

Anhang 5A

Technisches Lastenheft
Postbeamtenkrankenkasse (PBeaKK)

Anlage 5A – Technisches Lastenheft

Dieser Anhang fasst die technischen Anforderungen der Leistungsbeschreibung zusammen und bildet eine Grundlage für die Bewertung.

0 Hinweise zur Nutzung dieses Dokuments

Der Bieter kennzeichnet jede Anforderung mit:

- ✓ – vollständig erfüllt
- ~ – teilweise erfüllt (Einschränkungen zwingend erläutern)
- X – nicht erfüllt

Für jede mit ~ oder X gekennzeichnete Anforderung ist eine schriftliche Erläuterung mit Verweis auf das entsprechende Konzeptkapitel des Angebots erforderlich.

1 Systemarchitektur und Allgemeine Anforderungen

Nr.	Anforderung	✓/~/X	Erläuterung / Verweis
1.1	Die Lösung beantwortet Fragen auf Basis indexierter interner Dokumente. Jede Antwort weist die verwendeten Quelldokumente explizit aus. Antworten, die nicht durch Quellinhalte belegt sind, dürfen nicht als gesicherte Aussagen formuliert werden. Der Bieter legt das eingesetzte Retrieval-Verfahren offen und begründet dessen Eignung hinsichtlich Qualität und Nachvollziehbarkeit. Verschiedene Architekturansätze (z. B. vektorbasiertes Retrieval, Graph-basiertes Retrieval, agentic Retrieval oder Kombinationen) sind zulässig.		
1.2	Die Architektur trennt klar zwischen den Komponenten: Benutzeroberfläche (Frontend), Anwendungslogik (Backend), Wissensindex, Ingestion-Pipeline und Modell-Gateway. Jede Komponente ist über definierte Schnittstellen angebunden und einzeln austauschbar.		
1.3	Das eingesetzte Sprachmodell ist über ein Modell-Gateway angebunden. Das Gateway stellt eine standardisierte Schnittstelle bereit, sodass das Modell ohne Umbau der Anwendungslogik ausgetauscht werden kann. Der Bieter weist dies durch ein Architekturkonzept nach.		
1.4	Die Lösung wird in drei getrennten Umgebungen betrieben: Entwicklungsumgebung (DEV), Testumgebung (TEST), Produktionsumgebung (PROD). Die Umgebungen teilen keine gemeinsamen Laufzeitressourcen.		

2 Frontend-Anforderungen

Nr.	Anforderung	✓/I/X	Erläuterung / Verweis
2.1	Die Benutzeroberfläche ist als webbasierte Anwendung umzusetzen, die ohne Installation auf dem Arbeitsplatz-PC über einen aktuellen Browser aufrufbar ist. Unterstützte Browser: mindestens Microsoft Edge (aktuell), Google Chrome (aktuell).		
2.2	Die Oberfläche wird im Corporate Design der PBeaKK umgesetzt. Logo, Primärfarbe und Schriftart werden vom Auftraggeber vorgegeben. Der Bieter integriert diese Vorgaben vollständig.		
2.3	Die Startseite enthält mindestens: kurze Systembeschreibung, Hinweis auf erforderliche fachliche Prüfung von Antworten, Bereichsauswahl, häufig genutzte Prompt-Vorlagen, Zugang zu Chatverlauf und Upload-Funktion sowie Link zum Hilfe-/FAQ-Bereich.		
2.4	Die Chatansicht enthält mindestens: Freitexteingabefeld, Anzeige des aktiven Bereichs, Antwortbereich mit Quellenanzeige, Kopier-Schaltfläche, Feedback-Schaltfläche (positiv/negativ), Möglichkeit zur Folgefrage im selben Gesprächsstrang sowie Anzeige hochgeladener Dokumente.		
2.5	Die Quellenanzeige weist je verwendeter Quelle mindestens aus: Titel der Quelle, Art der Quelle (z. B. MAP, Dokument, Internetseite), Datum bzw. Stand (sofern vorhanden) sowie einen direkt klickbaren Link zur Originalquelle (sofern technisch möglich).		
2.6	Die Benutzeroberfläche enthält eine konfigurierbare Bereichsstruktur. Bereiche werden vom Product Owner gepflegt. Die konkrete Bereichsstruktur (z. B. Leistungen, Pflege, Kundenberatung, Mitgliedschaft) wird in der Initialisierungsphase gemeinsam festgelegt.		
2.7	Die Oberfläche enthält eine Prompt-Vorlagen-Bibliothek. Vorlagen sind nach Bereichen gegliedert und vom Product Owner ohne Programmierkenntnisse pflegbar. Mindestens folgende Vorlagentypen sind vorkonfiguriert bereitzustellen: Kundenantwort erstellen, Sachverhalt zusammenfassen, Regelung erklären, Unterlagen benennen.		
2.8	Die Oberfläche zeigt dem Nutzenden transparent an, wenn keine ausreichende Quellengrundlage für eine Antwort vorliegt. Die Anzeige erfolgt als eindeutiger Hinweistext, nicht als scheinbar fundierte Antwort.		

Nr.	Anforderung	✓/I/X	Erläuterung / Verweis
2.9	Chatverläufe werden je Nutzenden gespeichert und sind aus der Oberfläche heraus abrufbar. Nutzende können einzelne Chats oder den gesamten Verlauf selbständig löschen.		
2.10	Die Oberfläche erfüllt die Anforderungen des § 4 BGG und muss im Sinne der DIN EN ISO 9241 nutzbar sein. Nachzuweisen ist mindestens (angelehnt an § 3 Abs. 1 BITV 2.0.): Tastaturbedienbarkeit aller Funktionen, Screenreader-Kompatibilität (getestet mit NVDA oder JAWS), Kontrastverhältnis $\geq 4,5:1$ für normalen Text, keine ausschließlich farbliche Informationsvermittlung sowie Alternativtexte für funktionale grafische Elemente.		
2.11	Die Oberfläche ist responsiv gestaltet und auf Desktop-Bildschirmen (≥ 1280 px Breite) vollständig nutzbar. Eine Optimierung für Tablet-Auflösungen ist wünschenswert. Mobile-First ist nicht gefordert.		
2.12	Wünschenswert ist, dass der Administrations- und Redaktionsbereich über dieselbe Webanwendung zugänglich und nur für berechtigte Rollen sichtbar. Er ermöglicht die Verwaltung von Datenquellen, Prompts, Benutzerrollen und Indexierungsaufträgen ohne Programmierkenntnisse.		

3 Backend-Anforderungen

Nr.	Anforderung	✓/I/X	Erläuterung / Verweis
3.1	Das Backend prüft bei jeder Anfrage die Berechtigung des anfragenden Nutzenden anhand der zugewiesenen Rolle. Anfragen ohne gültige Authentifizierung werden abgewiesen.		
3.2	Das Backend wählt für jede Anfrage die zugelassenen Datenquellen anhand des aktiven Bereichs und der Nutzerrolle aus. Es werden keine Quellen in die Antwortgenerierung einbezogen, für die der Nutzende keine Berechtigung besitzt.		
3.3	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 berechtigte Rollen		

Nr.	Anforderung	✓/~/X	Erläuterung / Verweis
	(Informationssicherheit, Datenschutz) lesbar und nicht manipulierbar. Er ermöglicht die nachträgliche Rekonstruktion, auf welcher Grundlage eine spezifische Antwort generiert wurde.		
3.4	Das Backend speichert technische Ereignisse (Fehler, Antwortzeiten, Anfragevolumen, Authentifizierungsvorgänge, Administrationszugriffe) in einem Betriebslog. Der Inhalt von Chatgesprächen wird vom technischen Betriebslog getrennt gehalten.		

4 Wissensindex und Suche

Nr.	Anforderung	✓/~/X	Erläuterung / Verweis
4.1	Das System liefert für fachliche Anfragen präzise, thematisch relevante Dokumentenabschnitte und erkennt dabei sowohl wörtliche als auch sinnverwandte Treffer. Der Bieter legt das eingesetzte Suchverfahren dar und weist die erzielte Suchqualität am Testfragenkatalog nach (vgl. 11.1). Die Relevanzgewichtung ist administrativ konfigurierbar. Verschiedene Suchverfahren (z. B. Vektorsuche, Volltextsuche, Late-Interaction-Modelle oder Kombinationen) sind zulässig.		
4.2	Der Wissensindex speichert je Dokument-Chunk mindestens folgende Metadaten: Quellbezeichnung, Quellentyp, Ursprungs-URL oder Dateipfad, Datum der Indexierung, Datum des letzten Quellinhalts (sofern verfügbar), zugeordneter Bereich und Freigabestatus.		
4.3	Der Wissensindex erkennt veraltete oder deaktivierte Inhalte und schließt diese von der Suche aus. Deaktivierung ist über die Administrationsoberfläche je Quelle und je Dokument möglich.		
4.4	Der Index unterstützt mindestens 50.000 Dokumente ohne messbare Performancedegradation. Die tatsächliche Dimensionierung wird in Initialisierungsphase festgelegt.		
4.5	Quellenpriorisierung ist konfigurierbar: Bestimmte Quellen (z. B. aktuelle Richtlinien) können gegenüber anderen (z. B. allgemeinen Webseiteninhalten) in der Suchreihenfolge bevorzugt werden.		

5 Ingestion-Pipeline

Nr.	Anforderung	✓/I-/X	Erläuterung / Verweis
5.1	Die Ingestion-Pipeline verarbeitet mindestens folgende Dokumentformate: PDF (text-basiert), Microsoft Word (.docx), Microsoft Excel (.xlsx), CSV, PPTX, HTML (Webseiten) und Nur-Text (.txt).		
5.2	Für bildbasierte PDFs (Scan-Dokumente) bietet der Bieter eine OCR-Verarbeitung an. Der Einsatz ist konfigurierbar (aktivierbar/deaktivierbar je Quelle).		
5.3	Die Dokumentenaufbereitung für den Wissensindex erhält relevante Sinnzusammenhänge und vermeidet das Zerschneiden zusammengehöriger Inhalte. Der Bieter beschreibt seine Aufbereitungsstrategie je Quellentyp (z. B. nach Absätzen, Sektionen oder semantischen Einheiten) und begründet sie hinsichtlich der erzielbaren Suchqualität. Die Aufbereitungsparameter sind durch den Betreiber anpassbar. Die Strategie ist im Betriebshandbuch dokumentiert.		
5.4	Die Pipeline reichert jeden Chunk mit Metadaten gemäß 4.2 an und erzeugt Embeddings, die in den Wissensindex geschrieben werden.		
5.5	Die Pipeline erkennt Dubletten auf Dokumentenebene (gleiche Datei in verschiedenen Quellen) und handhabt diese konfigurierbar (z. B. nur die aktuellere Version indexieren).		
5.6	Inkrementelle Aktualisierungen werden unterstützt: Nur neu hinzugekommene oder geänderte Dokumente werden erneut verarbeitet. Unveränderliche Dokumente werden nicht erneut indexiert.		
5.7	Fehler bei der Verarbeitung einzelner Dokumente (z. B. beschädigte Datei, Zugriffsfehler) führen nicht zum Abbruch des gesamten Indexierungsauftrags. Fehlgeschlagene Dokumente werden protokolliert und können gezielt erneut verarbeitet werden.		
5.8	Die Ingestion-Pipeline ist erweiterbar: Neue Datenquellen-Typen (Connector-Klassen) können ohne Änderung bestehender Pipeline-Komponenten hinzugefügt werden.		

6 Modell-Gateway

Nr.	Anforderung	✓/I/X	Erläuterung / Verweis
6.1	Das Modell-Gateway stellt eine vollständig dokumentierte, offene Schnittstelle bereit, über die verschiedene Sprachmodelle angebunden werden können. Ein Modellwechsel erfordert keine Änderung der Anwendungslogik oberhalb des Gateways. Die Schnittstellenspezifikation ist als Liefergegenstand zu übergeben. Verschiedene technische Umsetzungsformen der Gateway-Schnittstelle sind zulässig; der Bieter begründet seine Wahl und weist die Austauschbarkeit nach.		
6.2	Das Gateway unterstützt gleichzeitig mindestens zwei unterschiedliche Modelle (z. B. Haupt- und Hilfsmodell). Die Anwendung kann je Aufgabentyp das geeignete Modell ansprechen.		
6.3	Systemprompts sind über das Gateway zentral konfigurierbar und versioniert gespeichert. Änderungen an Systemprompts werden protokolliert.		
6.4	Das Gateway implementiert Sicherheitsregeln, insbesondere: Schutz vor Prompt-Injection-Angriffen durch Eingabeverarbeitung, Output-Filter zum Unterbinden unerwünschter Ausgaben sowie Limitierung von Anfragen je Nutzenden und Zeitfenster (Rate Limiting). Das Output-Filter-System unterbindet mindestens die Ausgabe von: (1) Aufforderungen zu rechtswidrigen Handlungen, (2) diskriminierenden oder beleidigenden Inhalten, (3) Inhalten, die dem DSGVO-Grundsatz der Zweckbindung widersprechen, (4) Antworten, die das System als System ohne KI-Charakter darstellen (KI-VO Art. 50). Der AN beschreibt im Angebot, wie das Filter-System konfiguriert und von berechtigten Rollen angepasst werden kann.		
6.5	Das Gateway protokolliert je Anfrage: Modellversion, Tokenverbrauch (Input/Output), Antwortzeit und Status (Erfolg/Fehler). Die Protokolldaten werden für Monitoring und Kostenauswertung bereitgestellt.		
6.6	Das Gateway unterstützt A/B-Tests zwischen Modellversionen: Anfragen können anteilig an unterschiedliche Modelle oder Modellversionen geleitet werden, um Antwortqualität zu vergleichen.		

7 Sprachmodell-Anforderungen

Nr.	Anforderung	✓/I/X	Erläuterung / Verweis
7.1	Das eingesetzte Sprachmodell erzeugt Antworten in sehr guter deutscher Sprachqualität (Grammatik, Rechtschreibung, Verwaltungssprache). Der Bieter weist dies durch Beispiellantworten auf praxisnahe Testfragen (durch den Bieter zu generieren) nach, die dem Angebot beizulegen sind.		
7.2	Das Modell greift ausschließlich auf bereitgestellte Kontextinhalte zurück und erfindet keine Informationen, die nicht aus den indexierten Quelldokumenten stammen. Es kann relevante Informationen aus dem Kontext extrahieren, korrekt zuordnen und zu einer kohärenten Antwort zusammenführen.		
7.3	Das Modell verarbeitet eine Anfrage einschließlich der zugehörigen Quelldokumentauszüge vollständig, ohne relevante Informationen durch Kontextbegrenzung zu verlieren. Der Bieter weist die maximale verarbeitbare Eingabelänge (in Zeichen oder Wörtern) nach und beschreibt das Verhalten bei Überschreitung dieser Grenze.		
7.4	Das Modell zeigt eine nachweislich geringe Halluzinationsrate in quellengestützten Fragestellungen. Der Bieter legt eine Einschätzung auf Basis bekannter Benchmarks (z. B. RAGAS, TruLens) oder eigener Evaluationsergebnisse vor.		
7.5	Das Modell kann im Rechenzentrum des Auftraggebers oder in einer durch den Auftraggeber kontrollierten Rechenzentrumsumgebung betrieben werden. Modelle, die ausschließlich über externe Cloud-APIs verfügbar sind, erfüllen diese Anforderung nur in Verbindung mit den vertraglich festgelegten Anforderungen (siehe v.a. Anlagen 6 und 6A)		
7.6	Modellversionen sind dokumentiert und nachvollziehbar versioniert. Modellwechsel werden dem Auftraggeber mindestens 6 Wochen im Voraus angekündigt.		
7.7	Der Bieter legt dar, wie das System Ressourceneinsatz und Antwortqualität optimiert – insbesondere wie einfache Klassifikations- und Vorverarbeitungsaufgaben effizient behandelt werden, ohne die Qualität komplexer Fachfragen zu beeinträchtigen. Verschiedene Architekturansätze (ein Modell, Mehrmodell-Konzept, spezialisierte Modelle je Aufgabentyp) sind zulässig und müssen hinsichtlich Qualität, Performance und Betriebskosten begründet werden.		

Nr.	Anforderung	✓/!✗	Erläuterung / Verweis
7.8	Das Modell erzeugt strukturierte Ausgaben (JSON, Markdown) auf Anforderung durch das Backend.		

8 Authentifizierung und Berechtigungen

Nr.	Anforderung	✓/!✗	Erläuterung / Verweis
8.1	Die Berechtigungsprüfung erfolgt serverseitig. Clientseitige Sichtbarkeitsbeschränkungen allein sind nicht ausreichend. Umsetzung von SSO gemäß LB.		
8.2	Das Berechtigungskonzept ist auf Quellenebene erweiterbar: Zukünftige Datenquellen mit erhöhtem Schutzbedarf können so konfiguriert werden, dass nur bestimmte Rollen oder Nutzende Zugriff erhalten, ohne die Gesamtarchitektur zu ändern.		
8.3	Alle administrativen Zugriffe (Konfigurationsänderungen, Rollenzuweisungen, Indexierungsaufträge, Quellenverwaltung) werden mit Zeitstempel und Aktion protokolliert. Hinweis: Eine direkte, proprietäre Anbindung an ein lokales Active Directory (AD) der Auftraggeberin (AG) ist aus Sicherheits- und Infrastrukturgründen nicht zulässig		

9 Upload-Funktion

Nr.	Anforderung	✓/!✗	Erläuterung / Verweis
9.1	Alle Nutzenden mit der Rolle Standardnutzer oder höher können einzelne Dokumente hochladen, die im Kontext einer Chatanfrage verarbeitet werden. Der Upload ist auf die aktive Sitzung oder auf eine konfigurierbare Aufbewahrungsdauer von maximal 10 Wochen begrenzt.		
9.2	Upload-Dokumente werden nicht automatisch in den allgemeinen Wissensbestand aufgenommen. Eine Aufnahme in den Wissensindex erfordert eine gesonderte Freigabeaktion durch eine berechtigte Rolle.		
9.3	Jeder Upload wird im technischen Log protokolliert (Nutzende, Zeitstempel, Dateiname, Dateigröße, Verarbeitungsstatus).		

Nr.	Anforderung	✓/I-/X	Erläuterung / Verweis
9.4	Nutzende können hochgeladene Dokumente vor Ablauf der maximalen Aufbewahrungsfrist eigenständig löschen.		
9.5	Die Oberfläche zeigt beim Upload einen Hinweis, dass personenbezogene oder besonders sensible Informationen verantwortungsbewusst behandelt werden sollen und nicht ohne Prüfung hochgeladen werden dürfen.		
9.6	Unterstützte Uploadformate: PDF (text-basiert), DOCX, XLSX, TXT, PPTX, CSV. Maximale Dateigröße je Upload: 20 MB Mindestens 2 Dokumente pro Prompt		
9.7	Upload-Funktion für dauerhaften, freigegebenen Upload in den Wissensbestand Diese Funktion erfordert: Zuordnung zu einem Bereich, Benennung eines fachlichen Owners, Freigabeprozess durch berechtigte Rolle, Versionierung und Gültigkeitsdatum.		

10 Betrieb und Monitoring

Nr.	Anforderung	✓/I-/X	Erläuterung / Verweis
10.1	Die Lösung ist für einen 24/7-Betrieb ausgelegt. Die fachlich relevanten Kernzeiten sind Montag bis Freitag, 07:00 bis 18:00 Uhr. In diesen Zeiten gelten besondere SLA-Reaktionszeiten (vgl. Leistungsbeschreibung).		
10.2	Verfügbarkeitsziel: $\geq 99,0\%$ im Regelbetrieb (rollierend über 30 Tage, ohne angekündigte Wartungsfenster).		
10.3	Ein aktives Monitoring-System überwacht kontinuierlich: Systemverfügbarkeit (HTTP-Endpunkte), Antwortzeiten (P50, P95, P99), Fehlerrate, Indexierungsstatus, GPU-Auslastung, RAM-Auslastung und Tokenverbrauch. Alerting bei Überschreitung definierter Schwellenwerte ist konfigurierbar.		
10.4	Backup: Vollständige Sicherung des Wissensindex, der Systemkonfiguration, der Prompt-Bibliothek und der Chatverläufe mindestens täglich. Aufbewahrungsdauer: mindestens 30 Tage.		
10.5	Ein Restore-Test wird vor Go-live durchgeführt und dokumentiert. Recovery Time Objective (RTO): ≤ 24 Stunden. Recovery Point Objective (RPO): ≤ 24 Stunden.		

Nr.	Anforderung	✓/I-/X	Erläuterung / Verweis
10.6	Ein Notfall- und Wiederanlaufkonzept ist als Teil des technischen Konzepts zu erstellen. Es beschreibt Eskalationswege, Verantwortlichkeiten und Wiederanlaufschritte für die Szenarien: Totalausfall, Teilausfall Modell-Gateway, Teilausfall Wissensindex.		
10.7	Updates und Patches werden in einem definierten Wartungsfenster außerhalb der Kernzeiten eingespielt. Kritische Sicherheitsupdates werden gemäß Kriterienkatalog für Cloudleistungen nach Bekanntwerden eingespielt. Nach jedem Update besteht die Möglichkeit zum Rollback auf die vorherige Version.		

11 Qualitätssicherung und Abnahmevoraussetzungen

Nr.	Anforderung	✓/I-/X	Erläuterung / Verweis
11.1	Der Auftragnehmer erstellt als Liefergegenstand von Phase 2 einen Testfragenkatalog mit mindestens 50 praxisnahen Fragen. Der Katalog enthält Fragen mit eindeutiger Antwort, Fragen mit widersprüchlichen Quellen, Fragen zu veralteten Informationen sowie Fragen je priorisierter Datenquelle. Für jede Testfrage ist eine Referenzantwort und ein Bewertungskriterium definiert.		
11.2	Die Antwortqualität wird vor Abnahme am Testfragenkatalog gemessen. Mindestanforderungen: ≥ 85 % der Testfragen erhalten die Bewertung „hilfreich“, ≥ 90 % der Antworten weisen korrekte Quellenangaben auf, 0 Antworten enthalten kritische Falschaussagen (Informationen, die fachlich nachweislich falsch und ohne Quellenbasis sind).		
11.3	Vor Go-live ist ein Lasttest durchzuführen, der die gleichzeitige Nutzung durch mindestens 100 Nutzende simuliert. Bei diesem Test müssen Antwortzeiten unter der Anforderung aus 11.4 bleiben und die Fehlerrate unter 1 % liegen. Lasttest-Ergebnis ist Liefergegenstand.		
11.4	Antwortzeit: Median-Antwortzeit ≤ 10 Sekunden bei Normallast (≤ 30 gleichzeitige Anfragen). P95-Antwortzeit ≤ 20 Sekunden bei Normallast.		
11.5	Vor jeder Produktivübernahme neuer Modellversionen oder wesentlicher Prompt-Änderungen ist ein Regressionstest durchzuführen.		
11.6	Eine Sicherheitsprüfung (Penetrationstest oder strukturiertes Security Review) wird vor Go-live		

Nr.	Anforderung	✓/!-/X	Erläuterung / Verweis
	durch den Auftragnehmer oder einen beauftragten Dritten durchgeführt. Ergebnis und Maßnahmenplan sind als Liefergegenstand zu übergeben. Kritische und hohe Schwachstellen sind vor Go-live zu beheben.		