicht, über einzelne Konversationen hinaus relevante Informationen unter Berücksichtigung von Datenschutz, Berechtigungen und Mandantenkontext wiederzuverwenden. Das Memory-Konzept muss mindestens folgende Ebenen berücksichtigen: Conversation Memory zur Nutzung des aktuellen Gesprächskontexts innerhalb einer Konversation, User Memory zur nutzerbezogenen Speicherung wiederkehrender Präferenzen und Arbeitskontexte, Agent Memory zur Speicherung agentenspezifischer Zustände und Konfigurationen, Tenant Memory zur mandantenbezogenen Verwaltung gemeinsamer organisatorischer Wissens- und Kontextinformationen sowie temporäres Laufzeit-Memory für einzelne Agenten- oder Workflow-Ausführungen. Die Nutzung von Memory-Funktionen muss rollen- und mandantenabhängig steuerbar, transparent nachvollziehbar sowie durch Nutzer und Administratoren konfigurierbar sein. Memory-Inhalte müssen den geltenden Datenschutz- und Löschanforderungen entsprechen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Im Rahmen der produktiven Erstimplementierung ist mindestens Conversation Memory für den aktuellen Gesprächskontext bereitzustellen. Die weiteren Memory-Ebenen sind architektonisch vorzusehen, über geeignete Datenmodelle, Berechtigungen, Schnittstellen und Löschmechanismen vorzubereiten und können in späteren Ausbaustufen vollständig umgesetzt werden.
Begründung	Ein differenziertes Memory-Konzept ist für eine zukunftsfähige KI-Plattform wichtig, erhöht jedoch den Umsetzungs- und Datenschutzaufwand erheblich. Für die Erstimplementierung genügt eine belastbare Architekturvorbereitung mit produktivem Conversation Memory.

FA-010 – Spracheingabe

Anforderungs- beschreibung	Die Plattform muss eine Spracheingabefunktion (Speech-to-Text) unterstützen, über die Nutzer per Mikrofon mit dem KI-Assistenten interagieren können. Die Spracherkennung muss über DSGVO-konforme Sprachmodelle oder funktional gleichwertige, datenschutzkonforme Verfahren erfolgen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die technische Anbindung an Speech-to-Text-Dienste ist architektonisch vorzubereiten. Eine vollständige Integration der Spracheingabe in die Benutzeroberfläche des KI-Dialogs ist nicht zwingender Bestandteil der produktiven Erstimplementierung, sofern die entsprechenden Schnittstellen und Erweiterungspunkte vorgesehen sind.
Begründung	Spracheingabe ist eine sinnvolle Erweiterung der Bedienbarkeit, jedoch für den initialen produktiven Kernbetrieb des KI-Dialogs nicht zwingend erforderlich. Die spätere Integration muss vorbereitet bleiben.

FA-011 – Sprachausgabe

Anforderungs- beschreibung	Die Plattform soll eine Sprachausgabefunktion (Text-to-Speech) unterstützen, über die KI-Antworten vorgelesen werden können.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Architektur muss die spätere Anbindung von Text-to-Speech-Diensten ermöglichen. Eine vollständige produktive Umsetzung innerhalb der Chatoberfläche ist nicht Bestandteil der Erstimplementierung.
Begründung	Sprachausgabe verbessert Barrierefreiheit und Nutzungskomfort, ist jedoch keine zwingende Kernfunktion der Erstimplementierung. Eine spätere Integration soll ohne grundlegende Architekturänderung möglich sein.

FA-012 – Bildanalyse

Anforderungs- beschreibung	Die Plattform muss die Analyse hochgeladener Bilddateien durch multimodale Modelle oder angebundene ML-Services unterstützen, insbesondere für Bildbeschreibung, Texterkennung aus Screenshots und die Analyse von Fotodokumentationen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Architektur muss die Verarbeitung und Übergabe von Bilddateien an geeignete multimodale Modelle oder ML-Services unterstützen. Eine vollständige fachliche Umsetzung aller Bildanalyse-Szenarien ist nicht Bestandteil der Erstimplementierung.
Begründung	Bildanalyse ist für spätere verwaltungsnahe Anwendungsfälle relevant, gehört jedoch nicht zum zwingenden Kernumfang des initialen KI-Dialogs. Die Schnittstellen und Datenflüsse müssen vorbereitet werden.

FA-013 – Datei-Upload im Chat

Anforderungs- beschreibung	Die Plattform muss den Upload und die Verarbeitung von Dateien im Chat unterstützen. Zu berücksichtigen sind insbesondere DOCX, DOC, TXT, PDF, XLSX, XLS, JSON, PPTX, XML, HTML, MSG, MD, CSV, PNG, JPEG, TIFF, ODP, ODS und ODT.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist im Rahmen der produktiven Erstimplementierung bereitzustellen. Die Umsetzung kann stufenweise erfolgen, sofern die für die Erstimplementierung relevanten Dokumentformate für die Dokumentenverarbeitung, den Dokumentenchat und RAG-Prozesse unterstützt werden.
Begründung	Datei-Upload und Dokumentenverarbeitung sind zentrale Funktionen der Erstimplementierung. Die breite Formatunterstützung muss wirtschaftlich umgesetzt und mit der Dokumentenpipeline abgestimmt werden.

FA-014 – Konversationshistorie

Anforderungs- beschreibung	Die Plattform muss die Konversationshistorie der letzten 90 Tage speichern. Die Aufbewahrungsdauer muss durch Administratoren zentral konfigurierbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist im Rahmen der produktiven Erstimplementierung bereitzustellen. Speicherung, Aufbewahrung und Löschung müssen mit Datenschutz-, Rollen- und Mandantenanforderungen abgestimmt sein.
Begründung	Eine konfigurierbare Konversationshistorie ist erforderlich, um die Nutzung des KI-Dialogs im Verwaltungsalltag nachvollziehbar und praktikabel zu gestalten.

FA-015 – Löschung von Chat-Verläufen

Anforderungs- beschreibung	Nutzer müssen ihre eigenen Chat-Verläufe selbstständig löschen können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist im Rahmen der produktiven Erstimplementierung vollständig bereitzustellen. Die Löschung muss datenschutzkonform erfolgen und mit Aufbewahrungs-, Protokollierungs- und Nachweispflichten abgestimmt sein.
Begründung	Die selbstständige Löschung eigener Chat-Verläufe ist für Datenschutz, Nutzerkontrolle und Akzeptanz der Plattform erforderlich.

FA-016 – Chat-Organisation

Anforderungs- beschreibung	Nutzer sollen ihre Chats organisieren können, insbesondere durch Ordner, Favoriten, Archivieren und Umbenennen.
Kategorie	E – Erweiterung der Zielarchitektur
Umsetzung	Die Funktion ist nicht Bestandteil der produktiven Erstimplementierung. Die Architektur und Datenmodelle sollen jedoch so gestaltet werden, dass eine spätere Erweiterung um Organisationsfunktionen möglich ist.
Begründung	Chat-Organisation ist eine Komfortfunktion. Sie verbessert die Nutzerführung, ist jedoch für die produktive Grundausbaustufe nicht zwingend erforderlich und kann wirtschaftlich sinnvoll in einer späteren Ausbaustufe umgesetzt werden.

FA-017 – Responsive Design für KI-Chat

Anforderungs- beschreibung	Der KI-Chat soll ein Responsive Design bieten, das eine Nutzung auf Desktop-Computern, Tablets und Smartphones unterstützt.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist im Rahmen der produktiven Erstimplementierung auf Basis der vorhandenen Frontend-Architektur und UI-Komponenten des Urban Stack bereitzustellen. Der Schwerpunkt liegt auf einer praxistauglichen Nutzung für Desktop und Tablet; Smartphone-Optimierungen können angemessen reduziert umgesetzt werden, sofern die grundsätzliche Nutzbarkeit gewährleistet bleibt.
Begründung	Ein responsives Layout ist für eine moderne Verwaltungsanwendung erforderlich. Durch Nutzung vorhandener UI-Komponenten kann die Umsetzung wirtschaftlich erfolgen.

4.2 KI-Agenten-Plattform

Die KI-Agenten-Plattform ermöglicht die Erstellung, Konfiguration, Verwaltung und Nutzung spezialisierter KI-Agenten innerhalb des Urban KI Hub. Die Erstimplementierung umfasst die Bereitstellung einer produktiv nutzbaren Agentenplattform mit zwei Referenzagenten. Erweiterte Komfortfunktionen und organisatorische Funktionen werden architektonisch vorbereitet oder in spätere Ausbaustufen eingeordnet.

FA-018 – Agentenerstellung

Anforderungsbeschreibung	Die Plattform muss die Erstellung und Konfiguration von KI-Agenten ermöglichen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist im Rahmen der produktiven Erstimplementierung vollständig bereitzustellen. Agenten müssen über die Administrationsoberfläche erstellt und verwaltet werden können.
Begründung	Die Erstellung eigener KI-Agenten gehört zu den Kernfunktionen der Plattform und ist Voraussetzung für den produktiven Einsatz in den Fachbereichen.

FA-019 – Konfigurierbarkeit

Anforderungsbeschreibung	KI-Agenten müssen hinsichtlich verwendeter Modelle, Prompts, Wissensquellen und Werkzeuge konfigurierbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Konfiguration erfolgt über die bereitgestellte Administrationsoberfläche. Die Architektur muss eine spätere Erweiterung um zusätzliche Konfigurationsparameter ermöglichen.
Begründung	Erst durch die flexible Konfiguration können Agenten an unterschiedliche Verwaltungsaufgaben angepasst werden.

FA-020 – Agent Runtime

Anforderungs- beschreibung	Die Ausführung sämtlicher KI-Agenten muss über eine zentrale Agent Runtime erfolgen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Agent Runtime ist als Bestandteil des AI Harness bereitzustellen und übernimmt die Ausführung, Steuerung und Protokollierung sämtlicher Agentenprozesse.
Begründung	Eine zentrale Agent Runtime gewährleistet einheitliche Abläufe, Governance, Sicherheit und Skalierbarkeit.

FA-021 – Mandantenfähigkeit

Anforderungs- beschreibung	KI-Agenten müssen mandantenfähig verwaltet und ausgeführt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Mandanteninformationen müssen bei Konfiguration, Ausführung, Berechtigungen und Protokollierung vollständig berücksichtigt werden.
Begründung	Die Mandantenfähigkeit gehört zu den grundlegenden Architekturprinzipien des Urban Stack und muss bereits in der Erstimplementierung vollständig umgesetzt werden.

FA-022 – Wiederverwendbarkeit

Anforderungs- beschreibung	KI-Agenten sollen organisationsweit wiederverwendbar sein.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Architektur muss die spätere organisationsweite Bereitstellung gemeinsamer Agenten unterstützen. In der Erstimplementierung genügt die Nutzung innerhalb des jeweiligen Mandanten.
Begründung	Die organisationsweite Wiederverwendung erhöht langfristig die Effizienz, ist jedoch für den initialen Produktivbetrieb nicht zwingend erforderlich.

FA-023 – Agenten-Metadaten

Anforderungsbeschreibung	Jeder KI-Agent muss über einen konfigurierbaren Namen, eine Beschreibung, ein Icon oder einen Avatar sowie Kategorien und Tags zur besseren Auffindbarkeit verfügen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Metadaten sind über die Administrationsoberfläche pflegbar bereitzustellen und bei der Auswahl von Agenten anzuzeigen.
Begründung	Einheitliche Metadaten erleichtern die Verwaltung und Nutzung von Agenten und schaffen eine konsistente Benutzerführung.

FA-024 – Rollen- und Berechtigungskonzept

Anforderungsbeschreibung	Die Plattform muss die rollen- und berechtigungsabhängige Nutzung, Verwaltung und Konfiguration von KI-Agenten unterstützen. Es muss konfigurierbar sein, welche Nutzer oder Nutzergruppen Agenten erstellen, bearbeiten, veröffentlichen oder verwenden dürfen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Das Rollen- und Berechtigungskonzept ist auf Basis der vorhandenen Authentifizierungs- und Autorisierungsinfrastruktur des Urban Stack umzusetzen.
Begründung	KI-Agenten können unterschiedliche Berechtigungen und Wissensquellen nutzen. Eine rollenbasierte Steuerung ist daher Voraussetzung für einen sicheren Produktivbetrieb.

FA-025 – Veröffentlichung von Agenten

Anforderungsbeschreibung	Die Plattform muss die Veröffentlichung von KI-Agenten ermöglichen. Agenten müssen als Entwurf, freigegeben oder deaktiviert verwaltet werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Veröffentlichung erfolgt über definierte Freigabeprozesse innerhalb der Agentenverwaltung.
Begründung	Durch definierte Veröffentlichungsprozesse wird sichergestellt, dass ausschließlich geprüfte Agenten produktiv genutzt werden.

FA-026 – Agentenversionierung

Anforderungs- beschreibung	Änderungen an KI-Agenten müssen versioniert und nachvollziehbar dokumentiert werden. Frühere Versionen müssen wiederhergestellt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform muss Versionen einschließlich Änderungszeitpunkt, Bearbeiter und Freigabestatus verwalten.
Begründung	Versionierung gewährleistet Nachvollziehbarkeit, Auditierbarkeit und eine kontrollierte Weiterentwicklung produktiver Agenten.

FA-027 – Tool-Nutzung

Anforderungs- beschreibung	KI-Agenten müssen definierte Werkzeuge (Tools) zur Bearbeitung ihrer Aufgaben nutzen können. Die Nutzung erfolgt ausschließlich über den AI Harness.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Tool-Anbindung erfolgt über standardisierte Schnittstellen. Zugriffe sind rollen- und richtlinienbasiert zu steuern.
Begründung	Die kontrollierte Nutzung von Werkzeugen erweitert den Funktionsumfang der Agenten und gewährleistet gleichzeitig Sicherheit und Governance.

FA-028 – Wissensquellen

Anforderungs- beschreibung	KI-Agenten müssen auf definierte Wissensquellen zugreifen können. Die Zuordnung erfolgt agentenspezifisch und unter Berücksichtigung von Rollen, Mandanten und Berechtigungen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Wissensquellen werden über die Retrieval-Komponente verwaltet und dem Agenten konfigurierbar zugeordnet.
Begründung	Unterschiedliche Agenten benötigen unterschiedliche Wissensgrundlagen. Eine getrennte Verwaltung erhöht Sicherheit und Antwortqualität.

FA-029 – Prompt-Konfiguration

Anforderungsbeschreibung	Jedem KI-Agenten muss mindestens ein konfigurierbarer System-Prompt zugeordnet werden können. Änderungen an Prompts müssen versioniert und nachvollziehbar dokumentiert werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Prompt-Verwaltung erfolgt zentral über die Prompt Registry des AI Harness.
Begründung	System-Prompts bestimmen das Verhalten eines Agenten maßgeblich und müssen daher zentral verwaltet und versioniert werden.

FA-030 – Modellauswahl

Anforderungsbeschreibung	Für jeden KI-Agenten muss das zu verwendende Sprachmodell konfigurierbar sein. Die Auswahl erfolgt über die Modellabstraktions- und Routing-Schicht.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform muss die Auswahl unterschiedlicher Modelle ermöglichen, ohne dass Agenten angepasst werden müssen.
Begründung	Die modellunabhängige Architektur ist ein zentrales Ziel der Plattform und ermöglicht zukünftige Modellwechsel ohne Anpassung der Agentenlogik.

FA-031 – Referenzagenten

Anforderungsbeschreibung	Im Rahmen der Erstimplementierung sind mindestens zwei produktiv nutzbare Referenzagenten bereitzustellen. Die Referenzagenten dienen als Grundlage für die spätere Entwicklung weiterer Fachagenten.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Referenzagenten werden durch den Auftragnehmer implementiert und vollständig dokumentiert.
Begründung	Die Konzentration auf zwei Referenzagenten reduziert den Projektumfang erheblich und schafft gleichzeitig eine belastbare Grundlage für spätere Erweiterungen.

FA-032 – Erweiterbarkeit der Agentenplattform

Anforderungs- beschreibung	Die Agentenplattform muss so aufgebaut sein, dass zusätzliche Agenten, Werkzeuge, Wissensquellen und Ausführungslogiken ohne grundlegende Änderungen an der Gesamtarchitektur integriert werden können.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Architektur muss standardisierte Erweiterungspunkte, dokumentierte Schnittstellen und modulare Konfigurationsmechanismen bereitstellen.
Begründung	Ziel der Ausschreibung ist eine langfristig erweiterbare Plattform. Die vollständige Umsetzung weiterer Agententypen erfolgt in späteren Ausbaustufen, die Architektur muss dies jedoch bereits unterstützen.

4.3 Retrieval, RAG und Wissensmanagement

Zielsetzung

Die Retrieval- und Wissensmanagement-Komponente bildet die Grundlage für eine vertrauenswürdige Nutzung verwaltungsinterner Informationen. Ziel ist die Bereitstellung einer leistungsfähigen Retrieval-Augmented-Generation-Architektur (RAG), die Informationen aus freigegebenen Wissensquellen auffindbar macht und KI-Modelle mit belastbaren Kontextinformationen versorgt.

Die Erstimplementierung konzentriert sich auf die Integration der vorhandenen OpenSearch-Infrastruktur als zentrale Retrieval-Komponente sowie auf die Anbindung der Dokumentenpipeline. Erweiterte Retrieval-Verfahren und zusätzliche Wissensquellen werden architektonisch vorbereitet.

Architekturgrundsatz

Die Plattform nutzt OpenSearch als zentrale Retrieval- und Vektordatenbank. Dokumente werden ausschließlich über die definierte Dokumentenpipeline in den Wissensbestand übernommen. Die Context Engine stellt sicher, dass ausschließlich berechnete und relevante Informationen an KI-Modelle übergeben werden. Sämtliche Retrieval-Prozesse werden durch den AI Harness orchestriert.

FA-033 – Retrieval-Augmented Generation (RAG)

Anforderungsbeschreibung	Die Plattform muss Retrieval-Augmented Generation (RAG) unterstützen und externe Wissensquellen vor der Antwortgenerierung in den Modellkontext integrieren können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die RAG-Funktion ist vollständig bereitzustellen und auf Basis der vorhandenen OpenSearch-Infrastruktur umzusetzen.
Begründung	RAG ist die zentrale Funktion zur Nutzung verwaltungsinterner Wissensbestände und reduziert das Risiko unbelegter Antworten erheblich.

FA-034 – OpenSearch als Retrieval-Komponente

Anforderungsbeschreibung	Die Plattform muss OpenSearch als zentrale Retrieval-, Such- und Vektordatenbank integrieren und für semantische Suche, Volltextsuche, hybride Suche sowie Metadatenfilterung nutzen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die vorhandene OpenSearch-Installation ist zu integrieren. Die Leistung umfasst ausschließlich Konfiguration, Datenmodellierung und Anbindung; Aufbau oder Betrieb der Infrastruktur sind nicht Bestandteil der Leistung.
Begründung	Die Nutzung der vorhandenen Infrastruktur reduziert den Implementierungsaufwand und stellt eine wirtschaftliche Umsetzung sicher.

FA-035 – Dokumentenpipeline

Anforderungsbeschreibung	Dokumente müssen über die definierte Dokumentenpipeline verarbeitet, aufbereitet und für Retrieval-Prozesse verfügbar gemacht werden. In der Erstimplementierung sind ausschließlich Dokumente aus dem vorhandenen S3-Dokumentenspeicher sowie über definierte REST-Schnittstellen zu berücksichtigen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Dokumentenpipeline ist für S3 und REST bereitzustellen. Weitere Dokumentenquellen müssen architektonisch vorbereitet, jedoch nicht produktiv umgesetzt werden.
Begründung	Die Begrenzung auf S3 und REST reduziert den Projektumfang erheblich, ohne die spätere Erweiterbarkeit einzuschränken.

FA-036 – Chunking und Dokumentstruktur

Anforderungs- beschreibung	Die Plattform muss Dokumente automatisiert strukturieren, in geeignete Verarbeitungseinheiten (Chunks) zerlegen und Metadaten für Retrieval-Prozesse bereitstellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Chunking, Metadatenextraktion und Indexierung sind Bestandteil der Dokumentenpipeline.
Begründung	Eine geeignete Dokumentenstruktur ist Voraussetzung für qualitativ hochwertige Retrieval-Ergebnisse.

FA-037 – Berechtigungsabhängiges Retrieval

Anforderungs- beschreibung	Die Plattform muss sicherstellen, dass ausschließlich Dokumente und Informationen berücksichtigt werden, für die der jeweilige Nutzer über die erforderlichen Berechtigungen verfügt.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Berechtigungen sind während des Retrieval-Prozesses mandanten-, rollen- und nutzerbezogen zu berücksichtigen.
Begründung	Datenschutz und Informationssicherheit erfordern eine konsequente Zugriffskontrolle bereits während des Retrievals.

FA-038 – Hybrid Search

Anforderungs- beschreibung	Die Plattform muss semantische Suche und klassische Volltextsuche kombinieren können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Hybrid Search ist auf Basis der vorhandenen OpenSearch-Funktionalitäten bereitzustellen.
Begründung	Die Kombination beider Suchverfahren verbessert die Qualität der Suchergebnisse erheblich.

FA-039 – Metadatenfilter

Anforderungs- beschreibung	Retrieval-Prozesse müssen Dokumente anhand von Metadaten filtern können, insbesondere nach Quelle, Dokumenttyp, Organisationseinheit, Mandant, Datum und Berechtigungen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die erforderlichen Metadaten sind im Dokumentenindex vorzuhalten und für Suchanfragen nutzbar zu machen.
Begründung	Metadaten verbessern die Präzision der Suchergebnisse und ermöglichen eine zielgerichtete Nutzung verwaltungsinterner Informationen.

FA-040 – Erweiterbare Wissensquellen

Anforderungs- beschreibung	Die Architektur muss die spätere Anbindung zusätzlicher Wissensquellen unterstützen, insbesondere Fachverfahren, SharePoint, Dateisysteme, DMS oder weitere Dokumentenplattformen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die erforderlichen Schnittstellen und Erweiterungspunkte sind vorzusehen. Eine produktive Anbindung zusätzlicher Quellen ist nicht Bestandteil der Erstimplementierung.
Begründung	Die Konzentration auf S3 und REST reduziert den Projektumfang erheblich und erhält gleichzeitig die langfristige Erweiterbarkeit.

FA-041 – Dokumentenchat

Anforderungs- beschreibung	Die Plattform muss die dialogbasierte Interaktion mit einzelnen Dokumenten unterstützen. Nutzer sollen Fragen zu hochgeladenen Dokumenten stellen können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Funktion ist auf Basis der Dokumentenpipeline und der Retrieval-Komponente bereitzustellen.
Begründung	Der Dokumentenchat gehört zu den wichtigsten fachlichen Anwendungsfällen der Erstimplementierung.

FA-042 – Erweiterbare Retrieval-Architektur

Anforderungs- beschreibung	Die Retrieval-Architektur muss modular aufgebaut sein und zukünftige Erweiterungen, insbesondere GraphRAG, Knowledge Graphs, zusätzliche Ranking-Verfahren oder weitere Retrieval-Technologien unterstützen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Architektur muss standardisierte Erweiterungspunkte und modulare Retrieval-Komponenten bereitstellen. Eine vollständige Umsetzung dieser Verfahren ist nicht Bestandteil der Erstimplementierung.
Begründung	Die modulare Architektur gewährleistet Investitionssicherheit und ermöglicht die schrittweise Weiterentwicklung der Plattform ohne grundlegende Architekturänderungen.

4.4 Erweiterte Recherche (Deep Research)

Zielsetzung

Die Plattform soll die strukturierte Recherche über interne Wissensquellen sowie externe Informationsquellen unterstützen. Ziel ist die schrittweise Bereitstellung einer leistungsfähigen Recherchefunktion, die Informationen automatisiert sammelt, bewertet, zusammenführt und nachvollziehbar dokumentiert.

Die produktive Erstimplementierung konzentriert sich auf einen einfachen Recherche-Agenten mit dokumentierter Quellenverarbeitung. Erweiterte autonome Recherchefunktionen, mehrstufige Analyseprozesse und komplexe Recherche-Workflows werden architektonisch vorbereitet und können in späteren Ausbaustufen ergänzt werden.

Architekturgrundsatz

Deep Research ist Bestandteil des AI Harness und wird über die Agent Runtime ausgeführt. Rechercheprozesse nutzen die Context Engine, Retrieval-Komponenten, angebundene Werkzeuge sowie definierte Wissensquellen. Sämtliche Rechercheergebnisse müssen nachvollziehbar dokumentiert und mit Quellen versehen werden.

FA-043 – Recherche-Agent

Anforderungs- beschreibung	Die Plattform muss mindestens einen KI-Agenten bereitstellen, der strukturierte Recherchen über freigegebene Wissensquellen durchführen und die Ergebnisse zusammenfassen kann.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Im Rahmen der Erstimplementierung ist mindestens ein produktiv nutzbarer Recherche-Agent bereitzustellen. Die Recherche erfolgt auf Basis der angebundenen Wissensquellen und Retrieval-Komponenten.
Begründung	Ein Recherche-Agent stellt einen wesentlichen Mehrwert für die Verwaltung dar und bildet die Grundlage für spätere Deep-Research-Funktionen.

FA-044 – Quellenbasierte Recherche

Anforderungs- beschreibung	Rechercheergebnisse müssen die verwendeten Quellen nachvollziehbar dokumentieren und referenzieren.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Sämtliche Rechercheergebnisse sind mit den zugrunde liegenden Dokumenten oder Informationsquellen zu verknüpfen.
Begründung	Quellenangaben sind Voraussetzung für nachvollziehbare und belastbare Rechercheergebnisse im Verwaltungskontext.

FA-045 – Mehrstufige Recherche

Anforderungs- beschreibung	Die Architektur muss mehrstufige Rechercheprozesse unterstützen, bei denen mehrere Recherche-, Analyse- und Zusammenführungsschritte automatisiert ausgeführt werden können.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Agent Runtime und der AI Harness müssen entsprechende Erweiterungspunkte bereitstellen. Eine vollständige Umsetzung ist nicht Bestandteil der Erstimplementierung.
Begründung	Mehrstufige Deep-Research-Prozesse erhöhen den Implementierungsaufwand erheblich und können wirtschaftlich sinnvoll in einer späteren Ausbaustufe umgesetzt werden.

FA-046 – Externe Informationsquellen

Anforderungs- beschreibung	
Kategorie	A – Architekturvorbereitung
Umsetzung	Die erforderlichen Schnittstellen und Erweiterungspunkte sind vorzusehen. Eine produktive Nutzung externer Informationsquellen ist nicht Bestandteil der Erstimplementierung.
Begründung	Die Konzentration auf interne Wissensquellen reduziert Komplexität, Datenschutzrisiken und Projektkosten, ohne die langfristige Erweiterbarkeit einzuschränken.

FA-047 – Rechercheprotokollierung

Anforderungs- beschreibung	Rechercheprozesse müssen nachvollziehbar protokolliert werden. Insbesondere sind verwendete Quellen, durchgeführte Verarbeitungsschritte und erzeugte Ergebnisse revisionssicher zu dokumentieren.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Protokollierung erfolgt über die Audit- und Monitoring-Komponenten des AI Harness.
Begründung	Die Nachvollziehbarkeit von Rechercheprozessen ist für Transparenz, Governance und Revisionssicherheit erforderlich.

FA-048 – Erweiterbare Deep-Research-Architektur

Anforderungs- beschreibung	Die Architektur muss so gestaltet sein, dass zukünftige Deep-Research-Funktionen, autonome Recherche-Agenten, Planungsalgorithmen sowie komplexe Recherche-Workflows ohne grundlegende Änderungen an der Gesamtarchitektur integriert werden können.
Kategorie	A – Architekturvorbereitung
Umsetzung	Der AI Harness, die Agent Runtime und die Context Engine müssen modulare Erweiterungspunkte für zukünftige Deep-Research-Funktionen bereitstellen.
Begründung	Die modulare Architektur gewährleistet Investitionssicherheit und ermöglicht den späteren Ausbau der Recherchefunktionen entsprechend dem Stand der Technik.

4.5 Bildgenerierung

Zielsetzung

Die Plattform soll perspektivisch die Erstellung und Bearbeitung von Bildern durch generative KI-Modelle unterstützen. Ziel ist eine modulare Architektur, die zukünftig die Nutzung verschiedener Bildgenerierungsmodelle ermöglicht, ohne den zentralen AI Harness oder die Gesamtarchitektur grundlegend anpassen zu müssen.

Die Bildgenerierung gehört nicht zum verpflichtenden Funktionsumfang der produktiven Erstimplementierung. Die Architektur ist jedoch so vorzubereiten, dass entsprechende Modelle und Dienste später integriert werden können.

Architekturgrundsatz

Die Bildgenerierung ist als optionaler KI-Service innerhalb der Gesamtarchitektur vorzusehen. Die Anbindung erfolgt über den AI Harness sowie die Modellabstraktions- und Routing-Schicht. Bildgenerierungsmodelle müssen wie Sprachmodelle über standardisierte Schnittstellen integriert werden können.

FA-049 – Architektur für Bildgenerierung

Anforderungsbeschreibung	Die Plattform muss die spätere Integration generativer Bildmodelle über standardisierte Schnittstellen unterstützen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die erforderlichen Schnittstellen und Erweiterungspunkte innerhalb des AI Harness sowie der Modellabstraktionsschicht sind bereitzustellen. Eine produktive Bildgenerierung ist nicht Bestandteil der Erstimplementierung.
Begründung	Die Architektur bleibt zukunftssicher, ohne zusätzliche GPU-Ressourcen oder Modellbetrieb bereits in der Erstimplementierung zu erfordern.

FA-050 – Austauschbare Bildmodelle

Anforderungsbeschreibung	Die Plattform muss den modellunabhängigen Betrieb unterschiedlicher Bildgenerierungsmodelle unterstützen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Modellabstraktionsschicht muss die spätere Einbindung unterschiedlicher Bildmodelle ermöglichen.
Begründung	Die Plattform soll langfristig unabhängig von einzelnen Herstellern oder Modellen bleiben.

FA-051 – Bildbearbeitung

Anforderungsbeschreibung	Die Architektur soll die spätere Unterstützung KI-gestützter Bildbearbeitungsfunktionen, insbesondere Inpainting, Outpainting, Hintergrundentfernung und Bildvariationen ermöglichen.
Kategorie	E – Erweiterung der Zielarchitektur
Umsetzung	Die notwendigen Erweiterungspunkte sind vorzusehen. Eine produktive Umsetzung ist nicht Bestandteil der Erstimplementierung.
Begründung	Diese Funktionen stellen Komfort- und Spezialfunktionen dar und sind für den initialen Produktivbetrieb nicht erforderlich.

FA-052 – Governance für Bildgenerierung

Anforderungsbeschreibung	Die Plattform muss sicherstellen, dass zukünftig integrierte Bildgenerierungsmodelle in die bestehenden Governance-, Audit-, Policy- und Sicherheitsmechanismen des AI Harness eingebunden werden können.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Architektur muss eine einheitliche Protokollierung, Richtlinienprüfung und Zugriffskontrolle auch für Bildgenerierungsdienste unterstützen.
Begründung	Die Einbindung in bestehende Governance-Strukturen gewährleistet einen konsistenten und sicheren Betrieb zukünftiger Bildgenerierungsfunktionen.

4.6 Prompt Management (Prompt Registry)

Zielsetzung

Das Prompt Management dient der zentralen Verwaltung, Versionierung und Bereitstellung von Prompts innerhalb der KI-Plattform. Ziel ist eine einheitliche, nachvollziehbare und wiederverwendbare Verwaltung von System-, Benutzer- und Agenten-Prompts.

Die produktive Erstimplementierung konzentriert sich auf die Bereitstellung einer zentralen Prompt Registry einschließlich Versionierung, Freigabe und Rollensteuerung. Erweiterte Funktionen wie Prompt-Optimierung oder automatische Prompt-Generierung werden architektonisch vorbereitet.

Architekturgrundsatz

Die Prompt Registry ist Bestandteil des AI Harness. Sämtliche KI-Agenten, Dialoge und Workflows greifen ausschließlich auf zentral verwaltete Prompt-Versionen zu. Änderungen an Prompts müssen versioniert, nachvollziehbar dokumentiert und kontrolliert freigegeben werden.

FA-053 – Zentrale Prompt Registry

Anforderungsbeschreibung	Die Plattform muss eine zentrale Prompt Registry zur Verwaltung aller System-, Agenten- und Standard-Prompts bereitstellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Prompt Registry ist als Bestandteil des AI Harness bereitzustellen und über die Administrationsoberfläche verwaltbar auszuführen.
Begründung	Eine zentrale Verwaltung verhindert redundante Prompt-Versionen und gewährleistet konsistente KI-Ergebnisse.

FA-054 – Versionierung von Prompts

Anforderungsbeschreibung	Änderungen an Prompts müssen versioniert, historisiert und nachvollziehbar dokumentiert werden. Frühere Versionen müssen wiederhergestellt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Versionen einschließlich Änderungsdatum, Bearbeiter und Freigabestatus sind dauerhaft zu verwalten.
Begründung	Versionierung ist Voraussetzung für Governance, Auditierung und reproduzierbare KI-Ergebnisse.

FA-055 – Rollen- und Freigabekonzept

Anforderungsbeschreibung	Die Plattform muss ein rollenbasiertes Freigabeverfahren für Prompts unterstützen. Es muss steuerbar sein, welche Nutzer Prompts erstellen, bearbeiten, freigeben oder produktiv verwenden dürfen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Freigabe erfolgt über definierte Rollen und Berechtigungen unter Nutzung der vorhandenen Authentifizierungs- und Autorisierungsinfrastruktur.
Begründung	Freigabeprozesse verhindern unkontrollierte Änderungen produktiver Prompts und erhöhen die Qualität der Plattform.

FA-056 – Kategorien und Metadaten

Anforderungs- beschreibung	Prompts müssen mit Metadaten versehen werden können, insbesondere Name, Beschreibung, Kategorie, Anwendungsbereich, Version und Verantwortlichkeit.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Metadaten sind zentral zu verwalten und für Suche sowie Wiederverwendung bereitzustellen.
Begründung	Einheitliche Metadaten erleichtern die Verwaltung wachsender Prompt-Bestände erheblich.

FA-057 – Parametrisierbare Prompts

Anforderungs- beschreibung	Die Plattform muss die Verwendung parametrisierbarer Prompt-Vorlagen unterstützen. Variablen sollen zur Laufzeit durch den AI Harness ersetzt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Prompt-Vorlagen müssen Platzhalter und strukturierte Variablen unterstützen. Die Ersetzung erfolgt innerhalb des AI Harness unmittelbar vor dem Modellaufruf.
Begründung	Parametrisierte Prompts vermeiden Redundanzen und ermöglichen die Wiederverwendung standardisierter Prompt-Strukturen.

FA-058 – Wiederverwendung von Prompts

Anforderungs- beschreibung	Prompts müssen mehreren KI-Agenten, Workflows und Dialogen gemeinsam zugeordnet und wiederverwendet werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Prompt Registry muss referenzierte Prompts unterstützen, sodass Änderungen zentral wirksam werden können.
Begründung	Wiederverwendbare Prompts reduzieren Pflegeaufwand und gewährleisten einheitliches Verhalten verschiedener KI-Komponenten.

FA-059 – Erweiterbarkeit der Prompt Registry

Anforderungs- beschreibung	Die Architektur muss zukünftige Erweiterungen des Prompt Managements unterstützen, insbesondere automatische Prompt-Optimierung, Prompt-Tests, Prompt-Evaluierung sowie KI-gestützte Prompt-Generierung.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die erforderlichen Erweiterungspunkte sind innerhalb der Prompt Registry vorzusehen. Eine produktive Umsetzung dieser Funktionen ist nicht Bestandteil der Erstimplementierung.
Begründung	Moderne Prompt-Engineering-Verfahren entwickeln sich schnell. Eine vorbereitete Architektur ermöglicht zukünftige Erweiterungen ohne grundlegende Änderungen der Plattform.

4.7 Zeitgesteuerte Aufgaben

Zielsetzung

Die Plattform soll die zeitgesteuerte Ausführung von KI-Funktionen unterstützen. Ziel ist die automatische Ausführung wiederkehrender KI-Aufgaben, Agenten oder Prompt-basierter Prozesse zu definierten Zeitpunkten oder nach einem festgelegten Zeitplan.

Die produktive Erstimplementierung konzentriert sich auf einfache zeitgesteuerte Ausführungen einzelner KI-Agenten und Prompts. Komplexe Workflow-Orchestrierungen sowie ereignisgesteuerte Prozessketten werden durch bestehende Komponenten des Urban Stack oder City Tools bereitgestellt und sind nicht Bestandteil dieser Leistung.

Architekturgrundsatz

Zeitgesteuerte Aufgaben werden über den AI Harness ausgeführt. Die Planung und Ausführung erfolgt über standardisierte Scheduler-Schnittstellen. Externe Workflow- oder Automatisierungsplattformen können diese Funktionen über definierte APIs ansteuern.

FA-060 – Zeitgesteuerte Ausführung

Anforderungs- beschreibung	Die Plattform muss die zeitgesteuerte Ausführung von KI-Agenten und Prompt-basierten Aufgaben unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Es sind einfache einmalige und wiederkehrende Ausführungen auf Basis definierter Zeitpläne bereitzustellen.
Begründung	Zeitgesteuerte Aufgaben ermöglichen die Automatisierung wiederkehrender Verwaltungsprozesse ohne zusätzliche Workflow-Logik.

FA-061 – Wiederkehrende Aufgaben

Anforderungsbeschreibung	Die Plattform muss wiederkehrende Ausführungen nach konfigurierbaren Intervallen unterstützen, insbesondere täglich, wöchentlich, monatlich oder nach Cron-ähnlichen Zeitplänen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Zeitplanung erfolgt über den AI Harness oder eine gleichwertige Scheduler-Komponente.
Begründung	Wiederkehrende Aufgaben gehören zu den grundlegenden Automatisierungsfunktionen einer modernen KI-Plattform.

FA-062 – Benachrichtigungen

Anforderungsbeschreibung	Nach Abschluss einer zeitgesteuerten Aufgabe muss die Plattform Ergebnisse oder Statusinformationen bereitstellen können. Die Benachrichtigung kann über die Benutzeroberfläche oder angebundene Systeme erfolgen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die erforderlichen Schnittstellen für Benachrichtigungen sind vorzusehen. Die Integration weiterer Kommunikationskanäle erfolgt in späteren Ausbaustufen.
Begründung	Benachrichtigungen erhöhen den Nutzen zeitgesteuerter Aufgaben, sind jedoch für den initialen Produktivbetrieb nicht zwingend erforderlich.

FA-063 – Integration in Workflow- und Event-Systeme

Anforderungsbeschreibung	Die Plattform muss standardisierte Schnittstellen bereitstellen, über die zeitgesteuerte Aufgaben durch externe Workflow- oder Event-Systeme ausgelöst werden können. Eigene Workflow- oder Event-Engines sind nicht Bestandteil der Leistung.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Integration erfolgt über dokumentierte REST-APIs, Webhooks oder funktional gleichwertige Schnittstellen. Die vorhandenen Workflow-Komponenten des Urban Stack bzw. City Tools sind zu nutzen.
Begründung	Die Trennung zwischen KI-Plattform und Workflow-Automatisierung reduziert den Projektumfang, vermeidet Doppelentwicklungen und unterstützt die modulare Gesamtarchitektur.

4.8 Personalisierung

Zielsetzung

Die Plattform soll eine personalisierte Nutzung der KI-Funktionen ermöglichen. Ziel ist es, Benutzereinstellungen, Präferenzen und individuelle Arbeitskontexte zu berücksichtigen, ohne die Mandantenfähigkeit, Datenschutzanforderungen oder organisatorische Vorgaben zu beeinträchtigen.

Die produktive Erstimplementierung konzentriert sich auf grundlegende Personalisierungsfunktionen. Erweiterte Personalisierungsmechanismen werden über das Memory-Konzept sowie die Context Engine architektonisch vorbereitet.

Architekturgrundsatz

Personalisierungsinformationen werden zentral über User Memory, Tenant Memory und die Context Engine verwaltet. Die Plattform muss sicherstellen, dass personenbezogene Einstellungen ausschließlich unter Berücksichtigung von Rollen, Berechtigungen, Datenschutz und Mandantentrennung genutzt werden.

FA-064 – Benutzerpräferenzen

Anforderungs- beschreibung	Die Plattform muss die Speicherung und Nutzung grundlegender Benutzereinstellungen unterstützen. Hierzu gehören insbesondere bevorzugte Sprache, Standardmodell, bevorzugte Antwortformate sowie weitere persönliche Einstellungen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Benutzereinstellungen sind mandantenbezogen zu speichern und bei der Nutzung der Plattform automatisch zu berücksichtigen.
Begründung	Benutzerpräferenzen verbessern die Nutzerfreundlichkeit und reduzieren wiederkehrende Konfigurationsaufwände im täglichen Einsatz.

FA-065 – Personalisierte Startansicht

Anforderungs- beschreibung	Die Plattform soll personalisierte Inhalte auf der Startseite bereitstellen können, insbesondere zuletzt genutzte Chats, Agenten, Prompts und Dokumente.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Startansicht nutzt vorhandene Nutzungsinformationen sowie Benutzerpräferenzen.
Begründung	Eine personalisierte Startseite erleichtert den Einstieg in häufig genutzte Arbeitsabläufe und erhöht die Benutzerfreundlichkeit.

FA-066 – Nutzung von User Memory

Anforderungs- beschreibung	Die Plattform muss User Memory zur Speicherung wiederkehrender Präferenzen und Arbeitskontexte unterstützen. Die Nutzung personenbezogener Informationen muss transparent erfolgen und durch den Nutzer steuerbar sein.
Kategorie	A – Architekturvorbereitung
Umsetzung	Für die Erstimplementierung genügt die Vorbereitung der erforderlichen Datenmodelle und Schnittstellen. Eine vollständige Umsetzung erfolgt in Verbindung mit dem differenzierten Memory-Konzept.
Begründung	User Memory stellt einen wesentlichen Baustein moderner KI-Systeme dar, erfordert jedoch erhöhte Anforderungen an Datenschutz, Transparenz und Nutzerkontrolle.

FA-067 – Personalisierte Kontextbildung

Anforderungs- beschreibung	Die Context Engine muss in der Lage sein, freigegebene Benutzerpräferenzen, Rollen, Mandanteninformationen und weitere zulässige Kontextinformationen bei der Zusammenstellung des Modellkontexts zu berücksichtigen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die erforderlichen Erweiterungspunkte innerhalb der Context Engine sind vorzusehen. Eine vollständige Nutzung aller Memory-Ebenen ist nicht Bestandteil der Erstimplementierung.
Begründung	Personalisierte Kontextbildung verbessert die Qualität der KI-Ausgaben erheblich und stellt gleichzeitig sicher, dass ausschließlich zulässige Informationen verwendet werden.

FA-068 – Erweiterbare Personalisierungsarchitektur

Anforderungs- beschreibung	Die Architektur muss zukünftige Personalisierungsfunktionen unterstützen, insbesondere adaptive Benutzeroberflächen, lernende Präferenzen sowie organisationsspezifische Personalisierungsmechanismen.
Kategorie	E – Erweiterung der Zielarchitektur
Umsetzung	Die erforderlichen Erweiterungspunkte innerhalb des User Memory und der Context Engine sind vorzusehen. Eine produktive Umsetzung ist nicht Bestandteil der Erstimplementierung.
Begründung	Erweiterte Personalisierungsfunktionen bieten langfristig zusätzlichen Mehrwert, gehören jedoch nicht zum verpflichtenden Leistungsumfang der Erstimplementierung.

4.9 ML-Services API und Umfang des Loses ML-Services

Zielsetzung

Die ML-Services ergänzen den zentralen KI-Stack um spezialisierte Machine-Learning-Funktionen. Sie werden als eigenständige, lose gekoppelte Dienste bereitgestellt und ausschließlich über standardisierte Schnittstellen durch den AI Harness genutzt.

Die ML-Services stellen keine eigenständige Plattform dar. Sie übernehmen ausschließlich spezialisierte KI- und ML-Aufgaben, die von Sprachmodellen oder der Retrieval-Komponente nicht oder nur eingeschränkt erfüllt werden können.

Die produktive Erstimplementierung konzentriert sich auf die Bereitstellung weniger, fachlich relevanter ML-Services mit klar definierten REST-APIs.

Architekturgrundsatz

Die ML-Services werden ausschließlich über den AI Harness angesprochen. Sie besitzen keine eigene Benutzeroberfläche, keine eigene Workflow-Orchestrierung und keine eigenständige Agentenlogik.

Die Steuerung erfolgt vollständig durch den AI Harness oder angebundene Workflow-Systeme des Urban Stack. Die Kommunikation erfolgt ausschließlich über dokumentierte REST-APIs oder funktional gleichwertige standardisierte Schnittstellen.

FA-069 – API-basierte Bereitstellung

Anforderungsbeschreibung	Sämtliche ML-Services müssen als eigenständige API-Dienste bereitgestellt werden und ausschließlich über standardisierte Schnittstellen angesprochen werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Kommunikation erfolgt über REST-APIs oder funktional gleichwertige Schnittstellen mit strukturierten Datenformaten.
Begründung	Eine API-basierte Architektur ermöglicht die lose Kopplung der ML-Services sowie deren Wiederverwendung innerhalb des gesamten Urban Stack.

FA-070 – Dokumentenanalyse

Anforderungsbeschreibung	Die Plattform muss einen ML-Service zur Analyse und Strukturierung von Dokumenten bereitstellen. Hierzu gehören insbesondere Dokumentenstrukturierung, Layout-Erkennung und Inhaltsaufbereitung.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Umsetzung soll bevorzugt unter Verwendung etablierter Open-Source-Komponenten wie Docling oder funktional gleichwertiger Lösungen erfolgen.
Begründung	Die Dokumentenanalyse bildet die Grundlage der Dokumentenpipeline und des Dokumentenchats.

FA-071 – OCR-Service

Anforderungsbeschreibung	Die Plattform muss einen OCR-Service zur Texterkennung in gescannten Dokumenten und Bildern bereitstellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Umsetzung soll bevorzugt unter Nutzung etablierter Open-Source-Komponenten wie PaddleOCR oder Tesseract erfolgen.
Begründung	OCR gehört zu den wichtigsten Verwaltungsanwendungen und stellt eine Kernfunktion der Dokumentenverarbeitung dar.

FA-072 – Speech-to-Text

Anforderungsbeschreibung	Die Plattform muss einen Service zur automatischen Sprachtranskription bereitstellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Umsetzung soll bevorzugt unter Verwendung datenschutzkonformer Open-Source-Modelle wie faster-whisper oder funktional gleichwertiger Lösungen erfolgen.
Begründung	Speech-to-Text erweitert die Plattform um wichtige sprachbasierte Verwaltungsprozesse und kann sowohl im KI-Dialog als auch in Fachanwendungen genutzt werden.

FA-073 – Datenschutz- und PII-Erkennung

Anforderungsbeschreibung	Die Plattform muss einen ML-Service zur Erkennung personenbezogener Informationen bereitstellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Umsetzung soll bevorzugt unter Nutzung von Microsoft Presidio oder funktional gleichwertigen Open-Source-Komponenten erfolgen.
Begründung	Datenschutz ist ein wesentlicher Bestandteil kommunaler KI-Anwendungen und unterstützt die Umsetzung der DSGVO sowie des AI Acts.

FA-074 – Bildanalyse

Anforderungsbeschreibung	Die Architektur muss die spätere Integration spezialisierter Bildanalyse-Services unterstützen, insbesondere für Objekterkennung, Bildklassifikation oder Dokumentenbilder.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die erforderlichen API-Schnittstellen sind vorzusehen. Eine produktive Umsetzung ist nicht Bestandteil der Erstimplementierung.
Begründung	Bildanalyse stellt eine sinnvolle Erweiterung dar, ist jedoch für die Erstimplementierung nicht zwingend erforderlich.

FA-075 – Erweiterbare ML-Service-Architektur

Anforderungsbeschreibung	Die Architektur muss die spätere Integration weiterer spezialisierter ML-Services unterstützen, ohne Änderungen am AI Harness oder an den Schnittstellen erforderlich zu machen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die ML-Services sind modular und lose gekoppelt bereitzustellen.
Begründung	Eine modulare Service-Architektur gewährleistet langfristige Erweiterbarkeit und reduziert spätere Integrationsaufwände.

FA-076 – Keine eigenständige ML-Plattform

Anforderungs- beschreibung	Los 3 umfasst ausschließlich die Bereitstellung spezialisierter ML-Services. Eigene Benutzeroberflächen, Agentenplattformen, Workflow-Engines, Orchestrierungsfunktionen oder Governance-Komponenten sind nicht Bestandteil dieses Loses. Diese Funktionen werden durch den AI Harness beziehungsweise andere Komponenten des Urban Stack bereitgestellt.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die ML-Services sind ausschließlich als technische Backend-Komponenten bereitzustellen und über standardisierte APIs in den KI-Stack zu integrieren.
Begründung	Die klare Abgrenzung verhindert Doppelentwicklungen, reduziert den Projektumfang und stellt eine wirtschaftliche Aufteilung der Verantwortlichkeiten zwischen den Losen sicher.

4.10 Prompt Governance

Zielsetzung

Die Plattform soll einen nachvollziehbaren, kontrollierten und qualitätsgesicherten Umgang mit Prompts gewährleisten. Ziel ist es, produktive Prompts strukturiert zu entwickeln, zu testen, freizugeben und über ihren gesamten Lebenszyklus nachvollziehbar zu verwalten.

Während die Prompt Registry die technische Verwaltung von Prompts übernimmt, beschreibt dieses Kapitel die organisatorischen und fachlichen Anforderungen an deren Qualitätssicherung und Governance.

Architekturgrundsatz

Prompt Governance ist Bestandteil des AI Harness. Sämtliche produktiv eingesetzten Prompts unterliegen einer kontrollierten Verwaltung, Versionierung und Freigabe. Änderungen müssen nachvollziehbar dokumentiert werden und dürfen ausschließlich über definierte Rollen und Berechtigungen erfolgen.

FA-077 – Prompt-Lebenszyklus

Anforderungs- beschreibung	Die Plattform muss den vollständigen Lebenszyklus produktiver Prompts unterstützen. Hierzu gehören Erstellung, Test, Freigabe, Versionierung, Veröffentlichung, Änderung und Außerbetriebnahme.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Der Lebenszyklus ist über die zentrale Prompt Registry sowie die Administrationsoberfläche abzubilden.
Begründung	Ein definierter Lebenszyklus erhöht die Qualität produktiver KI-Anwendungen und reduziert Fehlkonfigurationen.

FA-078 – Freigabeverfahren

Anforderungs- beschreibung	Produktive Prompts müssen vor ihrer Veröffentlichung kontrolliert freigegeben werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Freigaben erfolgen rollenbasiert unter Nutzung der vorhandenen Berechtigungsstrukturen.
Begründung	Freigabeprozesse verhindern unkontrollierte Änderungen produktiver KI-Anwendungen.

FA-079 – Nachvollziehbarkeit

Anforderungsbeschreibung	Änderungen an produktiven Prompts müssen revisionssicher dokumentiert werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Änderungen sind mit Zeitstempel, Version, Bearbeiter und Änderungsbeschreibung zu protokollieren.
Begründung	Revisionssicherheit unterstützt Governance, Auditierung und den sicheren Betrieb produktiver KI-Anwendungen.

FA-080 – Erweiterbare Prompt Governance

Anforderungsbeschreibung	Die Architektur muss zukünftige Verfahren zur automatisierten Bewertung, Optimierung und Qualitätssicherung von Prompts unterstützen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die erforderlichen Erweiterungspunkte sind innerhalb der Prompt Registry und des AI Harness vorzusehen. Eine produktive Umsetzung automatisierter Prompt-Optimierungen ist nicht Bestandteil der Erstimplementierung.
Begründung	Moderne Verfahren des Prompt Engineering entwickeln sich dynamisch. Eine vorbereitete Architektur gewährleistet Zukunftssicherheit, ohne den Umfang der Erstimplementierung unnötig zu erweitern.

4.11 Tenant Context

Zielsetzung

Der Tenant Context dient der Bereitstellung gemeinsamer organisatorischer Kontextinformationen innerhalb eines Mandanten. Ziel ist es, KI-Dialoge, Agenten, Retrieval-Prozesse und Workflows um mandantenbezogene Informationen zu ergänzen und dadurch konsistente sowie organisationsspezifische Ergebnisse zu ermöglichen.

Die produktive Erstimplementierung konzentriert sich auf die Verwaltung grundlegender organisatorischer Kontextinformationen. Erweiterte Organisationsmodelle und semantische Wissensstrukturen werden architektonisch vorbereitet.

Architekturgrundsatz

Der Tenant Context ist Bestandteil der Context Engine und des AI Harness. Er ergänzt den Modellkontext um organisationsbezogene Informationen, ohne personenbezogene Daten oder mandantenübergreifende Informationen einzubeziehen.

Die Verwaltung erfolgt zentral und unterliegt den bestehenden Rollen-, Berechtigungs- und Governance-Regelungen.

FA-081 – Verwaltung des Tenant Context

Anforderungs- beschreibung	Die Plattform muss einen Tenant Context bereitstellen, in dem organisationsbezogene Informationen zentral verwaltet und für KI-Funktionen bereitgestellt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Der Tenant Context ist als Bestandteil der Context Engine bereitzustellen und mandantenbezogen zu verwalten.
Begründung	Organisationsbezogene Kontextinformationen erhöhen die Konsistenz der KI-Ergebnisse und vermeiden redundante Konfigurationen einzelner Agenten.

FA-082 – Organisationsbezogene Kontextinformationen

Anforderungs- beschreibung	Der Tenant Context muss mindestens Organisationsinformationen, organisatorische Vorgaben, allgemeine Arbeitskontexte, Standardbegriffe und weitere gemeinsam nutzbare Kontextinformationen verwalten können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Informationen sind zentral administrierbar und für berechtigte KI-Agenten sowie Dialoge nutzbar bereitzustellen.
Begründung	Gemeinsame Kontextinformationen gewährleisten ein einheitliches Verhalten der Plattform innerhalb eines Mandanten.

FA-083 – Berücksichtigung im Modellkontext

Anforderungs- beschreibung	Die Context Engine muss den Tenant Context bei der Zusammenstellung des Modellkontexts berücksichtigen, sofern dies für den jeweiligen Anwendungsfall erforderlich und zulässig ist.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Einbindung erfolgt regelbasiert unter Berücksichtigung von Rollen, Berechtigungen und Datenschutzvorgaben.
Begründung	Der Tenant Context entfaltet seinen Nutzen erst durch die kontrollierte Einbindung in den Gesamtkontext der KI-Verarbeitung.

FA-084 – Mandantentrennung

Anforderungsbeschreibung	Die Plattform muss sicherstellen, dass Tenant-Context-Informationen ausschließlich innerhalb des jeweiligen Mandanten verwendet werden und nicht mandantenübergreifend verfügbar sind.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Mandantentrennung ist über die bestehende Rollen-, Berechtigungs- und Sicherheitsarchitektur sicherzustellen.
Begründung	Eine konsequente Trennung organisatorischer Informationen ist Voraussetzung für Datenschutz, Informationssicherheit und Mandantenfähigkeit.

FA-085 – Erweiterbarkeit des Tenant Context

Anforderungsbeschreibung	Die Architektur muss die spätere Erweiterung des Tenant Context um zusätzliche organisatorische Wissensmodelle, Regelwerke oder semantische Organisationsstrukturen unterstützen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die erforderlichen Erweiterungspunkte innerhalb der Context Engine und des AI Harness sind vorzusehen. Eine vollständige Umsetzung ist nicht Bestandteil der Erstimplementierung.
Begründung	Die vorbereitete Architektur ermöglicht den späteren Ausbau organisationsspezifischer Wissensmodelle, ohne den Umfang der Erstimplementierung zu erhöhen.

4.12 Citation Validator

Zielsetzung

Der Citation Validator dient der Qualitätssicherung KI-generierter Antworten. Ziel ist es, die Nachvollziehbarkeit von Aussagen zu erhöhen, indem Quellenangaben geprüft, validiert und transparent dargestellt werden.

Die produktive Erstimplementierung konzentriert sich auf die Validierung von Quellen aus Retrieval-Prozessen (RAG) sowie der Dokumentenverarbeitung. Erweiterte Verfahren zur automatisierten Faktenprüfung oder externen Quellenvalidierung werden architektonisch vorbereitet.

Architekturgrundsatz

Der Citation Validator ist Bestandteil des AI Harness und arbeitet eng mit der Retrieval-Komponente sowie der Context Engine zusammen.

Er prüft, ob die von der KI verwendeten Quellen tatsächlich zur erzeugten Antwort passen und stellt diese nachvollziehbar dar. Der Citation Validator ersetzt keine fachliche Prüfung durch den Nutzer.

FA-086 – Quellenvalidierung

Anforderungsbeschreibung	Die Plattform muss Quellenangaben aus Retrieval-Prozessen auf ihre Zuordnung zu den erzeugten Antworten prüfen und nachvollziehbar darstellen können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Validierung erfolgt für Quellen aus RAG-Prozessen und der Dokumentenverarbeitung. Die Ergebnisse sind dem Nutzer gemeinsam mit der Antwort bereitzustellen.
Begründung	Die Validierung von Quellen erhöht die Nachvollziehbarkeit und das Vertrauen in KI-generierte Antworten im Verwaltungskontext.

FA-087 – Kennzeichnung unzureichender Quellen

Anforderungsbeschreibung	Die Plattform muss Antworten kennzeichnen können, wenn keine ausreichende oder belastbare Quellenbasis vorhanden ist.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Der Citation Validator arbeitet mit den Guardrails des AI Harness zusammen und kann Antworten entsprechend markieren oder deren Ausgabe unterbinden.
Begründung	Die transparente Kennzeichnung unzureichender Quellen reduziert das Risiko fachlich nicht belastbarer KI-Antworten.

FA-088 – Darstellung der Quellen

Anforderungsbeschreibung	Die Plattform muss die für eine Antwort verwendeten Quellen nachvollziehbar darstellen und eine Zuordnung zu Dokumenten oder Dokumentabschnitten ermöglichen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Quellen sind mindestens auf Dokument- oder Abschnittsebene anzuzeigen. Die Darstellung erfolgt innerhalb der Benutzeroberfläche des Urban KI Hub.
Begründung	Die transparente Darstellung der Quellen ermöglicht eine fachliche Prüfung der KI-Ergebnisse durch die Nutzenden.

FA-089 – Erweiterbare Validierungsverfahren

Anforderungs- beschreibung	Die Architektur muss die spätere Integration zusätzlicher Validierungsverfahren unterstützen, insbesondere semantische Quellenbewertung, Faktenprüfung, Confidence Scores oder externe Validierungsdienste.
Kategorie	A – Architekturvorbereitung
Umsetzung	Der Citation Validator ist modular aufzubauen und über definierte Schnittstellen erweiterbar zu gestalten. Eine vollständige Umsetzung zusätzlicher Validierungsverfahren ist nicht Bestandteil der Erstimplementierung.
Begründung	Moderne Verfahren zur Qualitätsbewertung entwickeln sich dynamisch. Eine vorbereitete Architektur gewährleistet Zukunftssicherheit ohne zusätzlichen Initialaufwand.

4.13 Entscheidungslogik & Human Approval

Zielsetzung

Die Plattform soll sicherstellen, dass KI-gestützte Entscheidungen und automatisierte Agentenausführungen nachvollziehbar, kontrollierbar und bei Bedarf durch menschliche Freigaben abgesichert werden können. Ziel ist die sichere Unterstützung von Verwaltungsprozessen unter Berücksichtigung von Governance, Datenschutz und regulatorischen Anforderungen.

Die produktive Erstimplementierung konzentriert sich auf regelbasierte Entscheidungslogiken sowie Human-Approval-Prozesse für definierte Anwendungsfälle. Erweiterte Entscheidungsmodelle oder Business-Rule-Engines werden architektonisch vorbereitet.

Architekturgrundsatz

Die Entscheidungslogik ist Bestandteil des AI Harness und arbeitet mit der Policy Engine, den Guardrails sowie der Context Engine zusammen. Vor der Ausführung definierter Aktionen kann eine menschliche Freigabe (Human Approval) erforderlich sein. Die Entscheidung darüber erfolgt auf Basis konfigurierbarer Regeln. Die Plattform stellt keine eigenständige Workflow- oder Business-Rule-Engine bereit. Komplexe Prozesslogiken werden durch angebundene Workflow-Systeme des Urban Stack oder vergleichbare Lösungen umgesetzt.

FA-090 – Regelbasierte Entscheidungslogik

Anforderungsbeschreibung	Die Plattform muss regelbasierte Entscheidungslogiken unterstützen, um KI-Aktionen abhängig von definierten Richtlinien, Berechtigungen und Kontextinformationen zu steuern.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Entscheidungslogik ist innerhalb des AI Harness bereitzustellen und mit Policy Engine, Guardrails und Context Engine zu integrieren.
Begründung	Regelbasierte Entscheidungen gewährleisten einen kontrollierten und nachvollziehbaren Betrieb der KI-Plattform.

FA-091 – Human Approval

Anforderungsbeschreibung	Die Plattform muss Human-Approval-Prozesse unterstützen. Für definierte Aktionen oder Agentenausführungen muss vor der Durchführung eine Freigabe durch berechtigte Nutzer eingeholt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Human Approval ist für ausgewählte Anwendungsfälle bereitzustellen. Die Freigabe erfolgt über die Benutzeroberfläche des Urban KI Hub oder über angebundene Fachanwendungen.
Begründung	Human Approval reduziert Risiken bei sensiblen oder kritischen KI-gestützten Entscheidungen und unterstützt die Anforderungen des AI Acts.

FA-092 – Nachvollziehbarkeit von Entscheidungen

Anforderungsbeschreibung	Entscheidungen, Freigaben und Ablehnungen müssen revisionssicher protokolliert und nachvollziehbar dokumentiert werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Sämtliche Entscheidungs- und Freigabeprozesse sind über die Audit- und Monitoring-Komponenten des AI Harness zu protokollieren.
Begründung	Die Nachvollziehbarkeit von Entscheidungen ist Voraussetzung für Governance, Compliance und Revisionssicherheit.

FA-093 – Konfigurierbare Freigaberegeln

Anforderungsbeschreibung	Die Plattform muss konfigurierbare Regeln bereitstellen, mit denen festgelegt werden kann, für welche Aktionen oder Agenten Human Approval erforderlich ist.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Regeln werden zentral administriert und können abhängig von Agent, Werkzeug, Mandant, Rolle oder Aktion definiert werden.
Begründung	Flexible Freigaberegeln ermöglichen eine risikoorientierte Steuerung der Plattform, ohne sämtliche Prozesse manuell freigeben zu müssen.

FA-094 – Erweiterbare Entscheidungsarchitektur

Anforderungsbeschreibung	Die Architektur muss die spätere Integration erweiterter Entscheidungsverfahren unterstützen, insbesondere mehrstufige Genehmigungsprozesse, zusätzliche Regelwerke oder externe Entscheidungsdienste.
Kategorie	A – Architekturvorbereitung
Umsetzung	Der AI Harness muss standardisierte Erweiterungspunkte für zusätzliche Entscheidungs- und Genehmigungsmechanismen bereitstellen. Eine vollständige Umsetzung dieser Funktionen ist nicht Bestandteil der Erstimplementierung.
Begründung	Die vorbereitete Architektur ermöglicht den späteren Ausbau der Entscheidungslogik, ohne den Umfang der Erstimplementierung unnötig zu vergrößern.

4.14 Token- und Kostenmonitoring

Zielsetzung

Die Plattform soll die Nutzung von KI-Ressourcen transparent überwachen und auswerten. Ziel ist es, den Verbrauch von Modellen, Token, ML-Services sowie weiteren KI-Komponenten nachvollziehbar darzustellen und eine wirtschaftliche Nutzung der Plattform zu unterstützen.

Die produktive Erstimplementierung konzentriert sich auf ein zentrales Monitoring von Tokenverbrauch, Modellnutzung und geschätzten Kosten. Erweiterte FinOps-Funktionen werden architektonisch vorbereitet.

Architekturgrundsatz

Das Token- und Kostenmonitoring ist Bestandteil des AI Harness.

Sämtliche Modellaufrufe, Agentenausführungen, Retrieval-Prozesse sowie ML-Service-Aufrufe werden zentral protokolliert und für Monitoring- und Auswertungszwecke bereitgestellt.

Zur Speicherung und Analyse von Nutzungsdaten soll die bereits vorhandene ClickHouse-Infrastruktur genutzt werden. Eine eigenständige Monitoring- oder Reporting-Datenbank ist nicht Bestandteil der Leistung.

FA-095 – Tokenmonitoring

Anforderungsbeschreibung	Die Plattform muss den Tokenverbrauch sämtlicher Modellaufrufe erfassen und auswerten können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Der Tokenverbrauch ist modell-, mandanten-, agenten- und nutzerbezogen zu protokollieren und zentral auszuwerten.
Begründung	Die Transparenz des Tokenverbrauchs ist Grundlage für einen wirtschaftlichen und nachvollziehbaren Betrieb der Plattform.

FA-096 – Kostenmonitoring

Anforderungsbeschreibung	Die Plattform muss die geschätzten oder tatsächlich anfallenden Nutzungskosten für Modellaufrufe und ML-Services erfassen und auswerten können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Kosten sind auf Basis der verfügbaren Modellinformationen oder hinterlegter Kostenparameter zu berechnen und mandantenbezogen auszuwerten.
Begründung	Ein transparentes Kostenmonitoring unterstützt den wirtschaftlichen Betrieb der KI-Plattform und erleichtert die Budgetplanung.

FA-097 – Nutzungsstatistiken

Anforderungsbeschreibung	Die Plattform muss Statistiken über die Nutzung von Modellen, Agenten, Retrieval-Prozessen und ML-Services bereitstellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Auswertungen sind zentral bereitzustellen und mindestens nach Zeitraum, Mandant, Modell, Agent und Organisationseinheit filterbar auszuführen.
Begründung	Nutzungsstatistiken ermöglichen die Optimierung der Plattform sowie die Identifikation häufig genutzter KI-Funktionen.

FA-098 – Nutzung vorhandener Monitoring-Infrastruktur

Anforderungs- beschreibung	Die Plattform muss vorhandene Monitoring- und Analysekomponenten der Auftraggeberin nutzen können. Insbesondere ist die vorhandene ClickHouse-Infrastruktur für die Speicherung und Analyse von Monitoringdaten einzubinden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Leistung umfasst ausschließlich die Integration in die vorhandene Infrastruktur. Aufbau oder Betrieb zusätzlicher Monitoring-Datenbanken sind nicht Bestandteil der Leistung.
Begründung	Die Nutzung bestehender Infrastruktur reduziert Implementierungsaufwand, Betriebskosten und vermeidet redundante Systemkomponenten.

FA-099 – Budget- und Verbrauchsgrenzen

Anforderungs- beschreibung	Die Plattform soll konfigurierbare Budget- oder Verbrauchsgrenzen unterstützen und bei deren Überschreitung entsprechende Hinweise oder Warnungen bereitstellen können.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die erforderlichen Erweiterungspunkte für Budgetüberwachung und Benachrichtigungen sind vorzusehen. Eine vollständige Budgetsteuerung ist nicht Bestandteil der Erstimplementierung.
Begründung	Budgetgrenzen unterstützen den wirtschaftlichen Betrieb, können jedoch in einer späteren Ausbaustufe umgesetzt werden.

FA-100 – Erweiterbares KI-FinOps

Anforderungs- beschreibung	Die Architektur muss die spätere Integration weitergehender FinOps-Funktionen unterstützen, insbesondere Kostenprognosen, Optimierungsempfehlungen, Ressourcenanalysen oder automatisierte Kostensteuerung.
Kategorie	A – Architekturvorbereitung
Umsetzung	Der AI Harness muss standardisierte Erweiterungspunkte für zukünftige FinOps-Komponenten bereitstellen. Eine vollständige Umsetzung ist nicht Bestandteil der Erstimplementierung.
Begründung	KI-FinOps entwickelt sich zu einem zentralen Bestandteil moderner KI-Plattformen. Die vorbereitete Architektur gewährleistet langfristige Erweiterbarkeit bei begrenztem Initialaufwand.

4.15 Performance-Monitoring

Zielsetzung

Die Plattform soll die technische und fachliche Performance der KI-Komponenten kontinuierlich überwachen. Ziel ist es, die Verfügbarkeit, Stabilität und Leistungsfähigkeit der KI-Plattform transparent darzustellen sowie die Grundlage für Betrieb, Fehleranalyse und Optimierung bereitzustellen.

Die produktive Erstimplementierung konzentriert sich auf die Überwachung KI-spezifischer Komponenten. Das allgemeine Infrastruktur-, Netzwerk- und Cluster-Monitoring ist nicht Bestandteil der Leistung und wird durch die bestehende Betriebsumgebung bereitgestellt.

Architekturgrundsatz

Das Performance-Monitoring ist Bestandteil des AI Harness.

Es überwacht insbesondere Modellaufrufe, Agentenausführungen, Retrieval-Prozesse, ML-Services sowie Schnittstellenaufrufe.

Zur Speicherung und Auswertung von Performance-Daten soll die vorhandene ClickHouse-Infrastruktur genutzt werden. Bereits vorhandene Monitoring-Lösungen des Urban Stack sind einzubinden. Eine eigenständige Monitoring-Plattform ist nicht Bestandteil der Leistung.

FA-101 – Überwachung der KI-Komponenten

Anforderungs- beschreibung	Die Plattform muss die Performance zentraler KI-Komponenten überwachen. Hierzu gehören insbesondere Modellaufrufe, Agentenausführungen, Retrieval-Prozesse, ML-Services und API-Aufrufe.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die relevanten Performance-Kennzahlen sind durch den AI Harness zu erfassen und für Monitoring- und Betriebszwecke bereitzustellen.
Begründung	Die Überwachung der KI-Komponenten ermöglicht einen stabilen und nachvollziehbaren Betrieb der Plattform.

FA-102 – Performance-Kennzahlen

Anforderungs- beschreibung	Die Plattform muss zentrale Leistungskennzahlen bereitstellen. Hierzu gehören insbesondere Antwortzeiten, Laufzeiten von Agenten, Dauer von Retrieval-Prozessen, Verfügbarkeit von Diensten sowie Fehlerraten.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Kennzahlen sind mandantenübergreifend erfassbar und für Administratoren in geeigneter Form bereitzustellen.
Begründung	Einheitliche Leistungskennzahlen erleichtern den Betrieb, die Fehleranalyse und die Optimierung der KI-Plattform.

FA-103 – Nutzung vorhandener Monitoring-Infrastruktur

Anforderungs- beschreibung	Die Plattform muss vorhandene Monitoring- und Observability-Komponenten der Auftraggeberin integrieren können. Insbesondere sind bestehende Monitoring-Dienste des Urban Stack sowie die vorhandene ClickHouse-Infrastruktur für die Speicherung und Auswertung von Performance-Daten zu nutzen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Leistung umfasst ausschließlich die Integration in vorhandene Monitoring-Systeme. Aufbau und Betrieb zusätzlicher Monitoring-Infrastrukturen sind nicht Bestandteil der Leistung.
Begründung	Die Nutzung vorhandener Infrastruktur reduziert Implementierungs- und Betriebskosten und vermeidet redundante Systemkomponenten.

FA-104 – Alarmierung

Anforderungs- beschreibung	Die Plattform soll definierbare Schwellwerte unterstützen und bei kritischen Ereignissen oder Grenzwertüberschreitungen Warnungen oder Benachrichtigungen bereitstellen können.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die erforderlichen Schnittstellen für Alarmierungen sind vorzusehen. Die konkrete Anbindung an bestehende Monitoring- oder Benachrichtigungssysteme kann in einer späteren Ausbaustufe erfolgen.
Begründung	Alarmierungsfunktionen erhöhen die Betriebssicherheit, sind jedoch für die Erstimplementierung nicht zwingend erforderlich.

FA-105 – Erweiterbare Observability

Anforderungs- beschreibung	Die Architektur muss die spätere Integration weiterer Observability-Funktionen unterstützen, insbesondere verteiltes Tracing, Service-Abhängigkeiten, KI-spezifische Qualitätsmetriken sowie automatisierte Performanceanalysen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Der AI Harness muss standardisierte Erweiterungspunkte für zusätzliche Monitoring- und Observability-Komponenten bereitstellen. Eine vollständige Umsetzung dieser Funktionen ist nicht Bestandteil der Erstimplementierung.
Begründung	Eine vorbereitete Observability-Architektur ermöglicht den späteren Ausbau des Plattformbetriebs, ohne den Umfang der Erstimplementierung unnötig zu erhöhen.

4.16 Datenschutz und Administration

Zielsetzung

Die Plattform muss eine sichere, datenschutzkonforme und administrierbare Nutzung sämtlicher KI-Funktionen ermöglichen. Ziel ist die Bereitstellung zentraler Administrations-, Sicherheits- und Datenschutzfunktionen, die einen rechtskonformen, nachvollziehbaren und wirtschaftlichen Betrieb der KI-Plattform gewährleisten.

Die produktive Erstimplementierung konzentriert sich auf die Integration der bestehenden Authentifizierungs-, Autorisierungs- und Administrationsinfrastruktur der Auftraggeberin. Eigene Identity-Management-, Benutzerverwaltungs- oder Infrastrukturkomponenten sind nicht Bestandteil der Leistung.

Architekturgrundsatz

Die Administrationsfunktionen sind Bestandteil des AI Harness und des Urban KI Hub.

Die Plattform integriert sich in die vorhandene Authentifizierungs- und Autorisierungsinfrastruktur der Auftraggeberin. Rollen, Berechtigungen, Mandanten sowie Datenschutzvorgaben sind bei sämtlichen KI-Funktionen konsistent zu berücksichtigen.

Datenschutz, Governance, Policy Engine, Guardrails, Context Engine und Auditierung arbeiten dabei als integrierte Bestandteile des AI Harness zusammen.

FA-106 – Rollen- und Benutzerverwaltung

Anforderungs- beschreibung	Die Plattform muss die vorhandene Rollen- und Benutzerverwaltung der Auftraggeberin integrieren und für sämtliche KI-Funktionen nutzen. Eine eigenständige Benutzerverwaltung ist nicht Bestandteil der Leistung.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Integration erfolgt über die vorhandenen Authentifizierungs- und Autorisierungssysteme (z. B. SSO, OpenID Connect, OAuth 2.0, SAML 2.0 oder funktional gleichwertige Verfahren).
Begründung	Die Nutzung der vorhandenen Identitätsinfrastruktur reduziert Implementierungsaufwand und gewährleistet eine konsistente Benutzerverwaltung.

FA-107 – Rollenbasierte Administration

Anforderungs- beschreibung	Die Plattform muss eine rollenbasierte Administration sämtlicher KI-Komponenten unterstützen. Administratoren müssen abhängig von ihrer Rolle unterschiedliche Verwaltungsfunktionen nutzen können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Rollen und Berechtigungen sind zentral zu verwalten und für sämtliche Komponenten des AI Harness anzuwenden.
Begründung	Eine differenzierte Administration erhöht Sicherheit, Nachvollziehbarkeit und Wartbarkeit der Plattform.

FA-108 – Datenschutzkonforme Verarbeitung

Anforderungs- beschreibung	Die Plattform muss personenbezogene Daten datenschutzkonform verarbeiten und dabei die Vorgaben der DSGVO sowie weiterer einschlägiger rechtlicher Anforderungen berücksichtigen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Datenschutzanforderungen sind innerhalb des AI Harness, der Context Engine, des Memory-Konzepts sowie sämtlicher Modell- und Tool-Aufrufe zu berücksichtigen.
Begründung	Datenschutz ist eine grundlegende Voraussetzung für den produktiven Einsatz von KI-Systemen in der öffentlichen Verwaltung.

FA-109 – Lösch- und Aufbewahrungskonzepte

Anforderungsbeschreibung	Die Plattform muss konfigurierbare Aufbewahrungs- und Löschrregeln für Chatverläufe, Memory-Inhalte, Protokolle und weitere KI-bezogene Daten unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Lösch- und Aufbewahrungsregeln sind mandanten- und rollenbezogen administrierbar bereitzustellen.
Begründung	Konfigurierbare Löschkonzepte unterstützen Datenschutz, Compliance und die Einhaltung gesetzlicher Aufbewahrungspflichten.

FA-110 – Audit und Nachvollziehbarkeit

Anforderungsbeschreibung	Sämtliche administrativen Änderungen sowie sicherheitsrelevanten Aktionen müssen revisionssicher protokolliert und nachvollziehbar dokumentiert werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Protokollierung erfolgt zentral über die Audit-Funktionen des AI Harness und ist für berechtigte Administratoren auswertbar bereitzustellen.
Begründung	Eine vollständige Nachvollziehbarkeit administrativer Änderungen ist Voraussetzung für Governance, Compliance und Revisionssicherheit.

FA-111 – Datenschutz- und Governance-Einstellungen

Anforderungsbeschreibung	Die Plattform muss zentrale Konfigurationsmöglichkeiten für Datenschutz-, Governance-, Policy- und Guardrail-Einstellungen bereitstellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Konfiguration erfolgt zentral über die Administrationsoberfläche und wirkt konsistent auf sämtliche Komponenten des AI Harness.
Begründung	Zentrale Einstellungen vereinfachen den Betrieb und gewährleisten eine konsistente Umsetzung organisatorischer Vorgaben.

FA-112 – Erweiterbare Administrationsarchitektur

Anforderungs- beschreibung	Die Architektur muss die spätere Integration zusätzlicher Administrations-, Datenschutz- und Governance-Funktionen unterstützen, insbesondere KI-Inventare, AI-Act-Compliance-Funktionen, Risikoanalysen oder weitere Verwaltungswerkzeuge.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die erforderlichen Erweiterungspunkte sind innerhalb des AI Harness und der Administrationsoberfläche vorzusehen. Eine vollständige Umsetzung dieser Funktionen ist nicht Bestandteil der Erstimplementierung.
Begründung	Die vorbereitete Architektur ermöglicht die schrittweise Erweiterung der Plattform entsprechend zukünftiger regulatorischer und organisatorischer Anforderungen.

5 Nicht-funktionale Anforderungen

Zielsetzung

Die nachfolgenden nicht-funktionalen Anforderungen beschreiben die qualitativen, technischen und betrieblichen Eigenschaften der KI-Plattform. Sie definieren die Anforderungen an Leistungsfähigkeit, Skalierbarkeit, Sicherheit, Datenschutz, Erweiterbarkeit und Betrieb.

Die Anforderungen beziehen sich ausschließlich auf die im Rahmen dieser Ausschreibung zu entwickelnden bzw. zu integrierenden Komponenten. Bereits vorhandene Infrastrukturkomponenten der Auftraggeberin sind zu nutzen und nicht erneut bereitzustellen.

Die nicht-funktionalen Anforderungen sind unter Berücksichtigung der vorhandenen Systemlandschaft, der Wirtschaftlichkeit sowie der modularen Zielarchitektur umzusetzen.

5.1 Performance

Zielsetzung

Die Plattform muss eine leistungsfähige und stabile Verarbeitung von KI-Anfragen gewährleisten. Dabei stehen insbesondere kurze Antwortzeiten, parallele Verarbeitung sowie ein stabiler Betrieb unter Last im Vordergrund.

Die Anforderungen beziehen sich ausschließlich auf die durch den Auftragnehmer bereitgestellten Komponenten. Antwortzeiten externer Modellanbieter oder bereits vorhandener Infrastrukturkomponenten sind hiervon ausgenommen.

Architekturgrundsatz

Die Performance der Plattform ist durch den AI Harness sicherzustellen. Vorhandene Infrastruktur des Urban Stack sowie des IONOS Model Hub sind zu nutzen. Performanceoptimierungen sollen vorrangig durch horizontale Skalierung, Caching und effiziente Nutzung vorhandener Ressourcen erfolgen.

NFA-001 – Antwortzeit (Time-to-First-Token)

Anforderungs- beschreibung	Die Plattform soll bei Standardanfragen eine Time-to-First-Token von höchstens fünf Sekunden erreichen. Bei komplexen Retrieval-, Agenten- oder Dokumentenprozessen darf die Antwortzeit entsprechend der Verarbeitungskomplexität höher ausfallen. Die Plattform muss Anfragen innerhalb technisch angemessener Zeiträume verarbeiten und darf keine systembedingten Blockaden verursachen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anforderungen beziehen sich ausschließlich auf die durch den Auftragnehmer bereitgestellten Komponenten. Verzögerungen externer Modellanbieter oder der bereitgestellten Infrastruktur sind hiervon ausgenommen.
Begründung	Kurze Antwortzeiten sind entscheidend für die Akzeptanz und Nutzbarkeit der Plattform im Verwaltungsalltag.

NFA-002 – Parallelverarbeitung

Anforderungs- beschreibung	Die Plattform muss mehrere gleichzeitige Anfragen eines Nutzers sowie parallele Anfragen mehrerer Nutzer zuverlässig verarbeiten können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Parallelisierung erfolgt über die vorhandene Container- und Kubernetes-Infrastruktur.
Begründung	Mehrere parallele KI-Anfragen gehören zum typischen Nutzungsverhalten moderner KI-Anwendungen.

NFA-003 – Verarbeitungsleistung

Anforderungs- beschreibung	Die Plattform soll eine durchschnittliche Generierungsleistung von mindestens 15 Tokens pro Sekunde erreichen, sofern dies durch das jeweils eingesetzte Modell unterstützt wird.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Architektur muss eine optimale Nutzung leistungsfähiger Modelle ermöglichen. Eine feste Tokenrate ist aufgrund unterschiedlicher Modellanbieter nicht verbindlich.
Begründung	Die tatsächliche Generierungsleistung hängt wesentlich vom eingesetzten Modell und der vorhandenen Hardware ab und kann daher nicht vollständig durch den Auftragnehmer beeinflusst werden.

NFA-004 – Lastverhalten

Anforderungs- beschreibung	Die Plattform muss auch bei erhöhter Last stabil erreichbar bleiben und Anfragen ohne systembedingte Fehler oder Abbrüche verarbeiten können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Lastverteilung erfolgt über die vorhandene Infrastruktur. Die KI-Komponenten müssen zustandslos und horizontal skalierbar ausgeführt werden können.
Begründung	Ein stabiles Lastverhalten ist Voraussetzung für den produktiven Betrieb einer zentralen KI-Plattform.

5.2 Skalierbarkeit

Zielsetzung

Die Plattform muss für einen langfristigen Ausbau der KI-Nutzung innerhalb der Stadtverwaltung ausgelegt sein. Ziel ist eine skalierbare Architektur, die steigende Nutzerzahlen, zusätzliche KI-Dienste sowie weitere Anwendungsfälle unterstützen kann, ohne grundlegende Änderungen an der Gesamtarchitektur zu erfordern.

Die Anforderungen beziehen sich ausschließlich auf die durch den Auftragnehmer bereitgestellten Komponenten. Die zugrundeliegende Container-, Speicher- und Infrastrukturplattform wird durch die Auftraggeberin bereitgestellt und ist über standardisierte Schnittstellen zu nutzen.

Architekturgrundsatz

Sämtliche Komponenten des AI Harness, des Urban KI Hub sowie der ML-Services sind zustandslos oder entsprechend containerfähig zu entwickeln und müssen den Betrieb innerhalb der vorhandenen Kubernetes-Infrastruktur unterstützen.

Die Skalierung der Infrastruktur selbst ist nicht Bestandteil der Leistung.

NFA-005 – Horizontale Skalierbarkeit

Anforderungs- beschreibung	Die durch den Auftragnehmer bereitgestellten Komponenten müssen horizontal skalierbar ausgelegt sein und den parallelen Betrieb mehrerer Instanzen unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Komponenten sind containerfähig und zustandsarm zu entwickeln. Die eigentliche Skalierung erfolgt über die vorhandene Kubernetes-Infrastruktur der Auftraggeberin.
Begründung	Eine horizontale Skalierbarkeit gewährleistet den langfristigen Ausbau der Plattform, ohne dass Anwendungen grundlegend angepasst werden müssen.

NFA-006 – Nutzerkapazität

Anforderungs- beschreibung	Die Architektur muss den Betrieb mit mindestens 1.000 gleichzeitig aktiven Nutzern unterstützen und eine spätere Erweiterung auf höhere Nutzerzahlen ermöglichen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Architektur ist entsprechend auszulegen. Ein Lasttest mit 1.000 gleichzeitig aktiven Nutzern ist nicht Bestandteil der Erstimplementierung, sofern die Skalierbarkeit architektonisch nachgewiesen wird.
Begründung	Die tatsächliche Nutzerkapazität hängt wesentlich von der bereitgestellten Infrastruktur und den eingesetzten KI-Modellen ab. Die Architektur muss diese Skalierung unterstützen, ohne unrealistische Leistungszusagen zu verlangen.

NFA-007 – Elastizität

Anforderungs- beschreibung	Die Plattform muss den Betrieb innerhalb einer Infrastruktur unterstützen, die Lastspitzen automatisch erkennt und zusätzliche Ressourcen bereitstellt.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Komponenten müssen Auto-Scaling unterstützen. Die Bereitstellung und Konfiguration der Auto-Scaling-Infrastruktur ist nicht Bestandteil der Leistung.
Begründung	Auto-Scaling ist Aufgabe der Betriebsplattform. Die Anwendungen müssen diese Funktion unterstützen, jedoch nicht selbst bereitstellen.

NFA-008 – Mandantenskalierung

Anforderungs- beschreibung	Die Plattform muss mehrere Organisationseinheiten (Mandanten) parallel unterstützen, ohne dass Daten, Konfigurationen oder Verarbeitungskontexte zwischen Mandanten vermischt werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Mandantentrennung ist innerhalb sämtlicher Komponenten des AI Harness, des Urban KI Hub sowie der ML-Services konsequent umzusetzen.
Begründung	Die Mandantenfähigkeit ist ein grundlegendes Architekturprinzip des Urban Stack und Voraussetzung für den produktiven Einsatz der Plattform.

5.3 Verfügbarkeit

Zielsetzung

Die Plattform muss einen zuverlässigen und stabilen Betrieb der bereitgestellten KI-Komponenten ermöglichen. Ziel ist eine hohe Verfügbarkeit der Anwendungen unter Nutzung der bestehenden Betriebs- und Infrastrukturmgebung der Auftraggeberin.

Die Anforderungen beziehen sich ausschließlich auf die durch den Auftragnehmer bereitgestellten Komponenten. Anforderungen an Rechenzentrum, Netzwerk, Kubernetes-Cluster, Speicher oder sonstige Infrastruktur sind nicht Bestandteil dieser Leistung.

Architekturgrundsatz

Die Anwendungen des AI Harness, des Urban KI Hub sowie der ML-Services sind so zu entwickeln, dass sie innerhalb der vorhandenen Container- und Betriebsplattform hochverfügbar betrieben werden können.

Die Bereitstellung redundanter Infrastrukturkomponenten erfolgt durch die Auftraggeberin.

NFA-009 – Verfügbarkeit der Plattformkomponenten

Anforderungs- beschreibung	Die bereitgestellten Softwarekomponenten müssen für einen hochverfügbaren Betrieb innerhalb der vorhandenen Infrastruktur geeignet sein. Die angestrebte Gesamtverfügbarkeit der Plattform beträgt 99,5 % im Jahresmittel, soweit diese durch die bereitgestellten Komponenten beeinflussbar ist.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen sind zustandsarm, containerfähig und für den redundanten Betrieb innerhalb der vorhandenen Infrastruktur auszulegen.
Begründung	Eine hohe Verfügbarkeit wird durch das Zusammenspiel von Anwendung und vorhandener Infrastruktur erreicht. Der Auftragnehmer kann ausschließlich den Anwendungsanteil beeinflussen.

NFA-010 – Fehlertoleranz

Anforderungsbeschreibung	Die bereitgestellten Anwendungen müssen so entwickelt werden, dass der Ausfall einzelner Instanzen nicht zum Ausfall der gesamten Anwendung führt.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen müssen den parallelen Betrieb mehrerer Instanzen unterstützen und dürfen keine unnötigen Single Points of Failure innerhalb der Anwendung enthalten.
Begründung	Fehlertolerante Anwendungen bilden die Grundlage für einen stabilen Betrieb innerhalb einer containerbasierten Infrastruktur.

NFA-011 – Wiederherstellbarkeit

Anforderungsbeschreibung	Die bereitgestellten Anwendungen müssen nach einem Fehler oder Neustart ohne manuelle Eingriffe wieder betriebsfähig sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen sind so zu entwickeln, dass sie automatisiert neu gestartet und erneut in Betrieb genommen werden können.
Begründung	Moderne containerbasierte Anwendungen müssen automatisch wiederherstellbar sein und dürfen keine manuellen Wiederanlaufprozesse erfordern.

NFA-012 – Datensicherung

Anforderungsbeschreibung	Die Plattform muss bestehende Sicherungs- und Wiederherstellungsverfahren der Auftraggeberin unterstützen. Eigene Backup- oder Restore-Infrastrukturen sind nicht Bestandteil der Leistung.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Persistente Daten sind so abzulegen, dass sie durch die vorhandenen Sicherungsverfahren der Betriebsplattform gesichert werden können.
Begründung	Die Datensicherung ist Aufgabe der bestehenden Betriebsumgebung. Die Anwendungen müssen diese Verfahren unterstützen, jedoch keine eigenen Backup-Lösungen bereitstellen.

NFA-013 – Wartbarkeit

Anforderungs- beschreibung	Die bereitgestellten Komponenten müssen so entwickelt werden, dass Updates, Fehlerbehebungen und neue Versionen mit möglichst geringer Beeinträchtigung des laufenden Betriebs eingespielt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen müssen Rolling Updates, Container-Restarts sowie den parallelen Betrieb unterschiedlicher Versionen unterstützen, soweit dies durch die vorhandene Betriebsplattform ermöglicht wird.
Begründung	Wartbare Anwendungen reduzieren Betriebsunterbrechungen und ermöglichen eine kontinuierliche Weiterentwicklung der Plattform.

5.4 Sicherheit

Zielsetzung

Die Plattform muss einen sicheren Betrieb sämtlicher KI-Komponenten gewährleisten. Ziel ist der Schutz von Daten, Modellen, KI-Prozessen und Schnittstellen vor unbefugtem Zugriff sowie vor klassischen und KI-spezifischen Angriffsformen.

Die Anforderungen beziehen sich ausschließlich auf die durch den Auftragnehmer bereitgestellten Komponenten. Allgemeine Infrastruktur-, Netzwerk- und Identitätssicherheitsfunktionen werden durch die bestehende Betriebsumgebung der Auftraggeberin bereitgestellt und sind zu integrieren.

Architekturgrundsatz

Die Sicherheitsmechanismen sind als Bestandteil des AI Harness umzusetzen und arbeiten mit der Policy Engine, den Guardrails, der Context Engine sowie den Audit- und Monitoring-Funktionen zusammen.

Vorhandene Authentifizierungs- und Autorisierungssysteme der Auftraggeberin sind zu nutzen. Eigene Identity-Management- oder Netzwerk-Sicherheitslösungen sind nicht Bestandteil der Leistung.

NFA-014 – Verschlüsselung

Anforderungs- beschreibung	Die Plattform muss die Verschlüsselung sämtlicher Datenübertragungen sowie gespeicherter Daten unterstützen. Hierbei sind etablierte und dem Stand der Technik entsprechende kryptografische Verfahren zu verwenden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen müssen die Verschlüsselungsmechanismen der vorhandenen Infrastruktur nutzen und unterstützen. Eigene Verschlüsselungsinfrastrukturen sind nicht Bestandteil der Leistung.
Begründung	Verschlüsselung ist eine grundlegende Voraussetzung für den sicheren Betrieb der Plattform und den Schutz sensibler Verwaltungsdaten.

NFA-015 – Rollenbasierte Zugriffskontrolle

Anforderungs- beschreibung	Die Plattform muss eine rollen- und berechtigungsbasierte Zugriffskontrolle (RBAC) unterstützen und konsistent auf sämtliche KI-Komponenten anwenden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Rollen- und Berechtigungsinformationen sind aus der vorhandenen Authentifizierungs- und Autorisierungsinfrastruktur zu übernehmen.
Begründung	Eine konsistente Zugriffskontrolle gewährleistet den sicheren und mandantenfähigen Betrieb der Plattform.

NFA-016 – Integration der Authentifizierung

Anforderungs- beschreibung	Die Plattform muss die vorhandenen Authentifizierungsverfahren der Auftraggeberin integrieren. Unterstützt werden müssen insbesondere OpenID Connect, OAuth 2.0, SAML 2.0 oder funktional gleichwertige Verfahren.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Leistung umfasst ausschließlich die Integration vorhandener Authentifizierungsdienste. Die Bereitstellung eines eigenen Identity-Management-Systems ist nicht Bestandteil der Leistung.
Begründung	Die Nutzung der vorhandenen Identitätsinfrastruktur reduziert Implementierungsaufwand und gewährleistet eine einheitliche Benutzerverwaltung.

NFA-017 – KI-spezifische Sicherheitsmechanismen

Anforderungs- beschreibung	Die Plattform muss Schutzmechanismen gegen KI-spezifische Angriffe bereitstellen. Hierzu gehören insbesondere Maßnahmen gegen Prompt Injection, indirekte Prompt Injection, Prompt Leakage, Jailbreak-Versuche, unautorisierte Tool-Aufrufe, Tool Poisoning, Datenexfiltration, missbräuchliche Modellnutzung sowie Data Poisoning. Sicherheitsrelevante Ereignisse müssen erkannt, nachvollziehbar protokolliert und administrativ konfigurierbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Schutzmechanismen sind innerhalb des AI Harness über Guardrails, Policy Engine und weitere Sicherheitskomponenten umzusetzen. Die Konfiguration erfolgt zentral über die Administrationsoberfläche.
Begründung	KI-spezifische Angriffe stellen ein wesentliches Risiko moderner KI-Plattformen dar und erfordern integrierte Schutzmechanismen.

NFA-018 – Sicherheitsprotokollierung

Anforderungs- beschreibung	Sicherheitsrelevante Ereignisse müssen revisionssicher protokolliert und für berechtigte Administratoren nachvollziehbar bereitgestellt werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Protokollierung erfolgt zentral über die Audit- und Monitoring-Komponenten des AI Harness. Vorhandene Logging- und Monitoring-Infrastrukturen der Auftraggeberin sind einzubinden.
Begründung	Eine vollständige Sicherheitsprotokollierung ist Voraussetzung für Incident Response, Compliance und den sicheren Betrieb der Plattform.

NFA-019 – Erweiterbare Sicherheitsarchitektur

Anforderungs- beschreibung	Die Architektur muss die spätere Integration weiterer Sicherheitsmechanismen unterstützen, insbesondere zusätzliche Guardrails, Bedrohungsanalysen, Sicherheitsbewertungen von Modellen und Tools sowie neue regulatorische Anforderungen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Der AI Harness muss standardisierte Erweiterungspunkte für zusätzliche Sicherheitskomponenten bereitstellen. Eine vollständige Umsetzung ist nicht Bestandteil der Erstimplementierung.
Begründung	Sicherheitsanforderungen für KI-Systeme entwickeln sich kontinuierlich weiter. Eine erweiterbare Architektur gewährleistet langfristige Zukunftssicherheit.

5.5 Datenschutz

Zielsetzung

Die Plattform muss die datenschutzkonforme Verarbeitung personenbezogener und sensibler Informationen unterstützen. Ziel ist es, sämtliche KI-Funktionen unter Einhaltung der Datenschutz-Grundverordnung (DSGVO), der Anforderungen des AI Acts sowie weiterer einschlägiger rechtlicher Vorgaben bereitzustellen.

Die Anforderungen beziehen sich ausschließlich auf die durch den Auftragnehmer bereitgestellten Komponenten. Allgemeine organisatorische Datenschutzmaßnahmen sowie die Datenschutzorganisation der Auftraggeberin sind nicht Bestandteil der Leistung.

Architekturgrundsatz

Datenschutz ist als integraler Bestandteil des AI Harness umzusetzen und arbeitet mit der Context Engine, dem Memory-Konzept, der Policy Engine, den Guardrails sowie den ML-Services zusammen.

Die Plattform muss datenschutzfreundliche Voreinstellungen (Privacy by Default) sowie datenschutzgerechte Systemgestaltung (Privacy by Design) unterstützen.

NFA-020 – Datenminimierung

Anforderungsbeschreibung	Die Plattform muss sicherstellen, dass ausschließlich die für die jeweilige Verarbeitung erforderlichen Informationen an KI-Modelle, Tools oder externe Dienste übermittelt werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Context Engine muss ausschließlich zulässige und erforderliche Kontextinformationen an Modelle oder Tools übergeben.
Begründung	Datenminimierung ist ein Grundprinzip der DSGVO und reduziert gleichzeitig Datenschutz- und Sicherheitsrisiken.

NFA-021 – Pseudonymisierung und Anonymisierung

Anforderungsbeschreibung	Personenbezogene Daten müssen vor der Übermittlung an externe Modelle oder Dienste automatisiert erkannt und – sofern erforderlich – pseudonymisiert oder anonymisiert werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Umsetzung erfolgt unter Nutzung der bereitgestellten PII-Erkennungsdienste aus Los 3 sowie der Datenschutzfunktionen des AI Harness.
Begründung	Die automatisierte Erkennung personenbezogener Daten unterstützt die datenschutzkonforme Nutzung externer KI-Modelle.

NFA-022 – Lösch- und Aufbewahrungskonzept

Anforderungs- beschreibung	Die Plattform muss konfigurierbare Lösch- und Aufbewahrungsregeln für Chatverläufe, Protokolle, Dokumente, Memory-Inhalte und weitere KI-bezogene Daten unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Löschfristen sind mandantenbezogen administrierbar bereitzustellen und auf sämtliche relevanten Datenbestände anzuwenden.
Begründung	Einheitliche Lösch- und Aufbewahrungskonzepte unterstützen Datenschutz, Compliance und gesetzliche Aufbewahrungspflichten.

NFA-023 – Verwaltung von Memory-Daten

Anforderungs- beschreibung	Nutzer müssen nachvollziehen können, welche personenbezogenen Informationen im User Memory gespeichert sind. Im Rahmen ihrer Berechtigungen müssen sie Memory-Einträge einsehen, löschen oder deaktivieren können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Verwaltung erfolgt über die Benutzeroberfläche des Urban KI Hub sowie die Administrationsfunktionen des AI Harness.
Begründung	Transparenz und Kontrolle über personenbezogene Memory-Daten sind wesentliche Anforderungen der DSGVO und stärken das Vertrauen der Nutzenden.

NFA-024 – Datenlokation

Anforderungs- beschreibung	Die Verarbeitung und Speicherung sensibler Daten muss innerhalb der Europäischen Union oder in nachweislich DSGVO-konformen Betriebsumgebungen erfolgen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform muss ausschließlich entsprechend konfigurierte Modell- und Serviceanbieter nutzen beziehungsweise deren Auswahl unterstützen.
Begründung	Die Einhaltung europäischer Datenschutzanforderungen ist Voraussetzung für den Einsatz der Plattform in der öffentlichen Verwaltung.

NFA-025 – Trennung von Nutzerdaten und Modelltraining

Anforderungsbeschreibung	Die Plattform muss sicherstellen, dass Nutzerdaten nicht ohne ausdrückliche Freigabe für das Training oder die Optimierung externer KI-Modelle verwendet werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die eingesetzten Modellanbieter und Schnittstellen sind entsprechend zu konfigurieren oder auszuwählen.
Begründung	Die Trennung zwischen Nutzungsdaten und Modelltraining ist eine wesentliche Voraussetzung für Datenschutz und Informationssicherheit.

NFA-026 – Erweiterbare Datenschutzarchitektur

Anforderungsbeschreibung	Die Architektur muss die spätere Integration zusätzlicher Datenschutzfunktionen unterstützen, insbesondere Einwilligungsmanagement, Datenschutz-Folgenabschätzungen, automatisierte Compliance-Prüfungen oder weitere regulatorische Anforderungen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Der AI Harness muss standardisierte Erweiterungspunkte für zusätzliche Datenschutz- und Compliance-Funktionen bereitstellen. Eine vollständige Umsetzung ist nicht Bestandteil der Erstimplementierung.
Begründung	Datenschutzanforderungen entwickeln sich kontinuierlich weiter. Eine erweiterbare Architektur gewährleistet langfristige Rechtssicherheit und Zukunftsfähigkeit.

5.6 Erweiterbarkeit & Modularität

Zielsetzung

Die Plattform muss langfristig erweiterbar, modular und anpassungsfähig ausgelegt sein. Ziel ist es, zukünftige Anforderungen, zusätzliche KI-Funktionen, neue Modelle sowie weitere Integrationen ohne grundlegende Änderungen an der Gesamtarchitektur umsetzen zu können. Die Erstimplementierung konzentriert sich auf die Bereitstellung einer modularen und schnittstellenbasierten Architektur. Erweiterungen sollen durch Konfiguration, standardisierte Schnittstellen oder den Austausch einzelner Komponenten ermöglicht werden.

Architekturgrundsatz

Die Plattform ist nach dem Prinzip einer lose gekoppelten Service-Architektur aufzubauen. Der AI Harness bildet die zentrale Integrations- und Orchestrierungsschicht. Frontend, KI-Dienste, ML-Services, Retrieval-Komponenten sowie externe Systeme kommunizieren ausschließlich über definierte und dokumentierte Schnittstellen.

Bereits vorhandene Infrastrukturkomponenten der Auftraggeberin, insbesondere Kubernetes, OpenSearch, ClickHouse, S3, Authentifizierungssysteme sowie der IONOS Model Hub, sind einzubinden und dürfen nicht durch proprietäre oder redundante Komponenten ersetzt werden.

NFA-027 – Modulare Architektur

Anforderungsbeschreibung	Die Plattform muss modular aufgebaut sein, sodass einzelne Komponenten unabhängig voneinander entwickelt, betrieben, aktualisiert oder ersetzt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Komponenten sind als lose gekoppelte Dienste mit klar definierten Verantwortlichkeiten und Schnittstellen umzusetzen.
Begründung	Eine modulare Architektur reduziert Abhängigkeiten, vereinfacht Wartung und ermöglicht den schrittweisen Ausbau der Plattform.

NFA-028 – Austauschbarkeit von Komponenten

Anforderungsbeschreibung	KI-Modelle, Retrieval-Komponenten, ML-Services sowie weitere technische Komponenten müssen ohne grundlegende Änderungen an der Gesamtarchitektur austauschbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anbindung erfolgt ausschließlich über standardisierte Schnittstellen sowie die Modellabstraktions- und Routing-Schicht des AI Harness.
Begründung	Die Austauschbarkeit verhindert Herstellerabhängigkeiten und unterstützt die langfristige Weiterentwicklung der Plattform.

NFA-029 – Schnittstellenbasierte Integration

Anforderungsbeschreibung	Sämtliche Komponenten müssen ausschließlich über standardisierte, dokumentierte und versionierte Schnittstellen miteinander kommunizieren. Direkte Kopplungen zwischen einzelnen Komponenten sind zu vermeiden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Für die Kommunikation sind REST-APIs oder funktional gleichwertige Schnittstellen zu verwenden. Änderungen an einer Komponente dürfen keine Anpassungen an anderen Komponenten erfordern, sofern die Schnittstellen unverändert bleiben.
Begründung	Standardisierte Schnittstellen erhöhen die Interoperabilität und erleichtern den Austausch einzelner Komponenten.

NFA-030 – Erweiterbarkeit durch Drittsysteme

Anforderungsbeschreibung	Die Plattform muss die Integration zusätzlicher Dienste, Fachverfahren oder externer Systeme über standardisierte Schnittstellen unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die erforderlichen Schnittstellen sind vollständig zu dokumentieren und für Drittanbieter nutzbar bereitzustellen.
Begründung	Die Integration weiterer Systeme ist ein wesentlicher Bestandteil einer offenen und interoperablen Plattformarchitektur.

NFA-031 – Releasefähigkeit

Anforderungsbeschreibung	Neue Versionen und Fehlerkorrekturen müssen ohne grundlegende Unterbrechung des laufenden Betriebs eingespielt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen müssen containerfähig sein und Rolling Updates oder funktional gleichwertige Verfahren unterstützen. Die Umsetzung erfolgt unter Nutzung der vorhandenen Kubernetes-Plattform.
Begründung	Eine moderne Releasefähigkeit reduziert Betriebsunterbrechungen und erleichtert die kontinuierliche Weiterentwicklung der Plattform.

NFA-032 – Erweiterbare Zielarchitektur

Anforderungsbeschreibung	Die Architektur muss die spätere Integration zusätzlicher KI-Modelle, Agenten, ML-Services, Datenquellen, Sicherheitsmechanismen und weiterer Plattformkomponenten unterstützen, ohne grundlegende Änderungen an der Gesamtarchitektur erforderlich zu machen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Der AI Harness sowie sämtliche Schnittstellen sind so zu gestalten, dass zukünftige Erweiterungen überwiegend durch Konfiguration oder die Ergänzung zusätzlicher Komponenten erfolgen können.
Begründung	Die vorbereitete Zielarchitektur ermöglicht eine langfristige Weiterentwicklung der Plattform und schützt die Investition der Auftraggeberin.

6 Architektur- und Integrationsanforderungen

Die Architektur des KI-Stacks ist modular, komponentenbasiert und lose gekoppelt auszugestalten. Ziel ist die Bereitstellung einer langfristig erweiterbaren, interoperablen und herstellerunabhängigen Plattform, die sich nahtlos in den bestehenden Urban Stack integriert.

Der AI Harness bildet die zentrale Integrations-, Orchestrierungs- und Laufzeitumgebung des KI-Stacks. Er stellt die gemeinsame Ausführungs- und Steuerungsschicht für sämtliche KI-Funktionen bereit und übernimmt die Koordination der Modellanbindung, Agent Runtime, Context Engine, Retrieval-Prozesse, Tool-Integration, Prompt-Verwaltung, Policy Enforcement, Guardrails, Monitoring sowie Auditierungsfunktionen.

Sämtliche KI-Aufrufe, Agentenausführungen, Retrieval-Prozesse und Tool-Aufrufe sind über den AI Harness abzuwickeln.

Die Architektur ist nach dem Prinzip klar abgegrenzter fachlicher und technischer Verantwortlichkeiten aufzubauen. Benutzerinteraktion, Orchestrierung, Retrieval, Modellinferenz, Dokumentenverarbeitung, ML-Services sowie Werkzeugintegration sind als logisch getrennte Komponenten auszugestalten.

Die Komponenten müssen unabhängig voneinander betrieben, erweitert, aktualisiert, ausgetauscht und skaliert werden können. Die Kommunikation zwischen den Komponenten erfolgt ausschließlich über standardisierte, dokumentierte und versionierte Schnittstellen.

Die Orchestrierungsschicht, die Dokumentenverarbeitung sowie die ML-Services sind als eigenständige und unabhängig skalierbare Komponenten bereitzustellen. Monolithische Implementierungen sind nur zulässig, sofern die geforderte Austauschbarkeit, Erweiterbarkeit und Skalierbarkeit uneingeschränkt gewährleistet bleibt.

Die KI-Funktionen des KI-Stacks sind als backendseitige Dienste bereitzustellen. Die Nutzung durch den Urban KI Hub sowie weitere berechnete Anwendungen oder Komponenten des Urban Stack erfolgt ausschließlich über die bereitgestellten standardisierten Programmierschnittstellen. Der KI-Stack darf keine technische Abhängigkeit zu einer bestimmten Benutzeroberfläche voraussetzen.

Die Plattform muss die Integration einer Dokumentenverarbeitungspipeline unterstützen. Diese dient der Übernahme, Aufbereitung und Bereitstellung von Dokumenten für Retrieval-, RAG- und weitere KI-Prozesse. Die konkrete funktionale Ausgestaltung der Dokumentenverarbeitung richtet sich nach den funktionalen Anforderungen dieser Leistungsbeschreibung.

Die Architektur muss Human-in-the-Loop-Prozesse als integralen Bestandteil der KI-Verarbeitung unterstützen. Entsprechende Entscheidungs- und Freigabeschritte müssen Bestandteil des orchestrierten Verarbeitungsflusses sein und dürfen nicht ausschließlich auf Anwendungsebene umgesetzt werden.

Monitoring- und Observability-Funktionen sind komponentenübergreifend vorzusehen und müssen sämtliche wesentlichen Verarbeitungsabläufe des KI-Stacks unterstützen. Die konkreten Anforderungen an Monitoring, Protokollierung, Governance und Policy Management ergeben sich aus den funktionalen und nichtfunktionalen Anforderungen dieser Leistungsbeschreibung.

6.1 Trennung von Retrieval, Modell und Orchestrierung

Die Architektur des KI-Stacks muss eine klare fachliche und technische Trennung der zentralen Verarbeitungsschichten gewährleisten. Ziel ist eine modulare, austauschbare und herstellerunabhängige Architektur, in der einzelne Komponenten unabhängig voneinander weiterentwickelt und ersetzt werden können.

Die Plattform muss mindestens folgende logisch getrennte Architekturkomponenten unterstützen:

- Retrieval-Schicht zur Suche, Filterung und Bereitstellung relevanter Informationen,
- Modellschicht zur Ausführung von Sprachmodellen und weiteren KI-Modellen,
- Orchestrierungsschicht (AI Harness) zur Steuerung von KI-Prozessen, Agenten, Modellaufrufen, Retrieval, Tool-Integration, Prompt-Ausführung sowie der Zusammenstellung des Verarbeitungskontexts.

Die Context Engine ist Bestandteil des AI Harness. Sie übernimmt die kontrollierte und regelbasierte Zusammenstellung des für eine KI-Ausführung erforderlichen Gesamtkontexts. Hierbei können insbesondere Nutzerprofil, Rolle, Mandant, Konversationsverlauf, Agentenkonfiguration, Prompt-Version, Memory, Wissensquellen, Retrieval-Ergebnisse, Tool-Kontext, Policies und Berechtigungen berücksichtigt werden.

Die Auswahl der tatsächlich an Modelle oder Tools übergebenen Kontextinformationen erfolgt ausschließlich auf Grundlage fachlicher Relevanz, bestehender Berechtigungen sowie geltender Datenschutz-, Sicherheits- und Governance-Vorgaben.

Die Retrieval-Schicht muss unabhängig von den eingesetzten KI-Modellen betrieben werden können. Such-, Indexierungs- und Abrufvorgänge dürfen nicht an einen bestimmten Modellanbieter oder eine bestimmte Modelltechnologie gebunden sein.

Ebenso muss die Orchestrierungsschicht modellunabhängig ausgelegt sein. Die Auswahl geeigneter Modelle, Retrieval-Verfahren und Werkzeuge erfolgt durch den AI Harness auf Basis des jeweiligen Anwendungsfalls sowie der konfigurierten Richtlinien.

Die Kommunikation zwischen Retrieval, Modellschicht und Orchestrierung erfolgt ausschließlich über standardisierte und dokumentierte Schnittstellen. Dadurch muss sichergestellt werden, dass einzelne Komponenten ohne grundlegende Änderungen an der Gesamtarchitektur ausgetauscht oder erweitert werden können.

6.2 OpenSearch als zentrale Suchkomponente innerhalb der Hybrid-Retrieval-Architektur

Die Retrieval-Schicht ist als eigenständige und modellunabhängige Architekturkomponente auszugestalten. Ziel ist die Bereitstellung einer flexiblen Hybrid-Retrieval-Architektur, welche unterschiedliche Suchverfahren kombiniert und unabhängig von den eingesetzten Sprachmodellen betrieben werden kann.

Für die semantische Suche, die Vektorsuche sowie die klassische dokumentenbasierte Suche ist die bereits vorhandene OpenSearch-Instanz der Auftraggeberin als zentrale Suchkomponente zu integrieren. Die Leistung umfasst die Integration und Konfiguration der bestehenden OpenSearch-Infrastruktur. Aufbau, Betrieb oder Erweiterung der OpenSearch-Plattform sind nicht Bestandteil der Leistung.

Die Retrieval-Schicht muss mindestens folgende Verfahren unterstützen:

semantische Vektorsuche,

lexikalische Volltextsuche,

hybride Suche,

Metadatenfilterung,

berechtigungsabhängige Filterung (Security Trimming),

Bereitstellung relevanter Kontexte für Retrieval-Augmented Generation (RAG).

Die Auswahl und Kombination der Retrieval-Verfahren erfolgt durch den AI Harness und muss abhängig vom jeweiligen Anwendungsfall konfigurierbar sein.

Die Architektur muss die spätere Integration zusätzlicher Verfahren zur Anfrageoptimierung unterstützen, insbesondere Query Rewriting, Query Expansion oder funktional gleichwertige Verfahren. Eine vollständige Umsetzung dieser Funktionen ist nicht Bestandteil der produktiven Erstimplementierung.

Ebenso muss die Architektur die spätere Integration von Reranking-Verfahren unterstützen. Hierzu gehören insbesondere Cross-Encoder, neuronale Reranker oder funktional gleichwertige Verfahren. Die Architektur ist so auszulegen, dass entsprechende Komponenten ohne Änderungen der Gesamtarchitektur ergänzt werden können.

Die Plattform muss den Austausch von Embedding-Modellen unterstützen. Embedding-Modelle dürfen nicht zu einer technischen Abhängigkeit der Retrieval-Architektur von bestimmten Modellanbietern führen.

Beim Wechsel eines Embedding-Modells muss eine kontrollierte Neuindexierung unterstützt werden. Für erzeugte Embeddings sind das verwendete Embedding-Modell sowie dessen Version nachvollziehbar zu dokumentieren.

Die Retrieval-Schicht muss sicherstellen, dass Berechtigungen und Zugriffsrechte über den gesamten Retrieval-Prozess hinweg konsistent berücksichtigt werden. Dies umfasst insbesondere Dokumente, Dokumentversionen, Dokumentsegmente (Chunks), Metadaten sowie weitere durch die Dokumentenverarbeitung erzeugte Informationsobjekte.

Die Architektur muss die spätere Erweiterung um zusätzliche Wissensrepräsentationen, insbesondere graphbasierte Wissensstrukturen oder weitere Retrieval-Verfahren, unterstützen, ohne grundlegende Änderungen an der Gesamtarchitektur zu erfordern.

Die Gewichtung einzelner Datenquellen oder Quellgruppen soll administrativ konfigurierbar sein. Die Architektur muss diese Funktion unterstützen; eine vollständige Umsetzung einer dynamischen Relevanzsteuerung ist nicht Bestandteil der produktiven Erstimplementierung.

6.3 Flexible Modellanbindung und Modell-Routing

Die Plattform muss eine modellunabhängige Modellabstraktions- und Routing-Schicht bereitstellen. Ziel ist die flexible Nutzung unterschiedlicher KI-Modelle sowie die langfristige Vermeidung von Herstellerabhängigkeiten.

Die Modellanbindung erfolgt über den AI Harness. Dieser übernimmt die Auswahl, Ansteuerung und Verwaltung der angebundenen Modelle und stellt eine einheitliche Schnittstelle für sämtliche KI-Funktionen bereit.

Für die Erstimplementierung ist die Integration des durch die Auftraggeberin bereitgestellten **IONOS Model Hub** vorzusehen. Die Architektur muss darüber hinaus die spätere Anbindung weiterer Modellanbieter oder lokal betriebener Open-Weight-Modelle unterstützen, ohne grundlegende Änderungen an der Gesamtarchitektur zu erfordern.

Die Plattform muss die parallele Nutzung mehrerer KI-Modelle (Multi-LLM) unterstützen. Die Auswahl des jeweils eingesetzten Modells muss administrativ konfigurierbar sein und nachvollziehbar dokumentiert werden.

Das Modell-Routing ist Bestandteil des AI Harness und muss regelbasiert erfolgen. Die Routing-Regeln müssen zentral administrierbar sein und insbesondere folgende Kriterien berücksichtigen können:

Datenschutzanforderungen,
Modellverfügbarkeit,
Wirtschaftlichkeit,
fachlicher Anwendungsfall,
benötigte Modellfähigkeiten.

Die Architektur muss die spätere Erweiterung um zusätzliche Routing-Kriterien unterstützen, insbesondere Anfragekomplexität, Antwortlatenz, Kostenoptimierung, Qualitätsbewertungen oder weitere fachliche Entscheidungsregeln. Eine vollständig automatisierte Modellauswahl anhand sämtlicher Kriterien ist nicht Bestandteil der produktiven Erstimplementierung.

Die Plattform ist so auszulegen, dass bevorzugt das für den jeweiligen Anwendungsfall geeignete und wirtschaftlichste Modell eingesetzt werden kann. Der Einsatz leistungsstärkerer oder externer Modelle soll administrativ steuerbar und auf fachlich erforderliche Anwendungsfälle begrenzt sein.

Für die Nutzenden muss transparent nachvollziehbar sein, welches Modell zur Bearbeitung einer Anfrage verwendet wurde. Die konkrete Darstellung erfolgt über den Urban KI Hub.

6.4 Modellunabhängigkeit

Die Architektur des KI-Stacks ist konsequent modellunabhängig auszugestalten. Ziel ist es, die langfristige Austauschbarkeit und Erweiterbarkeit von KI-Modellen sicherzustellen sowie Herstellerabhängigkeiten zu vermeiden.

Die Anwendungslogik, Agentensteuerung, Orchestrierung, Retrieval-Prozesse sowie sämtliche Fachfunktionen dürfen nicht direkt an ein bestimmtes Sprachmodell oder einen bestimmten Modellanbieter gekoppelt sein.

Die Modellanbindung erfolgt ausschließlich über die Modellabstraktions- und Routing-Schicht des AI Harness. Ein Austausch bestehender Modelle sowie die Integration zusätzlicher Modelle müssen ohne grundlegende Änderungen an der Gesamtarchitektur möglich sein.

Die Architektur muss unterschiedliche Modelltypen unterstützen. Hierzu gehören insbesondere:

- große Sprachmodelle (Large Language Models),
- ressourceneffiziente Small Language Models (SLMs),
- spezialisierte KI-Modelle für einzelne Fachaufgaben,
- Open-Weight-Modelle,
- API-basierte Modelle externer Anbieter.

Für die produktive Erstimplementierung ist die Nutzung der durch die Auftraggeberin bereitgestellten Modellinfrastruktur vorgesehen. Die Architektur muss darüber hinaus den späteren parallelen Betrieb unterschiedlicher Modelltypen und Modellanbieter unterstützen.

Die Auswahl geeigneter Modelle erfolgt über den AI Harness. Die Architektur muss dabei die Berücksichtigung fachlicher, technischer, wirtschaftlicher sowie datenschutzrechtlicher Anforderungen unterstützen, ohne die Anwendungslogik an einzelne Modelle zu koppeln.

Die Plattform muss so ausgelegt sein, dass zukünftige Modellgenerationen oder zusätzliche Modellanbieter durch Konfiguration oder Erweiterung der Modellanbindung integriert werden können. Anpassungen an Fachanwendungen, Agenten oder Retrieval-Komponenten dürfen hierfür grundsätzlich nicht erforderlich sein.

6.5 RAG als Standardmechanismus

Retrieval-Augmented Generation (RAG) bildet den Standardmechanismus für dokumenten- und wissensbasierte Anfragen innerhalb der Plattform. Ziel ist die Generierung nachvollziehbarer und auf aktuellen Wissensquellen basierender Antworten.

Für wissensbasierte Anwendungsfälle muss vor der Generierung einer Antwort ein Retrieval-Prozess durchgeführt werden. Hierbei sind relevante Informationen über die Retrieval-Schicht zu ermitteln und der weiteren Verarbeitung durch den AI Harness als Kontext bereitzustellen.

Die Retrieval-Schicht nutzt hierfür die durch die Auftraggeberin bereitgestellte OpenSearch-Infrastruktur sowie gegebenenfalls weitere über die Hybrid-Retrieval-Architektur angebundene Wissensquellen.

Die Architektur muss RAG als regulären Verarbeitungsmechanismus für dokumenten- und wissensbasierte Anfragen unterstützen. Für Anwendungsfälle, bei denen kein externer Wissensbezug erforderlich ist, insbesondere allgemeine Sprachassistentz, Texttransformationen oder Übersetzungen, muss die Architektur eine Verarbeitung ohne vorgeschalteten Retrieval-Prozess ermöglichen.

Die Einbindung der Retrieval-Ergebnisse in den Modellkontext erfolgt durch die Context Engine des AI Harness unter Berücksichtigung von Berechtigungen, Datenschutz-, Governance- und Sicherheitsvorgaben.

Die Architektur muss die spätere Erweiterung des RAG-Mechanismus um zusätzliche Retrieval-Verfahren, Wissensquellen oder Optimierungsverfahren unterstützen, ohne grundlegende Änderungen an der Gesamtarchitektur zu erfordern.

6.6 Token- und Ressourcenoptimierung

Die Plattform muss Mechanismen zur effizienten Nutzung von KI-Modellen, Retrieval-Verfahren und Systemressourcen unterstützen. Ziel ist die wirtschaftliche Verarbeitung von KI-Anfragen unter Berücksichtigung von Kosten, Antwortqualität und Performance.

Die Optimierung erfolgt durch den AI Harness in Zusammenarbeit mit der Context Engine, der Retrieval-Schicht sowie dem Modell-Routing.

Die Architektur muss insbesondere folgende Optimierungsmechanismen unterstützen:

Begrenzung der an Modelle übergebenen Kontextinformationen auf den jeweils erforderlichen Umfang,

Übergabe ausschließlich relevanter Retrieval-Ergebnisse und Dokumentsegmente,

Vermeidung redundanter Kontextübertragungen,

Wiederverwendung vorhandener Embeddings, Retrieval-Ergebnisse oder weiterer zwischengespeicherter Informationen, soweit dies fachlich und datenschutzrechtlich zulässig ist,

Reduzierung unnötiger Modellaufrufe durch vorgelagerte Retrieval-, Filter-, Policy- oder Regelmechanismen,

Unterstützung der Kontextverdichtung oder Zusammenfassung längerer Gesprächsverläufe zur Begrenzung des Kontextumfangs.

Die Architektur muss darüber hinaus die bevorzugte Nutzung geeigneter und wirtschaftlicher KI-Modelle unterstützen. Hierzu können abhängig vom jeweiligen Anwendungsfall kleinere oder spezialisierte Modelle eingesetzt werden, sofern diese die fachlichen Anforderungen erfüllen.

Die konkrete Auswahl von Optimierungsverfahren erfolgt administrativ konfigurierbar über den AI Harness. Die Architektur muss die spätere Integration weiterer Verfahren zur Kosten-, Ressourcen- und Performanceoptimierung unterstützen, ohne grundlegende Änderungen an der Gesamtarchitektur zu erfordern.

6.7 Agent Runtime

Die Agent Runtime ist Bestandteil des zentralen AI Harness und bildet die einheitliche Laufzeitumgebung für sämtliche KI-Agenten innerhalb der Plattform.

Ziel der Agent Runtime ist die standardisierte, nachvollziehbare und kontrollierbare Ausführung agentenbasierter Prozesse. Sie übernimmt die Orchestrierung von Modellaufrufen, Retrieval-Prozessen, Tool-Interaktionen sowie mehrstufigen Bearbeitungsabläufen und stellt sicher, dass sämtliche Agenten nach denselben Governance-, Sicherheits- und Datenschutzvorgaben ausgeführt werden.

Die Agent Runtime muss insbesondere folgende Funktionen unterstützen:

Ausführung agentenspezifischer Konfigurationen und Instruktionen,

Anbindung der über den AI Harness bereitgestellten KI-Modelle,

Integration der Retrieval-Schicht und weiterer Wissensquellen,

Steuerung von Tool-Aufrufen,

Unterstützung mehrstufiger Bearbeitungsabläufe,

Nutzung mehrerer Werkzeuge und Datenquellen innerhalb eines Bearbeitungsvorgangs,

Verwaltung von Konversations- und Laufzeitkontexten,

Nutzung der Context Engine zur Zusammenstellung des für die jeweilige Verarbeitung erforderlichen Gesamtkontexts,

Berücksichtigung von Rollen, Berechtigungen, Mandantenkontext, Policies, Guardrails sowie Datenschutzvorgaben,

Unterstützung unterschiedlicher Memory-Ebenen entsprechend den funktionalen Anforderungen dieser Leistungsbeschreibung,

Nachvollziehbare Protokollierung der Ausführungsschritte, verwendeten Konfigurationen sowie aufgerufenen Modelle und Werkzeuge,

Behandlung von Fehler-, Abbruch- und Eskalationsfällen.

Die Agent Runtime muss mandantenfähig ausgelegt sein und Berechtigungen sowie Security-Trimming-Regeln über den gesamten Verarbeitungsablauf hinweg konsistent berücksichtigen.

Die Architektur muss die schrittweise Erweiterung der Agent Runtime um zusätzliche Agentenfähigkeiten, Planungsstrategien, Entscheidungslogiken oder weitere Ausführungsmechanismen unterstützen, ohne grundlegende Änderungen an der Gesamtarchitektur zu erfordern.

Die konkrete technische Umsetzung der Agent Runtime bleibt dem Auftragnehmer überlassen, sofern die funktionalen und architektonischen Anforderungen dieser Leistungsbeschreibung erfüllt werden.

6.8 Tool-Integration und ereignisgesteuerte Verarbeitung

Die Plattform muss standardisierte Mechanismen zur Integration externer Werkzeuge, Dienste und Fachsysteme bereitstellen. Ziel ist die kontrollierte und nachvollziehbare Nutzung externer Funktionen innerhalb der KI-Plattform unter Beibehaltung einer lose gekoppelten Architektur.

Die Tool-Integration erfolgt über den AI Harness. Externe Werkzeuge und Dienste sind ausschließlich über standardisierte, dokumentierte und versionierte Schnittstellen einzubinden. Die konkrete technische Umsetzung der jeweiligen Tools bleibt unabhängig von den eingesetzten KI-Modellen.

Die Plattform muss sicherstellen, dass Tool-Aufrufe unter Berücksichtigung von Rollen, Berechtigungen, Mandantenkontext, Guardrails, Policies sowie Datenschutzvorgaben kontrolliert ausgeführt und revisionssicher protokolliert werden.

Für die produktive Erstimplementierung ist die Integration der im Rahmen dieser Leistungsbeschreibung vorgesehenen Werkzeuge sowie der definierten REST-Schnittstellen ausreichend. Die Architektur muss darüber hinaus die spätere Anbindung weiterer Werkzeuge, Fachverfahren oder externer Dienste unterstützen, ohne grundlegende Änderungen an der Gesamtarchitektur zu erfordern.

Die Architektur muss die Integration ereignisgesteuerter Verarbeitungsprozesse unterstützen. Ereignisse aus internen oder externen Systemen können als Auslöser für Agenten, Retrieval-Prozesse, Dokumentenverarbeitung, ML-Services oder weitere KI-Funktionen dienen.

Die vollständige Umsetzung einer allgemeinen Event-Processing- oder Workflow-Plattform ist nicht Bestandteil der produktiven Erstimplementierung. Die Architektur muss jedoch die spätere Integration entsprechender Komponenten über standardisierte Schnittstellen unterstützen.

Sämtliche ereignisgesteuerten Verarbeitungsabläufe müssen nachvollziehbar protokolliert sowie rollen-, mandanten- und berechtigungsabhängig steuerbar sein.

6.9 Protokollierung und Nachvollziehbarkeit

Die Plattform muss eine durchgängige und nachvollziehbare Protokollierung sämtlicher wesentlicher Verarbeitungsschritte innerhalb des KI-Stacks unterstützen. Ziel ist die technische und fachliche Nachvollziehbarkeit von KI-Ausführungen sowie die Unterstützung von Betrieb, Fehleranalyse, Auditierung und Governance.

Die Protokollierung erfolgt zentral über den AI Harness und umfasst insbesondere die Verarbeitungsschritte innerhalb der Orchestrierung, der Retrieval-Schicht, der Modellanbindung, der Tool-Integration sowie der Agent Runtime.

Die Plattform muss mindestens folgende Informationen nachvollziehbar protokollieren:

- Eingang und Verarbeitung von Anfragen,
- Auswahl des verwendeten KI-Modells,
- Durchführung von Retrieval-Prozessen,
- verwendete Wissensquellen und Metadatenfilter,
- Tool-Aufrufe und deren Ergebnisse,
- wesentliche Verarbeitungsschritte innerhalb agentenbasierter Abläufe,

Fehler-, Abbruch- und Eskalationsereignisse, sicherheits- und governancebezogene Entscheidungen.

Die Protokollierung muss so ausgestaltet sein, dass technische Abläufe, Fehlerursachen sowie die wesentlichen fachlichen Entscheidungen nachvollzogen werden können. Die Speicherung und Verarbeitung der Protokolldaten hat unter Berücksichtigung der geltenden Datenschutz- und Sicherheitsanforderungen zu erfolgen.

Jede Agenten-, Retrieval- oder RAG-Ausführung muss über eine eindeutige Ausführungskennung (Trace-ID) identifizierbar sein. Die Trace-ID dient der korrelierbaren Zuordnung sämtlicher zugehöriger Verarbeitungsschritte innerhalb eines Ausführungspfades.

Die Architektur muss die spätere Erweiterung um weitergehende Observability- und Analysefunktionen unterstützen. Hierzu gehören insbesondere die vollständige Nachvollziehbarkeit komplexer Agentenausführungen, die Dokumentation verwendeter Modell- und Prompt-Versionen, Laufzeitparameter, Retrieval-Konfigurationen sowie weiterer technischer oder fachlicher Ausführungsdetails. Eine vollständige Reproduzierbarkeit sämtlicher KI-Ausführungen ist nicht Bestandteil der produktiven Erstimplementierung.

Die Protokollierung ist so auszugestalten, dass bestehende Logging-, Monitoring- und Observability-Lösungen der Auftraggeberin eingebunden werden können. Der Aufbau einer eigenständigen Logging- oder Monitoring-Infrastruktur ist nicht Bestandteil der Leistung.

6.10 Wissensgraph und Graph-RAG

Die Architektur des KI-Stacks muss die spätere Integration graphbasierter Wissensstrukturen unterstützen. Ziel ist die Möglichkeit, Beziehungen zwischen Dokumenten, Entitäten, Organisationseinheiten, Prozessen oder weiteren Wissensobjekten zukünftig strukturiert abzubilden und für KI-Anwendungen nutzbar zu machen.

Die produktive Erstimplementierung basiert auf der Hybrid-Retrieval-Architektur unter Nutzung der durch die Auftraggeberin bereitgestellten OpenSearch-Infrastruktur. Ein produktiver Wissensgraph oder eine Graphdatenbank sind nicht Bestandteil der Erstimplementierung.

Die Architektur muss jedoch so ausgestaltet sein, dass graphbasierte Wissenskomponenten zu einem späteren Zeitpunkt ohne grundlegende Änderungen an der Gesamtarchitektur integriert werden können.

Hierzu gehören insbesondere:

- Integration einer Graphdatenbank oder funktional gleichwertigen Wissenskomponente,
- Speicherung von Entitäten und Relationen,
- Verknüpfung von Dokumenten, Dokumentsegmenten und Wissensobjekten,
- Nutzung graphbasierter Retrieval-Verfahren,
- Integration von Graph-RAG als ergänzendes Retrieval-Verfahren.

Die Architektur muss sicherstellen, dass zukünftige graphbasierte Komponenten dieselben Sicherheits-, Datenschutz-, Mandanten- und Governance-Anforderungen erfüllen wie sämtliche übrigen Komponenten des KI-Stacks.

Insbesondere müssen Berechtigungen, Mandanteninformationen sowie Security-Trimming-Regeln auch bei zukünftigen graphbasierten Retrieval-Verfahren konsistent berücksichtigt werden.

Die Integration graphbasierter Wissensstrukturen erfolgt ausschließlich über standardisierte Schnittstellen des AI Harness. Dadurch muss gewährleistet werden, dass Graph-RAG zukünftig

als ergänzendes Retrieval-Verfahren neben Vektor-, Hybrid- und Volltextsuche genutzt werden kann, ohne bestehende Komponenten der Plattform grundlegend anzupassen.

6.11 Dokumentenverarbeitungspipeline

Die Plattform muss eine strukturierte und orchestrierte Dokumentenverarbeitungspipeline bereitstellen. Ziel ist die kontrollierte Übernahme, Aufbereitung und Bereitstellung von Dokumenten für Retrieval-, RAG-, Agenten- und weitere KI-Prozesse.

Die Dokumentenverarbeitungspipeline ist als eigenständige Verarbeitungskomponente innerhalb des KI-Stacks umzusetzen und über den AI Harness in die Gesamtarchitektur einzubinden. Sie ist logisch von allgemeinen ML-Services zu trennen. ML-Services können ergänzende Funktionen, insbesondere OCR, Dokumentenanalyse oder PII-Erkennung, bereitstellen, ersetzen jedoch nicht die strukturierte Dokumentenverarbeitungspipeline.

Für die produktive Erstimplementierung ist die Dokumentenverarbeitung auf Dokumente aus dem bereitgestellten S3-kompatiblen Objektspeicher sowie auf Dokumente aus definierten REST-Schnittstellen zu begrenzen. Weitere Dokumentenquellen sind architektonisch vorzubereiten, jedoch nicht Bestandteil der produktiven Erstimplementierung.

Die Dokumentenverarbeitungspipeline muss mindestens folgende Verarbeitungsschritte unterstützen:

- Übernahme von Dokumenten aus den definierten Quellen,

- Extraktion von Inhalten und Strukturinformationen,

- Segmentierung der Inhalte in geeignete Verarbeitungseinheiten,

- Anreicherung mit Metadaten,

- Erzeugung oder Anbindung von Embeddings für semantische Verarbeitung,

- Indexierung der aufbereiteten Inhalte in der Retrieval-Schicht, insbesondere in OpenSearch,

- Bereitstellung der aufbereiteten Inhalte für RAG-, Agenten- und nachgelagerte KI-Prozesse.

Die Segmentierung muss so erfolgen, dass fachlich zusammenhängende Inhalte möglichst erhalten bleiben und für Retrieval- und RAG-Prozesse nutzbar gemacht werden können. Die Architektur muss die spätere Integration erweiterter Segmentierungsverfahren unterstützen, insbesondere hierarchisches Chunking, Parent-Child-Retrieval, kontextangereicherte Segmentierung sowie strukturorientierte Verarbeitung anhand von Überschriften, Kapiteln, Tabellen oder vergleichbaren Dokumentstrukturen.

Die einzelnen Verarbeitungsschritte der Pipeline müssen modular aufgebaut sein, sodass sie unabhängig voneinander erweitert, ersetzt oder skaliert werden können.

Die Dokumentenverarbeitungspipeline muss Berechtigungs- und Mandanteninformationen aus den angebundenen Quellsystemen übernehmen oder verarbeiten können. Diese Informationen müssen an nachgelagerte Artefakte, insbesondere Dokumente, Dokumentversionen, Dokumentsegmente, Metadaten und Embeddings, weitergegeben werden.

Die Architektur muss sicherstellen, dass Berechtigungsinformationen auch bei Neuindexierungen, Re-Chunking-Prozessen oder Versionierungsereignissen erhalten bleiben und durch die Retrieval-Schicht sowie nachgelagerte RAG- und Agentenprozesse berücksichtigt werden können.

Die spätere Erweiterung der Dokumentenverarbeitungspipeline um zusätzliche Verarbeitungsstufen, insbesondere Entitäts- und Relationsextraktion, Graph-Extraktion oder weitere wissensstrukturierende Verfahren, muss architektonisch möglich sein. Eine produktive Umsetzung graphbasierter Wissensverarbeitung ist nicht Bestandteil der Erstimplementierung.

7 Betrieb & Infrastruktur

Zielsetzung

Die nachfolgenden Anforderungen definieren die Rahmenbedingungen für den Betrieb der im Rahmen dieser Ausschreibung bereitgestellten Softwarekomponenten. Ziel ist die Bereitstellung einer containerfähigen, wartbaren und skalierbaren KI-Plattform, die sich nahtlos in die bestehende Betriebs- und Infrastrukturmgebung der Auftraggeberin integriert.

Die Auftraggeberin stellt wesentliche Infrastrukturkomponenten bereits bereit. Die Leistung des Auftragnehmers umfasst daher die Integration und Nutzung dieser Infrastruktur sowie die Bereitstellung der hierfür erforderlichen Anwendungen, Konfigurationen und Deployments.

Architekturgrundsatz

Sämtliche Komponenten des KI-Stacks sind cloud-native und containerfähig auszugestalten. Sie müssen innerhalb der vorhandenen Kubernetes-Infrastruktur der Auftraggeberin betrieben werden können.

Der Aufbau oder Betrieb einer eigenständigen Infrastruktur ist nicht Bestandteil der Leistung.

Vorhandene Plattformdienste, insbesondere Kubernetes, OpenSearch, ClickHouse, S3-kompatibler Objektspeicher, Authentifizierungsdienste sowie der IONOS Model Hub, sind zu nutzen und über standardisierte Schnittstellen einzubinden.

Die Architektur muss den unabhängigen Betrieb einzelner Komponenten sowie deren schrittweise Erweiterung unterstützen.

7.1 Cloud-native Architektur

Zielsetzung

Die bereitgestellten Komponenten des KI-Stacks sind nach den Prinzipien moderner cloud-nativer Softwarearchitekturen zu entwickeln. Ziel ist eine wartbare, skalierbare und containerfähige Anwendung, die sich nahtlos in die bestehende Betriebsumgebung der Auftraggeberin integriert. Die Anforderungen beziehen sich ausschließlich auf die durch den Auftragnehmer bereitgestellten Softwarekomponenten. Der Aufbau und Betrieb der zugrundeliegenden Cloud- und Kubernetes-Infrastruktur sind nicht Bestandteil der Leistung.

Architekturgrundsatz

Sämtliche Komponenten des KI-Stacks sind containerfähig, lose gekoppelt und unabhängig voneinander betreibbar auszugestalten. Die Anwendungen müssen den Betrieb innerhalb der vorhandenen Kubernetes-Plattform der Auftraggeberin unterstützen.

BA-001 – Cloud-native Umsetzung

Anforderungsbeschreibung	Die bereitgestellten Softwarekomponenten müssen nach den Prinzipien einer cloud-nativen Architektur entwickelt werden und für den Betrieb innerhalb der vorhandenen Kubernetes-Infrastruktur geeignet sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Leistung umfasst die Entwicklung und Bereitstellung cloud-nativer Anwendungen. Der Aufbau oder Betrieb einer Cloud-Infrastruktur ist nicht Bestandteil der Leistung.
Begründung	Cloud-native Anwendungen ermöglichen einen wartbaren, skalierbaren und langfristig erweiterbaren Betrieb innerhalb der bestehenden Infrastruktur der Auftraggeberin.

BA-002 – Containerisierung

Anforderungsbeschreibung	Sämtliche Komponenten des KI-Stacks müssen containerisiert bereitgestellt werden und den Betrieb innerhalb der vorhandenen Kubernetes-Plattform unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen sind als OCI-kompatible Container bereitzustellen. Alternative cloud-native Bereitstellungsmodelle sind zulässig, sofern sie funktional gleichwertig sind und sich in die bestehende Betriebsumgebung integrieren lassen.
Begründung	Containerisierung gewährleistet einen einheitlichen, portablen und automatisierbaren Betrieb der Plattform.

BA-003 – Lose gekoppelte Komponenten

Anforderungsbeschreibung	Die Komponenten des KI-Stacks müssen lose gekoppelt aufgebaut sein, sodass sie unabhängig voneinander entwickelt, betrieben, aktualisiert, skaliert und ausgetauscht werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Kommunikation zwischen den Komponenten erfolgt ausschließlich über standardisierte und dokumentierte Schnittstellen. Direkte technische Abhängigkeiten zwischen einzelnen Komponenten sind zu vermeiden.
Begründung	Eine lose gekoppelte Architektur erhöht Wartbarkeit, Erweiterbarkeit und Herstellerunabhängigkeit der Plattform.

7.2 Nutzung von Managed Services

Zielsetzung

Die Plattform soll vorhandene Infrastruktur- und Plattformdienste der Auftraggeberin konsequent nutzen. Ziel ist die Reduzierung des Implementierungs- und Betriebsaufwands durch die Wiederverwendung bereits bereitgestellter Dienste. Die Leistung des Auftragnehmers umfasst die Integration und Nutzung dieser Dienste. Die Bereitstellung oder der Betrieb eigener Plattform- oder Infrastrukturservices ist nicht Bestandteil der Leistung, sofern dies nicht ausdrücklich gefordert wird.

Architekturgrundsatz

Vorhandene Plattformdienste sind über standardisierte Schnittstellen in den KI-Stack einzubinden. Die Architektur muss sicherstellen, dass einzelne Plattformdienste zukünftig ohne grundlegende Änderungen an der Gesamtarchitektur ersetzt oder erweitert werden können.

BA-004 – Nutzung vorhandener Plattformdienste

Anforderungs- beschreibung	Die bereitgestellten Komponenten müssen die vorhandenen Plattform- und Infrastrukturdienste der Auftraggeberin nutzen. Hierzu zählen insbesondere Kubernetes, OpenSearch, ClickHouse, S3-kompatibler Objektspeicher, Authentifizierungsdienste sowie der IONOS Model Hub.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Integration erfolgt über standardisierte Schnittstellen und vorhandene Betriebsmechanismen. Eine Bereitstellung gleichwertiger Infrastrukturkomponenten durch den Auftragnehmer ist nicht Bestandteil der Leistung.
Begründung	Die Nutzung vorhandener Plattformdienste reduziert Implementierungsaufwand, vermeidet Doppelstrukturen und unterstützt einen wirtschaftlichen Betrieb.

BA-005 – Integration vorhandener Dienste

Anforderungs- beschreibung	Der Auftragnehmer ist für die fachliche und technische Integration der bereitgestellten Plattformdienste in die Komponenten des KI-Stacks verantwortlich. Die Betriebsverantwortung für die zugrunde liegende Infrastruktur verbleibt bei der Auftraggeberin.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen müssen die vorhandenen Dienste unterstützen und deren Schnittstellen nutzen. Änderungen an der Infrastruktur sind nicht Bestandteil der Leistung.
Begründung	Eine klare Trennung zwischen Infrastruktur und Anwendung reduziert Projektrisiken und schafft eindeutige Verantwortlichkeiten.

BA-006 – Austauschbarkeit von Plattformdiensten

Anforderungs- beschreibung	Die Architektur muss so gestaltet sein, dass einzelne Plattformdienste oder Infrastrukturkomponenten zukünftig ohne grundlegende Änderungen an den Anwendungen ersetzt oder erweitert werden können.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Anbindung erfolgt ausschließlich über standardisierte und dokumentierte Schnittstellen. Proprietäre Abhängigkeiten sind auf das technisch notwendige Maß zu begrenzen.
Begründung	Die Austauschbarkeit einzelner Plattformdienste gewährleistet langfristige Flexibilität und unterstützt die digitale Souveränität der Auftraggeberin.

3.8 Verzicht auf eigene KI-Infrastruktur

Zielsetzung

Die Plattform soll vorhandene Modell- und KI-Dienste der Auftraggeberin nutzen und integrieren. Ziel ist die Bereitstellung einer flexiblen und wirtschaftlichen Modellanbindung, ohne den Aufbau einer eigenen Infrastruktur für Training oder Betrieb großer KI-Modelle. Die Leistung umfasst die Integration der bereitgestellten Modellinfrastruktur sowie die Entwicklung einer modellunabhängigen Anbindung innerhalb des AI Harness.

Architekturgrundsatz

Die Modellanbindung erfolgt ausschließlich über die Modellabstraktions- und Routing-Schicht des AI Harness. Die Plattform darf nicht von einem einzelnen Modellanbieter oder einer bestimmten Modelltechnologie abhängig sein.

BA-007 – Nutzung vorhandener Modellinfrastruktur

Anforderungsbeschreibung	Die Plattform muss die durch die Auftraggeberin bereitgestellte Modellinfrastruktur integrieren. Hierzu gehört insbesondere der IONOS Model Hub. Ergänzend können weitere DSGVO-konforme Modellanbieter oder lokal betriebene Open-Weight-Modelle angebunden werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Integration erfolgt über die Modellabstraktions- und Routing-Schicht des AI Harness. Der Betrieb der Modellinfrastruktur ist nicht Bestandteil der Leistung.
Begründung	Die Nutzung vorhandener Modellinfrastruktur reduziert Implementierungs- und Betriebsaufwand und unterstützt die digitale Souveränität der Auftraggeberin.

BA-008 – Keine eigene Trainingsinfrastruktur

Anforderungsbeschreibung	Der Aufbau und Betrieb einer eigenen Trainingsinfrastruktur für Foundation Models oder vergleichbare KI-Modelle sind nicht Bestandteil der Leistung.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Leistung beschränkt sich auf die Nutzung, Integration und Orchestrierung vorhandener KI-Modelle. Fine-Tuning oder Training eigener Foundation Models werden nicht gefordert.
Begründung	Der Verzicht auf eine Trainingsinfrastruktur reduziert den Projektumfang erheblich und entspricht dem vorgesehenen Einsatz vorhandener Modellplattformen.

BA-009 – Modellunabhängige Anbindung

Anforderungsbeschreibung	Die Architektur muss die Anbindung unterschiedlicher Modellanbieter sowie weiterer KI-Modelle unterstützen, ohne die Anwendungslogik oder den AI Harness an einzelne Anbieter zu koppeln.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Integration erfolgt ausschließlich über standardisierte Schnittstellen der Modellabstraktions- und Routing-Schicht. Zusätzliche Modellanbieter müssen zukünftig ohne grundlegende Änderungen an der Gesamtarchitektur ergänzt werden können.
Begründung	Eine modellunabhängige Architektur gewährleistet langfristige Flexibilität und vermeidet Herstellerabhängigkeiten.

3.9 Skalierung und Verfügbarkeit

Zielsetzung

Die bereitgestellten Komponenten des KI-Stacks müssen für einen skalierbaren und hochverfügbaren Betrieb innerhalb der vorhandenen Infrastruktur der Auftraggeberin ausgelegt sein. Die Anforderungen beziehen sich auf die Anwendungen und deren Betriebsfähigkeit, nicht auf die Bereitstellung der zugrundeliegenden Infrastruktur.

Architekturgrundsatz

Die Anwendungen müssen den Betrieb innerhalb der vorhandenen Kubernetes-Plattform unterstützen und so gestaltet sein, dass Skalierungs- und Hochverfügbarkeitsmechanismen der Infrastruktur genutzt werden können.

BA-010 – Skalierungsfähigkeit

Anforderungsbeschreibung	Die bereitgestellten Anwendungen müssen so entwickelt werden, dass sie innerhalb der vorhandenen Kubernetes-Infrastruktur horizontal skaliert werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen müssen zustandsarm (stateless), containerfähig und für den parallelen Betrieb mehrerer Instanzen geeignet sein.
Begründung	Eine skalierbare Anwendungsarchitektur ermöglicht den wirtschaftlichen Betrieb bei steigenden Nutzerzahlen und Lastspitzen.

BA-011 – Unterstützung automatisierter Skalierung

Anforderungsbeschreibung	Die Anwendungen müssen den Einsatz automatisierter Skalierungsmechanismen der vorhandenen Kubernetes-Plattform unterstützen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die konkrete Konfiguration und Steuerung des Auto-Scalings erfolgt durch die Betriebsumgebung der Auftraggeberin und ist nicht Bestandteil der Leistung.
Begründung	Die Architektur soll zukünftige automatische Skalierungsmechanismen unterstützen, ohne dass hierfür Anpassungen an den Anwendungen erforderlich sind.

BA-012 – Hochverfügbarkeit

Anforderungs- beschreibung	Die bereitgestellten Anwendungen müssen den hochverfügbaren Betrieb innerhalb der vorhandenen Infrastruktur unterstützen. Einzelne Komponenten dürfen keine unnötigen Single Points of Failure verursachen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Umsetzung erfolgt durch eine geeignete Anwendungsarchitektur. Die Bereitstellung redundanter Infrastrukturkomponenten ist nicht Bestandteil der Leistung.
Begründung	Hochverfügbare Anwendungen erhöhen die Betriebssicherheit und ermöglichen die Nutzung der vorhandenen Infrastrukturmechanismen.

BA-013 – Lastverteilung

Anforderungs- beschreibung	Die Anwendungen müssen den Betrieb hinter den vorhandenen Lastverteilungsmechanismen der Infrastruktur unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen müssen parallel auf mehreren Instanzen betrieben werden können. Die Bereitstellung und Konfiguration von Load-Balancing-Komponenten ist nicht Bestandteil der Leistung.
Begründung	Die Trennung zwischen Infrastruktur und Anwendung ermöglicht eine wirtschaftliche Nutzung der bereits vorhandenen Betriebsplattform und vermeidet redundante Infrastrukturleistungen.

3.4 Monitoring Integration

BA-014 – Systemmonitoring

Anforderungs- beschreibung	Die Plattform muss eine kontinuierliche Überwachung des Betriebszustands ermöglichen, insbesondere hinsichtlich Verfügbarkeit, Performance und Fehlerzuständen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendung muss geeignete Metriken und Statusinformationen bereitstellen, damit Betriebsparameter kontinuierlich überwacht werden können.
Begründung	Ein kontinuierliches Monitoring ermöglicht die frühzeitige Erkennung von Störungen und unterstützt einen stabilen und zuverlässigen Betrieb.

BA-015 – Schnittstellen für Monitoring

Anforderungsbeschreibung	Das System muss standardisierte Schnittstellen zur Integration in bestehende Monitoring- und Observability-Systeme bereitstellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendung muss geeignete Integrationsmöglichkeiten (z. B. APIs, Webhooks oder standardisierte Metrik-Endpunkte) bereitstellen. Die Auswahl und Konfiguration der Monitoring-Lösung ist nicht Bestandteil der Leistung.
Begründung	Die Integration in vorhandene Monitoring-Infrastrukturen vermeidet redundante Betriebswerkzeuge und ermöglicht eine einheitliche Überwachung.

BA-016 – Alarmierung

Anforderungsbeschreibung	Bei kritischen Systemzuständen müssen automatisierte Benachrichtigungen ausgelöst werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendung muss kritische Ereignisse und Zustände so bereitstellen, dass diese durch angeschlossene Monitoring-Systeme für Benachrichtigungen und Eskalationen genutzt werden können. Die Einrichtung konkreter Alarmierungsregeln ist nicht Bestandteil der Leistung.
Begründung	Eine automatisierte Alarmierung verkürzt Reaktionszeiten bei Störungen und erhöht die Betriebssicherheit.

BA-017 – Transparenz

Anforderungsbeschreibung	Der Systemzustand muss für Administratoren über geeignete Dashboards nachvollziehbar dargestellt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendung muss die erforderlichen Betriebsdaten für die Visualisierung in Dashboards bereitstellen. Die Bereitstellung oder Konfiguration eines Dashboard-Systems ist nicht Bestandteil der Leistung.
Begründung	Eine transparente Darstellung des Systemzustands erleichtert die Betriebsführung, Fehleranalyse und Kapazitätsplanung.

7.6 Sicherheitsstandards

Zielsetzung

Die bereitgestellten Komponenten des KI-Stacks müssen sich in die bestehende Sicherheitsarchitektur der Auftraggeberin integrieren und den sicheren Betrieb innerhalb der vorhandenen Infrastruktur unterstützen.

Die Anforderungen beziehen sich auf die durch den Auftragnehmer bereitgestellten Anwendungen und deren Integration. Der Betrieb der zugrunde liegenden Infrastruktur verbleibt bei der Auftraggeberin.

Architekturgrundsatz

Die Anwendungen sind so zu entwickeln, dass sie bestehende Sicherheitsmechanismen der Betriebsplattform nutzen und keine zusätzlichen Sicherheitsrisiken verursachen. Hierzu gehören insbesondere Authentifizierung, Autorisierung, Verschlüsselung, Protokollierung sowie die Trennung unterschiedlicher Betriebsumgebungen.

BA-018 – Einhaltung von Sicherheitsstandards

Anforderungs- beschreibung	Die bereitgestellten Anwendungen müssen den anerkannten Sicherheitsstandards entsprechen und sich in die Sicherheitsarchitektur der Auftraggeberin integrieren lassen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Sicherheitsrelevante Funktionen sind nach anerkannten Best Practices zu entwickeln. Die Integration erfolgt in die vorhandenen Sicherheitsmechanismen der Betriebsplattform.
Begründung	Einheitliche Sicherheitsstandards gewährleisten einen sicheren und wartbaren Betrieb der Plattform.

BA-019 – Verschlüsselung

Anforderungs- Beschreibung	Die Anwendungen müssen die Verschlüsselung von Daten während der Übertragung sowie die Nutzung verschlüsselter Speicherdienste der vorhandenen Infrastruktur unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen nutzen sichere Kommunikationsprotokolle sowie die bereitgestellten Verschlüsselungsmechanismen der Infrastruktur. Die Bereitstellung eigener Verschlüsselungsdienste ist nicht Bestandteil der Leistung.
Begründung	Die Nutzung vorhandener Sicherheitsmechanismen reduziert Komplexität und gewährleistet ein einheitliches Sicherheitsniveau.

BA-020 – Zugriffsschutz

Anforderungs- beschreibung	Die Anwendungen müssen die vorhandenen Authentifizierungs- und Autorisierungsmechanismen der Auftraggeberin unterstützen und sämtliche Zugriffe nachvollziehbar protokollieren.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Integration erfolgt über die vorhandenen Identity- und Access-Management-Systeme. Eigene Benutzer- oder Berechtigungsverwaltungen sind nur zulässig, soweit sie für die Anwendung erforderlich sind.
Begründung	Die Nutzung einer zentralen Zugriffssteuerung erhöht Sicherheit und vereinfacht den Betrieb.

BA-021 – Trennung von Betriebsumgebungen

Anforderungs- beschreibung	Die Anwendungen müssen den getrennten Betrieb in Entwicklungs-, Test- und Produktivumgebungen unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen sind so bereitzustellen, dass sie in unterschiedlichen Umgebungen mit separaten Konfigurationen betrieben werden können. Die Bereitstellung dieser Umgebungen erfolgt durch die Auftraggeberin.
Begründung	Die Trennung der Betriebsumgebungen reduziert Risiken und unterstützt einen geordneten Entwicklungs- und Freigabeprozess.

BA-022 – Patch- und Updatefähigkeit

Anforderungs- beschreibung	Die bereitgestellten Anwendungen müssen so entwickelt werden, dass Sicherheitsupdates und neue Versionen ohne grundlegende Änderungen an der Gesamtarchitektur eingespielt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Der Auftragnehmer stellt regelmäßig Sicherheitsupdates für die gelieferten Softwarekomponenten bereit. Das Einspielen von Betriebssystem- oder Infrastruktur-Patches ist nicht Bestandteil der Leistung.
Begründung	Eine klare Trennung zwischen Anwendungs- und Infrastrukturverantwortung ermöglicht einen sicheren und wirtschaftlichen Betrieb.

8 Datenschutz & Informationssicherheit

Zielsetzung

Die nachfolgenden Anforderungen definieren die datenschutzrechtlichen, informationssicherheitsrelevanten sowie regulatorischen Rahmenbedingungen für die Bereitstellung und den Betrieb des KI-Stacks.

Ziel ist die Entwicklung einer vertrauenswürdigen, nachvollziehbaren und datenschutzkonformen KI-Plattform, welche die Anforderungen der Datenschutz-Grundverordnung (DSGVO), des Bundesdatenschutzgesetzes (BDSG), der einschlägigen Informationssicherheitsstandards sowie – soweit anwendbar – des europäischen AI Acts unterstützt.

Die Anforderungen dieses Kapitels ergänzen die funktionalen, nichtfunktionalen und architektonischen Anforderungen der Leistungsbeschreibung. Sie begründen keine darüber hinausgehenden eigenständigen Funktionen, sondern konkretisieren die Anforderungen an Datenschutz, Informationssicherheit, Governance und Compliance.

Grundsätze

Die Plattform ist nach den Prinzipien **Privacy by Design** und **Privacy by Default** auszugestalten. Sämtliche Komponenten des KI-Stacks müssen den Schutz personenbezogener Daten, die Nachvollziehbarkeit von KI-Verarbeitungen sowie die Einhaltung geltender Sicherheits- und Compliance-Anforderungen unterstützen.

Bereits vorhandene Sicherheits- und Datenschutzmechanismen der Auftraggeberin sind zu nutzen und in die Plattform zu integrieren. Eigene Sicherheits- oder Datenschutzinfrastrukturen sind nur bereitzustellen, soweit dies zur Erfüllung der funktionalen Anforderungen erforderlich ist. Die Umsetzung erfolgt unter Berücksichtigung der bestehenden Betriebs- und Sicherheitsarchitektur der Auftraggeberin.

8.1 Einhaltung der DSGVO-Anforderungen

Zielsetzung

Die Plattform muss den datenschutzkonformen Umgang mit personenbezogenen Daten unterstützen und die Anforderungen der Datenschutz-Grundverordnung (DSGVO) sowie weiterer einschlägiger datenschutzrechtlicher Vorschriften erfüllen.

Die Anforderungen dieses Abschnitts beziehen sich auf die datenschutzkonforme Gestaltung der bereitgestellten Softwarekomponenten sowie deren Integration in die Betriebsumgebung der Auftraggeberin.

Datenschutzgrundsatz

Die Plattform ist nach den Grundsätzen **Privacy by Design** und **Privacy by Default** auszulegen. Sämtliche Verarbeitungen personenbezogener Daten müssen nachvollziehbar, zweckgebunden und auf das erforderliche Maß beschränkt erfolgen.

DA-001 – DSGVO-Konformität

Anforderungs- beschreibung	Die Plattform muss die datenschutzkonforme Verarbeitung personenbezogener Daten entsprechend den Anforderungen der Datenschutz-Grundverordnung (DSGVO) unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die bereitgestellten Anwendungen sind so zu entwickeln, dass die datenschutzrechtlichen Anforderungen der Auftraggeberin umgesetzt werden können.
Begründung	Die DSGVO-Konformität ist grundlegende Voraussetzung für den produktiven Einsatz einer KI-Plattform in der öffentlichen Verwaltung.

DA-002 – Zweckbindung

Anforderungs- beschreibung	Personenbezogene Daten dürfen ausschließlich für den jeweils vorgesehenen und zulässigen Verarbeitungszweck verarbeitet werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform muss geeignete Mechanismen zur Steuerung und Begrenzung der Datenverarbeitung entsprechend dem jeweiligen Anwendungsfall unterstützen.
Begründung	Die Zweckbindung ist ein zentrales datenschutzrechtliches Prinzip der DSGVO.

DA-003 – Datenminimierung

Anforderungs- beschreibung	Die Plattform darf nur diejenigen personenbezogenen Daten verarbeiten oder an externe Dienste übermitteln, die für die jeweilige Verarbeitung erforderlich sind.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Datenminimierung ist insbesondere bei Modellaufrufen, Retrieval-Prozessen sowie der Nutzung externer KI-Dienste zu berücksichtigen.
Begründung	Die Beschränkung auf notwendige Daten reduziert Datenschutzrisiken und entspricht den Vorgaben der DSGVO.

DA-004 – Transparenz der Datenverarbeitung

Anforderungsbeschreibung	Die Verarbeitung personenbezogener Daten muss nachvollziehbar dokumentiert und für berechtigte Administratoren sowie Nutzer transparent sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform muss die Nachvollziehbarkeit der Datenverarbeitung durch Protokollierung, Dokumentation und geeignete Administrationsfunktionen unterstützen.
Begründung	Transparenz schafft Vertrauen und unterstützt die Erfüllung gesetzlicher Nachweis- und Rechenschaftspflichten.

DA-005 – Datenschutzkonforme Einbindung externer Dienste

Anforderungsbeschreibung	Werden externe Dienste oder Modellanbieter genutzt, muss deren Einbindung datenschutzkonform erfolgen. Erforderliche Vereinbarungen zur Auftragsverarbeitung sowie weitere rechtliche Voraussetzungen sind zu berücksichtigen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform muss den Einsatz DSGVO-konformer Dienste unterstützen. Die Auswahl konkreter Dienstleister erfolgt in Abstimmung mit der Auftraggeberin.
Begründung	Der datenschutzkonforme Einsatz externer KI-Dienste ist Voraussetzung für den rechtskonformen Betrieb der Plattform.

8.2 PII-Erkennung und –Verarbeitung

Zielsetzung

Die Plattform muss personenbezogene Daten (Personally Identifiable Information – PII) automatisiert erkennen und entsprechend den Datenschutzvorgaben verarbeiten können. Ziel ist die Unterstützung datenschutzkonformer KI-Verarbeitungen sowie die Vermeidung einer unbeabsichtigten Übermittlung personenbezogener Daten an externe KI-Dienste.

Die technische Umsetzung der PII-Erkennung erfolgt über die im Rahmen von Los 3 bereitgestellten ML-Services oder funktional gleichwertige Komponenten.

Datenschutzgrundsatz

Die PII-Erkennung ist Bestandteil der Datenschutz- und Governance-Architektur des KI-Stacks. Sie muss vor der Weitergabe von Daten an externe Modelle oder Dienste eingesetzt werden können und sich in die bestehende Orchestrierungs-, Policy- und Berechtigungslogik des AI Harness integrieren.

DA-006 – Erkennung personenbezogener Daten

Anforderungs- beschreibung	Die Plattform muss personenbezogene Daten in Eingaben und – soweit erforderlich – auch in Ausgaben automatisiert erkennen können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Erkennung erfolgt über den bereitgestellten ML-Service für PII-Erkennung oder eine funktional gleichwertige Komponente.
Begründung	Die automatisierte Erkennung personenbezogener Daten ist Grundlage für eine datenschutzkonforme Nutzung von KI-Diensten.

DA-007 – Unterstützte Datenkategorien

Anforderungs- beschreibung	Die PII-Erkennung muss mindestens Personennamen, Adressen, Telefonnummern, E-Mail-Adressen, Identifikationsnummern sowie Bankverbindungen erkennen können. Weitere Kategorien müssen ergänzbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform muss die Konfiguration zusätzlicher PII-Kategorien unterstützen, ohne Änderungen an der Gesamtarchitektur zu erfordern.
Begründung	Unterschiedliche Verwaltungsprozesse erfordern eine anpassbare Erkennung personenbezogener Daten.

DA-008 – Betriebsmodi der PII-Verarbeitung

Anforderungs- beschreibung	Die Plattform muss mindestens einen Hinweismodus sowie einen Modus zur automatisierten Maskierung oder Pseudonymisierung personenbezogener Daten unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Betriebsart muss administrativ konfigurierbar und abhängig vom jeweiligen Anwendungsfall auswählbar sein.
Begründung	Unterschiedliche Schutzanforderungen erfordern unterschiedliche Strategien für den Umgang mit personenbezogenen Daten.

DA-009 – Konfigurierbarkeit

Anforderungsbeschreibung	Die Regeln zur PII-Erkennung und -Verarbeitung müssen administrativ konfigurierbar sein und sich für unterschiedliche Anwendungsfälle, Agenten oder Modellklassen anpassen lassen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Konfiguration erfolgt zentral über die Administrations- und Governance-Funktionen des KI-Stacks.
Begründung	Eine zentrale Konfiguration ermöglicht die einheitliche Umsetzung unterschiedlicher Datenschutzerfordernungen innerhalb der Plattform.

DA-010 – Erweiterbarkeit der Erkennungsregeln

Anforderungsbeschreibung	Die Plattform muss die Erweiterung bestehender Erkennungsregeln und Datenkategorien unterstützen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Zusätzliche Muster, Regelwerke oder spezialisierte Erkennungsmodelle müssen zukünftig ergänzt werden können, ohne grundlegende Änderungen an der Gesamtarchitektur vorzunehmen.
Begründung	Datenschutzerfordernungen entwickeln sich kontinuierlich weiter. Eine erweiterbare Architektur gewährleistet die langfristige Anpassungsfähigkeit der Plattform.

8.3 Datenschutz-Ampelsystem

Zielsetzung

Die Plattform muss das Datenschutzniveau der jeweils verwendeten KI-Verarbeitung für die Nutzenden transparent und nachvollziehbar darstellen. Ziel ist die Unterstützung datenschutzbewusster Entscheidungen sowie die Erhöhung der Transparenz bei der Nutzung unterschiedlicher KI-Modelle und Verarbeitungsumgebungen.

Datenschutzgrundsatz

Die Datenschutzklassifizierung ist Bestandteil der Benutzeroberfläche des Urban KI Hub und wird durch den AI Harness auf Grundlage der jeweils verwendeten Modell- und Verarbeitungsumgebung bereitgestellt. Die Klassifizierung soll das Datenschutzniveau für die Nutzenden einfach verständlich darstellen. Hierfür kann beispielsweise ein Datenschutz-Ampelsystem verwendet werden.

Eine mögliche Ausgestaltung ist:

- **Grün** – Verarbeitung ausschließlich innerhalb der freigegebenen Infrastruktur der Auftraggeberin oder unter Nutzung freigegebener DSGVO-konformer Open-Weight-Modelle. Personenbezogene Daten verlassen die kontrollierte Infrastruktur der Auftraggeberin nicht.
- **Gelb** – Verarbeitung unter Nutzung freigegebener externer KI-Dienste oder Modellanbieter, für die geeignete datenschutzrechtliche und organisatorische Maßnahmen bestehen (z. B. Auftragsverarbeitungsvertrag, Verarbeitung innerhalb der EU oder vergleichbare Datenschutzgarantien). Die Verarbeitung ist zulässig, sollte den Nutzenden jedoch transparent angezeigt werden.
- **Rot** – Verarbeitung unter Nutzung externer Dienste oder Modellanbieter, bei denen personenbezogene Daten die kontrollierte Infrastruktur der Auftraggeberin verlassen oder erhöhte datenschutzrechtliche Risiken bestehen. Die Nutzung solcher Dienste kann administrativ eingeschränkt, genehmigungspflichtig oder ausgeschlossen werden.

Die konkrete Ausgestaltung der Visualisierung bleibt dem Auftragnehmer überlassen, sofern eine gleichwertige Transparenz und Nachvollziehbarkeit gewährleistet wird. Die Architektur muss eine spätere Erweiterung um zusätzliche Klassifizierungsstufen oder differenziertere Bewertungsverfahren unterstützen.

DA-011 – Visualisierung des Datenschutzniveaus

Anforderungs- beschreibung	Die Plattform muss das Datenschutzniveau der jeweils verwendeten KI-Verarbeitung visuell darstellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Darstellung erfolgt über ein Datenschutz-Ampelsystem oder eine funktional gleichwertige Visualisierung innerhalb des Urban KI Hub.
Begründung	Eine verständliche Visualisierung unterstützt die datenschutzbewusste Nutzung der Plattform und erhöht die Transparenz für die Anwender.

DA-012 – Datenschutzklassifizierung

Anforderungs- beschreibung	Das System muss das Datenschutzniveau der jeweils verwendeten KI-Verarbeitung automatisiert klassifizieren und mindestens drei Datenschutzstufen unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Klassifizierung erfolgt automatisiert auf Basis der jeweils verwendeten Modell- und Verarbeitungsumgebung. Die Architektur muss die spätere Erweiterung um zusätzliche Klassifizierungsstufen unterstützen.
Begründung	Die Einordnung unterstützt die transparente Bewertung unterschiedlicher Verarbeitungswege und schafft eine Grundlage für datenschutzbewusste Entscheidungen.

DA-013 – Transparenz für Nutzende

Anforderungsbeschreibung	Die Datenschutzklassifizierung muss den Nutzenden während der Verwendung eines KI-Modells jederzeit eindeutig und gut sichtbar angezeigt werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Darstellung erfolgt unmittelbar innerhalb der Benutzeroberfläche des Urban KI Hub und wird bei einem Wechsel des Modells oder der Verarbeitungsumgebung automatisch aktualisiert.
Begründung	Eine kontinuierliche Sichtbarkeit erhöht die Transparenz und unterstützt einen verantwortungsvollen Umgang mit KI-Diensten.

DA-014 – Automatische Klassifizierung

Anforderungsbeschreibung	Die Datenschutzklassifizierung muss automatisiert anhand der jeweils verwendeten Modell- und Verarbeitungsumgebung erfolgen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Einstufung erfolgt durch den AI Harness auf Grundlage administrativ definierter Regeln und wird automatisch an den Urban KI Hub übermittelt. Die Klassifizierungsregeln müssen zentral administrierbar sein.
Begründung	Eine automatisierte und zentral verwaltete Klassifizierung gewährleistet eine konsistente, nachvollziehbare und revisionssichere Bewertung der Datenschutzsituation über sämtliche KI-Anwendungsfälle hinweg.

8.4 Trennung von Daten und Modellen

Zielsetzung

Die Plattform muss sicherstellen, dass personenbezogene und fachliche Daten unabhängig von den eingesetzten KI-Modellen verarbeitet werden. Ziel ist die Wahrung der Datenhoheit der Auftraggeberin sowie die Vermeidung einer unkontrollierten Nutzung von Daten durch interne oder externe KI-Dienste.

Datenschutzgrundsatz

Die Architektur ist nach dem Prinzip der Trennung von Daten, Verarbeitung und Modellinferenz auszulegen. Nutzerdaten verbleiben grundsätzlich unter der Kontrolle der Auftraggeberin und dürfen ausschließlich im Rahmen des jeweils zulässigen Verarbeitungszwecks genutzt werden.

DA-015 – Keine Nutzung für Modelltraining

Anforderungsbeschreibung	Personenbezogene oder fachliche Daten der Auftraggeberin dürfen nicht zum Training, Fine-Tuning oder zur Verbesserung interner oder externer KI-Modelle verwendet werden, sofern dies nicht ausdrücklich durch die Auftraggeberin freigegeben wurde.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform muss sicherstellen, dass Modellanbieter oder nachgelagerte Dienste keine Daten für Trainings- oder Optimierungszwecke verwenden können.
Begründung	Die Wahrung der Datenhoheit ist eine zentrale Voraussetzung für den rechtskonformen Einsatz von KI-Systemen in der öffentlichen Verwaltung.

DA-016 – Trennung von Daten, Verarbeitung und Modellinferenz

Anforderungsbeschreibung	Die Architektur muss eine klare Trennung zwischen Nutzerdaten, Verarbeitungslogik und Modellinferenz gewährleisten.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Trennung erfolgt durch den AI Harness, die Context Engine sowie die modellunabhängige Modellabstraktionsschicht.
Begründung	Die Trennung der Verarbeitungsebenen erhöht Datenschutz, Sicherheit und Herstellerunabhängigkeit der Plattform.

DA-017 – Minimierung externer Datenübertragungen

Anforderungsbeschreibung	Die Übermittlung von Daten an externe Dienste ist auf das für den jeweiligen Verarbeitungsvorgang erforderliche Mindestmaß zu beschränken.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform muss Mechanismen zur Datenminimierung, Pseudonymisierung oder Maskierung unterstützen und vorhandene Datenschutzmechanismen des AI Harness berücksichtigen.
Begründung	Die Reduzierung externer Datenübertragungen minimiert Datenschutzrisiken und unterstützt die Einhaltung der DSGVO.

DA-018 – Kontrollierte Datenweitergabe

Anforderungsbeschreibung	Die Weitergabe von Daten an externe Dienste oder KI-Modelle darf ausschließlich auf Grundlage administrativ definierter Richtlinien, Berechtigungen und Datenschutzvorgaben erfolgen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Entscheidung über eine Datenweitergabe erfolgt durch den AI Harness unter Berücksichtigung von Policies, Rollen, Mandantenkontext, Datenschutzklassifizierung und Freigaberegeln. Sämtliche Datenübertragungen sind nachvollziehbar zu protokollieren.
Begründung	Eine kontrollierte und nachvollziehbare Datenweitergabe stellt sicher, dass personenbezogene und fachliche Informationen ausschließlich im zulässigen Rahmen verarbeitet werden.

8.5 KI-Inventar und EU AI Act

Zielsetzung

Die Plattform muss die transparente Verwaltung der eingesetzten KI-Komponenten unterstützen und die Auftraggeberin bei der Erfüllung regulatorischer Anforderungen, insbesondere des europäischen AI Acts, unterstützen.

Hierzu ist ein zentrales KI-Inventar bereitzustellen, das die eingesetzten Modelle und wesentlichen KI-Komponenten dokumentiert und eine nachvollziehbare Bewertung ihrer Nutzung ermöglicht.

Compliance-Grundsatz

Das KI-Inventar dient der Transparenz, Nachvollziehbarkeit und Governance der eingesetzten KI-Komponenten. Es bildet die Grundlage für Dokumentations-, Auditierungs- und Compliance-Prozesse und unterstützt die Umsetzung der Anforderungen des EU AI Acts.

DA-019 – Zentrales KI-Inventar

Anforderungsbeschreibung	Die Plattform muss ein zentrales KI-Inventar bereitstellen, in dem die eingesetzten KI-Modelle und wesentlichen KI-Komponenten nachvollziehbar dokumentiert werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Das KI-Inventar ist zentral über den AI Harness zu verwalten und für berechtigte Administratoren einsehbar bereitzustellen.
Begründung	Ein zentrales KI-Inventar schafft Transparenz und bildet die Grundlage für Governance-, Audit- und Compliance-Anforderungen.

DA-020 – Dokumentation der KI-Komponenten

Anforderungs- beschreibung	Für jedes eingesetzte KI-Modell müssen mindestens Anbieter, Modellbezeichnung, Modellversion, Einsatzbereich sowie Datenschutzklassifizierung dokumentiert werden. Weitere Metadaten müssen ergänzbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Informationen sind zentral im KI-Inventar zu verwalten und bei Änderungen nachvollziehbar zu aktualisieren.
Begründung	Eine strukturierte Dokumentation unterstützt den sicheren und nachvollziehbaren Betrieb der KI-Plattform.

DA-021 – Unterstützung regulatorischer Klassifizierungen

Anforderungs- beschreibung	Das KI-Inventar muss die Zuordnung regulatorischer Informationen, insbesondere einer Risikoklassifizierung gemäß den Vorgaben des EU AI Acts, unterstützen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Architektur muss die spätere Ergänzung weiterer regulatorischer Informationen und Compliance-Merkmale ermöglichen, ohne grundlegende Änderungen an der Gesamtarchitektur zu erfordern.
Begründung	Die Anforderungen des EU AI Acts entwickeln sich weiter. Eine erweiterbare Architektur gewährleistet die langfristige Anpassungsfähigkeit der Plattform.

DA-022 – Nachvollziehbarkeit der KI-Nutzung

Anforderungs- beschreibung	Die Nutzung von KI-Modellen muss nachvollziehbar dokumentiert werden. Hierzu gehören insbesondere Informationen über das verwendete Modell, die Modellversion, den Zeitpunkt der Nutzung sowie der jeweilige Anwendungsfall.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Dokumentation erfolgt über die Protokollierungs- und Monitoring-Funktionen des AI Harness und ist mit dem KI-Inventar verknüpft.
Begründung	Die nachvollziehbare Dokumentation der KI-Nutzung unterstützt Auditierungen, Fehleranalysen und regulatorische Nachweispflichten.

8.6 Informationssicherheit

Zielsetzung

Die Plattform muss den sicheren Betrieb der bereitgestellten KI-Komponenten unterstützen und sich in die bestehende Informationssicherheitsarchitektur der Auftraggeberin integrieren. Ziel ist der Schutz von Informationen, Systemen und KI-Verarbeitungen vor unberechtigtem Zugriff, Manipulation, Datenverlust sowie sonstigen sicherheitsrelevanten Ereignissen.

Informationssicherheitsgrundsatz

Die Plattform ist unter Berücksichtigung anerkannter Grundsätze der Informationssicherheit zu entwickeln. Hierzu gehören insbesondere die Gewährleistung von Vertraulichkeit, Integrität, Verfügbarkeit, Nachvollziehbarkeit sowie die Nutzung bestehender Sicherheitsmechanismen der Auftraggeberin.

DA-023 – Zugriffsschutz

Anforderungs- beschreibung	Zugriffe auf Daten, Funktionen und Systemkomponenten müssen durch geeignete Authentifizierungs- und Autorisierungsmechanismen geschützt werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform integriert sich in die vorhandenen Identity- und Access-Management-Systeme der Auftraggeberin und unterstützt rollen- sowie berechtigungsbasierte Zugriffssteuerungen.
Begründung	Eine kontrollierte Zugriffssteuerung schützt sensible Informationen und verhindert unberechtigte Zugriffe auf KI-Funktionen und Daten.

DA-024 – Protokollierung sicherheitsrelevanter Ereignisse

Anforderungs- beschreibung	Sicherheitsrelevante Ereignisse müssen nachvollziehbar protokolliert und für Auditierungs- und Analysezwecke bereitgestellt werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Protokollierung erfolgt über die zentralen Logging- und Monitoring-Mechanismen der Plattform und muss sich in die vorhandenen Observability-Lösungen der Auftraggeberin integrieren lassen.
Begründung	Eine revisionssichere Protokollierung unterstützt die Erkennung sicherheitsrelevanter Vorfälle sowie deren spätere Analyse.

DA-025 – Integrität und Verfügbarkeit

Anforderungsbeschreibung	Die bereitgestellten Anwendungen müssen Maßnahmen unterstützen, welche die Integrität und Verfügbarkeit der verarbeiteten Daten und KI-Funktionen sicherstellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen müssen die vorhandenen Mechanismen der Betriebsplattform zur Sicherstellung von Datenintegrität, Fehlertoleranz und Wiederherstellbarkeit unterstützen.
Begründung	Die Sicherstellung der Integrität und Verfügbarkeit ist eine wesentliche Voraussetzung für einen sicheren und zuverlässigen Betrieb der KI-Plattform.

DA-026 – Trennung von Betriebsumgebungen

Anforderungsbeschreibung	Die Anwendungen müssen den getrennten Betrieb in Entwicklungs-, Test- und Produktivumgebungen unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Anwendungen sind so bereitzustellen, dass sie mit getrennten Konfigurationen und Berechtigungen in unterschiedlichen Betriebsumgebungen betrieben werden können. Die Bereitstellung der Umgebungen erfolgt durch die Auftraggeberin.
Begründung	Die Trennung von Betriebsumgebungen reduziert Sicherheitsrisiken und unterstützt einen geordneten Entwicklungs-, Test- und Freigabeprozess.

8.7 Sicherheit von KI-Systemen (LLM Security)

Zielsetzung

Die Plattform muss geeignete technische und organisatorische Maßnahmen zum Schutz von Sprachmodellen, Agenten, Retrieval-Prozessen, ML-Services sowie KI-gestützten Workflows bereitstellen. Ziel ist der Schutz der Plattform vor bekannten Angriffsmustern, Missbrauchsszenarien sowie unbeabsichtigten oder manipulierten KI-Ausführungen.

Sicherheitsgrundsatz

Die Sicherheitsmechanismen müssen sämtliche Verarbeitungsschritte des KI-Stacks berücksichtigen. Hierzu gehören insbesondere Eingaben, Dokumentenverarbeitung, Retrieval-Prozesse, Tool-Interaktionen, Modellaufrufe sowie die erzeugten Ausgaben.

Die Plattform ist nach dem Prinzip **Secure by Design** auszulegen. Sicherheitsmechanismen sind zentral über den AI Harness zu steuern und müssen mit den Governance-, Policy- und Monitoring-Funktionen der Plattform zusammenwirken.

Sicherheitsrelevante Ereignisse sind nachvollziehbar zu protokollieren und für Monitoring-, Auditierungs- und Compliance-Zwecke bereitzustellen.

NFA-SEC-01 – Schutz vor Prompt Injection

Anforderungsbeschreibung	Die Plattform muss Maßnahmen zum Schutz vor direkter und indirekter Prompt Injection unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Eingaben sowie über Retrieval oder externe Quellen bereitgestellte Inhalte müssen vor ihrer Verwendung innerhalb der KI-Verarbeitung bewertet und geeigneten Sicherheitsmechanismen unterzogen werden können.
Begründung	Prompt Injection zählt zu den wesentlichen Angriffsszenarien moderner KI-Systeme und muss bereits in der Erstimplementierung berücksichtigt werden.

NFA-SEC-02 – Validierung von Eingaben und Kontexten

Anforderungsbeschreibung	Eingaben, Dokumente, Retrieval-Ergebnisse sowie Tool-Rückgaben müssen vor ihrer Weiterverarbeitung validiert und sicher verarbeitet werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Validierung erfolgt innerhalb des AI Harness unter Berücksichtigung definierter Sicherheits- und Governance-Regeln.
Begründung	Die Validierung reduziert Manipulationsrisiken und verbessert die Sicherheit der gesamten Verarbeitungskette.

NFA-SEC-03 – Schutz vor Data Poisoning

Anforderungsbeschreibung	Die Plattform muss Mechanismen unterstützen, um manipulierte oder kompromittierte Wissensquellen zu erkennen oder deren Auswirkungen auf Retrieval- und RAG-Prozesse zu minimieren.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform muss die Validierung von Wissensquellen sowie geeignete Governance- und Freigabemechanismen unterstützen.
Begründung	Manipulierte Wissensquellen können die Qualität und Vertrauenswürdigkeit KI-generierter Antworten erheblich beeinträchtigen.

NFA-SEC-04 – Schutz vor Tool Poisoning

Anforderungs- beschreibung	Tool-Aufrufe müssen kontrolliert, autorisiert und gegen missbräuchliche oder manipulierte Nutzung abgesichert werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Tool-Aufrufe erfolgen ausschließlich über den AI Harness unter Berücksichtigung von Rollen, Berechtigungen, Policies und Guardrails.
Begründung	Kontrollierte Tool-Interaktionen verhindern den Missbrauch angebundener Systeme und Dienste.

NFA-SEC-05 – Schutz vor missbräuchlicher Modellnutzung

Anforderungs- beschreibung	Die Plattform muss Mechanismen zur Begrenzung missbräuchlicher oder ressourcenintensiver Modellaufrufe unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform muss insbesondere Rate-Limits, Quoten, Kostenlimits sowie weitere Schutzmechanismen gegen übermäßige Modellnutzung unterstützen.
Begründung	Der Schutz vor missbräuchlicher Nutzung erhöht die Verfügbarkeit der Plattform und begrenzt Betriebskosten.

NFA-SEC-06 – Validierung KI-generierter Ausgaben

Anforderungs- beschreibung	KI-generierte Ausgaben müssen vor einer automatisierten Weiterverarbeitung oder externen Weitergabe validiert und bei Bedarf durch Guardrails oder Human-in-the-Loop-Prozesse abgesichert werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Validierung erfolgt entsprechend den im AI Harness definierten Sicherheits-, Governance- und Freigaberegeln.
Begründung	Die Validierung reduziert Risiken fehlerhafter oder unerwünschter automatisierter Entscheidungen.

NFA-SEC-07 – Unterstützung isolierter Ausführungsumgebungen

Anforderungsbeschreibung	Die Architektur muss den späteren Einsatz logisch getrennter Ausführungsumgebungen (Sandboxing) für Agenten oder Tool-Ausführungen unterstützen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Erstimplementierung muss die spätere Integration entsprechender Sicherheitsmechanismen ermöglichen, ohne deren vollständige Umsetzung zu erfordern.
Begründung	Sandboxing erhöht die Sicherheit komplexer Agentensysteme, ist jedoch nicht zwingend Bestandteil der produktiven Erstimplementierung.

NFA-SEC-08 – Auditierung und Compliance

Anforderungsbeschreibung	Sicherheitsrelevante Ereignisse müssen zentral protokolliert, überwacht und für Auditierungs- und Compliance-Zwecke nachvollziehbar bereitgestellt werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Protokollierung erfolgt über die zentralen Logging-, Monitoring- und Observability-Komponenten der Plattform und ist mit den Governance-Funktionen des AI Harness zu verknüpfen.
Begründung	Eine vollständige Nachvollziehbarkeit sicherheitsrelevanter Ereignisse ist Voraussetzung für einen sicheren und revisionssicheren Betrieb der KI-Plattform.

8.8 Agentenverwaltung und Agentenkonfiguration

Zielsetzung

Die Plattform muss die zentrale Verwaltung und Konfiguration von KI-Agenten unterstützen. Ziel ist es, berechtigten Administratoren die Pflege, Konfiguration und Bereitstellung von KI-Agenten zu ermöglichen, ohne Änderungen am Quellcode vornehmen zu müssen.

Die Agentenverwaltung dient der Administration der im KI-Stack bereitgestellten Agenten und ergänzt die Agent Runtime des AI Harness.

Governance-Grundsatz

Die fachliche Konfiguration der Agenten erfolgt über den Urban KI Hub. Die technische Ausführung der Agenten übernimmt die Agent Runtime des AI Harness.

Die Plattform muss eine nachvollziehbare und kontrollierte Verwaltung von Agentenkonfigurationen unterstützen. Änderungen an Agenten müssen versionierbar und administrativ nachvollziehbar sein.

DA-027 – Agentenkonfiguration

Anforderungs- beschreibung	Berechtigte Administratoren müssen KI-Agenten konfigurieren und verwalten können, ohne Änderungen am Quellcode vornehmen zu müssen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Für die Erstimplementierung genügt die Konfiguration der im Leistungsumfang vorgesehenen Agenten über den Urban KI Hub.
Begründung	Eine grafische Konfiguration reduziert den Administrationsaufwand und vereinfacht den Betrieb der Plattform.

DA-028 – Zuordnung von Modellen und Wissensquellen

Anforderungs- beschreibung	Agenten müssen konfigurierbaren Zugriff auf Modelle, Wissensquellen, Retrieval-Räume sowie freigegebene Werkzeuge erhalten können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Zuordnung erfolgt über die Administrationsoberfläche des Urban KI Hub und wird durch den AI Harness zur Laufzeit berücksichtigt.
Begründung	Die flexible Zuordnung unterstützt unterschiedliche fachliche Einsatzszenarien und reduziert den Konfigurationsaufwand.

DA-029 – Versionierung und Freigabe

Anforderungs- beschreibung	Die Plattform muss die Versionierung sowie die kontrollierte Freigabe und Deaktivierung von Agentenkonfigurationen unterstützen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Für die Erstimplementierung genügt eine einfache Versionierung und Bereitstellung der vorgesehenen Agenten. Erweiterte Freigabe- und Veröffentlichungsprozesse müssen architektonisch vorbereitet sein.
Begründung	Umfangreiche Release- und Freigabeworkflows verursachen einen hohen Implementierungsaufwand und können schrittweise ergänzt werden.

DA-030 – Wiederverwendung von Konfigurationen

Anforderungs- beschreibung	Die Plattform muss die Wiederverwendung vorhandener Modelle, Wissensquellen, Prompt-Konfigurationen und Werkzeuge innerhalb verschiedener Agenten unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Konfiguration erfolgt über wiederverwendbare Referenzen innerhalb der Agentenkonfiguration.
Begründung	Wiederverwendbare Konfigurationen reduzieren den Pflegeaufwand und fördern eine konsistente Administration der Plattform.

9 Schnittstellen

Zielsetzung

Die Plattform muss über standardisierte, dokumentierte und abgesicherte Schnittstellen verfügen, um eine interoperable Integration in die bestehende Systemlandschaft der Auftraggeberin sowie die zukünftige Erweiterung des KI-Stacks zu ermöglichen.

Ziel ist eine offene, modulare und herstellerunabhängige Integrationsarchitektur, bei der sämtliche Komponenten des KI-Stacks sowie angebundene Fachverfahren, Datenquellen und externe Dienste über definierte Schnittstellen miteinander kommunizieren.

Integrationsgrundsatz

Die Kommunikation zwischen den Komponenten des KI-Stacks erfolgt ausschließlich über standardisierte, dokumentierte und versionierte Schnittstellen.

Die Schnittstellen müssen so ausgestaltet sein, dass einzelne Komponenten unabhängig voneinander entwickelt, betrieben, erweitert oder ausgetauscht werden können, ohne dass grundlegende Änderungen an anderen Komponenten erforderlich sind.

Die Plattform muss bestehende Integrationsmechanismen der Auftraggeberin unterstützen. Proprietäre oder nicht dokumentierte Schnittstellen sind nur zulässig, sofern keine funktional gleichwertige offene Schnittstelle verfügbar ist.

Die technische Umsetzung der Schnittstellen bleibt dem Auftragnehmer überlassen, sofern die funktionalen und nichtfunktionalen Anforderungen dieser Leistungsbeschreibung erfüllt werden.

9.1 API-Schnittstellen

Zielsetzung

Die Plattform muss standardisierte, dokumentierte und abgesicherte Programmierschnittstellen (APIs) bereitstellen, über die sämtliche Funktionen des KI-Stacks durch den Urban KI Hub, Fachverfahren sowie weitere berechnete Systeme genutzt werden können.

Ziel ist eine offene, interoperable und herstellerunabhängige Integrationsarchitektur, welche die lose Kopplung der Plattformkomponenten sowie deren zukünftige Erweiterbarkeit unterstützt.

Integrationsgrundsatz

Sämtliche Funktionen des AI Harness sind ausschließlich über definierte und versionierte Schnittstellen bereitzustellen.

Die APIs müssen konsistent aufgebaut, maschinenlesbar dokumentiert und langfristig stabil nutzbar sein. Änderungen an bestehenden Schnittstellen dürfen die Nutzung bereits integrierter Systeme nicht unnötig beeinträchtigen.

IA-001 – Bereitstellung standardisierter APIs

Anforderungsbeschreibung	Die Plattform muss standardisierte Programmierschnittstellen bereitstellen, über welche die zentralen Funktionen des KI-Stacks programmatisch genutzt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Sämtliche Schnittstellen werden zentral durch den AI Harness bereitgestellt.
Begründung	Standardisierte APIs ermöglichen die Integration weiterer Anwendungen sowie die lose Kopplung sämtlicher Plattformkomponenten.

IA-002 – Standardisierte API-Implementierung

Anforderungsbeschreibung	Die APIs müssen auf allgemein anerkannten Standards basieren, beispielsweise REST oder funktional gleichwertigen Schnittstellentechnologien.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die konkrete technische Umsetzung bleibt dem Auftragnehmer überlassen, sofern Interoperabilität, Sicherheit und Dokumentierbarkeit gewährleistet sind.
Begründung	Standardisierte Schnittstellen erleichtern die Integration bestehender und zukünftiger Systeme.

IA-003 – Umfang der bereitgestellten APIs

Anforderungs- beschreibung	Die Plattform muss mindestens Schnittstellen für KI-Dialoge, Agentenausführungen, Wissens- und Retrieval-Funktionen, ML-Services sowie den Abruf von Ergebnissen und Statusinformationen bereitstellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die APIs bilden die zentrale Integrationsschnittstelle des AI Harness für den Urban KI Hub sowie weitere berechnigte Anwendungen.
Begründung	Der definierte Funktionsumfang stellt die vollständige Nutzbarkeit des KI-Stacks über standardisierte Schnittstellen sicher.

IA-004 – Authentifizierung und Autorisierung

Anforderungs- beschreibung	Sämtliche APIs müssen durch geeignete Authentifizierungs- und Autorisierungsmechanismen abgesichert sein und die bestehende Identity- und Access-Management-Infrastruktur der Auftraggeberin unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Absicherung erfolgt unter Berücksichtigung der zentralen Authentifizierungs-, Rollen- und Berechtigungskonzepte der Plattform.
Begründung	Eine zentrale Zugriffskontrolle gewährleistet den sicheren und nachvollziehbaren Zugriff auf sämtliche KI-Funktionen.

IA-005 – API-Dokumentation und Versionierung

Anforderungs- beschreibung	Sämtliche APIs müssen vollständig dokumentiert, versioniert und maschinenlesbar beschrieben werden (z. B. OpenAPI oder funktional gleichwertig). Änderungen an bestehenden Schnittstellen müssen nachvollziehbar dokumentiert werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Dokumentation muss Entwicklern eine eigenständige Integration ermöglichen und Informationen zu Versionen, Authentifizierung, Datenformaten sowie Fehlercodes enthalten.
Begründung	Eine vollständige und versionierte API-Dokumentation gewährleistet einen langfristig stabilen und wartbaren Integrationsbetrieb.

3.4 Integration in den Urban Stack

Zielsetzung

Die Plattform muss sich über standardisierte Schnittstellen in die bestehende Systemlandschaft des Urban Stack integrieren lassen. Ziel ist eine lose gekoppelte Integrationsarchitektur, welche den Austausch von Daten und Diensten zwischen dem KI-Stack und den vorhandenen Plattformkomponenten ermöglicht.

Die Integration erfolgt über den AI Harness und die definierten Programmierschnittstellen der Plattform.

Integrationsgrundsatz

Die Plattform darf keine direkte technische Abhängigkeit zu einzelnen Komponenten des Urban Stack aufweisen. Sämtliche Integrationen erfolgen über standardisierte, dokumentierte und versionierte Schnittstellen.

Vorhandene Plattformdienste und Integrationsmechanismen der Auftraggeberin sind zu nutzen. Die Entwicklung individueller Punkt-zu-Punkt-Integrationen ist auf das technisch erforderliche Maß zu beschränken.

IA-006 – Integration in den Urban Stack

Anforderungs- beschreibung	Die Plattform muss sich über standardisierte Schnittstellen in die bestehende Systemlandschaft des Urban Stack integrieren lassen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Integration erfolgt über den AI Harness unter Nutzung der vorhandenen Integrationsmechanismen und Plattformdienste der Auftraggeberin.
Begründung	Die Integration in den Urban Stack ermöglicht die Nutzung vorhandener Daten, Dienste und Fachverfahren ohne redundante Systemlandschaften.

IA-007 – Datenintegration

Anforderungs- beschreibung	Die Plattform muss Daten aus angebundenen Fachverfahren und Plattformdiensten lesen sowie – soweit fachlich erforderlich – Verarbeitungsergebnisse über definierte Schnittstellen bereitstellen können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Für die Erstimplementierung erfolgt die Integration über die definierten REST-Schnittstellen sowie den bereitgestellten S3-kompatiblen Objektspeicher. Weitere Datenquellen müssen architektonisch vorbereitbar sein.
Begründung	Eine standardisierte Datenintegration ermöglicht die effiziente Nutzung vorhandener Informationen innerhalb des KI-Stacks.

IA-008 – Standardisierte Datenformate

Anforderungs- beschreibung	Die Kommunikation zwischen dem KI-Stack und angebundenen Systemen muss über standardisierte und maschinenlesbare Datenformate erfolgen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Es sind insbesondere JSON oder funktional gleichwertige Formate zu unterstützen.
Begründung	Standardisierte Datenformate erleichtern die Interoperabilität und reduzieren den Integrationsaufwand.

IA-009 – Lose Kopplung der Integration

Anforderungs- beschreibung	Die Integration des KI-Stacks muss so erfolgen, dass Änderungen an einzelnen Komponenten des Urban Stack keine grundlegenden Änderungen an der Architektur oder den Schnittstellen des KI-Stacks erfordern.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Anbindung erfolgt ausschließlich über standardisierte und versionierte Schnittstellen. Direkte technische Abhängigkeiten zu einzelnen Plattformkomponenten sind zu vermeiden.
Begründung	Eine lose gekoppelte Integrationsarchitektur erhöht die Wartbarkeit, Erweiterbarkeit und Zukunftssicherheit der Plattform.

9.3 Einbindung externer Systeme

Zielsetzung

Die Plattform muss die standardisierte und sichere Einbindung externer Systeme, Fachverfahren und Dienste unterstützen. Ziel ist die Integration zusätzlicher Datenquellen, Werkzeuge und Fachanwendungen in den KI-Stack, ohne die lose Kopplung der Gesamtarchitektur zu beeinträchtigen.

Die Einbindung erfolgt über den AI Harness und dessen standardisierte Integrationsmechanismen.

Integrationsgrundsatz

Externe Systeme sind ausschließlich über definierte, dokumentierte und abgesicherte Schnittstellen anzubinden. Die Ausführung externer Funktionen erfolgt kontrolliert durch den AI Harness unter Berücksichtigung von Rollen, Berechtigungen, Governance-Richtlinien und Sicherheitsmechanismen.

IA-010 – Integration externer Systeme

Anforderungsbeschreibung	Die Plattform muss die Einbindung externer Systeme, Fachverfahren und Dienste über standardisierte Schnittstellen unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Integration erfolgt über den AI Harness und dessen standardisierte Integrationsmechanismen.
Begründung	Die Anbindung externer Systeme erweitert den Funktionsumfang der Plattform und ermöglicht die Integration in bestehende Verwaltungsprozesse.

IA-011 – Unterstützte Integrationsmechanismen

Anforderungsbeschreibung	Die Plattform muss mindestens API-basierte Integrationen, datenbasierte Import- und Exportverfahren sowie ereignisgesteuerte Integrationen unterstützen. Darüber hinaus soll die Anbindung standardisierter Tool-Protokolle, insbesondere MCP (Model Context Protocol) oder funktional gleichwertiger Verfahren, unterstützt werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die konkrete technische Umsetzung bleibt dem Auftragnehmer überlassen, sofern die definierten Integrationsmechanismen unterstützt werden.
Begründung	Unterschiedliche Integrationsmechanismen ermöglichen die flexible Anbindung verschiedenster Systeme und Dienste.

IA-012 – Integration als KI-Werkzeug

Anforderungsbeschreibung	Externe Systeme und Dienste müssen als Werkzeuge (Tools) innerhalb der Agent Runtime eingebunden und durch KI-Agenten kontrolliert genutzt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Tool-Aufrufe erfolgen ausschließlich über den AI Harness unter Berücksichtigung der definierten Sicherheits-, Governance- und Berechtigungsmechanismen.
Begründung	Die standardisierte Tool-Integration ermöglicht die kontrollierte Nutzung externer Funktionen innerhalb von KI-gestützten Prozessen.

IA-013 – Kontrollierte Nutzung externer Systeme

Anforderungsbeschreibung	Die Nutzung externer Systeme, Dienste und Werkzeuge muss rollen-, mandanten- und berechtigungsabhängig steuerbar sein. Zugriffe sowie der Austausch von Daten müssen nachvollziehbar protokolliert und unter Berücksichtigung der geltenden Sicherheits-, Datenschutz- und Governance-Richtlinien erfolgen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Steuerung erfolgt zentral über den AI Harness. Hierbei sind insbesondere Rollen und Berechtigungen, Mandantenkontext, Security Trimming, Datenschutzklassifizierung, Policy- und Guardrail-Regeln sowie Freigabe- und Human-in-the-Loop-Prozesse zu berücksichtigen. Sämtliche externen Interaktionen müssen revisionssicher protokolliert und über die zentralen Monitoring- und Observability-Funktionen nachvollziehbar sein.
Begründung	Eine zentrale Steuerung und Nachvollziehbarkeit der Nutzung externer Systeme gewährleistet einen sicheren, datenschutzkonformen und revisionssicheren Betrieb des KI-Stacks. Gleichzeitig wird sichergestellt, dass externe Integrationen denselben Governance-, Sicherheits- und Compliance-Anforderungen unterliegen wie alle übrigen Komponenten der Plattform.

9.4 Webhooks und event-basierte Integration

Zielsetzung

Die Plattform muss ereignisgesteuerte Integrationsmechanismen unterstützen, um den automatisierten Austausch von Informationen zwischen dem KI-Stack und angebundenen Systemen zu ermöglichen.

Ziel ist die Auslösung, Verarbeitung und Weitergabe definierter Ereignisse zur Unterstützung automatisierter Geschäftsprozesse, Workflows sowie KI-gestützter Verarbeitungsketten.

Integrationsgrundsatz

Die Ereignisverarbeitung erfolgt zentral über den AI Harness. Ereignisse müssen nachvollziehbar verarbeitet, protokolliert und unter Berücksichtigung von Rollen, Berechtigungen sowie Governance-Richtlinien weitergeleitet werden.

Die konkrete technische Umsetzung bleibt dem Auftragnehmer überlassen. Hierzu können beispielsweise Webhooks, Event-Broker, Messaging-Systeme oder funktional gleichwertige Mechanismen eingesetzt werden.

IA-014 – Unterstützung ereignisgesteuerter Integrationen

Anforderungs- beschreibung	Die Plattform muss Mechanismen bereitstellen, über die externe Systeme automatisiert über definierte Ereignisse informiert werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Benachrichtigung kann beispielsweise über Webhooks oder funktional gleichwertige ereignisgesteuerte Integrationsmechanismen erfolgen.
Begründung	Ereignisgesteuerte Integrationen ermöglichen die Automatisierung von Geschäftsprozessen und reduzieren den Integrationsaufwand.

IA-015 – Ereignisgesteuerte Kommunikation

Anforderungs- beschreibung	Die Plattform muss definierte Ereignisse erzeugen, weiterverarbeiten und für nachgelagerte Systeme bereitstellen können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Ereignisverarbeitung erfolgt zentral über den AI Harness und muss sich in bestehende Integrations- und Prozesslandschaften einfügen.
Begründung	Ereignisgesteuerte Kommunikation ermöglicht lose gekoppelte und skalierbare Integrationsprozesse.

IA-016 – Konfiguration von Ereignissen

Anforderungs- beschreibung	Berechtigte Administratoren müssen Ereignisse, Benachrichtigungen und deren Zielsysteme konfigurieren können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Konfiguration erfolgt über die Administrationsoberfläche des Urban KI Hub oder eine funktional gleichwertige Administrationsoberfläche.
Begründung	Eine zentrale Konfiguration ermöglicht die flexible Anpassung ereignisgesteuerter Integrationen an unterschiedliche fachliche Anforderungen.

IA-017 – Unterstützte Ereignistypen

Anforderungs- beschreibung	Die Plattform muss mindestens Ereignisse zu abgeschlossenen KI-Anfragen, Fehlerzuständen, Workflow-Ausführungen sowie Änderungen an relevanten Konfigurationen bereitstellen. Weitere Ereignistypen müssen zukünftig ergänzt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Ereignisse müssen eindeutig definiert, dokumentiert und über die bereitgestellten Integrationsmechanismen verfügbar sein.
Begründung	Standardisierte Ereignistypen ermöglichen die Integration des KI-Stacks in bestehende Automatisierungs- und Monitoringprozesse sowie die spätere Erweiterung um zusätzliche Ereignisse.

9.5 Interoperabilität und Erweiterbarkeit

Zielsetzung

Die Plattform muss durch standardisierte, dokumentierte und versionierte Schnittstellen eine langfristig interoperable und erweiterbare Integrationsarchitektur unterstützen.

Ziel ist die herstellerunabhängige Integration neuer Komponenten, Fachverfahren und Dienste, ohne grundlegende Änderungen an bestehenden Systemen oder laufenden Betriebsprozessen zu verursachen.

Integrationsgrundsatz

Die Schnittstellen des KI-Stacks sind nach dem Prinzip der losen Kopplung zu gestalten. Änderungen oder Erweiterungen einzelner Komponenten dürfen die Funktion anderer Komponenten nicht unnötig beeinträchtigen.

Neue Funktionen und Integrationen müssen über definierte Schnittstellen ergänzt werden können, ohne bestehende Integrationen grundlegend anzupassen.

IA-018 – Standardisierte Schnittstellen

Anforderungsbeschreibung	Sämtliche Schnittstellen müssen auf offenen und standardisierten Technologien basieren, sodass die Integration weiterer Systeme ohne proprietäre Abhängigkeiten möglich ist.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform nutzt dokumentierte und allgemein anerkannte Schnittstellenstandards. Proprietäre Schnittstellen sind nur zulässig, sofern keine funktional gleichwertige offene Alternative verfügbar ist.
Begründung	Offene Schnittstellen fördern Interoperabilität, Herstellerunabhängigkeit und langfristige Investitionssicherheit.

IA-019 – Versionierung und Kompatibilität

Anforderungsbeschreibung	Sämtliche APIs und Integrationsschnittstellen müssen versioniert werden. Änderungen an bestehenden Schnittstellen sind nachvollziehbar zu dokumentieren und sollen bestehende Integrationen möglichst nicht beeinträchtigen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform muss ein nachvollziehbares Versionierungskonzept bereitstellen. Nicht rückwärtskompatible Änderungen sind zu kennzeichnen und mit angemessenen Übergangsfristen bereitzustellen.
Begründung	Eine kontrollierte Versionierung gewährleistet einen stabilen Integrationsbetrieb und erleichtert die Weiterentwicklung der Plattform.

IA-020 – Erweiterbarkeit der Schnittstellen

Anforderungsbeschreibung	Die Schnittstellen müssen so gestaltet sein, dass zusätzliche Funktionen, Integrationen und Dienste ohne Unterbrechung des laufenden Betriebs ergänzt werden können.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Architektur muss die schrittweise Erweiterung der Schnittstellen sowie die Integration zusätzlicher Komponenten ermöglichen, ohne grundlegende Änderungen an bestehenden Schnittstellen vorzunehmen.
Begründung	Eine erweiterbare Schnittstellenarchitektur erhöht die Zukunftssicherheit des KI-Stacks und ermöglicht dessen kontinuierliche Weiterentwicklung.

9.6 Integration von MCP-Servern (Model Context Protocol)

Zielsetzung

Die Plattform soll standardisierte Mechanismen zur Integration externer Werkzeuge, Datenquellen und Dienste unterstützen. Ziel ist die herstellerunabhängige, interoperable und erweiterbare Einbindung zusätzlicher Funktionen in den KI-Stack.

Hierzu soll insbesondere das **Model Context Protocol (MCP)** als standardisierter Integrationsmechanismus unterstützt werden. Funktional gleichwertige Verfahren bleiben ausdrücklich zulässig.

Integrationsgrundsatz

Die Einbindung externer Werkzeuge und Dienste erfolgt über den AI Harness. Die Architektur muss sicherstellen, dass Werkzeuge unabhängig von einzelnen Sprachmodellen genutzt werden können und zentral durch die Agent Runtime orchestriert werden.

Die konkrete technische Umsetzung bleibt dem Auftragnehmer überlassen, sofern die Anforderungen an Interoperabilität, Sicherheit, Erweiterbarkeit und Governance erfüllt werden.

IA-021 – Unterstützung standardisierter Tool-Protokolle

Anforderungsbeschreibung	Die Plattform soll standardisierte Mechanismen zur Integration externer Werkzeuge, Datenquellen und Dienste unterstützen. Hierzu soll insbesondere das Model Context Protocol (MCP) oder ein funktional gleichwertiger Standard genutzt werden können.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Architektur muss die Integration standardisierter Tool-Protokolle ermöglichen, ohne sich auf ein bestimmtes Verfahren festzulegen.
Begründung	Standardisierte Integrationsmechanismen fördern Interoperabilität, Herstellerunabhängigkeit und Zukunftssicherheit.

IA-022 – Nutzung standardisierter Integrationsmechanismen

Anforderungsbeschreibung	Für die Anbindung externer Werkzeuge, Datenquellen und Dienste sollen bevorzugt standardisierte Integrationsmechanismen, insbesondere MCP oder funktional gleichwertige Verfahren, eingesetzt werden.
Kategorie	Soll-Anforderung
Umsetzung	Abweichende Integrationsmechanismen sind zulässig, sofern sie keine Einschränkungen hinsichtlich Interoperabilität, Erweiterbarkeit oder Wartbarkeit verursachen.
Begründung	Standardisierte Verfahren reduzieren Integrationsaufwand und erleichtern die spätere Erweiterung der Plattform.

IA-023 – Entkopplung von Modellen und Werkzeugen

Anforderungsbeschreibung	Die Architektur muss sicherstellen, dass Werkzeuge unabhängig vom jeweils eingesetzten Sprachmodell genutzt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Tool-Orchestrierung erfolgt zentral über den AI Harness und die Agent Runtime.
Begründung	Die Trennung von Modell und Werkzeugen erhöht die Flexibilität und ermöglicht den Austausch einzelner Komponenten ohne Auswirkungen auf die Gesamtarchitektur.

IA-024 – Integration externer Tool-Server

Anforderungsbeschreibung	Die Plattform soll externe MCP-Server oder funktional gleichwertige Tool-Server anbinden und deren Funktionen innerhalb der KI-Verarbeitung nutzen können.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Architektur muss die spätere Integration zusätzlicher Tool-Server unterstützen, ohne grundlegende Änderungen an der Plattform vorzunehmen.
Begründung	Eine offene Integrationsarchitektur ermöglicht die schrittweise Erweiterung des KI-Stacks um zusätzliche Werkzeuge und Dienste.

IA-025 – Berechtigungs- und Zugriffskontrolle

Anforderungsbeschreibung	Der Zugriff auf externe Tool-Server sowie deren Funktionen muss rollen-, mandanten- und berechtigungsabhängig steuerbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Steuerung erfolgt zentral über den AI Harness unter Berücksichtigung von Rollen, Mandantenkontext, Policies, Guardrails und Security Trimming. Sämtliche Zugriffe sind nachvollziehbar zu protokollieren.
Begründung	Eine zentrale Zugriffskontrolle gewährleistet einen sicheren und datenschutzkonformen Einsatz externer Werkzeuge.

IA-026 – Protokollierung von Tool-Aufrufen

Anforderungsbeschreibung	Aufrufe externer Tool-Server sollen vollständig protokolliert und den jeweiligen Agenten- und Verarbeitungsausführungen zugeordnet werden können.
Kategorie	Soll-Anforderung
Umsetzung	Die Protokollierung erfolgt über die zentralen Logging-, Monitoring- und Observability-Funktionen des AI Harness.
Begründung	Die Nachvollziehbarkeit von Tool-Aufrufen unterstützt Monitoring, Auditierung und Fehleranalyse.

IA-027 – Erweiterbarkeit

Anforderungsbeschreibung	Die Plattform soll so gestaltet sein, dass zusätzliche MCP-Server oder funktional gleichwertige Tool-Server ohne grundlegende Änderungen an der bestehenden Systemarchitektur integriert werden können.
Kategorie	A – Architekturvorbereitung
Umsetzung	Neue Tool-Server müssen über definierte Schnittstellen ergänzt werden können, ohne bestehende Integrationen zu beeinträchtigen.
Begründung	Eine erweiterbare Integrationsarchitektur erhöht die Zukunftssicherheit und ermöglicht die kontinuierliche Erweiterung des KI-Stacks.

10. Qualitätssicherung & Monitoring

Zielsetzung

Die Plattform muss Mechanismen zur Qualitätssicherung, Überwachung, Nachvollziehbarkeit und kontinuierlichen Verbesserung sämtlicher Komponenten des KI-Stacks bereitstellen.

Ziel ist die Sicherstellung eines stabilen, transparenten, wirtschaftlichen und fachlich verlässlichen Betriebs der Plattform. Hierzu sind sämtliche wesentlichen Verarbeitungsschritte, Komponenten und Integrationen übergreifend zu überwachen und nachvollziehbar auszuwerten.

Qualitätsgrundsatz

Das Monitoring muss komponentenübergreifend erfolgen und den gesamten Verarbeitungsprozess innerhalb des KI-Stacks abdecken. Hierzu gehören insbesondere Benutzerinteraktionen, Agentenausführungen, Orchestrierung, Retrieval-Prozesse, Dokumentenverarbeitung, Modellaufrufe, Tool-Interaktionen, ML-Services sowie externe Integrationen.

Die Plattform muss zentrale Mechanismen zur Observability bereitstellen und insbesondere Logging, Metriken, Tracing sowie Auditinformationen miteinander verknüpfen können. Ziel ist eine vollständige Nachvollziehbarkeit technischer Abläufe, fachlicher Entscheidungen und sicherheitsrelevanter Ereignisse.

Die Qualitätssicherung umfasst neben der technischen Überwachung auch die kontinuierliche Bewertung der Antwortqualität, der Retrieval-Ergebnisse, der Modellnutzung sowie der Einhaltung definierter Governance-, Datenschutz- und Sicherheitsrichtlinien.

Die konkrete technische Umsetzung bleibt dem Auftragnehmer überlassen, sofern die Anforderungen an Monitoring, Nachvollziehbarkeit und Qualitätssicherung erfüllt werden.

10.1 KI-Qualität

Zielsetzung

Die Plattform muss Mechanismen zur kontinuierlichen Bewertung, Überwachung und Verbesserung der Qualität von Sprachmodellen, Retrieval-Verfahren, RAG-Konfigurationen, Agenten sowie Prompt-Konfigurationen unterstützen.

Ziel ist die objektive, reproduzierbare und nachvollziehbare Bewertung der Qualität sämtlicher KI-Komponenten über den gesamten Lebenszyklus der Plattform.

Qualitätsgrundsatz

Die Plattform soll einen zentralen **Evaluations-Harness** bereitstellen oder integrieren. Dieser dient der reproduzierbaren Bewertung und Vergleichbarkeit unterschiedlicher Modelle, Agenten, Retrieval-Strategien, Prompt-Versionen sowie weiterer KI-Komponenten.

Die Evaluierung ist Bestandteil der Qualitätssicherungs- und Governance-Architektur des KI-Stacks und soll über den AI Harness mit den Monitoring-, Observability- und Protokollierungsfunktionen der Plattform zusammenwirken.

QM-001 – Evaluations-Harness

Anforderungsbeschreibung	Die Plattform soll einen Evaluations-Harness bereitstellen oder integrieren, der eine reproduzierbare Bewertung von Sprachmodellen, Retrieval-Verfahren, RAG-Konfigurationen, Agenten und Prompt-Konfigurationen ermöglicht.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Architektur muss die Integration eines Evaluations-Harness unterstützen. Die konkrete technische Umsetzung bleibt dem Auftragnehmer überlassen.
Begründung	Ein Evaluations-Harness ermöglicht eine kontinuierliche Qualitätssicherung und objektive Vergleichbarkeit unterschiedlicher KI-Konfigurationen.

QM-002 – Reproduzierbare Evaluierungen

Anforderungsbeschreibung	Evaluierungen müssen reproduzierbar durchgeführt werden können. Hierzu sind identische Eingaben, Konfigurationen und Bewertungsgrundlagen nachvollziehbar wiederverwendbar bereitzustellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform muss die Dokumentation der verwendeten Modelle, Prompt-Versionen, Retrieval-Konfigurationen sowie relevanter Laufzeitparameter unterstützen.
Begründung	Reproduzierbare Evaluierungen bilden die Grundlage für objektive Qualitätsvergleiche und kontinuierliche Verbesserungen.

QM-003 – Benchmark- und Referenzdatensätze

Anforderungsbeschreibung	Die Plattform soll Benchmark-Datensätze, Referenzfragen (Golden Questions) sowie vergleichbare Evaluierungsszenarien unterstützen.
Kategorie	Soll-Anforderung
Umsetzung	Unterschiedliche Benchmark-Sammlungen müssen für wiederkehrende Qualitätsbewertungen genutzt werden können.
Begründung	Referenzdatensätze ermöglichen objektive Qualitätsvergleiche über verschiedene Modelle und Konfigurationen hinweg.

QM-004 – Vergleichbarkeit von KI-Konfigurationen

Anforderungsbeschreibung	Unterschiedliche Modelle, Prompt-Versionen, Retrieval-Strategien, Agentenkonfigurationen sowie Modell-Routing-Strategien sollen vergleichbar bewertet werden können.
Kategorie	Soll-Anforderung
Umsetzung	Die Evaluierung muss einheitliche Qualitätskennzahlen sowie vergleichbare Bewertungsverfahren unterstützen.
Begründung	Die Vergleichbarkeit unterstützt die Auswahl fachlich, wirtschaftlich und datenschutzrechtlich geeigneter KI-Konfigurationen.

QM-005 – Historisierung von Evaluierungen

Anforderungsbeschreibung	Evaluierungsergebnisse müssen versioniert und historisch nachvollziehbar gespeichert werden.
Kategorie	Soll-Anforderung
Umsetzung	Die Historisierung umfasst insbesondere Modelle, Modellversionen, Prompt-Versionen, Agentenkonfigurationen sowie die jeweiligen Bewertungsergebnisse.
Begründung	Historische Evaluierungen unterstützen Trendanalysen und kontinuierliche Qualitätsverbesserungen.

QM-006 – Nachvollziehbarkeit von Antworten

Anforderungsbeschreibung	KI-generierte Antworten müssen nachvollziehbar sein und auf den verwendeten Wissensquellen sowie den zugrunde liegenden Verarbeitungsschritten basieren.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform nutzt die Protokollierungs-, Tracing- und Citation-Mechanismen des AI Harness.
Begründung	Nachvollziehbare Antworten erhöhen Vertrauen, Transparenz und fachliche Verlässlichkeit.

QM-007 – Quellenbasierte Antworten

Anforderungsbeschreibung	Wissensbasierte Antworten müssen auf den durch die Retrieval-Schicht bereitgestellten Informationen basieren. Quellenangaben müssen für geeignete Anwendungsfälle dargestellt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Quelleninformationen werden über die Retrieval-Schicht sowie den Citation Validator bereitgestellt.
Begründung	Quellenbasierte Antworten erhöhen die Nachvollziehbarkeit und reduzieren das Risiko unbelegter Aussagen.

QM-008 – Reduktion fehlerhafter Antworten

Anforderungsbeschreibung	Die Plattform muss Mechanismen unterstützen, um fehlerhafte, unbelegte oder halluzinierte Antworten zu reduzieren.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Hierzu werden insbesondere Retrieval, Quellenvvalidierung, Guardrails, Citation Validator sowie Human-in-the-Loop-Prozesse genutzt.
Begründung	Die Reduktion fehlerhafter Antworten verbessert die Qualität und Vertrauenswürdigkeit der Plattform.

QM-009 – Kontrollierte Antwortverweigerung

Anforderungsbeschreibung	Das System muss bei unzureichender Datenlage oder unzureichender Antwortsicherheit auf eine Antwort verzichten oder die Unsicherheit transparent darstellen können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Entscheidung erfolgt auf Grundlage definierter Qualitäts- und Governance-Regeln innerhalb des AI Harness.
Begründung	Eine kontrollierte Antwortverweigerung ist fachlich sinnvoller als die Ausgabe unbelegter oder potenziell fehlerhafter Antworten.

QM-010 – Qualitätskennzahlen

Anforderungsbeschreibung	Die Plattform soll Qualitätskennzahlen zur Bewertung eingesetzter Modelle und Konfigurationen bereitstellen. Hierzu gehören insbesondere Antwortqualität, Antwortlatenz, Halluzinationsrate, Kosten pro Anfrage, Tokenverbrauch, Retrieval-Erfolgsquote sowie Erfolgsquote von Agenten- und Tool-Aufrufen.
Kategorie	Soll-Anforderung
Umsetzung	Die Kennzahlen werden über den Evaluations-Harness sowie die Monitoring- und Observability-Komponenten bereitgestellt.
Begründung	Qualitätskennzahlen unterstützen die kontinuierliche Optimierung der Plattform hinsichtlich Qualität, Wirtschaftlichkeit und Leistungsfähigkeit.

QM-011 – Vergleichbarkeit von Modellen

Anforderungsbeschreibung	Die Plattform soll die vergleichende Bewertung unterschiedlicher Modelle, Modellversionen und Modell-Routing-Strategien unterstützen, um fachlich, wirtschaftlich und datenschutzrechtlich geeignete Modelle identifizieren zu können.
Kategorie	Soll-Anforderung
Umsetzung	Die Bewertung erfolgt auf Grundlage einheitlicher Qualitätsmetriken und nachvollziehbarer Evaluierungsergebnisse.
Begründung	Eine systematische Modellbewertung unterstützt die langfristige Optimierung des KI-Stacks und reduziert Kosten bei gleichbleibender Qualität.

10.1.1 Evaluation von Retrieval-, RAG- und Agentensystemen

Zielsetzung

Die Plattform soll Mechanismen zur kontinuierlichen Evaluierung und Qualitätssicherung von Retrieval-, Retrieval-Augmented-Generation-(RAG)- sowie agentenbasierten Verarbeitungsvorgängen unterstützen.

Ziel ist die objektive Bewertung, der Vergleich sowie die kontinuierliche Verbesserung der Qualität wissensbasierter KI-Anwendungen über den gesamten Lebenszyklus der Plattform.

Evaluierungsgrundsatz

Die Plattform soll wiederkehrende Evaluierungen von Retrieval-, RAG- und Agentensystemen ermöglichen. Hierzu sind Testdatensätze, Referenzfragen (Golden Questions), Benchmark-Szenarien sowie automatisierte Evaluierungsverfahren nutzbar zu machen.

Die Evaluierung ist Bestandteil des Evaluations-Harness und soll mit den Monitoring-, Observability- und Governance-Funktionen des AI Harness zusammenwirken.

QM-012 – Evaluation von Retrieval- und Agentensystemen

Anforderungsbeschreibung	Die Plattform soll Mechanismen zur kontinuierlichen Evaluierung von Retrieval-, RAG- und Agentensystemen unterstützen.
Kategorie	Soll-Anforderung
Umsetzung	Die Evaluierung erfolgt auf Grundlage reproduzierbarer Testdatensätze, Benchmark-Szenarien sowie Referenzfragen.
Begründung	Eine kontinuierliche Evaluierung unterstützt die langfristige Sicherung und Verbesserung der Antwortqualität.

QM-013 – Qualitätskriterien

Anforderungsbeschreibung	Die Evaluierung soll insbesondere Retrieval-Qualität, Quellenrelevanz, Antworttreue (Faithfulness), Kontextnutzung, Halluzinationsrate, Agentenqualität, Tool-Nutzung, Modellvergleich sowie die Qualität des Modell-Routings bewerten können.
Kategorie	Soll-Anforderung
Umsetzung	Die Qualitätsbewertung erfolgt anhand definierter Evaluierungskriterien und nachvollziehbarer Bewertungsverfahren.
Begründung	Einheitliche Qualitätskriterien ermöglichen objektive Vergleiche unterschiedlicher Konfigurationen.

QM-014 – Wiederkehrende Evaluierungen und Regressionstests

Anforderungsbeschreibung	Die Plattform soll regelmäßige Evaluierungen sowie Regressionstests unterstützen, um Qualitätsveränderungen von Retrieval-, RAG- und Agentensystemen über mehrere Zeitpunkte hinweg nachvollziehbar erkennen und bewerten zu können.
Kategorie	Soll-Anforderung
Umsetzung	Evaluierungsergebnisse müssen historisch vergleichbar gespeichert und automatisiert gegenüber früheren Ergebnissen ausgewertet werden können.
Begründung	Regressionstests ermöglichen die frühzeitige Erkennung von Qualitätsverschlechterungen nach Änderungen an Modellen, Prompts, Retrieval-Konfigurationen oder Agenten.

QM-015 – Dokumentation und Qualitätsverbesserung

Anforderungs- beschreibung	Die Ergebnisse der Evaluierungen sollen nachvollziehbar dokumentiert und für die kontinuierliche Verbesserung von Modellen, Retrieval-Verfahren, Agenten und Prompt-Konfigurationen nutzbar gemacht werden.
Kategorie	Soll-Anforderung
Umsetzung	Die Evaluierungsergebnisse werden versioniert gespeichert und mit den Monitoring- und Observability-Daten der Plattform verknüpft.
Begründung	Die systematische Dokumentation bildet die Grundlage für datengetriebene Qualitätsverbesserungen und langfristige Optimierungen des KI-Stacks.

10.2 Systemmonitoring

Zielsetzung

Die Plattform muss ein zentrales Monitoring aller wesentlichen Komponenten des KI-Stacks unterstützen. Ziel ist die frühzeitige Erkennung von Störungen, Leistungsproblemen und Qualitätsabweichungen sowie die Sicherstellung eines stabilen und nachvollziehbaren Betriebs.

Monitoring-Grundsatz

Das Monitoring muss den gesamten KI-Stack umfassen und insbesondere Infrastruktur, AI Harness, Agent Runtime, Dokumentenverarbeitung, Retrieval, ML-Services, Modellaufrufe sowie externe Integrationen berücksichtigen.

Monitoring-Daten sind Bestandteil der zentralen Observability-Architektur und sollen mit Logging-, Tracing- und Auditinformationen verknüpft werden können.

QM-012 – Überwachung des Systemzustands

Anforderungsbeschreibung	Die Plattform muss hinsichtlich Verfügbarkeit, Performance und Fehlerzuständen überwacht werden. Die Überwachung muss darüber hinaus insbesondere Dokumenteneingänge, Dokumentenverarbeitung, Extraktion, Chunking, Embedding-Erzeugung, Indexierung, Re-Chunking, Neuindexierungen, Wissensdatenbanken, Retrieval-Indizes sowie die Agent Runtime und Modellaufrufe umfassen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Das Monitoring erfolgt komponentenübergreifend über die Observability-Funktionen des AI Harness und umfasst sämtliche wesentlichen Verarbeitungsschritte des KI-Stacks.
Begründung	Eine umfassende Überwachung ermöglicht die frühzeitige Erkennung technischer Störungen und unterstützt einen stabilen Betrieb der Plattform.

QM-013 – Monitoring-Dashboard

Anforderungsbeschreibung	Der aktuelle Zustand der Plattform muss über ein zentrales Monitoring-Dashboard für berechtigte Administratoren nachvollziehbar dargestellt werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Das Dashboard soll insbesondere den Status von Infrastruktur, AI Harness, Agent Runtime, Dokumentenverarbeitung, Retrieval-Komponenten, ML-Services, Modellaufrufen sowie externen Integrationen visualisieren.
Begründung	Ein zentrales Dashboard erleichtert die Überwachung, Fehleranalyse und Administration der Plattform.

QM-014 – Alarmierung und Ereignismanagement

Anforderungsbeschreibung	Bei kritischen Systemzuständen, Sicherheitsereignissen oder Qualitätsabweichungen müssen automatische Benachrichtigungen ausgelöst werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Schwellwerte, Benachrichtigungen und Eskalationsmechanismen müssen administrativ konfigurierbar sein und sich in bestehende Monitoring- und Incident-Management-Prozesse integrieren lassen.
Begründung	Eine automatisierte Alarmierung ermöglicht eine schnelle Reaktion auf kritische Ereignisse und unterstützt einen sicheren und hochverfügbaren Betrieb der Plattform.

10.3 KI- und Nutzungsmonitoring

Zielsetzung

Die Plattform muss die Nutzung der KI-Komponenten transparent überwachen und auswertbar machen. Ziel ist die kontinuierliche Analyse der Modellnutzung, des Ressourcenverbrauchs, der Wirtschaftlichkeit sowie der Ausführung von Retrieval-, Agenten- und Tool-Prozessen.

Monitoring-Grundsatz

Das KI- und Nutzungsmonitoring ist Bestandteil der Observability-Architektur des AI Harness. Es muss eine nachvollziehbare Bewertung der Nutzung von Sprachmodellen, Agenten, Retrieval-Komponenten, ML-Services und externen Werkzeugen ermöglichen. Die gewonnenen Informationen dienen insbesondere der Kostenkontrolle, der Optimierung des Modell-Routings, der Kapazitätsplanung sowie der kontinuierlichen Verbesserung der Plattform.

QM-015 – Modell- und Token-Nutzung

Anforderungsbeschreibung	Die Nutzung von Sprachmodellen, KI-Komponenten sowie der Tokenverbrauch müssen vollständig erfasst, ausgewertet und nachvollziehbar dargestellt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Erfassung erfolgt zentral über den AI Harness und umfasst insbesondere Modellaufrufe, Modellversionen, Tokenverbrauch, Antwortzeiten sowie weitere nutzungsbezogene Kennzahlen.
Begründung	Die transparente Erfassung der Modellnutzung bildet die Grundlage für Kostenkontrolle, Kapazitätsplanung und Optimierung der Plattform.

QM-016 – Kostenmonitoring

Anforderungsbeschreibung	Die Nutzungskosten externer KI-Dienste sowie weiterer kostenrelevanter Komponenten müssen transparent erfasst, ausgewertet und nachvollziehbar dargestellt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Das Monitoring umfasst insbesondere Kosten pro Anfrage, Kosten pro Modell, Kosten pro Mandant, Kosten pro Agent sowie aggregierte Nutzungsstatistiken. Administrativ konfigurierbare Budgets, Kostenlimits und Warnschwellen sollen unterstützt werden.
Begründung	Eine transparente Kostenübersicht unterstützt den wirtschaftlichen Betrieb sowie die Optimierung der Modellnutzung und Ressourcenplanung.

QM-017 – Monitoring von Retrieval-, Agenten- und Tool-Nutzung

Anforderungsbeschreibung	Die Nutzung der Retrieval-Komponenten, Agenten, ML-Services sowie externer Werkzeuge und Tool-Server muss nachvollziehbar überwacht und ausgewertet werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Das Monitoring umfasst insbesondere Retrieval-Anfragen, Agentenausführungen, Tool-Aufrufe, Erfolgs- und Fehlerraten, Antwortzeiten, Dokumentenverarbeitung sowie die Nutzung externer Dienste. Die Ergebnisse müssen über die zentralen Monitoring- und Observability-Funktionen des AI Harness auswertbar sein.
Begründung	Die Überwachung der KI-Verarbeitung ermöglicht die kontinuierliche Optimierung von Qualität, Performance, Sicherheit und Ressourceneinsatz.

QM-018 – Transparenz des Modell-Routings

Anforderungsbeschreibung	Die Plattform muss nachvollziehbar dokumentieren, welches Modell für eine Anfrage ausgewählt wurde und auf Grundlage welcher Routing-Entscheidung dies erfolgt ist. Die Entscheidungsparameter müssen administrativ nachvollziehbar und konfigurierbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Zu dokumentieren sind insbesondere das ausgewählte Modell, die Modellversion, die Routing-Strategie, relevante Entscheidungsparameter, Tokenverbrauch, Kosten, Antwortlatenz sowie die Begründung der Routing-Entscheidung. Die Informationen sind über die Monitoring-, Logging-, Tracing- und Observability-Funktionen des AI Harness bereitzustellen.
Begründung	Die transparente Nachvollziehbarkeit des Modell-Routings unterstützt die Optimierung von Qualität, Wirtschaftlichkeit, Datenschutz und Governance und ermöglicht eine fundierte Analyse von Modellentscheidungen und deren Auswirkungen auf den Plattformbetrieb.

10.4 Protokollierung

Zielsetzung

Die Plattform muss eine zentrale und revisionssichere Protokollierung aller wesentlichen Verarbeitungsvorgänge des KI-Stacks unterstützen. Ziel ist die Nachvollziehbarkeit technischer Abläufe, fachlicher Entscheidungen sowie sicherheits- und datenschutzrelevanter Ereignisse.

Protokollierungsgrundsatz

Die Protokollierung ist Bestandteil der Observability- und Governance-Architektur des AI Harness. Audit-Informationen müssen mit Logging-, Monitoring- und Tracing-Daten verknüpft werden können, um eine durchgängige Nachvollziehbarkeit sämtlicher Verarbeitungsschritte zu gewährleisten.

Die Protokollierung muss mandantenfähig erfolgen und den geltenden Datenschutz-, Sicherheits- und Aufbewahrungsvorgaben entsprechen.

QM-019 – Audit-Logging

Anforderungs- beschreibung	Die Plattform muss alle wesentlichen Verarbeitungsvorgänge revisionssicher protokollieren. Hierzu gehören insbesondere Modellaufrufe, Modellversionen, Agentenausführungen, Retrieval-Prozesse, verwendete Wissensquellen, Tool-Aufrufe, Modell-Routing-Entscheidungen, Human-in-the-Loop-Entscheidungen sowie sicherheits- und datenschutzrelevante Ereignisse.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Audit-Protokollierung erfolgt zentral über den AI Harness und muss mit Logging-, Monitoring- und Tracing-Funktionen verknüpft werden. Die Protokolle sind mandantenfähig, revisionssicher und gegen unbefugte Änderungen zu schützen.
Begründung	Eine revisionssichere Audit-Protokollierung bildet die Grundlage für Governance, Compliance, Datenschutz, Sicherheitsanalysen und spätere Nachweise.

QM-020 – Nachvollziehbarkeit von Verarbeitungsvorgängen

Anforderungs- beschreibung	Sämtliche wesentlichen Verarbeitungsschritte müssen so dokumentiert werden, dass technische Abläufe, fachliche Entscheidungen und erzeugte Ergebnisse im Nachhinein vollständig nachvollzogen werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Nachvollziehbarkeit umfasst insbesondere die verwendeten Modelle und Modellversionen, Prompt-Versionen, Retrieval-Ergebnisse, Wissensquellen, Agentenkonfigurationen, Tool-Aufrufe, Routing-Entscheidungen sowie die jeweilige Trace-ID der Ausführung. Die Informationen sind über die zentralen Observability- und Audit-Funktionen des AI Harness auswertbar bereitzustellen.
Begründung	Eine vollständige Nachvollziehbarkeit unterstützt Fehleranalysen, Qualitätsverbesserungen, Auditierungen sowie die Einhaltung regulatorischer Anforderungen.

10.5 KI-Governance & Vertrauenswürdigkeit

Zielsetzung

Die Plattform muss Mechanismen zur Steuerung, Qualitätssicherung, Nachvollziehbarkeit und kontinuierlichen Überwachung von KI-gestützten Prozessen bereitstellen. Ziel ist die Gewährleistung eines kontrollierten, transparenten und vertrauenswürdigen Einsatzes von KI-Systemen unter Berücksichtigung fachlicher, datenschutzrechtlicher, sicherheitsrelevanter und regulatorischer Anforderungen.

Governance-Grundsatz

Die Plattform muss zentrale Governance-Mechanismen bereitstellen, welche die Einhaltung definierter Richtlinien, Qualitätsanforderungen und Sicherheitsvorgaben während des gesamten Lebenszyklus einer KI-Verarbeitung unterstützen.

Die Governance-Funktionen sind Bestandteil des AI Harness und arbeiten mit der Policy Engine, der Agent Runtime, der Context Engine sowie den Monitoring-, Audit- und Observability-Komponenten zusammen.

Die Plattform soll insbesondere die Anforderungen des EU AI Acts sowie weiterer einschlägiger regulatorischer Vorgaben unterstützen, ohne sich auf eine bestimmte regulatorische Ausprägung festzulegen. Die Architektur muss zukünftige Erweiterungen regulatorischer Anforderungen ohne grundlegende Änderungen ermöglichen.

QM-021 – Zentrale KI-Governance

Anforderungsbeschreibung	Die Plattform muss zentrale Mechanismen zur Steuerung, Überwachung und Qualitätssicherung sämtlicher KI-Prozesse bereitstellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Governance erfolgt zentral über den AI Harness unter Einbindung der Policy Engine, Agent Runtime, Context Engine sowie der Monitoring- und Audit-Komponenten.
Begründung	Eine zentrale Governance gewährleistet einen einheitlichen, kontrollierten und nachvollziehbaren Betrieb sämtlicher KI-Komponenten.

QM-022 – Vertrauenswürdige KI-Verarbeitung

Anforderungsbeschreibung	Die Plattform muss Mechanismen unterstützen, welche Transparenz, Nachvollziehbarkeit, Quellenbezug, kontrollierte Entscheidungsprozesse sowie die Einhaltung definierter Governance-Richtlinien fördern.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Hierzu sind insbesondere Quellenangaben, Citation Validator, Guardrails, Human-in-the-Loop, Policy Engine sowie Audit- und Monitoring-Funktionen einzusetzen.
Begründung	Vertrauenswürdige KI-Verarbeitung erhöht die Akzeptanz der Plattform und unterstützt einen verantwortungsvollen Einsatz von KI.

QM-023 – Richtlinien- und Compliance-Überwachung

Anforderungsbeschreibung	Die Plattform muss die Einhaltung administrativ definierter Richtlinien, Sicherheitsvorgaben, Datenschutzanforderungen und Governance-Regeln überwachen sowie Verstöße nachvollziehbar dokumentieren können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Überwachung erfolgt durch die Policy Engine in Verbindung mit den Logging-, Monitoring- und Audit-Funktionen des AI Harness. Verstöße müssen nachvollziehbar protokolliert und für berechnigte Administratoren auswertbar sein.
Begründung	Die kontinuierliche Überwachung der Governance-Richtlinien unterstützt Compliance, Datenschutz und Informationssicherheit.

QM-024 – Unterstützung regulatorischer Anforderungen

Anforderungs- beschreibung	Die Architektur muss die Umsetzung und Erweiterung regulatorischer Anforderungen, insbesondere des EU AI Acts sowie vergleichbarer gesetzlicher Vorgaben, unterstützen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Plattform muss regulatorische Informationen, Dokumentationen und Klassifizierungen flexibel erweitern können, ohne grundlegende Änderungen an der Gesamtarchitektur vorzunehmen.
Begründung	Regulatorische Anforderungen entwickeln sich kontinuierlich weiter. Eine erweiterbare Governance-Architektur gewährleistet die langfristige Anpassungsfähigkeit der Plattform.

10.5.1 Prompt Governance

Zielsetzung

Die Plattform muss Mechanismen zur kontrollierten Verwaltung, Versionierung, Freigabe und Nachvollziehbarkeit von Prompt-Konfigurationen bereitstellen. Ziel ist die Sicherstellung einer nachvollziehbaren, qualitätsgesicherten und revisionssicheren Nutzung von Prompts innerhalb des gesamten KI-Stacks.

Governance-Grundsatz

Die Verwaltung von Prompts erfolgt zentral über die Prompt Registry des AI Harness. Änderungen an Prompt-Konfigurationen müssen nachvollziehbar dokumentiert, versioniert und administrativ steuerbar sein.

Die Plattform muss sicherstellen, dass ausschließlich freigegebene Prompt-Versionen für produktive KI-Prozesse verwendet werden.

KG-001 – Versionierung von Prompts

Anforderungsbeschreibung	Die Plattform muss die Versionierung von Prompt-Konfigurationen unterstützen. Änderungen müssen nachvollziehbar dokumentiert und frühere Versionen historisch abrufbar sein. Frühere Versionen müssen wiederhergestellt und erneut aktiviert werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Versionierung erfolgt zentral über die Prompt Registry. Jede Version erhält eine eindeutige Versionskennung und kann einzelnen Agenten, Modellen oder Anwendungen zugeordnet werden.
Begründung	Die Versionierung gewährleistet die Nachvollziehbarkeit von Änderungen und unterstützt eine kontrollierte Weiterentwicklung der Prompt-Konfigurationen.

KG-002 – Nachvollziehbarkeit von Änderungen

Anforderungsbeschreibung	Für jede Änderung an einer Prompt-Konfiguration müssen mindestens Autor, Zeitpunkt, Version, Änderungsbeschreibung sowie die Änderungshistorie nachvollziehbar gespeichert werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Sämtliche Änderungen werden revisionssicher protokolliert und mit den Audit- und Governance-Funktionen des AI Harness verknüpft.
Begründung	Eine vollständige Änderungshistorie unterstützt Qualitätssicherung, Auditierungen und die Einhaltung regulatorischer Anforderungen.

KG-003 – Freigabeprozesse

Anforderungsbeschreibung	Die Plattform muss konfigurierbare Freigabeprozesse für Prompt-Konfigurationen unterstützen. Für produktiv eingesetzte oder organisationsweit bereitgestellte Prompts muss mindestens ein Vier-Augen-Prinzip oder ein funktional gleichwertiger Freigabemechanismus unterstützt werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Freigaben erfolgen rollenbasiert und werden vollständig protokolliert. Die konkrete Ausgestaltung der Freigabeworkflows bleibt dem Auftragnehmer überlassen.
Begründung	Kontrollierte Freigabeprozesse reduzieren Fehlkonfigurationen und erhöhen die Qualität produktiver Prompt-Konfigurationen.

KG-004 – Lebenszyklus von Prompt-Konfigurationen

Anforderungs- beschreibung	Prompt-Konfigurationen müssen über definierte Lebenszykluszustände verwaltet werden können. Hierzu gehören mindestens die Zustände Entwurf, getestet, freigegeben und produktiv. Weitere Zustände müssen ergänzt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Verwaltung der Lebenszykluszustände erfolgt zentral über die Prompt Registry und wird bei der Ausführung von Agenten durch den AI Harness berücksichtigt.
Begründung	Ein definierter Lebenszyklus gewährleistet eine kontrollierte Bereitstellung und Weiterentwicklung produktiver Prompt-Konfigurationen.

KG-005 – Rollen- und mandantenbasierte Steuerung

Anforderungs- beschreibung	Die Nutzung, Bearbeitung, Freigabe und Veröffentlichung von Prompt-Konfigurationen muss rollen-, mandanten- und berechtigungsabhängig steuerbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Berechtigungssteuerung erfolgt zentral über den AI Harness unter Berücksichtigung von Rollen, Mandanten, Governance-Richtlinien sowie organisatorischen Freigaberegeln. Sämtliche Änderungen und Freigaben sind revisionssicher zu protokollieren.
Begründung	Eine differenzierte Berechtigungssteuerung gewährleistet einen sicheren und nachvollziehbaren Betrieb der Prompt Registry sowie eine konsistente Governance innerhalb des KI-Stacks.

10.5.2 Halluzinationsvermeidung und Antwortqualität

Zielsetzung

Die Plattform muss Mechanismen zur Sicherstellung einer hohen Antwortqualität sowie zur Reduzierung unbelegter oder fehlerhafter KI-Antworten bereitstellen. Ziel ist die nachvollziehbare, quellenbasierte und vertrauenswürdige Generierung von Antworten auf Grundlage der verfügbaren Wissensquellen.

Qualitätsgrundsatz

Für wissensbasierte Anfragen ist Retrieval-Augmented Generation (RAG) der Standardmechanismus zur Antwortgenerierung. Antworten müssen grundsätzlich auf den durch die Retrieval-Schicht bereitgestellten Informationen basieren und durch geeignete Qualitätsmechanismen abgesichert werden.

Die Plattform muss hierzu insbesondere Retrieval, Citation Validator, Guardrails, Policy Engine sowie Human-in-the-Loop-Verfahren unterstützen.

KG-006 – Quellenbasierte Antwortgenerierung

Anforderungsbeschreibung	Bei wissensbasierten Anfragen müssen Antworten auf Grundlage der durch die Retrieval-Schicht bereitgestellten Informationen generiert werden. Die Retrieval-Schicht umfasst insbesondere OpenSearch sowie gegebenenfalls weitere Hybrid-Retrieval-Komponenten, beispielsweise graphbasierte Wissensquellen. Die verwendeten Inhalte müssen referenzierbar und nachvollziehbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Antwortgenerierung erfolgt über den AI Harness unter Nutzung der Retrieval-Schicht sowie der definierten RAG-Prozesse.
Begründung	Quellenbasierte Antworten erhöhen die fachliche Qualität und reduzieren das Risiko unbelegter Aussagen.

KG-007 – Quellenangaben

Anforderungsbeschreibung	Die Plattform muss die Möglichkeit bieten, verwendete Quellen innerhalb der Antwort darzustellen oder referenzierbar bereitzustellen. Quellen sollen nach Möglichkeit bis auf Dokument-, Abschnitts- oder Chunk-Ebene nachvollziehbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Quellenreferenzierung erfolgt über den Citation Validator in Verbindung mit der Retrieval-Schicht.
Begründung	Transparente Quellenangaben erhöhen die Nachvollziehbarkeit und Vertrauenswürdigkeit der KI-Antworten.

KG-008 – Bewertung der Antwortqualität

Anforderungsbeschreibung	Die Plattform soll Mechanismen zur Einschätzung der Qualität oder Verlässlichkeit einer Antwort unterstützen, beispielsweise Confidence Scores oder funktional gleichwertige Verfahren.
Kategorie	Soll-Anforderung
Umsetzung	Die Bewertung kann auf Grundlage von Retrieval-Ergebnissen, Quellenqualität, Modellbewertungen oder weiteren Qualitätsmetriken erfolgen.
Begründung	Eine Einschätzung der Antwortqualität unterstützt die fachliche Bewertung und erhöht die Transparenz für Nutzende.

KG-009 – Reduktion unbelegter Antworten

Anforderungsbeschreibung	Die Plattform muss Mechanismen bereitstellen, um unbelegte, spekulative oder halluzinierte Antworten zu reduzieren oder entsprechend zu kennzeichnen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Hierzu sind insbesondere RAG, Citation Validator, Guardrails, Policy Engine sowie weitere Qualitätssicherungsmechanismen einzusetzen.
Begründung	Die Reduktion unbelegter Antworten verbessert die Qualität, Vertrauenswürdigkeit und Nachvollziehbarkeit der KI-Verarbeitung.

KG-010 – Kontrollierte Antwortverweigerung

Anforderungsbeschreibung	Bei unzureichender Datenlage, fehlendem Kontext oder unzureichender Antwortsicherheit muss die Plattform in der Lage sein, auf eine Antwort zu verzichten oder die Unsicherheit transparent zu kommunizieren.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Entscheidung erfolgt auf Grundlage definierter Qualitäts-, Governance- und Sicherheitsregeln innerhalb des AI Harness. Die Plattform muss nachvollziehbar dokumentieren, warum keine Antwort generiert wurde.
Begründung	Eine kontrollierte Antwortverweigerung ist einer potenziell fehlerhaften oder unbelegten Antwort vorzuziehen und erhöht die fachliche Verlässlichkeit der Plattform.

10.5.3 Human-in-the-Loop

Zielsetzung

Die Plattform muss Mechanismen zur Einbindung menschlicher Entscheidungen in KI-gestützte Prozesse bereitstellen. Ziel ist die Sicherstellung einer kontrollierten, nachvollziehbaren und verantwortungsvollen Nutzung von KI, insbesondere bei sensiblen, risikobehafteten oder fachlich kritischen Anwendungsfällen.

Governance-Grundsatz

Human-in-the-Loop-Prozesse sind Bestandteil der Governance-Architektur des AI Harness. Die Plattform muss die Einbindung menschlicher Freigaben, Prüfungen und Eskalationen innerhalb automatisierter KI-Prozesse unterstützen.

Die Notwendigkeit einer menschlichen Beteiligung muss administrativ konfigurierbar sein und sich an fachlichen, organisatorischen, sicherheitsrelevanten oder regulatorischen Anforderungen orientieren.

KG-011 – Freigabemechanismen für kritische Anwendungsfälle

Anforderungsbeschreibung	Die Plattform muss Freigabemechanismen für sensible oder besonders risikobehaftete Anwendungsfälle bereitstellen. Für definierte Prozesse muss eine menschliche Freigabe vor der Ausführung oder Veröffentlichung von KI-generierten Ergebnissen unterstützt werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Freigaben erfolgen über konfigurierbare Human-in-the-Loop-Prozesse innerhalb des AI Harness. Die konkrete Ausgestaltung der Freigabeworkflows bleibt dem Auftragnehmer überlassen.
Begründung	Menschliche Freigaben reduzieren Risiken und unterstützen einen verantwortungsvollen Einsatz von KI in sensiblen Anwendungsfällen.

KG-012 – Manuelle Validierung

Anforderungs- beschreibung	KI-generierte Inhalte müssen vor ihrer Weiterverwendung, Veröffentlichung oder automatisierten Weiterverarbeitung durch berechnete Nutzende geprüft, bearbeitet, freigegeben oder abgelehnt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Plattform stellt geeignete Funktionen zur Prüfung und Freigabe innerhalb der Benutzeroberfläche des Urban KI Hub oder einer funktional gleichwertigen Oberfläche bereit. Entscheidungen sind nachvollziehbar zu dokumentieren.
Begründung	Die manuelle Validierung erhöht die fachliche Qualität und reduziert das Risiko fehlerhafter oder unerwünschter KI-Ergebnisse.

KG-013 – Konfigurierbarkeit von Human-in-the-Loop-Prozessen

Anforderungs- beschreibung	Die Notwendigkeit einer menschlichen Freigabe muss abhängig von Anwendungsfall, Agent, Workflow, Modell, Datenklassifizierung oder definierten Governance-Regeln administrativ konfigurierbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Konfiguration erfolgt zentral über den AI Harness unter Berücksichtigung der Policy Engine, Rollen, Berechtigungen sowie organisatorischer Vorgaben.
Begründung	Eine flexible Konfiguration ermöglicht die Anpassung der Freigabeprozesse an unterschiedliche fachliche und regulatorische Anforderungen.

KG-014 – Eskalationsmechanismen

Anforderungs- beschreibung	Die Plattform muss Mechanismen bereitstellen, um KI-Prozesse bei definierten Regeln, Unsicherheiten oder Qualitätsabweichungen an menschliche Entscheidungsträger zu eskalieren.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Eskalationen können insbesondere bei unzureichender Datenlage, widersprüchlichen Retrieval-Ergebnissen, fehlender Antwortsicherheit, Policy-Verstößen oder sicherheitsrelevanten Ereignissen ausgelöst werden. Sämtliche Eskalationen und Entscheidungen sind revisionssicher zu protokollieren.
Begründung	Automatisierte Eskalationen stellen sicher, dass kritische Entscheidungen durch geeignete Personen überprüft werden und erhöhen die Transparenz sowie die Vertrauenswürdigkeit der KI-Verarbeitung.

10.5.4 Auditierbarkeit und KI-Logging (EU AI Act)

Zielsetzung

Die Plattform muss eine revisionssichere und nachvollziehbare Protokollierung sämtlicher wesentlicher KI-gestützter Verarbeitungsvorgänge unterstützen. Ziel ist die Erfüllung von Anforderungen an Transparenz, Auditierbarkeit, Governance und regulatorische Nachweispflichten, insbesondere im Kontext des EU AI Acts.

Governance-Grundsatz

Die Auditierung erfolgt zentral über den AI Harness und ist Bestandteil der Governance-, Monitoring- und Observability-Architektur der Plattform. Sämtliche wesentlichen Verarbeitungsschritte müssen nachvollziehbar dokumentiert und mit den Audit-, Logging- und Tracing-Funktionen verknüpft werden können.

Die Protokollierung muss mandantenfähig erfolgen und den Anforderungen des Datenschutzes sowie den geltenden Aufbewahrungs- und Löschkonzepten entsprechen.

KG-015 – Vollständige Protokollierung von KI-Prozessen

Anforderungs- beschreibung	Die Plattform muss sämtliche wesentlichen KI-gestützten Verarbeitungsvorgänge nachvollziehbar protokollieren. Hierzu gehören insbesondere Modellaufrufe, Retrieval-Prozesse, Agentenausführungen, Tool-Aufrufe, Modell-Routing-Entscheidungen, Human-in-the-Loop-Prozesse sowie sicherheits- und governance-relevante Ereignisse.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Protokollierung erfolgt zentral über den AI Harness und wird mit Logging-, Monitoring- und Tracing-Funktionen verknüpft.
Begründung	Eine vollständige Protokollierung ermöglicht Transparenz, Nachvollziehbarkeit und die Erfüllung regulatorischer Anforderungen.

KG-016 – Inhalt der Protokollierung

Anforderungs- beschreibung	Die Protokollierung muss – soweit technisch möglich und datenschutzrechtlich zulässig – mindestens Eingaben (Prompts oder Anfragen), verwendete Modelle und Modellversionen, relevante Modellparameter, Retrieval-Ergebnisse und Datenquellen, den übergebenen Kontext, erzeugte Antworten, Tool-Aufrufe, Routing-Entscheidungen, Prompt-Versionen sowie die jeweilige Trace-ID der Ausführung umfassen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Inhalte der Protokollierung müssen administrativ konfigurierbar sein und den geltenden Datenschutz- und Aufbewahrungsvorgaben entsprechen.
Begründung	Eine strukturierte Protokollierung bildet die Grundlage für Auditierungen, Fehleranalysen und die Reproduzierbarkeit von KI-Prozessen.

KG-017 – Nachvollziehbarkeit von Entscheidungen

Anforderungs- beschreibung	Die Verarbeitungsschritte müssen so dokumentiert werden, dass technische Abläufe, fachliche Entscheidungen und erzeugte Ergebnisse im Nachhinein nachvollzogen und bewertet werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Nachvollziehbarkeit umfasst insbesondere Modelle und Modellversionen, Prompt-Versionen, Retrieval-Ergebnisse, Wissensquellen, Agentenkonfigurationen, Tool-Aufrufe, Routing-Entscheidungen, Human-in-the-Loop-Prozesse sowie die zugehörige Trace-ID.
Begründung	Eine vollständige Nachvollziehbarkeit unterstützt Qualitätssicherung, Auditierung, Compliance und kontinuierliche Verbesserung der Plattform.

KG-018 – Revisionssichere Speicherung

Anforderungsbeschreibung	Audit- und Protokolldaten müssen gegen unbefugte Änderungen geschützt sowie entsprechend definierter Aufbewahrungs- und Löschfristen revisionssicher gespeichert werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Speicherung erfolgt unter Berücksichtigung gesetzlicher Vorgaben sowie administrativ konfigurierbarer Aufbewahrungsfristen.
Begründung	Eine revisionssichere Speicherung gewährleistet die Nachweisbarkeit von KI-Prozessen und unterstützt Compliance- sowie Prüfanforderungen.

KG-019 – Zugriffsschutz auf Protokolldaten

Anforderungsbeschreibung	Der Zugriff auf Audit- und Protokolldaten muss rollen-, mandanten- und berechtigungsabhängig erfolgen. Der Zugriff auf Protokolle mit potenziell personenbezogenen oder fachlichen Inhalten ist strikt zweckgebunden zu steuern. Die Plattform muss Mechanismen zur Trennung technischer Betriebsprotokolle und fachlicher Inhaltsprotokolle unterstützen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Zugriffskontrolle erfolgt zentral über den AI Harness unter Berücksichtigung von Rollen, Berechtigungen, Datenschutzvorgaben und Governance-Richtlinien. Sämtliche Zugriffe auf Auditdaten sind selbst revisionssicher zu protokollieren.
Begründung	Eine differenzierte Zugriffskontrolle schützt sensible Informationen und gewährleistet Datenschutz sowie Informationssicherheit.

KG-020 – Unterstützung regulatorischer Anforderungen

Anforderungsbeschreibung	Die Architektur muss die Umsetzung und Weiterentwicklung regulatorischer Anforderungen, insbesondere des EU AI Acts sowie vergleichbarer gesetzlicher Vorgaben, unterstützen.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Plattform muss Audit-, Dokumentations- und Governance-Funktionen bereitstellen und so ausgestaltet sein, dass zukünftige regulatorische Anforderungen ohne grundlegende Änderungen an der Gesamtarchitektur umgesetzt werden können.
Begründung	Regulatorische Anforderungen entwickeln sich kontinuierlich weiter. Eine erweiterbare Governance-Architektur gewährleistet die langfristige Compliance und Zukunftssicherheit der Plattform.

11. Schulung und Dokumentation

Zielsetzung

Der Auftragnehmer hat sicherzustellen, dass sämtliche bereitgestellten Komponenten des KI-Stacks nachvollziehbar dokumentiert sowie die für Betrieb, Administration, Konfiguration und Weiterentwicklung erforderlichen Kenntnisse an die Auftraggeberin vermittelt werden.

Ziel ist es, die Auftraggeberin in die Lage zu versetzen, die Plattform nach der Inbetriebnahme eigenständig zu administrieren, zu betreiben und im vorgesehenen Umfang fachlich zu konfigurieren.

Dokumentations- und Schulungsgrundsatz

Die Dokumentation muss vollständig, aktuell und verständlich sein und den gesamten KI-Stack umfassen. Hierzu gehören insbesondere die Architektur, Systemkomponenten, Schnittstellen, Konfigurationen, Betriebsprozesse, Sicherheitsmechanismen sowie Administrations- und Benutzerfunktionen.

Schulungen sind so durchzuführen, dass berechtigte Mitarbeitende der Auftraggeberin die Plattform entsprechend ihrer jeweiligen Rollen sicher und fachgerecht nutzen können.

Die konkrete Ausgestaltung der Dokumentation und der Schulungsmaßnahmen bleibt dem Auftragnehmer überlassen, sofern die nachfolgenden Anforderungen erfüllt werden.

11.1 Administratorhandbuch

Zielsetzung

Der Auftragnehmer muss ein vollständiges Administratorhandbuch in deutscher Sprache bereitstellen. Ziel ist es, Administratoren der Auftraggeberin in die Lage zu versetzen, den KI-Stack eigenständig zu installieren, zu konfigurieren, zu betreiben, zu überwachen und weiterzuentwickeln.

Dokumentationsgrundsatz

Das Administratorhandbuch muss sämtliche administrativen Funktionen des KI-Stacks nachvollziehbar beschreiben und den sicheren sowie wirtschaftlichen Betrieb der Plattform unterstützen.

Die Dokumentation ist aktuell zu halten und an neue Versionen der Plattform anzupassen.

SD-001 – Bereitstellung eines Administratorhandbuchs

Anforderungs- beschreibung	Der Auftragnehmer muss ein vollständiges Administratorhandbuch in deutscher Sprache bereitstellen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Das Administratorhandbuch ist in digitaler Form bereitzustellen und muss während der Vertragslaufzeit aktuell gehalten werden.
Begründung	Eine vollständige Dokumentation ermöglicht den eigenständigen Betrieb und die Administration des KI-Stacks durch die Auftraggeberin.

SD-002 – Inhalte des Administratorhandbuchs

Anforderungs- beschreibung	Das Administratorhandbuch muss mindestens folgende Inhalte umfassen: • Architektur des KI-Stacks einschließlich AI Harness und zentraler Komponenten • Systemvoraussetzungen sowie Betriebs- und Deploymentkonzepte • Installation, Konfiguration und Aktualisierung der Plattform • Administration von Nutzern, Rollen, Mandanten und Berechtigungen • Verwaltung von Sprachmodellen, Modell-Routing und Modellprofilen • Administration von Agenten, Prompt Registry und Prompt-Versionen • Verwaltung von Wissensquellen, Retrieval-Komponenten und Dokumentenverarbeitung • Konfiguration von Policies, Guardrails und Human-in-the-Loop-Prozessen • Integration externer Systeme, APIs, Tools und MCP-Server • Monitoring, Logging, Tracing, Auditierung und Fehleranalyse • Token- und Kostenmonitoring • Sicherheits- und Datenschutzkonfigurationen • Backup-, Wiederherstellungs- und Wartungsverfahren • Verfahren zur Fehlerbehebung und zum Support
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Dokumentation muss nachvollziehbar strukturiert, versioniert und bei Änderungen der Plattform aktualisiert werden.
Begründung	Ein umfassendes Administratorhandbuch unterstützt den sicheren, wirtschaftlichen und eigenständigen Betrieb des KI-Stacks durch die Auftraggeberin.

11.2 API-Dokumentation

Zielsetzung

Der Auftragnehmer muss eine vollständige, aktuelle und verständliche Dokumentation sämtlicher bereitgestellter Schnittstellen bereitstellen. Ziel ist es, der Auftraggeberin sowie berechtigten Dritten die eigenständige Integration des KI-Stacks in bestehende Systeme und Anwendungen zu ermöglichen.

Dokumentationsgrundsatz

Die API-Dokumentation muss sämtliche bereitgestellten Schnittstellen, Integrationsmechanismen und Kommunikationsverfahren vollständig beschreiben. Sie ist versioniert bereitzustellen und bei Änderungen der Plattform fortlaufend zu aktualisieren.

Die Dokumentation muss ohne zusätzliche Unterstützung des Auftragnehmers verständlich und nutzbar sein.

SD-003 – Dokumentation der Schnittstellen

Anforderungs- beschreibung	Sämtliche bereitgestellten APIs, Integrationsschnittstellen sowie ereignisgesteuerten Kommunikationsmechanismen müssen vollständig dokumentiert werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Dokumentation umfasst insbesondere REST-APIs oder funktional gleichwertige Schnittstellen, Authentifizierungsverfahren, Webhooks, Ereignisse, Datenformate, Versionen sowie Fehlercodes und wird während der Vertragslaufzeit aktuell gehalten.
Begründung	Eine vollständige Schnittstellendokumentation ermöglicht die eigenständige Integration und den sicheren Betrieb der Plattform.

SD-004 – Struktur und Verständlichkeit

Anforderungsbeschreibung	Die API-Dokumentation muss so strukturiert und verständlich sein, dass Integrationen durch Drittsysteme ohne zusätzliche Unterstützung des Auftragnehmers umgesetzt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Dokumentation ist vorzugsweise über einen offenen Standard wie OpenAPI oder eine funktional gleichwertige Spezifikation bereitzustellen. Sie muss mindestens Endpunkte, Parameter, Datenmodelle, Authentifizierung, Autorisierung, Versionierung, Beispielanfragen, Beispielfragen sowie Fehlerbehandlungen enthalten.
Begründung	Eine standardisierte und verständliche API-Dokumentation reduziert Integrationsaufwand, erhöht die Interoperabilität und unterstützt die langfristige Wartbarkeit der Plattform.

11.3 Schulung und Wissenstransfer

Zielsetzung

Der Auftragnehmer muss Schulungs- und Wissenstransfermaßnahmen bereitstellen, die Administratoren, fachliche Anwender sowie weitere berechnigte Nutzer in die Lage versetzen, den KI-Stack sicher, effizient und entsprechend ihrer jeweiligen Rollen zu nutzen und zu administrieren.

Schulungsgrundsatz

Die Schulungen müssen auf die unterschiedlichen Zielgruppen abgestimmt sein und sowohl technische als auch fachliche Inhalte vermitteln. Schulungsunterlagen sind in deutscher Sprache bereitzustellen und der Auftraggeberin dauerhaft zur Verfügung zu stellen.

SD-005 – Durchführung von Schulungen

Anforderungsbeschreibung	Der Auftragnehmer muss Schulungen für Administratoren, fachliche Administratoren sowie Endanwender durchführen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Schulungen können als Präsenz-, Online- oder Hybridveranstaltungen durchgeführt werden. Die Schulungsunterlagen sind in digitaler Form bereitzustellen.
Begründung	Zielgruppengerechte Schulungen unterstützen die sichere Einführung und den nachhaltigen Betrieb der Plattform.

SD-006 – Inhalte der Schulungen

Anforderungs- beschreibung	Die Schulungen müssen mindestens folgende Inhalte umfassen: • Nutzung des Urban KI Hub und des KI-Dialogsystems • Erstellung und Nutzung von KI-Agenten • Nutzung und Verwaltung von Wissensquellen und Dokumenten • Prompt-Bibliothek und Prompt Registry • Grundlagen von Retrieval-Augmented Generation (RAG) • Datenschutz, Informationssicherheit und verantwortungsvoller KI-Einsatz • Funktionen zur Quellenanzeige und Nachvollziehbarkeit • Grundlagen der Plattformarchitektur sowie der wichtigsten KI-Komponenten • Rollen- und Berechtigungskonzepte entsprechend der jeweiligen Zielgruppe
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Inhalte sind zielgruppengerecht aufzubereiten und regelmäßig an den Funktionsumfang der Plattform anzupassen.
Begründung	Ein strukturierter Wissenstransfer gewährleistet eine sichere, effiziente und verantwortungsvolle Nutzung des KI-Stacks.

11.4 Onboarding und Benutzerunterstützung

Zielsetzung

Die Plattform muss Funktionen bereitstellen, welche neue Nutzer beim Einstieg unterstützen und den selbstständigen Umgang mit den bereitgestellten KI-Funktionen erleichtern.

Onboarding-Grundsatz

Hilfefunktionen und Dokumentationen müssen leicht zugänglich sein und kontextbezogen innerhalb der Benutzeroberfläche bereitgestellt werden.

SD-007 – Unterstützung beim Einstieg

Anforderungs- beschreibung	Die Plattform muss Funktionen zur Unterstützung neuer Nutzer bereitstellen. Hierzu gehören insbesondere Einführungstexte, geführte Touren, kontextbezogene Hilfsfunktionen oder funktional gleichwertige Mechanismen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Onboarding-Funktionen werden innerhalb des Urban KI Hub oder einer funktional gleichwertigen Benutzeroberfläche bereitgestellt.
Begründung	Ein strukturiertes Onboarding erleichtert den Einstieg und fördert die sichere Nutzung der Plattform.

SD-008 – Zugang zur Dokumentation

Anforderungs- beschreibung	Benutzer- und Administrationsdokumentationen müssen für berechtigte Nutzer jederzeit leicht zugänglich und aktuell verfügbar sein.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Dokumentationen sind digital bereitzustellen und sollen über Such- und Navigationsfunktionen sowie Verlinkungen aus der Benutzeroberfläche erreichbar sein.
Begründung	Ein einfacher Zugang zur Dokumentation unterstützt die eigenständige Nutzung der Plattform und reduziert den Schulungs- und Supportaufwand.

11.5 Digitale Souveränität und Open Source

Zielsetzung

Die Auftraggeberin verfolgt das Ziel, digitale Souveränität, Transparenz, Interoperabilität sowie die langfristige Wartbarkeit und Weiterentwicklungsfähigkeit des KI-Stacks sicherzustellen. Hierzu sollen bevorzugt Open-Source-Komponenten, offene Standards sowie europäische und interoperable Technologien eingesetzt werden. Proprietäre Komponenten und Managed Services sind zulässig, sofern hierdurch keine unverhältnismäßigen Abhängigkeiten entstehen und die Anforderungen an Austauschbarkeit, Portabilität und Interoperabilität erfüllt bleiben.

Grundsatz der digitalen Souveränität

Die Architektur des KI-Stacks ist so zu gestalten, dass Herstellerabhängigkeiten auf das technisch und wirtschaftlich erforderliche Maß begrenzt werden.

Die konkrete Auswahl der eingesetzten Technologien bleibt dem Auftragnehmer überlassen, sofern die Anforderungen an Offenheit, Interoperabilität, Portabilität und Wartbarkeit erfüllt werden.

SD-009 – Einsatz von Open-Source-Komponenten

Anforderungs- beschreibung	Open-Source-Komponenten sowie europäische und offene Technologien sollen bevorzugt eingesetzt werden, sofern diese die funktionalen und nichtfunktionalen Anforderungen erfüllen.
Kategorie	Soll-Anforderung
Umsetzung	Die Auswahl erfolgt unter Berücksichtigung von Funktionalität, Sicherheit, Wartbarkeit, Wirtschaftlichkeit und langfristiger Weiterentwicklungsfähigkeit.
Begründung	Offene Technologien fördern digitale Souveränität, Transparenz und langfristige Unabhängigkeit.

SD-010 – Transparenz eingesetzter Komponenten

Anforderungs- beschreibung	Der Auftragnehmer muss sämtliche eingesetzten Softwarekomponenten dokumentieren und offenlegen. Hierbei ist insbesondere zwischen Open-Source-, proprietären sowie selbst entwickelten Komponenten zu unterscheiden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Dokumentation umfasst mindestens Hersteller oder Projekt, Version, Lizenzmodell, Einsatzbereich und Abhängigkeiten.
Begründung	Eine vollständige Komponentenübersicht unterstützt Betrieb, Wartung, Sicherheitsbewertungen und Compliance.

SD-011 – Einsatz proprietärer Komponenten

Anforderungs- beschreibung	Proprietäre Komponenten und Managed Services dürfen eingesetzt werden, sofern offene Standards, dokumentierte Schnittstellen sowie die Portabilität von Daten und Konfigurationen gewährleistet bleiben.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Proprietäre Komponenten dürfen die Austauschbarkeit wesentlicher Architekturkomponenten nicht unverhältnismäßig einschränken.
Begründung	Der Einsatz proprietärer Technologien kann fachlich sinnvoll sein, darf jedoch nicht zu langfristigen Herstellerabhängigkeiten führen.

SD-012 – Lizenzkonformität

Anforderungs- beschreibung	Die Nutzung sämtlicher Open-Source-Komponenten muss lizenzkonform erfolgen. Lizenzbedingungen sind einzuhalten und nachvollziehbar zu dokumentieren.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Dokumentation der eingesetzten Lizenzen ist Bestandteil der Projektdokumentation.
Begründung	Die Einhaltung von Lizenzbedingungen reduziert rechtliche Risiken und schafft Transparenz.

SD-013 – Begründung proprietärer Komponenten

Anforderungs- beschreibung	Für proprietäre Komponenten, die zentrale Architekturbausteine des KI-Stacks bilden, ist deren Einsatz nachvollziehbar zu begründen. Die Begründung soll insbesondere technische, wirtschaftliche, betriebliche, sicherheitsrelevante oder integrationsbezogene Aspekte sowie Auswirkungen auf Austauschbarkeit und Portabilität berücksichtigen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Begründung ist Bestandteil der technischen Projektdokumentation.
Begründung	Eine transparente Entscheidungsgrundlage unterstützt die Bewertung langfristiger Herstellerabhängigkeiten.

SD-014 – Vermeidung von Vendor Lock-in

Anforderungs- beschreibung	Die Architektur muss so gestaltet sein, dass keine unnötigen oder unverhältnismäßigen Herstellerabhängigkeiten entstehen und zentrale Komponenten austauschbar bleiben.
Kategorie	A – Architekturvorbereitung
Umsetzung	Die Plattform nutzt standardisierte Schnittstellen, offene Datenformate sowie modulare Architekturprinzipien zur Sicherstellung der Austauschbarkeit.
Begründung	Eine herstellerunabhängige Architektur erhöht die Zukunftssicherheit und reduziert langfristige Betriebsrisiken.

SD-015 – Einsatz von Managed Services

Anforderungs- beschreibung	Managed Services dürfen eingesetzt werden, sofern sie auf offenen Standards basieren oder funktional gleichwertige Integrationsmöglichkeiten bieten und keine unverhältnismäßigen proprietären Abhängigkeiten erzeugen.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Auswahl erfolgt unter Berücksichtigung von Portabilität, Interoperabilität und Austauschbarkeit.
Begründung	Managed Services können den Betriebsaufwand reduzieren, ohne die digitale Souveränität wesentlich einzuschränken.

SD-016 – Dokumentationsverfügbarkeit

Anforderungs- beschreibung	Für sämtliche eingesetzten Komponenten muss eine nachvollziehbare technische Dokumentation bereitgestellt werden.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Dokumentation umfasst insbesondere Installation, Konfiguration, Betrieb, Schnittstellen sowie Lizenzinformationen.
Begründung	Eine vollständige Dokumentation erleichtert Betrieb, Wartung und Weiterentwicklung der Plattform.

SD-017 – Quellcode und Repository

Anforderungs- beschreibung	Der im Rahmen des Auftrags individuell entwickelte oder angepasste Quellcode ist versioniert in einem für die Auftraggeberin zugänglichen Repository bereitzustellen. Soweit rechtlich und technisch zulässig, soll die Entwicklung nach anerkannten Standards der kollaborativen Softwareentwicklung erfolgen. Sofern keine rechtlichen, sicherheitsrelevanten oder lizenzrechtlichen Gründe entgegenstehen, sollen entwickelte Komponenten nach dem Grundsatz „Public Money, Public Code“ für eine Nachnutzung durch andere öffentliche Stellen bereitgestellt werden können.
Kategorie	P1 – Bestandteil der produktiven Erstimplementierung
Umsetzung	Die Versionsverwaltung erfolgt in einem geeigneten Repository. Änderungen sind nachvollziehbar zu dokumentieren.
Begründung	Eine transparente Quellcodeverwaltung unterstützt Wartbarkeit, Wissenstransfer und die langfristige Weiterentwicklung der Plattform sowie – soweit möglich – die Nachnutzung innerhalb der öffentlichen Verwaltung.

12. Support & Service Level (SLA)

Zielsetzung

Der Auftragnehmer hat einen zuverlässigen Support sowie definierte Service Levels für den Betrieb des KI-Stacks bereitzustellen. Ziel ist die Sicherstellung eines stabilen, sicheren und nachvollziehbaren Betriebs sowie eine zeitnahe Bearbeitung von Störungen, Serviceanfragen und sicherheitsrelevanten Ereignissen.

Support- und Servicegrundsatz

Der Support umfasst sämtliche produktiv bereitgestellten Komponenten des KI-Stacks. Hierzu gehören insbesondere der AI Harness, die Agent Runtime, Retrieval-Komponenten, Dokumentenverarbeitung, Sprachmodelle, ML-Services, Schnittstellen sowie weitere bereitgestellte Plattformdienste.

Störungen, Serviceanfragen und sicherheitsrelevante Ereignisse müssen nach definierten Prioritäten bearbeitet werden. Die Bearbeitung muss nachvollziehbar dokumentiert und für die Auftraggeberin transparent sein.

Die konkrete Organisation des Supports bleibt dem Auftragnehmer überlassen, sofern die nachfolgenden Serviceanforderungen eingehalten werden.

12.1 Supportorganisation

Zielsetzung

Der Auftragnehmer muss eine zentrale Supportorganisation für den Betrieb des KI-Stacks bereitstellen. Ziel ist die strukturierte, transparente und nachvollziehbare Bearbeitung von Supportanfragen, Störungen und Serviceanforderungen während der Vertragslaufzeit.

Supportgrundsatz

Die Supportorganisation muss als zentrale Anlaufstelle für sämtliche produktiv betriebenen Komponenten des KI-Stacks dienen. Supportanfragen sind nachvollziehbar zu erfassen, zu klassifizieren, zu priorisieren und bis zum Abschluss zu dokumentieren.

Die konkrete organisatorische Ausgestaltung des Supports bleibt dem Auftragnehmer überlassen, sofern die nachfolgenden Anforderungen erfüllt werden.

SLA-001 – Zentraler Ansprechpartner

Anforderungs- beschreibung	Der Auftragnehmer muss einen zentralen Ansprechpartner für Supportanfragen sowie die Koordination der Störungsbearbeitung bereitstellen.
Kategorie	P1 – Bestandteil der produktiven Betriebsphase
Umsetzung	Der Ansprechpartner dient als zentrale Kommunikationsschnittstelle für die Auftraggeberin und koordiniert bei Bedarf die Einbindung weiterer Fach- oder Supportebenen.
Begründung	Ein zentraler Ansprechpartner gewährleistet klare Kommunikationswege und eine effiziente Bearbeitung von Supportfällen.

SLA-002 – Supportkanäle

Anforderungs- beschreibung	Supportanfragen müssen mindestens über ein Ticketsystem und per E-Mail entgegengenommen werden können. Weitere Kommunikationskanäle können ergänzend bereitgestellt werden.
Kategorie	P1 – Bestandteil der produktiven Betriebsphase
Umsetzung	Sämtliche Supportkanäle müssen nachvollziehbar dokumentierte Supportprozesse unterstützen.
Begründung	Standardisierte Supportkanäle ermöglichen eine strukturierte und transparente Kommunikation zwischen Auftraggeberin und Auftragnehmer.

SLA-003 – Erfassung und Nachverfolgung von Supportfällen

Anforderungs- beschreibung	Alle Supportfälle müssen in einem Ticketsystem erfasst, klassifiziert, priorisiert und bis zum Abschluss nachvollziehbar dokumentiert werden. Der Bearbeitungsstatus muss für die Auftraggeberin jederzeit einsehbar oder auf Anfrage nachvollziehbar sein.
Kategorie	P1 – Bestandteil der produktiven Betriebsphase
Umsetzung	Das Ticketsystem muss mindestens die Erfassung von Priorität, Bearbeitungsstatus, Kommunikationsverlauf, Maßnahmen, Eskalationen sowie Abschlussinformationen unterstützen.
Begründung	Eine vollständige Dokumentation der Supportfälle erhöht die Transparenz, unterstützt die Einhaltung vereinbarter Service Levels und ermöglicht eine kontinuierliche Verbesserung des Supportprozesses.

12.2 Supportzeiten und Priorisierung

Zielsetzung

Der Auftragnehmer muss definierte Supportzeiten sowie ein transparentes Priorisierungskonzept für Supportanfragen und Störungen bereitstellen. Ziel ist eine schnelle, nachvollziehbare und risikoorientierte Bearbeitung von Incidents entsprechend ihrer fachlichen und technischen Auswirkungen.

Supportgrundsatz

Supportanfragen und Störungen sind nach einheitlichen Prioritätsklassen zu bewerten und entsprechend ihrer Kritikalität zu bearbeiten. Die Priorisierung erfolgt auf Grundlage der Auswirkungen auf den Betrieb des KI-Stacks und der betroffenen Geschäftsprozesse.

Die vereinbarten Service Levels beziehen sich auf die maximale Zeit bis zur ersten qualifizierten Reaktion. Weitergehende Lösungs- oder Wiederherstellungszeiten können vertraglich vereinbart werden.

SLA-004 – Supportzeiten

Anforderungs- beschreibung	Der Auftragnehmer muss mindestens werktags in der Zeit von 08:00 Uhr bis 17:00 Uhr Supportleistungen bereitstellen. Erweiterte Supportzeiten oder Rufbereitschaften können ergänzend vereinbart werden.
Kategorie	P1 – Bestandteil der produktiven Betriebsphase
Umsetzung	Die Supportzeiten sind Bestandteil des Servicekonzepts und gelten für die Annahme sowie Bearbeitung von Supportanfragen.
Begründung	Definierte Supportzeiten gewährleisten einen planbaren und verlässlichen Betrieb der Plattform.

SLA-005 – Priorisierung von Supportfällen

Anforderungs- beschreibung	Supportfälle und Störungen müssen anhand definierter Prioritätsklassen bewertet und entsprechend ihrer Auswirkungen auf den Betrieb klassifiziert werden.
Kategorie	P1 – Bestandteil der produktiven Betriebsphase
Umsetzung	Die Priorisierung erfolgt nach standardisierten Kriterien und wird im Ticketsystem dokumentiert.
Begründung	Eine einheitliche Priorisierung unterstützt eine risikoorientierte und effiziente Bearbeitung von Störungen.

SLA-006 – Prioritätsklassen

Anforderungs- beschreibung	Die Plattform muss mindestens folgende Prioritätsklassen unterstützen: P1 – Kritisch: Vollständiger Ausfall produktiver Kernfunktionen oder erhebliche Beeinträchtigung des Betriebs. P2 – Hoch: Wesentliche Einschränkung zentraler Funktionen mit erheblichen Auswirkungen auf den Betrieb. P3 – Mittel: Einschränkung einzelner Funktionen ohne vollständigen Betriebsausfall. P4 – Niedrig: Geringfügige Einschränkungen oder allgemeine Service- und Supportanfragen ohne wesentliche Auswirkungen auf den Betrieb.
Kategorie	P1 – Bestandteil der produktiven Betriebsphase
Umsetzung	Die Einstufung erfolgt im Rahmen der Ticketbearbeitung und kann bei geänderten Rahmenbedingungen angepasst werden.
Begründung	Einheitliche Prioritätsklassen schaffen Transparenz und bilden die Grundlage für die Einhaltung definierter Service Levels.

SLA-007 – Reaktionszeiten

Anforderungs- beschreibung	Für Supportfälle gelten folgende maximale Erstreaktionszeiten: • P1: ≤ 2 Stunden • P2: ≤ 4 Stunden • P3: ≤ 1 Werktag • P4: ≤ 2 Werktage
Kategorie	P1 – Bestandteil der produktiven Betriebsphase
Umsetzung	Die Reaktionszeit beginnt mit der ordnungsgemäßen Erfassung des Supportfalls im Ticketsystem und endet mit der ersten qualifizierten Rückmeldung an die Auftraggeberin. Abweichende Wiederherstellungs- oder Lösungszeiten können gesondert vereinbart werden.
Begründung	Verbindliche Reaktionszeiten schaffen Planungssicherheit und gewährleisten eine zeitnahe Bearbeitung von Störungen entsprechend ihrer Priorität.

12.4 Wartung und Betrieb

Zielsetzung

Der Auftragnehmer muss einen geordneten und sicheren Betrieb des KI-Stacks gewährleisten. Hierzu gehören insbesondere planbare Wartungsarbeiten, die zeitnahe Bereitstellung sicherheitsrelevanter Aktualisierungen sowie Maßnahmen zur Aufrechterhaltung der Betriebsfähigkeit.

Betriebsgrundsatz

Wartungs- und Betriebsmaßnahmen sind so durchzuführen, dass Beeinträchtigungen des produktiven Betriebs auf das notwendige Maß begrenzt werden. Sicherheitsrelevante Aktualisierungen sind zeitnah umzusetzen und die Auftraggeberin über geplante Wartungsmaßnahmen angemessen zu informieren.

SLA-008 – Wartungsarbeiten

Anforderungsbeschreibung	Geplante Wartungsarbeiten müssen der Auftraggeberin rechtzeitig angekündigt und nach Möglichkeit außerhalb der definierten Kernarbeitszeiten durchgeführt werden.
Kategorie	P1 – Bestandteil der produktiven Betriebsphase
Umsetzung	Wartungsmaßnahmen sind unter Angabe von Zeitraum, Auswirkungen und gegebenenfalls erforderlichen Einschränkungen zu kommunizieren.
Begründung	Eine frühzeitige Information reduziert betriebliche Auswirkungen und ermöglicht eine planbare Durchführung von Wartungsarbeiten.

SLA-009 – Sicherheitsupdates und Betriebsaktualisierungen

Anforderungsbeschreibung	Sicherheitsupdates sowie sicherheitsrelevante Aktualisierungen der eingesetzten Plattformkomponenten müssen zeitnah bewertet und entsprechend ihrer Kritikalität eingespielt werden.
Kategorie	P1 – Bestandteil der produktiven Betriebsphase
Umsetzung	Aktualisierungen erfolgen im Rahmen eines geregelten Patch- und Release-Managements unter Berücksichtigung von Stabilität, Kompatibilität und Sicherheitsanforderungen.
Begründung	Ein strukturiertes Update-Management reduziert Sicherheitsrisiken und gewährleistet einen stabilen Betrieb der Plattform.

12.5 Monitoring und Reporting

Zielsetzung

Der Auftragnehmer muss den Betrieb des KI-Stacks kontinuierlich überwachen und der Auftraggeberin transparente Informationen über Verfügbarkeit, Betriebszustand und Servicequalität bereitstellen.

Monitoring-Grundsatz

Monitoring und Reporting umfassen sämtliche produktiven Komponenten des KI-Stacks. Hierzu gehören insbesondere Infrastruktur, AI Harness, Agent Runtime, Retrieval-Komponenten, Dokumentenverarbeitung, Schnittstellen sowie weitere bereitgestellte Plattformdienste.

Die gewonnenen Informationen dienen der frühzeitigen Erkennung von Störungen, der Einhaltung der Service Levels sowie der kontinuierlichen Verbesserung des Betriebs.

SLA-010 – Kontinuierliches Monitoring

Anforderungsbeschreibung	Die Plattform muss kontinuierlich hinsichtlich Verfügbarkeit, Performance, Betriebszustand und wesentlicher Systemkomponenten überwacht werden.
Kategorie	P1 – Bestandteil der produktiven Betriebsphase
Umsetzung	Das Monitoring umfasst insbesondere Infrastruktur, KI-Komponenten, Schnittstellen, Agenten, Retrieval-Prozesse sowie Dokumentenverarbeitung und ist über zentrale Monitoring- und Observability-Funktionen bereitzustellen.
Begründung	Ein kontinuierliches Monitoring ermöglicht die frühzeitige Erkennung von Betriebsstörungen und unterstützt einen stabilen Plattformbetrieb.

SLA-011 – Erkennung kritischer Störungen

Anforderungsbeschreibung	Kritische Störungen und sicherheitsrelevante Ereignisse müssen automatisiert erkannt, protokolliert und entsprechend definierter Eskalationsprozesse gemeldet werden können.
Kategorie	P1 – Bestandteil der produktiven Betriebsphase
Umsetzung	Die Erkennung erfolgt über Monitoring-, Logging- und Alarmierungsmechanismen. Kritische Ereignisse sind entsprechend ihrer Priorität an die zuständigen Stellen zu eskalieren.
Begründung	Eine automatisierte Störungserkennung verkürzt Reaktionszeiten und erhöht die Verfügbarkeit der Plattform.

SLA-012 – Service Reporting

SLA-012	Service Reporting
Anforderungsbeschreibung	Der Auftragnehmer muss der Auftraggeberin regelmäßige Berichte über den Betrieb der Plattform bereitstellen. Die Berichte müssen mindestens Informationen zu Verfügbarkeit, Incidents, Service Levels, Reaktionszeiten sowie wesentlichen Betriebs- und Wartungsereignissen enthalten.
Kategorie	P1 – Bestandteil der produktiven Betriebsphase
Umsetzung	Umfang, Format und Bereitstellungsintervall der Berichte werden zwischen Auftraggeberin und Auftragnehmer abgestimmt. Die Berichte sollen digital bereitgestellt und historisch nachvollziehbar archiviert werden.
Begründung	Regelmäßige Serviceberichte schaffen Transparenz über den Betrieb der Plattform und unterstützen die kontinuierliche Bewertung der Servicequalität sowie die Einhaltung vereinbarter Service Levels.

13 Abnahme

Zielsetzung

Vor der produktiven Inbetriebnahme ist eine formale Abnahme der im Rahmen dieses Auftrags erbrachten Leistungen durch die Auftraggeberin durchzuführen. Ziel der Abnahme ist der Nachweis, dass die vereinbarten funktionalen und nichtfunktionalen Anforderungen erfüllt sind und der KI-Stack betriebsbereit ist.

Abnahmegrundsatz

Die Abnahme bezieht sich ausschließlich auf die im Rahmen dieses Auftrags neu entwickelten oder angepassten Komponenten des KI-Stacks. Bereits bestehende Systeme oder Komponenten der Auftraggeberin sind nicht Bestandteil der Abnahme, soweit sie nicht Gegenstand der vertraglich vereinbarten Leistungen sind.

Der Auftragnehmer hat geeignete Nachweise bereitzustellen, dass die gelieferten Komponenten entsprechend den Anforderungen dieser Leistungsbeschreibung implementiert, getestet und dokumentiert wurden.

Für KI-spezifische Funktionen sind nachvollziehbare Evaluierungen und Testverfahren bereitzustellen. Hierzu gehören insbesondere die Bewertung von Retrieval-, RAG- und Agentenfunktionen sowie weiterer KI-Komponenten, soweit diese Bestandteil der beauftragten Lösung sind.

Die konkreten Testverfahren, Referenzszenarien und Nachweise bleiben dem Auftragnehmer überlassen, sofern sie geeignet sind, die Erfüllung der Anforderungen nachvollziehbar nachzuweisen.

Abnahme von KI-Komponenten

Für die Abnahme der Retrieval-, RAG-, Agenten- und weiteren KI-Funktionen hat der Auftragnehmer geeignete Referenzdatensätze, Referenzfragen sowie nachvollziehbare Evaluierungsergebnisse bereitzustellen.

Der Auftragnehmer muss nachweisen, dass die implementierten KI-Komponenten anhand definierter Testszenarien reproduzierbar evaluiert werden können.

Die Nachweise sollen insbesondere folgende Aspekte umfassen:

- Qualität der Retrieval-Ergebnisse
- Relevanz der bereitgestellten Kontexte
- Nachvollziehbarkeit der Quellenverwendung
- Antworttreue gegenüber den verwendeten Quellen (Faithfulness)
- Qualität von Agentenabläufen
- Qualität von Tool-Aufrufen
- Funktionsfähigkeit des Modell-Routings (soweit Bestandteil der Lösung)
- Dokumentation der durchgeführten Evaluierungen

Die konkreten Bewertungsverfahren, Referenzszenarien und Testdatensätze sind durch den Auftragnehmer nachvollziehbar zu dokumentieren und im Rahmen der Abnahme vorzulegen.

Die Evaluierung der Retrieval-Komponenten soll anhand geeigneter Qualitätsmetriken erfolgen. Hierzu können insbesondere Precision, Recall, Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG) oder funktional gleichwertige Verfahren eingesetzt werden. Für RAG- und Agentensysteme sollen zusätzlich geeignete Bewertungsverfahren zur Beurteilung von Antworttreue (Faithfulness), Kontextnutzung, Quellenqualität, Agentenverhalten, Tool-Nutzung oder funktional gleichwertigen Qualitätsmerkmalen eingesetzt werden.

Die vorgelegten Evaluierungsergebnisse müssen auf den bereitgestellten Referenzdatensätzen und Referenzfragen reproduzierbar nachvollzogen werden können.

Die Evaluierungsergebnisse dienen dem Nachweis der Funktionsfähigkeit und Qualität der implementierten KI-Komponenten. Die Auftraggeberin behält sich vor, die bereitgestellten Referenzszenarien und Evaluierungsergebnisse im Rahmen der Abnahme stichprobenartig nachzuvollziehen.

13.1 Funktionale Tests

Zielsetzung

Vor der produktiven Inbetriebnahme muss nachgewiesen werden, dass sämtliche im Rahmen des Auftrags entwickelten oder angepassten Komponenten die vereinbarten funktionalen Anforderungen erfüllen.

Testgrundsatz

Die funktionalen Tests müssen nachvollziehbar dokumentiert werden und sämtliche wesentlichen Funktionen des KI-Stacks umfassen. Die Testergebnisse sind der Auftraggeberin im Rahmen der Abnahme bereitzustellen.

AB-001 – Funktionale Prüfung

Anforderungs- beschreibung	Im Rahmen der Abnahme ist nachzuweisen, dass alle vereinbarten funktionalen Anforderungen der neu entwickelten oder angepassten Komponenten vollständig umgesetzt wurden.
Kategorie	P1 – Bestandteil der Abnahme
Umsetzung	Die Prüfung umfasst insbesondere die Funktionen des AI Harness, der Agent Runtime, der Retrieval-Komponenten, der Dokumentenverarbeitung, der Modellanbindung, der Schnittstellen, der Sicherheits- und Governance-Funktionen sowie weiterer beauftragter Komponenten. Die Testergebnisse sind nachvollziehbar zu dokumentieren.
Begründung	Die funktionale Prüfung stellt sicher, dass die beauftragten Leistungen vollständig und vertragsgemäß umgesetzt wurden.

13.2 Last- und Performancetests

Zielsetzung

Vor der produktiven Inbetriebnahme muss die Leistungsfähigkeit und Stabilität der Plattform unter realistischen Betriebsbedingungen nachgewiesen werden.

Testgrundsatz

Die Last- und Performancetests müssen repräsentative Nutzungsszenarien berücksichtigen und die Einhaltung der vereinbarten nichtfunktionalen Anforderungen nachweisen.

AB-002 – Last- und Performancetests

Anforderungs- beschreibung	Der Auftragnehmer muss durch geeignete Tests nachweisen, dass die Plattform unter realistischer Last stabil betrieben werden kann und die vereinbarten Performanceanforderungen erfüllt.
Kategorie	P1 – Bestandteil der Abnahme
Umsetzung	Die Tests sollen insbesondere gleichzeitige Nutzeranfragen, Modellaufrufe, Retrieval-Prozesse, Agentenausführungen, Dokumentenverarbeitung sowie weitere produktive Nutzungsszenarien berücksichtigen. Die Ergebnisse sind nachvollziehbar zu dokumentieren.
Begründung	Last- und Performancetests belegen die Betriebsfähigkeit und Skalierbarkeit der Plattform unter realistischen Einsatzbedingungen.

13.3 Sicherheitstests

Zielsetzung

Vor der produktiven Inbetriebnahme muss nachgewiesen werden, dass die implementierten Sicherheitsmaßnahmen den vereinbarten Anforderungen entsprechen und die Plattform gegen wesentliche Bedrohungen angemessen geschützt ist.

Testgrundsatz

Die Sicherheitsprüfungen müssen risikoorientiert erfolgen und sowohl klassische IT-Sicherheitsaspekte als auch KI-spezifische Bedrohungen berücksichtigen.

AB-003 – Sicherheitsprüfung

Anforderungs- beschreibung	Vor der Abnahme sind geeignete Sicherheitsprüfungen durchzuführen, um die Einhaltung der vereinbarten Sicherheitsanforderungen nachzuweisen.
Kategorie	P1 – Bestandteil der Abnahme
Umsetzung	Die Sicherheitsprüfung soll insbesondere Authentifizierungs- und Autorisierungsmechanismen, Zugriffskontrollen, Datenschutzmaßnahmen, Schnittstellen, Infrastruktur sowie KI-spezifische Sicherheitsmechanismen – beispielsweise Schutz vor Prompt Injection, unautorisierten Tool-Aufrufen und Datenexfiltration – umfassen. Die Ergebnisse sind nachvollziehbar zu dokumentieren.
Begründung	Sicherheitsprüfungen dienen dem Nachweis eines sicheren und regelkonformen Betriebs der Plattform vor der produktiven Inbetriebnahme.

13.4 Dokumentations- und Betriebsabnahme

Zielsetzung

Vor der produktiven Inbetriebnahme ist nachzuweisen, dass sämtliche vertraglich geschuldeten Dokumentations-, Schulungs- und Betriebsleistungen vollständig erbracht wurden. Ziel ist die Sicherstellung, dass die Auftraggeberin den KI-Stack nach der Abnahme eigenständig betreiben, administrieren und im vorgesehenen Umfang fachlich konfigurieren kann.

9.2 Abnahmegrundsatz

Neben der technischen Funktionsfähigkeit sind auch die Dokumentation, Schulungsunterlagen und der Wissenstransfer Bestandteil der Abnahme.

Sämtliche Unterlagen müssen vollständig, aktuell und nachvollziehbar vorliegen.

AB-004 – Vollständigkeit der Dokumentation

Anforderungs- beschreibung	Der Auftragnehmer muss nachweisen, dass sämtliche im Rahmen des Auftrags geforderten Dokumentationen vollständig, aktuell und in deutscher Sprache bereitgestellt wurden.
Kategorie	P1 – Bestandteil der Abnahme
Umsetzung	Die Abnahme umfasst insbesondere das Administratorhandbuch, die API-Dokumentation, Benutzerdokumentationen, Betriebsdokumentationen sowie weitere vertraglich vereinbarte Unterlagen.
Begründung	Eine vollständige Dokumentation ist Voraussetzung für den eigenständigen Betrieb, die Administration und die Weiterentwicklung des KI-Stacks durch die Auftraggeberin.

AB-005 – Nachweis von Schulung und Wissenstransfer

Anforderungs- beschreibung	Der Auftragnehmer muss nachweisen, dass die vertraglich vereinbarten Schulungs- und Wissenstransfermaßnahmen durchgeführt sowie die entsprechenden Unterlagen bereitgestellt wurden.
Kategorie	P1 – Bestandteil der Abnahme
Umsetzung	Der Nachweis erfolgt beispielsweise durch Schulungsunterlagen, Teilnehmerlisten, Schulungsprotokolle oder vergleichbare Dokumentationen.
Begründung	Ein erfolgreicher Wissenstransfer stellt sicher, dass die Auftraggeberin den KI-Stack nach der Abnahme eigenständig betreiben, administrieren und fachlich weiterentwickeln kann.

13.4 Abnahmekriterien und Dokumentation

Zielsetzung

Die Abnahme setzt voraus, dass neben der technischen Funktionsfähigkeit sämtliche vertraglich geschuldeten Dokumentationsleistungen vollständig erbracht wurden. Ziel ist es, der Auftraggeberin alle für Betrieb, Administration, Nutzung und Weiterentwicklung erforderlichen Unterlagen bereitzustellen.

Abnahmegrundsatz

Die Dokumentation ist Bestandteil der Abnahme. Sämtliche Unterlagen müssen vollständig, aktuell, nachvollziehbar und in deutscher Sprache bereitgestellt werden. Die Dokumentationsanforderungen beziehen sich ausschließlich auf die im Rahmen dieses Auftrags neu entwickelten oder wesentlich erweiterten Komponenten.

AB-004 – Bereitstellung der Dokumentation

Anforderungs- beschreibung	Voraussetzung für die Abnahme ist die vollständige Bereitstellung der vertraglich vereinbarten Dokumentationen für die neu entwickelten oder wesentlich erweiterten Komponenten des KI-Stacks.
Kategorie	P1 – Bestandteil der Abnahme
Umsetzung	Hierzu gehören insbesondere das Administratorhandbuch, die Benutzerdokumentation für den Urban KI Hub, die API- und Entwicklerdokumentation, Betriebsdokumentationen sowie die Beschreibung der Systemarchitektur und weiterer vertraglich vereinbarter Unterlagen.
Begründung	Eine vollständige Dokumentation ist Voraussetzung für den eigenständigen Betrieb, die Administration und die Weiterentwicklung der Plattform durch die Auftraggeberin.

AB-005 – Technische Dokumentation

Anforderungs- Beschreibung	Für die wesentlichen Prozesse und Systembestandteile sind geeignete technische Dokumentationen und Darstellungen bereitzustellen.
Kategorie	P1 – Bestandteil der Abnahme
Umsetzung	Die Dokumentation umfasst insbesondere Architektur- und Systemübersichten, Komponentenbeschreibungen, UML- oder funktional gleichwertige Diagramme, Schnittstellenbeschreibungen sowie die wesentlichen Daten- und Verarbeitungsflüsse.
Begründung	Eine nachvollziehbare technische Dokumentation unterstützt Betrieb, Wartung, Fehleranalyse und Weiterentwicklung der Plattform.

AB-006 – Abgrenzung bestehender Systeme

Anforderungs- beschreibung	Die Dokumentations- und Abnahmeanforderungen gelten ausschließlich für die im Rahmen dieses Auftrags neu entwickelten oder wesentlich erweiterten Komponenten. Bereits bestehende Systeme oder Komponenten der Auftraggeberin sind hiervon ausgenommen, sofern sie nicht Gegenstand der beauftragten Leistungen sind.
Kategorie	P1 – Bestandteil der Abnahme
Umsetzung	Die Abgrenzung der betroffenen Komponenten ist nachvollziehbar zu dokumentieren.
Begründung	Eine eindeutige Abgrenzung schafft Transparenz über den Umfang der Abnahme und verhindert Missverständnisse hinsichtlich bestehender Systeme.

13.5 Abnahmeentscheidung

Zielsetzung

Die formale Abnahme bestätigt, dass die vertraglich vereinbarten Leistungen im Wesentlichen vollständig erbracht wurden und die Plattform für den produktiven Betrieb bereitsteht.

Abnahmegrundsatz

Die Entscheidung über die Abnahme erfolgt auf Grundlage der dokumentierten Prüfergebnisse, Testnachweise sowie der bereitgestellten Dokumentationen. Festgestellte Mängel sind hinsichtlich ihrer Auswirkungen auf den produktiven Betrieb zu bewerten.

AB-007 – Abnahmefreigabe

Anforderungs- beschreibung	Die Abnahme gilt als erfolgt, wenn die vereinbarten Anforderungen erfüllt sind oder lediglich unwesentliche Mängel vorliegen, die den produktiven Betrieb nicht wesentlich beeinträchtigen. Wesentliche Mängel stehen einer Abnahme bis zu ihrer Behebung entgegen.
Kategorie	P1 – Bestandteil der Abnahme
Umsetzung	Die Abnahme ist schriftlich zu dokumentieren. Festgestellte Mängel, vereinbarte Nachbesserungsmaßnahmen sowie gegebenenfalls verbleibende Restpunkte sind im Abnahmeprotokoll festzuhalten.
Begründung	Ein formalisierter Abnahmeprozess schafft Rechtssicherheit für Auftraggeberin und Auftragnehmer und dokumentiert den ordnungsgemäßen Übergang in den produktiven Betrieb.

13.6 Nachweis- und Abnahmeverfahren

Zielsetzung

Die Erfüllung der Anforderungen dieser Leistungsbeschreibung ist im Rahmen der Abnahme durch geeignete und nachvollziehbare Nachweise zu belegen. Ziel ist eine objektive und transparente Bewertung der vertragsgemäßen Leistungserbringung.

Nachweisgrundsatz

Der Auftragnehmer legt für jede Anforderung geeignete Nachweise vor. Art und Umfang der Nachweise richten sich nach der jeweiligen Anforderung und bleiben dem Auftragnehmer überlassen, sofern sie geeignet sind, die Erfüllung der Anforderungen nachvollziehbar zu belegen.

Geeignete Nachweise können insbesondere sein:

- Funktionstests
- Integrations- und Systemtests
- Last- und Performancetests
- Sicherheitsprüfungen
- Demonstrationen
- Evaluierungsergebnisse
- Testprotokolle
- Architektur- und Systemdokumentationen
- API-Spezifikationen
- Betriebs- und Administrationsdokumentationen

Die Auftraggeberin behält sich vor, die vorgelegten Nachweise im Rahmen der Abnahme stichprobenartig zu überprüfen und ergänzende Erläuterungen oder Demonstrationen anzufordern.

14 Exit-Strategie

Zielsetzung

Die Exit-Strategie dient der Sicherstellung digitaler Souveränität, langfristiger Betreiberunabhängigkeit sowie der nachhaltigen Weiterentwicklung des KI-Stacks durch die Auftraggeberin oder durch von ihr beauftragte Dritte.

Ziel ist es, den Wechsel des Betriebsmodells, einzelner Komponenten oder des Auftragnehmers ohne unverhältnismäßigen technischen, organisatorischen oder wirtschaftlichen Aufwand zu ermöglichen.

Grundsatz der Exit-Strategie

Die Architektur des KI-Stacks ist so auszugestalten, dass ein Übergang auf einen anderen Betreiber, einen anderen Infrastrukturanbieter oder einen anderen Auftragnehmer grundsätzlich möglich ist.

Der Auftragnehmer hat die Auftraggeberin beim Übergang in einen definierten Zielzustand zu unterstützen. Hierzu gehören insbesondere die Übergabe der erforderlichen Dokumentationen, Konfigurationen, Datenbestände sowie weiterer für den Betrieb erforderlicher Informationen.

Die Exit-Strategie umfasst insbesondere:

- Übergabe der System- und Betriebsdokumentation
- Übergabe von Konfigurationen und Parametrierungen
- Export von Daten und Metadaten
- Übergabe von Prompt-Konfigurationen und Prompt-Versionen
- Übergabe von Agentenkonfigurationen
- Übergabe von Policy- und Governance-Konfigurationen
- Übergabe von Retrieval- und Suchkonfigurationen
- Übergabe von Schnittstellen- und Integrationskonfigurationen
- Übergabe von Monitoring- und Audit-Konfigurationen
- Übergabe individuell entwickelter Softwarebestandteile entsprechend den vertraglichen Vereinbarungen

Die konkrete technische Umsetzung der Exit-Unterstützung bleibt dem Auftragnehmer überlassen, sofern die Anforderungen an Portabilität, Interoperabilität und Betreiberunabhängigkeit erfüllt werden.

14.1 Datenexport

Zielsetzung

Die Auftraggeberin muss sämtliche für den Betrieb und die Weiterentwicklung des KI-Stacks erforderlichen Daten und Konfigurationen in geeigneter Form exportieren können. Ziel ist die Sicherstellung der Datenportabilität sowie die Unterstützung eines Anbieter- oder Betreiberwechsels.

Exportgrundsatz

Der Datenexport muss vollständig, nachvollziehbar und ohne Verlust wesentlicher Informationen erfolgen. Exportierte Daten müssen einer erneuten Nutzung in einer funktional gleichwertigen Zielumgebung zugänglich sein.

EX-001 – Exportfähigkeit

Anforderungsbeschreibung	Alle im Rahmen des KI-Stacks gespeicherten oder verwalteten Daten und Konfigurationen müssen vollständig exportierbar sein.
Kategorie	P1 – Bestandteil der Exit-Strategie
Umsetzung	Der Export erfolgt über geeignete Exportmechanismen oder dokumentierte Schnittstellen.
Begründung	Eine vollständige Exportfähigkeit ist Voraussetzung für Datenportabilität und Betreiberunabhängigkeit.

EX-002 – Umfang des Datenexports

Anforderungsbeschreibung	Der Export muss insbesondere folgende Daten und Konfigurationen umfassen: • Wissensdatenbanken und Indizes • Dokumente, Metadaten und Embeddings (soweit technisch möglich) • Agenten- und Agentenkonfigurationen • Prompt-Versionen und Prompt-Konfigurationen • Policy- und Governance-Konfigurationen • Retrieval- und Suchkonfigurationen • Schnittstellen- und Integrationskonfigurationen • Monitoring- und Audit-Konfigurationen • Nutzerdaten und Memory-Daten, soweit rechtlich zulässig • Protokoll- und Metadaten
Kategorie	P1 – Bestandteil der Exit-Strategie
Umsetzung	Der Umfang des Exports richtet sich nach den tatsächlich eingesetzten Komponenten der Plattform.
Begründung	Ein vollständiger Export ermöglicht den Weiterbetrieb und die Migration des KI-Stacks in eine andere Betriebsumgebung.

14.2 Offene Formate

Zielsetzung

Exportierte Daten und Konfigurationen müssen unabhängig von proprietären Produkten weiterverarbeitet werden können.

Portabilitätsgrundsatz

Für Export und Übergabe sind offene, dokumentierte und maschinenlesbare Formate zu verwenden.

EX-003 – Verwendung offener Formate

Anforderungs- beschreibung	Der Datenexport muss in offenen, dokumentierten und maschinenlesbaren Formaten erfolgen oder funktional gleichwertige Exportformate unterstützen.
Kategorie	P1 – Bestandteil der Exit-Strategie
Umsetzung	Die verwendeten Formate sind nachvollziehbar zu dokumentieren.
Begründung	Offene Formate fördern Interoperabilität und langfristige Datenportabilität.

EX-004 – Weiterverarbeitbarkeit

Anforderungs- beschreibung	Exportierte Daten und Konfigurationen müssen ohne proprietäre Software oder proprietäre Datenformate weiterverarbeitet oder in andere Systeme übernommen werden können.
Kategorie	P1 – Bestandteil der Exit-Strategie
Umsetzung	Die Weiterverarbeitung muss anhand der bereitgestellten Dokumentation nachvollziehbar möglich sein.
Begründung	Die Weiterverwendbarkeit der Daten reduziert Herstellerabhängigkeiten und erleichtert Migrationen.

14.3 Unterstützung beim Anbieterwechsel

Zielsetzung

Der Auftragnehmer muss die Auftraggeberin beim Wechsel des Betreibers oder Auftragnehmers unterstützen und einen geordneten Übergang ermöglichen.

Übergangsgrundsatz

Der Übergang muss nachvollziehbar geplant und dokumentiert werden. Ziel ist die Fortführung des Betriebs mit möglichst geringer Unterbrechung.

EX-005 – Unterstützung beim Übergang

Anforderungs- beschreibung	Der Auftragnehmer muss einen strukturierten Übergang auf einen neuen Betreiber oder Auftragnehmer unterstützen.
Kategorie	P1 – Bestandteil der Exit-Strategie
Umsetzung	Die Unterstützung umfasst insbesondere Wissenstransfer, Dokumentation sowie die Übergabe der erforderlichen Daten und Konfigurationen.
Begründung	Ein geordneter Übergang reduziert Betriebsrisiken und erleichtert den Betreiberwechsel.

EX-006 – Herausgabe von Betriebsartefakten

Anforderungs- beschreibung	Sämtliche für den Betrieb erforderlichen Artefakte sind der Auftraggeberin bereitzustellen. Hierzu gehören insbesondere Quellcode, Konfigurationsdaten, Dokumentationen, Infrastrukturkonfigurationen sowie weitere individuell entwickelte oder angepasste Bestandteile der Plattform entsprechend den vertraglichen Vereinbarungen.
Kategorie	P1 – Bestandteil der Exit-Strategie
Umsetzung	Die Übergabe erfolgt in dokumentierter und nachvollziehbarer Form.
Begründung	Die Bereitstellung der Betriebsartefakte ermöglicht die Fortführung und Weiterentwicklung der Plattform unabhängig vom bisherigen Auftragnehmer.

14.4 Sicherstellung der Weiterverwendbarkeit

Zielsetzung

Die Plattform ist so auszugestalten, dass ein langfristiger Weiterbetrieb sowie ein Wechsel des Betreibers oder Auftragnehmers ohne unverhältnismäßigen Aufwand möglich sind.

Architekturgrundsatz

Die Architektur muss offene Standards, modulare Komponenten und dokumentierte Schnittstellen nutzen, um Herstellerabhängigkeiten auf das notwendige Maß zu begrenzen.

EX-007 – Vermeidung von Herstellerabhängigkeiten

Anforderungs- beschreibung	Die Plattform muss so gestaltet sein, dass keine unnötigen oder unverhältnismäßigen Herstellerabhängigkeiten entstehen, welche einen Anbieter- oder Betreiberwechsel wesentlich erschweren.
Kategorie	P1 – Bestandteil der Exit-Strategie
Umsetzung	Die Architektur basiert auf modularen Komponenten, offenen Standards und dokumentierten Schnittstellen.
Begründung	Eine herstellerunabhängige Architektur stärkt die digitale Souveränität und erleichtert zukünftige Migrationen.

EX-008 – Fortführbarkeit des Betriebs

Anforderungs- beschreibung	Der Betrieb der Plattform muss auf Grundlage der bereitgestellten Dokumentationen, Daten, Konfigurationen und Betriebsartefakte durch die Auftraggeberin oder einen von ihr beauftragten Dritten fortgeführt werden können.
Kategorie	P1 – Bestandteil der Exit-Strategie
Umsetzung	Die hierfür erforderlichen Unterlagen und Artefakte sind vollständig und nachvollziehbar bereitzustellen.
Begründung	Die Fortführbarkeit des Betriebs gewährleistet langfristige Betreiberunabhängigkeit und unterstützt die nachhaltige Weiterentwicklung des KI-Stacks.

15 Projektsteuerung und Projektmanagement

Zielsetzung

Die nachfolgenden Leistungen definieren die Mindestanforderungen an die Projektsteuerung und das Projektmanagement für die Einführung des KI-Stacks der Stadt Gütersloh.

Ziel ist eine strukturierte, transparente und nachvollziehbare Planung, Umsetzung, Integration, Inbetriebnahme und Abnahme der Plattform sowie die koordinierte Zusammenarbeit zwischen Auftragnehmer, Auftraggeberin und weiteren Projektbeteiligten. Aufgrund der Vergabe der ausgeschriebenen Leistungen in mehreren Fachlosen ist darüber hinaus eine enge und koordinierte Zusammenarbeit zwischen den Auftragnehmern der einzelnen Lose sicherzustellen, um eine reibungslose Umsetzung und Integration der Gesamtlösung zu gewährleisten.

Projektmanagementgrundsatz

Der Auftragnehmer hat während der gesamten Projektlaufzeit eine geeignete Projektorganisation sowie ein nachvollziehbares Projektmanagement sicherzustellen.

Das Projektmanagement umfasst sämtliche organisatorischen, fachlichen und technischen Leistungen zur Planung, Steuerung, Koordination, Dokumentation und Qualitätssicherung des Projekts.

Aufgrund der technischen Komplexität des Vorhabens – insbesondere hinsichtlich AI Harness, Agent Runtime, Retrieval-Architektur, Dokumentenverarbeitung, flexibler Modellanbindung, Agentensystemen, Governance, Datenschutz, Informationssicherheit sowie Integration in den Urban Stack – ist ein strukturiertes, risikoorientiertes und ergebnisorientiertes Projektmanagement erforderlich.

Die konkrete Projektmanagementmethode bleibt dem Auftragnehmer überlassen. Es können agile, hybride oder klassische Vorgehensmodelle (z. B. Scrum, Kanban, PRINCE2, PMI oder funktional gleichwertige Verfahren) eingesetzt werden, sofern diese eine transparente Projektsteuerung und die Erfüllung der Anforderungen dieser Leistungsbeschreibung gewährleisten.

Zusammenarbeit zwischen den Auftragnehmern der einzelnen Lose

Die ausgeschriebenen Leistungen werden in drei Fachlosen vergeben. Aufgrund der fachlichen, organisatorischen und technischen Abhängigkeiten zwischen den Losen ist eine enge Zusammenarbeit der jeweiligen Auftragnehmer erforderlich.

Die Auftragnehmer verpflichten sich, die für die Projektdurchführung erforderliche Abstimmung mit den Auftragnehmern der anderen Lose sicherzustellen. Hierzu gehören insbesondere:

- die regelmäßige Teilnahme an projektbezogenen Abstimmungs- und Koordinationsterminen,
- der zeitnahe Austausch relevanter fachlicher, organisatorischer und technischer Informationen,
- die gemeinsame Abstimmung von Schnittstellen, Datenformaten, Integrationsanforderungen und Abhängigkeiten,
- die kooperative Bearbeitung losübergreifender Fragestellungen,

- die Unterstützung bei der Durchführung von Tests, der Integration sowie der Inbetriebnahme der Gesamtlösung sowie
- die frühzeitige Information der übrigen Projektbeteiligten über Risiken, Änderungen oder Sachverhalte mit Auswirkungen auf andere Lose.

Die Auftragnehmer haben ihre Leistungen so zu erbringen, dass eine reibungslose Integration in die Gesamtarchitektur des KI-Stacks gewährleistet ist. Hierzu ist eine konstruktive, partnerschaftliche und lösungsorientierte Zusammenarbeit mit den Auftragnehmern der weiteren Lose sowie mit der Auftraggeberin sicherzustellen.

Die zur Koordination, Abstimmung und Zusammenarbeit mit den Auftragnehmern der anderen Lose erforderlichen Leistungen sind Bestandteil des Leistungsumfangs. Ein Anspruch auf gesonderte Vergütung hierfür besteht nicht; diese Leistungen sind mit den angebotenen Preisen abgegolten.

16 Zeitplan

Zielsetzung

Der Auftragnehmer hat während der gesamten Projektlaufzeit einen strukturierten Termin- und Meilensteinplan zu erstellen, fortzuschreiben und zu überwachen. Ziel ist die transparente Steuerung sämtlicher Projektphasen von der Projektinitialisierung über die Umsetzung und Integration bis zur Abnahme und Übergabe in den produktiven Betrieb.

Grundsatz der Terminplanung

Nach Zuschlagserteilung erstellt der Auftragnehmer in enger Abstimmung mit der Stadt Gütersloh einen belastbaren Termin- und Meilensteinplan für sämtliche im Projekt vorgesehenen Arbeitspakete und Aktivitäten.

Die Terminplanung muss den Gesamtprojektfortschritt nachvollziehbar abbilden und regelmäßig fortgeschrieben werden. Abhängigkeiten, Risiken und kritische Pfade sind transparent darzustellen. Terminabweichungen sind frühzeitig zu erkennen, zu kommunizieren und durch geeignete Steuerungsmaßnahmen zu adressieren.

Die konkrete Methodik der Terminplanung bleibt dem Auftragnehmer überlassen, sofern eine transparente Projektsteuerung und die Einhaltung der vertraglich vereinbarten Meilensteine gewährleistet werden.

Leistungsumfang

Die Leistungen des Auftragnehmers umfassen insbesondere:

- Schätzung des zeitlichen Aufwands für sämtliche Lose, Arbeitspakete und Projektaktivitäten
- Erstellung eines strukturierten Termin- und Meilensteinplans
- Definition nachvollziehbarer Meilensteine einschließlich der zugehörigen Projektergebnisse
- Kontinuierliche Fortschreibung und Detaillierung der Terminplanung während der gesamten Projektlaufzeit
- Abstimmung der Terminplanung mit der Stadt Gütersloh sowie weiteren Projektbeteiligten
- Identifikation und Dokumentation von Abhängigkeiten und kritischen Pfaden
- Regelmäßige Überprüfung des Projektfortschritts anhand der definierten Meilensteine
- Frühzeitige Identifikation von Terminabweichungen, Risiken und Verzögerungen einschließlich geeigneter Gegenmaßnahmen
- Berücksichtigung der Abhängigkeiten zwischen Architektur, Integration, Dokumentenverarbeitung, Retrieval-Komponenten, Agent Runtime, Modellintegration, Tests, Abnahme, Schulung und Betriebsaufnahme
- Regelmäßige Berichterstattung über den Terminstatus im Rahmen der Projektsteuerung

17 Schlussbestimmungen

Zielsetzung

Die nachfolgenden Bestimmungen dienen der Auslegung dieser Leistungsbeschreibung und legen grundlegende Anforderungen an die Umsetzung der ausgeschriebenen Leistungen fest.

Grundsatz

Die in dieser Leistungsbeschreibung beschriebenen Anforderungen stellen die Mindestanforderungen an die zu erbringenden Leistungen dar. Der Auftragnehmer ist für eine vollständige, funktionsfähige und betriebsbereite Umsetzung der ausgeschriebenen Leistungen verantwortlich.

Soweit in dieser Leistungsbeschreibung konkrete Produkte, Technologien, Standards oder Verfahren beispielhaft genannt werden, sind funktional gleichwertige Lösungen zulässig, sofern diese die beschriebenen Anforderungen mindestens in gleichem Umfang erfüllen.

Die konkrete technische Umsetzung bleibt – soweit nicht ausdrücklich anders geregelt – dem Auftragnehmer überlassen.

Der Auftragnehmer hat alle Leistungen zu erbringen, die für einen ordnungsgemäßen, sicheren und produktionsreifen Betrieb des KI-Stacks erforderlich sind, auch wenn diese im Einzelfall nicht ausdrücklich beschrieben sind, jedoch nach dem Stand der Technik üblicherweise zum Leistungsumfang gehören.

Änderungen oder Erweiterungen während der Projektlaufzeit erfolgen im gegenseitigen Einvernehmen der Vertragsparteien und nach den vertraglich vereinbarten Änderungsprozessen.