

Anlage 5 - Leistungsbeschreibung

KI-Wissensmanagement

Postbeamtenkrankenkasse (PBeaKK)

Inhaltsverzeichnis

1	Gegenstand und Ziel der Beschaffung	4
1.1	Gegenstand der Leistung	4
1.2	Zielsetzung	4
1.3	Zieltermin	4
1.4	Hintergrund & strategische Einordnung	5
1.4.1	Ausgangssituation der PBeaKK	5
1.4.2	Proof-of-Concept	5
2	Leistungsumfang – Funktionale Anforderungen	5
2.1	Chatfunktion (Kernleistung)	5
2.2	Antwortentwürfe für Kundenkommunikation	6
2.3	Bereichsstruktur	6
2.4	Prompt-Vorlagen-Bibliothek	6
2.5	Upload-Funktion für Dokumente	7
2.6	Administrations- und Redaktionsbereich	7
3	Technische Anforderungen & Architektur	8
3.1	Umgebungen	9
3.2	Modellstandard	9
3.3	Monitoring und Audit	9
4	Nicht-funktionale Anforderungen	10
4.1	Performance und Antwortzeiten	10
4.2	Skalierbarkeit	10
4.3	Barrierefreiheit	11
5	Datenschutz & Informationssicherheit	11
5.1	Grundsätze	11
5.2	Human-in-the-Loop (verbindliches Gestaltungsprinzip)	12
5.3	Datenkategorien und Schutzbedarfsabgrenzung	12
5.4	Umgang mit Chatverläufen	12
5.5	EU-KI-Verordnung	12
5.6	US-Hyperscaler und CLOUD Act	12
6	Schnittstellenanforderungen	12
6.1	Verbindlich anzubindende Systeme	12
6.2	Anbindungsarten	13
7	Implementierung, Einführung & Schulung	13
7.1	Projektvorgehen	13
7.2	Abnahme	13
7.3	Schulung, Change-Management und technischer Know-how-Transfer	14
8	Betrieb, Support & Wartung	14

8.1	Supportmodell	14
8.2	Updates und Releases	16
8.3	Notfall und Eskalation	16
8.4	Exit-Szenario, Anbieterwechsel und Datenportabilität	16
9	Anhang 1: Anforderungen an die Barrierefreiheit	17

1 Gegenstand und Ziel der Beschaffung

1.1 Gegenstand der Leistung

Die Postbeamtenkrankenkasse (PBeaKK) ist eine bundesunmittelbare rechtsfähige Körperschaft des öffentlichen Rechts mit Sitz in Stuttgart. Die PBeaKK ist eine Sozialeinrichtung der früheren Deutschen Bundespost. Sie erbringt nach Maßgabe des § 26d Bundesanstalt-Post-Gesetz (BAPostG) und ihrer Satzung für ihre Mitglieder Kranken- und Pflegeversicherungsleistungen. Darüber hinaus führt die PBeaKK im Auftrag der Bundesanstalt für Post und Telekommunikation (BAnst PT) die Beihilfebearbeitung für die aktiven und sich bereits im Ruhestand befindlichen Beamtinnen und Beamte der früheren Deutschen Bundespost und der Bundesanstalt durch, und erbringt nach Maßgabe öffentlich-rechtlicher Geschäftsbesorgungsverträge die Beihilfebearbeitung für weitere Körperschaften, Anstalten und Stiftungen des öffentlichen Rechts. Mit über 1.000 Beschäftigten (Beamtinnen und Beamte sowie Angestellte) sowie knapp 340.000 Versicherten ist sie an insgesamt 18 Standorten in ganz Deutschland vertreten.

Gegenstand dieser Ausschreibung ist die Bereitstellung eines produktionsfähigen, sicheren und dauerhaft betreibbaren internen KI-Wissensassistenten für die Mitarbeitenden der PBeaKK in Form einer Cloudleistung als Software-as-a-Service (SAAS).

Der Wissensassistent ermöglicht Mitarbeitenden, Fragen in natürlicher Sprache zu stellen und quellenbasierte, nachvollziehbare Antworten zu erhalten. Jede Antwort weist die verwendeten Quelldokumente explizit aus; Antworten ohne Quellenbasis dürfen nicht als gesicherte Aussagen formuliert werden bzw. es ist auf die fehlende Quellenbasis hinzuweisen. Darüber hinaus unterstützt er bei der Erstellung von Antwortentwürfen für die Kundenkommunikation.

Der Bieter legt das eingesetzte Retrieval- und Antwortgenerierungsverfahren offen und begründet dessen Eignung.

Der Auftragnehmer (AN) erbringt folgende Hauptleistungen:

- Bereitstellung einer produktionsfähigen, skalierbaren und sicheren Systemarchitektur
- Anbindung priorisierter Datenquellen (Wissensindex)
- Qualitätssicherung
- Schulung der Beschäftigten der PBeaKK

1.2 Zielsetzung

Mitarbeitende der PBeaKK – insbesondere in kundenbezogenen Bereichen (Kundenberatung, Schriftwechsel, Abhilfe- und Widerspruchsbearbeitung) – erhalten mit dem Wissensassistenten einen schnellen, komfortablen und nachvollziehbaren Zugang zu internem und extern freigegebenem Wissen. Wiederkehrende Such-, Zusammenfassungs- und Formulierungsaufgaben werden hierdurch reduziert. Qualität und Einheitlichkeit der Kundenkommunikation werden verbessert.

1.3 Zieltermin

Die produktive Nutzbarkeit des Wissensassistenten für die definierte Pilotnutzerguppe (Mitarbeitende mit Kundenbezug) muss spätestens vier Monate nach Beauftragung erreicht sein. Die vollständige Produktivsetzung für alle berechtigten Mitarbeitenden erfolgt im Anschluss an die Pilotphase.

1.4 Hintergrund & strategische Einordnung

1.4.1 Ausgangssituation der PBeaKK

Innerhalb der PBeaKK ist umfangreiches Fachwissen vorhanden. Dieses Wissen verteilt sich auf unterschiedliche Quellen, Systeme und Organisationsbereiche, das Mitarbeiterportal (MAP) auf Basis von Liferay und die öffentlichen Internetseiten der AG.

Für Mitarbeitende bedeutet dies im Alltag:

- Informationen müssen an mehreren Stellen gesucht werden, Suchprozesse dauern unnötig lange.
- Wissen ist teilweise abhängig von der Erfahrung einzelner Personen.
- Die Beantwortung von Kundenanfragen erfordert häufig eine manuelle Recherche.
- Unterschiedliche Wissensstände können zu uneinheitlichen Antworten gegenüber Kunden führen.
- Neue Mitarbeitende benötigen lange Einarbeitungszeiten.

1.4.2 Proof-of-Concept

Die PBeaKK hat einen internen Wissensassistenten als Minimum Viable Product (MVP) entwickelt. Dieser MVP:

- wird aktuell in einem externen Rechenzentrum betrieben,
- wird primär in der telefonischen Kundenberatung genutzt,
- greift auf eine asynchron bereitgestellte Datenquelle zu,
- ist nicht flächendeckend verfügbar,
- hat wesentliche Einschränkungen: manuelle Datenpflege, kein vollständiges Rollenkonzept, keine durchgängige Systemintegration.

Die grundsätzliche Akzeptanz des MVP bei den Nutzenden ist positiv. Der MVP belegt die prinzipielle Eignung des Ansatzes für die PBeaKK, ist jedoch nicht für einen dauerhaften, flächendeckenden Betrieb ausgelegt, weshalb nun eine externe Lösung in Form einer Cloudleistung gesucht wird.

2 Leistungsumfang – Funktionale Anforderungen

2.1 Chatfunktion (Kernleistung)

Der Wissensassistent stellt eine webbasierte Chatoberfläche bereit, über die Mitarbeitende Fragen in natürlicher Sprache eingeben können und verständliche, quellenbasierte Antworten erhalten.

Verbindliche Mindestanforderungen:

- Eingabe von Fragen / Aufgaben in natürlicher, deutscher Sprache ohne spezifisches Prompt-Training erforderlich
- Quellenbasierte Antworten: Jede fachliche Antwort verweist auf die verwendeten Quellen (Titel, Quelltyp, Datum/Stand, Link)
- Folgefragen im gleichen Gesprächsstrang möglich
- Das System muss so ausgelegt sein, dass fachliche Antworten auf den freigegebenen Wissensquellen basieren. Kann keine ausreichende Grundlage in den verfügbaren Quellen ermittelt werden, muss das System dies transparent anzeigen, anstatt unbelegte fachliche Aussagen zu generieren.
- Kopierfunktion: Antworten sind über einen Button in die Zwischenablage kopierbar

- Feedbackfunktion: Nutzende können jede Antwort als hilfreich oder nicht hilfreich bewerten
- Chatverlauf: Gesprächsverläufe werden je Nutzenden gespeichert und sind für diesen 6 Wochen wieder abrufbar.
- Upload von Dateien (siehe Kapitel 2.5)

2.2 Antwortentwürfe für Kundenkommunikation

Der Wissensassistent unterstützt Mitarbeitende bei der Formulierung von Antworten auf Kundenanfragen.

Verbindliche Mindestanforderungen:

- Erstellung sachlicher und kundenorientierter Antwortentwürfe auf Basis freigegebener Wissensquellen
- Möglichkeiten eines iterativen Chats mit dem KI-Wissensassistenten zur Präzisierung der Nutzerintention, bspw.
 - Auswahl der Tonalität (z. B. formell / freundlich)
 - Möglichkeit zur Kürzung oder ausführlicheren Ausgestaltung
- Klarer Hinweis in der Oberfläche, dass der Entwurf durch Mitarbeitende zu prüfen und freizugeben ist
- Keine automatische Versendung von Antworten

Typischer Anwendungsfall: Eine Kundenanfrage geht z.B. per Telefonanruf ein. Der/die Mitarbeitende fragt den Wissensassistenten über einen Prompt nach Antworten zu dem Sachverhalt. Der Assistent erstellt einen Antwortentwurf. Der/die Mitarbeitende prüft diesen, ergänzt und versendet die Antwort manuell.

2.3 Bereichsstruktur

Die Oberfläche enthält eine konfigurierbare Bereichsstruktur. Bereiche helfen Mitarbeitenden, Fragen thematisch einzugrenzen, und ermöglichen die bereichsspezifische Quellenpriorisierung.

Die konkrete Bereichsstruktur wird in Phase 1 gemeinsam mit Product Owner und Fachbereichen festgelegt.

2.4 Prompt-Vorlagen-Bibliothek

Der Wissensassistent stellt eine Bibliothek vordefinierter Prompt-Vorlagen bereit. Vorlagen sind nach Bereichen gegliedert und ermöglichen Mitarbeitenden ohne KI-Kenntnisse qualitativ hochwertige Ergebnisse.

Pflicht-Vorlagentypen bei Go-live:

Vorlage	Funktion
Kundenantwort erstellen	Formuliert Antwortentwurf auf Basis einer Kundenanfrage
Sachverhalt zusammenfassen	Fasst einen eingefügten Text verständlich zusammen
Regelung erklären	Erklärt eine Regelung einfach für Kunden oder Kollegen
Unterlagen benennen	Listet erforderliche Unterlagen für einen Sachverhalt auf
Antwort prüfen	Prüft einen Entwurf auf Verständlichkeit und Vollständigkeit
Führungskräfte-Zusammenfassung	Fasst ein Thema entscheidungsorientiert zusammen

Prompt-Vorlagen sind durch fachliche Experten ohne Programmierkenntnisse pfleg- und erweiterbar.

2.5 Upload-Funktion für Dokumente

Mitarbeitende können einzelne Dokumente hochladen, um diese im Kontext einer Chatanfrage zu verarbeiten (Analyse, Zusammenfassung, Antwortentwurf auf Basis des Dokuments).

Verbindliche Mindestanforderungen:

- Mindestens 2 Dokumente pro Prompt
- Min. 20 MB pro Upload ohne Performanceeinschränkung
- Unterstützte Formate: PDF (textbasiert; OCR-unterstützte Verarbeitung gescannter Dokumente optional), DOCX, XLSX, PPTX, TXT und CSV
- Upload-Dokumente werden maximal 10 Wochen gespeichert
- Keine automatische Aufnahme in den allgemeinen Wissensbestand
- Nutzende können hochgeladene Dokumente selbst löschen
- Hinweis auf verantwortungsvollen Umgang mit sensiblen Informationen
- Protokollierung jedes Uploads

2.6 Administrations- und Redaktionsbereich

Die Lösung umfasst einen administrierbaren Bereich für berechtigte Rollen (Product Owner, Administrator, Quellenverantwortliche). Dieser ermöglicht ohne Programmierkenntnisse:

- Verwaltung von Datenquellen und Freigabe-Status
- Steuerung von Indexierungsaufträgen
- Pflege von Prompt-Vorlagen
- Auswertung von Nutzerfeedback

- Verwaltung von Rollen und Benutzergruppen; Eine direkte, proprietäre Anbindung an ein lokales Active Directory (AD) der Auftraggeberin (AG) ist aus Sicherheits- und Infrastrukturgründen nicht zulässig
- Verwaltung und Veröffentlichung von Nutzungsrichtlinien einschließlich der Möglichkeit, deren erneute Bestätigung durch Nutzende bei Änderungen oder wesentlichen Funktionserweiterungen einzufordern
- Monitoring über Health Funktion (siehe Kapitel 3.3)

3 Technische Anforderungen & Architektur

Die Lösung ist modular aufgebaut mit klar definierten Schnittstellen zwischen den Komponenten:

Komponente	Funktion
Frontend	Webbasierte Benutzeroberfläche im PBeaKK-Corporate Design. Aktuell besteht keine Notwendigkeit, ein für mobile Endgeräte optimiertes Web-Frontend zur Verfügung zu stellen. Falls der AN eine entsprechende Lösung für mobile Geräte mit IOS zur Verfügung hat, kann diesem im Konzept „Bearbeitungs Oberfläche“ angemerkt werden. Unterstützung von Single Sign-On (SSO) für alle Benutzer, um eine zentrale Anmeldung zu ermöglichen.
Backend	Anwendungslogik, Berechtigungsprüfung, Orchestrierung
Wissensindex	Dokumentensuche und -indexierung, Metadaten, Filterung
Ingestion-Pipeline	Automatisierte Aufbereitung und Indexierung von Quellinhalten
Modell-Gateway	Modellagnostische Schnittstelle zum Sprachmodell
Sprachmodell (LLM)	Antwortgenerierung auf Basis von Kontext und Quelldaten

Modellagnostizität ist Pflicht: Das Sprachmodell ist über das Modell-Gateway so angebunden, dass es ohne Umbau der Anwendungslogik ausgetauscht werden kann. Der AN weist dies durch ein Architekturkonzept nach.

3.1 Umgebungen

Drei Betriebsumgebungen sind verpflichtend bereitzustellen:

Umgebung	Zweck
DEV (Entwicklung)	Technische Weiterentwicklung, keine Echtdaten
TEST ¹	Fachliche Abnahme, Qualitätssicherung, Lasttests
PROD (Produktion)	Regelbetrieb mit Mitarbeitenden

Darüber hinaus muss eine Möglichkeit zur Qualitätssicherung und Validierung von Wissensbeständen, Konfigurationen und Änderungen vor deren Übernahme in die Produktivumgebung vorhanden sein.

Die konkrete Ausgestaltung ist vom Bieter zu beschreiben und zu begründen.

3.2 Modellstandard

Das eingesetzte Sprachmodell erfüllt mindestens:

- Sehr gute deutsche Sprachqualität (grammatikalisch korrekt, fachlich präzise, Verwaltungssprache)
- Greift ausschließlich auf bereitgestellte Kontextinhalte zurück; erfindet keine Informationen ohne Quellenbasis
- Verarbeitet Anfragen einschließlich Quelldokumentauszüge vollständig ohne relevanten Informationsverlust; maximale Eingabelänge sowie Kontextfenster sind im Angebot auszuweisen
- Geringe Halluzinationsrate (Nachweis durch Benchmark oder eigene Evaluation)
- Betrieb in kontrollierter Rechenzentrumsumgebung möglich

Der AN legt dar, wie das System Ressourceneinsatz und Antwortqualität optimiert. Verschiedene Modellarchitekturansätze sind zulässig; das vorgeschlagene Konzept ist hinsichtlich Qualität, Performance und Betriebskosten zu begründen.

3.3 Monitoring und Audit

Der AN hat für die vertragsgegenständliche Lösung ein kontinuierliches, automatisiertes System- und Applikations-Performance-Monitoring (APM) zu implementieren, zu betreiben und aufrechtzuerhalten. Das Monitoring muss Ausfälle, Performance-Engpässe und Sicherheitsrisiken proaktiv erkennen und melden, bevor es zu einer Beeinträchtigung der Endnutzer kommt.

Die Anwendung und alle relevanten Microservices müssen dedizierte, gesicherte Endpunkte (z. B. /health) bereitstellen. Diese müssen den Zustand interner Abhängigkeiten (Datenbanken, Caches, Drittsysteme) in Echtzeit prüfen und standardisierte HTTP-Statuscodes zurückgeben.

¹ Diese Anforderung kann durch eine separate TWIN-Umgebung (Digitaler Wissenszwilling) oder durch eine andere geeignete technische Lösung (z. B. eine Testumgebung mit produktionsnahem Datenbestand) erfüllt werden.

Der AN stellt der AG ein Dashboard oder monatliche automatisierte Reports zur Verfügung, die die Einhaltung der Service Level Agreements (SLAs) auf Basis der Monitoring-Daten transparent nachweisen.

Protokolle (Logs) und Metriken zur Systemgesundheit müssen unter Einhaltung der Datenschutzvorgaben (§ 76 BDSG) gemäß den vertraglichen Vorgaben revisionssicher vorgehalten werden.

Der AN beschreibt seine Health-Function im Rahmen des technischen Konzepts.

Das Backend stellt den Kontext für das Sprachmodell aus den Suchergebnissen des Wissensindex zusammen. Die Kontexterstellung erfolgt nachvollziehbar und protokolliert. Der vollständige Audit-Trail je Anfrage umfasst: (1) pseudonymisierte Nutzer-ID, (2) Zeitstempel, (3) aktiver Bereich, (4) verwendete Quellen (Quelltitel, Chunk-IDs), (5) Modellversion, (6) Antwortzeit. Dieser Trail ist mindestens 60 Tage aufzubewahren, für berechnete Rollen (Informationssicherheit, Datenschutz) lesbar und nicht manipulierbar. Er ermöglicht die nachträgliche Rekonstruktion, auf welcher Grundlage eine spezifische Antwort generiert wurde.

Hinweis: Der Audit-Trail ist nicht Bestandteil der Vorgaben gemäß Punkt 2.3. des Kriterienkatalogs für Cloudleistungen (Anlage 6A).

4 Nicht-funktionale Anforderungen

4.1 Performance und Antwortzeiten

Kennzahl	Zielwert
	24/7 Betrieb
Verfügbarkeit im Regelbetrieb. Verfügbarkeitsziel: $\geq 99,0\%$ im Regelbetrieb (rollierend über 30 Tage, ohne angekündigte Wartungsfenster)	$\geq 99,0\%$ monatlich– entspricht VK 1 der AGB zum EVB-IT Cloud
Median-Antwortzeit (Normallast ≤ 30 gleichzeitige Nutzer) bei mindestens 100 Nutzenden	≤ 10 Sekunden
P95-Antwortzeit (Normallast)	≤ 20 Sekunden
Antwortzeit bei Spitzenlast (>200 gleichzeitige Nutzende)	Zu dokumentieren; kein Abbruch

Nachweis durch Lasttest vor Go-live.

4.2 Skalierbarkeit

- Auslegung für bis zu 1.000 potenzielle Nutzer (nicht alle gleichzeitig aktiv)
- Realistische gleichzeitige Nutzung: ca. 10–20 % (100–200 Nutzer gleichzeitig)
- Spitzenlast in Kernzeiten muss stabil verarbeitet werden
- Skalierung der Inferenz-Kapazität muss ohne vollständige Neuentwicklung möglich sein

4.3 Barrierefreiheit

Die zur Verfügung gestellte Lösung muss gemäß § 4 BGG für Menschen mit Behinderungen in der allgemein üblichen Weise, ohne besondere Erschwernis und grundsätzlich ohne fremde Hilfe auffindbar, zugänglich und im Sinne der DIN EN ISO 9241 nutzbar sein. Dies erfordert gemäß § 3 Abs. 1 BITV 2.0, dass sie wahrnehmbar, bedienbar, verständlich und robust ist. Hierbei ist die Nutzung behinderungsbedingt notwendiger Hilfsmittel zulässig. Barrieren sind lediglich zulässig, wenn der aktuelle Stand der Technik vollumfängliche Barrierefreiheit nicht ermöglicht. Die Barrierefreiheit der IT-Lösung ermöglicht eine Nutzung durch folgende Gruppen:

- Menschen ohne Sehvermögen sowie Menschen mit eingeschränktem Sehvermögen
- Menschen ohne Farbwahrnehmung sowie Menschen mit eingeschränkter Farbwahrnehmung
- Menschen ohne Hörvermögen sowie Menschen mit eingeschränktem Hörvermögen
- Menschen ohne Sprachvermögen
- Menschen mit eingeschränkter Handhabung oder Kraft
- Menschen mit eingeschränkter Reichweite
- Menschen mit Anfälligkeit durch Lichtreize (Photosensibilität)
- Menschen mit kognitiven Einschränkungen

Nähere Ausführungen sind Abschnitt 4 der DIN EN 301 549 V3.2.1 zu entnehmen. Barrierefreiheitsanforderungen gelten auch für Dokumente, die von der IT-Lösung automatisiert erzeugt werden, sowie für Hilfefunktionen, Benutzungshandbücher und Schulungsunterlagen. Die Anforderungen an die Barrierefreiheit an die IT-Lösung müssen über die gesamte Vertragslaufzeit erfüllt werden. Hierin eingeschlossen sind Pflege und Support der IT – Lösung, also insbesondere die Bereitstellung neuer Programmstände im Sinne der EVB-IT Verträge einschließlich Versionswechsel als auch die Störungsbeseitigung im Sinne der EVB-IT Verträge.

Hinweis: Diese Anforderungen gelten für die allgemeine Benutzeroberfläche. Für die Admin-Oberfläche ist eine Umsetzung wünschenswert, aber nicht notwendig.

Für konkrete Anforderungen an die Nutzeroberfläche für die Umsetzung der Barrierefreiheit siehe Kapitel 9.

Der Bieter stellt seine Umsetzung für Benutzer- und Admin-Oberfläche als Teil der Bewertung dar (siehe Zuschlagskriterien).

5 Datenschutz & Informationssicherheit

5.1 Grundsätze

Datenschutz und Informationssicherheit sind von Beginn an in der Systemarchitektur zu berücksichtigen (Privacy by Design, Security by Design). Eine nachträgliche Ergänzung von Sicherheitsmaßnahmen ist nicht ausreichend.

Es gelten die Vorgaben der vertraglichen Regelungen, insbesondere des Kriterienkatalogs zum EVB-IT-Cloud-Vertrag.

Nach Beauftragung des AN ist auf Grundlage der vertraglichen Vereinbarungen und der Leistungsbeschreibung, gemeinsam mit der AG, das IT-Sicherheitskonzept, in welchem Maßnahmen konkretisiert und Ansprechpartner namentlich festgelegt werden, aufeinander abzustimmen. Die AG hat das Recht, die Wirksamkeit der im Sicherheitskonzept beschriebenen Sicherheitsmaßnahmen selbst zu überprüfen

5.2 Human-in-the-Loop (verbindliches Gestaltungsprinzip)

Der Wissensassistent ist in der ersten Ausbaustufe ausschließlich als **Assistenzsystem** konzipiert. Er darf in dieser Ausbaustufe:

- **nicht** automatisch Kundenantworten versenden
- **nicht** eigenständige Leistungsentscheidungen treffen
- **nicht** personenbezogene Einzelfallentscheidungen abschließend bewerten

Alle Systemantworten und Antwortentwürfe werden von Mitarbeitenden geprüft, bevor sie weiterverwendet werden. Die fachliche Verantwortung verbleibt beim Menschen.

5.3 Datenkategorien und Schutzbedarfsabgrenzung

Der Wissensindex enthält in dieser Ausschreibung ausschließlich intern freigegebene Informationen. Es sind **keine personenbezogenen Daten** von der Verarbeitung betroffen. Es wird durch entsprechende Maßnahmen sichergestellt, dass durch den Wissensassistenten auch dann keine Verarbeitung personenbezogener Daten erfolgt, wenn diese entgegen den internen Vorgaben von Beschäftigten der AG (organisatorische Maßnahmen) eingegeben werden. Der AN beschreibt entsprechende Maßnahmen in seinem technischen Konzept.

5.4 Umgang mit Chatverläufen

Aus geeigneten Chatverläufen werden nach Ablauf der persönlichen Vorhaltefrist von 6 Wochen anonymisierte und fachlich verwertbare Wissensdatensätze erzeugt. Nach erfolgter Anonymisierung sind die Chatverläufe bzw. die daraus gewonnenen fachlich geeigneten Inhalte der Wissensbasis zur Verfügung zu stellen. Die Übernahme darf nur erfolgen, soweit die Inhalte für die Wissensbasis fachlich geeignet sind und keine sonstigen rechtlichen oder vertraglichen Gründe entgegenstehen.

5.5 EU-KI-Verordnung

Die angebotene Lösung muss die Anforderungen der EU-KI-Verordnung (Verordnung (EU) 2024/1689 oder KI-VO) vollständig erfüllen. Der AN legt im Angebot die Risikoklassifizierung und die eingehaltenen Transparenzpflichten dar. Hochrisiko-KI-Systeme gemäß KI-VO werden nicht akzeptiert.

5.6 US-Hyperscaler und CLOUD Act

Der Einsatz von Cloud-Diensten oder API-Diensten unter US-Jurisdiktion (CLOUD Act) für die Verarbeitung personenbezogener Daten oder vertraulicher PBeaKK-Inhalte ist ohne ausdrückliche schriftliche Genehmigung des AG unzulässig. Der AN legt im Angebot transparent offen, ob und wenn ja welche Komponenten (LLM-APIs, Embedding-Dienste, Hosting) unter nicht-europäische Jurisdiktion (außerhalb der EU und des EWR) fallen.

6 Schnittstellenanforderungen

Die Daten werden vom AG bestimmt und über die folgenden Systeme bereitgestellt.

6.1 Verbindlich anzubindende Systeme

Folgende Systeme sind Pflichtbestandteil des Angebotsumfangs:

Datenquelle	System	Anbindungsart
MAP – Mitarbeiterportal (Hauptquelle)	Liferay	REST-API (bevorzugt) / Web-Crawler
Öffentliche Internetseiten der PBeaKK (aktuell drei Seiten)	www.[TBD].de	Web-Crawler (vollständig)

Freigabestatus und Veröffentlichungsinformationen aus dem führenden Quellsystem sollen übernommen und berücksichtigt werden.

6.2 Anbindungsarten

Die Ingestion-Pipeline unterstützt mindestens:

- **API-Anbindung** (z. B. Liferay Headless Delivery API)
- **Datei-Crawler** (freigegebene Verzeichnisse, Netzlaufwerke)
- **Web-Crawler** (Internetseite, interne Webangebote)
- **Manueller Import** (Übergangslösung, einzelne Dokumente)
- Perspektivisch: Datenbankconnector und Fachsystem-API (nur lesend, nach gesonderter Prüfung)

7 Implementierung, Einführung & Schulung

7.1 Projektvorgehen

Das Projekt wird in vier Phasen durchgeführt. Alle Phasen sind inhaltlich verzahnt, nicht strikt sequenziell:

Phase	Bezeichnung	Kernergebnisse
Phase 1	Technischer Aufbau	DEV/TEST-Umgebung, Backend, Frontend, SSO, Modell-Gateway, Wissensindex-Aufbau, erste Quellen angebunden, Upload-Funktion
Phase 2	Wissensquellen & Fachlogik	Vollständige Anbindung priorisierter Quellen, Bereichsstruktur, Testfragenkatalog (Datenschutzprüfung I), Qualitätsmessung, Prompt-Vorlagen
Phase 3	Pilot & Abnahme	Pilotbetrieb mit Mitarbeitenden, Lasttest, Datenschutzprüfung II, Sicherheitsprüfung, fachliche/technische/betriebliche Abnahme
Phase 4	Produktivstart & Stabilisierung	Go-live, Monitoring der Systemlast, Fehlerbehebung, Regelbetrieb, Betriebsübergabe, Abschlussbericht

7.2 Abnahme

Fachliche Abnahme – Kriterien:

- MAP/Liferay, priorisierte Dokumente und Internetseiten vollständig angebunden

- Antwortqualitätsziele gemäß Testfragenkatalog erreicht: $\geq 85\%$ hilfreich, $\geq 90\%$ korrekte Quellenangabe, 0 kritische Falschaussagen
- Antwortentwürfe für Kundenkommunikation erzeugbar
- Prompt-Vorlagen verfügbar, Upload-Funktion nutzbar

Technische Abnahme – Kriterien:

- SSO funktioniert oder genehmigtes Fallback aktiv
- Rollenmodell vollständig implementiert
- DEV/TEST/PROD getrennt
- Monitoring und Logging aktiv
- Backup und getesteter Restore gemäß den Vorgaben der vertraglichen Regelungen
- Lasttest ohne kritische Auffälligkeiten abgeschlossen
- Sicherheitsprüfung ohne kritische oder hohe offene Befunde
- Nachweis der Funktionsfähigkeit datenschutzrechtlicher Vorgaben

7.3 Schulung, Change-Management und technischer Know-how-Transfer

Der AN konzipiert und führt zielgruppenspezifische Schulungsformate durch:

Zielgruppe	Format	Inhalte
Standardnutzende	Gruppenformat (Präsenz oder Webinar, bis 20 Personen)	Bedienung, Prompt-Nutzung, Quellenverständnis, Grenzen des Systems
Key-User	Vertiefte Schulung (Präsenz oder Webinar, bis 15 Personen)	Alle Nutzerfunktionen, Feedbackprozess, Qualitätsrückmeldung
Administratoren	Technische Übergabe (Präsenz oder Webinar, max. 5 Personen)	Konfiguration, Monitoring, Updateprozess, Troubleshooting
Product Owner	Dedizierte Einweisung (Präsenz oder Webinar, max. 3 Personen)	Prompt-Bibliothek pflegen, Backlog-Steuerung, Quellenfreigabe

Schulungsunterlagen in deutscher Sprache werden als Liefergegenstand übergeben und müssen ohne gesonderte Lizenz nutzbar und bearbeitbar sein.

8 Betrieb, Support & Wartung

8.1 Supportmodell

Das Supportmodell sieht wie folgt aus:

Stufe	Annahme	Leistung
First Level	Durch PBeaKK intern (Helpdesk)	Entgegennahme, erste Einschätzung, Weiterleitung
Fachliche Analyse	Produktverantwortlicher der AG in Abstimmung mit AN	Quellenprobleme, Prompt-Probleme, Qualitätsfragen
Technische Analyse	Durch AN	Systemfehler, Infrastruktur, Modell-Probleme

Servicezeiten

Der Support steht des AN während der folgenden Zeiten zur Verfügung (Servicezeitraum):

Tage	Uhrzeit	Zusatz:
Montag bis Freitag	07:00 – 18:00	(außer an bundeseinheitlichen Feiertagen)

Reaktionszeiten und Lösungszeiten

Priorität	Reaktionszeit	Wiederherstellung / Workaround	Endgültige Lösung
Schwerwiegende Störung ²			
Komplettausfall, keine Nutzung möglich	≤ 2 h	≤ 1 Arbeitstag	≤ 2 Arbeitstage
zentrale Funktion erheblich beeinträchtigt	≤ 4 h	≤ 1 Arbeitstag	≤ 3 Arbeitstage
erhebliche Störung ³			
einzelne Funktion gestört, Workaround vorhanden	≤ 1 AT	≤ 3 AT	≤ 10 AT
leichte Störung ³			
kleinere Fehler/Supportanfragen	≤ 3 AT	nicht zwingend	≤ 15 AT bzw. nächstes Release

RTO und RPO

- Recovery Time Objective (RTO): ≤ 24 Stunden
- Recovery Point Objective (RPO): ≤ 24 Stunden.

² Gemäß EVB-IT Cloud-AGB

8.2 Updates und Releases

- Angekündigte Updates mindestens 72 Stunden vor Einspielung
- Sicherheitspatches nach Bekanntwerden der Schwachstelle gemäß den Regelungen des Kriterienkataloges für Cloudleistungen, Punkt 1.18 (Anlage 6A) gehandhabt.
- Rollback-Möglichkeit nach jedem Update
- Modellwechsel werden mindestens 6 Wochen im Voraus angekündigt, sind durch den AG zu prüfen / genehmigen und werden vorab getestet

8.3 Notfall und Eskalation

In einem Notfall- und Wiederanlaufkonzept als Teil des technischen Konzeptes beschreibt der Bieter Eskalationswege, Verantwortlichkeiten und Wiederanlaufschritte für die Szenarien: Totalausfall, Teilausfall Modell-Gateway, Teilausfall Wissensindex.

8.4 Exit-Szenario, Anbieterwechsel und Datenportabilität

Bei Vertragsende übergibt der AN innerhalb von 30 Tagen vollständig:

- Vollständigen Wissensindex (Dokumente + Metadaten + Embeddings in offenem Format, mindestens JSON/CSV/Parquet)
- Alle Konfigurationsdaten (Prompts, Quellenstruktur, Rollen- und Rechtestruktur)
- Die innerhalb der letzten 12 Monate entstandenen Chatverläufe (JSON) und die daraus resultierenden, für die Wissensbasis bestimmten Inhalte. Dies gilt unabhängig davon, ob die Chatverläufe zum Zeitpunkt des Exports noch im persönlichen Zugriff (Vorhaltefrist) sind oder bereits gemäß den vorstehenden Anforderungen anonymisiert wurden (siehe Kapitel 5.4)
- Unterstützung bei Migration zu neuem Anbieter (mindestens 3 Monate Übergangssupport)

Proprietäre Formate, die eine Migration zu einem Nachfolger erheblich erschweren, sind zu dokumentieren; Migrationshilfe ist kostenfrei anzubieten. Nach Datenübergabe bestätigt der AN die vollständige Löschung aller AG-Daten auf AN-Systemen schriftlich.

9 Anhang 1: Anforderungen an die Barrierefreiheit

- Die Anwendung muss auch ohne Maus, also ausschließlich mit der Tastatur, zu benutzen sein. Alle Aktionen, die mit der Maus ausführbar sind, müssen auch mit der Tastatur ausführbar sein, mit Ausnahme von bewegungsabhängigen Funktionen wie Handschrift oder Zeichnungen.
- Verschiedene Funktionsbereiche der Anwendung wie Menü, Werkzeugleiste und Statuszeile sowie verschiedene Inhaltsbereiche wie Baumansicht und Detailansicht können durch einen Tastaturbefehl angesprungen werden.
- Anzeigen, die mit der Maus bedienbar sind, können auch mit der Tastatur bedient werden. Anzeigen sind Statusanzeigen, Anweisungen, Meldungen etc., die eine rein informative Funktion haben, ebenso nicht editierbare Daten. Solche Anzeigen können mit der Maus selektierbar sein, um Informationen daraus zu kopieren. Wenn dies der Fall ist, muss die Bedienung auch mit der Tastatur möglich sein.
- Der Tastaturfokus ist deutlich sichtbar, so dass er von Menschen mit leicht eingeschränkter Sehfähigkeit ohne Anstrengung gefunden werden kann.
- Beim Navigieren mit der Tastatur erhalten die Bedienelemente einer Anwendung den Fokus in einer Reihenfolge, die der logischen Struktur oder dem Bedienablauf entspricht. Dies ist in der Regel der Fall, wenn die Bedienelemente beim sequentiellen Navigieren in einer Reihenfolge fokussiert werden, die der sichtbaren Anordnung entspricht,
- wenn beim Aktivieren einer Funktion das Ziel der Aktion den Tastaturfokus erhält,
- wenn ein sich öffnender Dialog den Tastaturfokus erhält,
- wenn beim Schließen eines Dialogs der Fokus zu seiner Ausgangsposition oder zum logisch folgenden Element zurückkehrt.
- Die sinnvolle Reihenfolge muss nicht immer eindeutig sein, beispielsweise kann eine Tabelle sowohl zeilen- als auch spaltenweise gelesen werden.
- Wenn ein Bedienelement fokussiert wird oder eine Eingabe erhält, so löst dies nicht eine Änderung des Kontextes aus, es sei denn der Benutzer erhält einen Hinweis, wie er das Verhalten auf einfache Weise steuern kann.
- Eine Kontextänderung ist eine wesentliche Änderung des Inhalts, oder eine Änderung des Fokus, des Bildausschnitts oder der Bedienmodalität.
- Bei der Bedienung der Anwendung erhält der Benutzer fortlaufend Rückmeldung über alle von ihm vorgenommenen Änderungen, einschließlich der bearbeiteten Inhalte, der vorgenommenen Selektionen, der aktuellen Position des Fokus und dem Status von Bedienelementen und Anzeigen.
- Kontraste müssen gut erkennbar sein.
- Die Schriftgröße muss anpassbar sein.
- Prägnante Beschriftungen mit Alternativtexten - Wenn ein Element Benutzereingaben erwartet, werden Beschriftungen, Anweisungen oder Hilfetexte bereitgestellt, die dem Benutzer ausreichend Information geben, um Fehler zu vermeiden. Darunter versteht man z. B., dass Symbole für Eingabefehler erklärt werden. dass Label für Dialogelemente vorhanden sind und dass erforderliche Eingabeformate erklärt werden. Bedienung mit Zeigegeräten (Maus, Touch).
- Der Benutzer wird benachrichtigt über Ereignisse, die außerhalb seines aktuellen Fokus und ohne seine unmittelbare Veranlassung geschehen, wie etwa das Erscheinen von Systemmeldungen, die automatische Aktualisierung von Inhalten, etc. Der Benutzer muss nicht benachrichtigt werden über Änderungen, die durch seine explizite Aktion geschehen und sich bei sequentieller Abarbeitung nahe unterhalb seiner aktuellen Position befinden.

- Rückmeldungen und Benachrichtigungen müssen für die Benutzerinteraktion relevant sein und dürfen den Benutzer nicht stören.
- Die Software muss Screenreader kompatibel sein.