Konzept EMRGPT – Zielarchitektur mit Patienten-RAG, Dokumenten-RAG und GraphRAG (FHIR/SNOMED)

v2.1 – Besprechungsergebnisse 27.02.2026 eingearbeitet: Averbis-Architektur detailliert (Medical Summary + M-KIS), ADT-basierte Archivstrategie, Kostenindikatoren (90 k€ Teststellung, ~200 k€/a Lizenz), Cloud-/Datenschutzrisiken Sachsen, OCR-/KDL-Qualitätsrisiken, Variantenvergleich Kap. 16 aktualisiert, Kostengerüst Kap. 20 konkretisiert (Stand: 2026-02-27)

v2.0 – Umbenennung zu Konzept EMRGPT, Konsolidierung der Projektdokumentation (Stand: 2026-02-27)

v1.6 – Kapitelreihenfolge korrigiert, Inhaltsverzeichnis vervollständigt, SAP IS-H/i.s.h.med KI-Konnektor (Stand: 2026-02-27) Basis: v1 Björn | v1.1: QA-Review, PSP-Zuordnung, Variantenvergleich | v1.2: Dokumentenpipeline, FHIR Conformance, SAP-Berechtigungsmodell, Datenqualität, Preisreferenz | v1.3: Umnummerierung Kap. 2/3/14/15, Umsetzungsfahrplan, RACI, Kommunikationsplan, Anforderungsliste | v1.4: Betriebskonzept (Kap. 18), Incident-Response (Kap. 19), Kostengerüst-Template (Kap. 20), Change-Management (Kap. 21) | v1.4.1: QA-Korrekturen | v1.4.2: Terminologie, Struktur, Redundanzen | v1.5: Europe PMC-Anbindung (Kap. 7.2.1.1), automatisches Evidence-Matching (Kap. 7.2.1.2), SNOMED-Autotagging-Pipeline (Kap. 14.2.2), Snowstorm FHIR-API (Kap. 14.2.1), Use Case 4 (Kap. 3.3.4), FA-13 bis FA-16, UI Evidence-Panel (Kap. 4.2) | v1.5.1: Evidenz- und Studien-Domainservice als vierte Wissensschicht (Kap. 2.8.6), Studiendatenbanken (Kap. 7.2.1.3), LLM-Regelableitung (Kap. 7.2.1.4), Kap. 7.1 drei→vier Domänen, QA-Bereinigung (40+ Befunde: Querverweise, FHIR-Pfade, Terminologie, Grammatik, Markdown) | v1.6: Kapitelreihenfolge korrigiert (Kap. 14/15 nach Kap. 13 verschoben, Kap. 99 ans Ende), Inhaltsverzeichnis vollständig neu erstellt (inkl. Kap. 16–21), SAP IS-H/i.s.h.med KI-Konnektor-Analyse (Kap. 22)

---

0. Management Summary und Entscheidungsbedarf

Ausgangslage

Das Universitätsklinikum Leipzig (UKL) schaltet im **Oktober 2026** das KIS **SAP i.s.h.med** ab. Danach steht nur noch **M-KIS (Meierhofer)** als führendes System zur Verfügung. Die medizinischen Altdaten (ca. 21 Mio. Dokumente) müssen rechtssicher archiviert und dem klinischen Personal effizient zugänglich gemacht werden.

Entscheidungsbedarf

Der Lenkungsausschuss wird gebeten, über folgende Punkte zu entscheiden:

1. **Umsetzungsvariante:** Eigenlösung UKLGPT (KI-gestützter Informationsassistent mit RAG-Architektur) vs. integrierte Meierhofer/Averbis-Lösung (Medical Summary im M-KIS) – siehe Variantenvergleich in Kapitel 16.
2. **Freigabe der Ressourcen** für das Hauptprojekt (Projektleiter, Budget, Infrastruktur).
3. **Installation eines Projektleiters** mit Mandat für PM-Exzellenz im Umsetzungsprojekt.

Empfehlung

Die Eigenlösung UKLGPT bietet signifikant höhere Datenhoheit, Unabhängigkeit und Innovationstiefe. Die Meierhofer/Averbis-Variante muss als Vergleichsoption bewertet werden, ist aber aufgrund von Vendor Lock-in, eingeschränkter Archivdatennutzung und fehlender Kontrolle über die KI-Pipeline strategisch nachteilig. Empfohlen wird die **Eigenlösung mit MVP-Ansatz** und stufenweiser Umsetzung.

Konsequenzen bei Nicht-Entscheidung

- Medizinische Altdaten sind nach SAP-Abschaltung nur über manuelle Archivrecherche zugänglich (HYDMedia Viewer).
- Effizienzgewinne und KI-Readiness werden nicht realisiert.
- Strategischer Nachteil gegenüber anderen Universitätskliniken, die KI-Assistenzsysteme einführen.

0.1 Glossar und Produktbezeichnung

Begriff	Definition
EMRGPT	Kanonischer Projektname (ab v2.0). Ersetzt den bisherigen Produktnamen **UKLGPT** in allen neuen

Begriff	Definition
	Projektdokumenten (PSP, Projektplan, Präsentationen). In diesem Konzeptdokument wird aus Gründen der Konsistenz mit früheren Versionen weiterhin **UKLGPT** als Synonym verwendet. Arbeitstitel: hAIppokrates.
UKLGPT	Bisheriger Produktname des KI-gestützten klinischen Informationsassistenten des UKL. Ab v2.0 durch **EMRGPT** als kanonischen Projektnamen abgelöst. Im vorliegenden Dokument wird UKLGPT aus Versionskonsistenz beibehalten – alle Referenzen auf UKLGPT sind identisch mit EMRGPT.
RAG	Retrieval-Augmented Generation – Methode, bei der ein KI-Sprachmodell mit spezifischen, abgerufenen Fakten angereichert wird, bevor es eine Antwort generiert.
GraphRAG	RAG auf Basis eines Wissensgraphen (Netzwerkstruktur von Fakten und Beziehungen).
FHIR	Fast Healthcare Interoperability Resources – internationaler Standard für den Austausch medizinischer Daten.
SNOMED CT	Systematized Nomenclature of Medicine – klinische Referenzterminologie zur einheitlichen Kodierung medizinischer Begriffe.
M-KIS	Meierhofer Klinisches Informationssystem – neues KIS, das SAP IS-H/i.s.h.med ablöst.
HYDMedia	Dokumentenmanagementsystem (DMS) von Dedalus zur Archivierung medizinischer Dokumente.
LLM	Large Language Model – großes Sprachmodell (KI).
OCR	Optical Character Recognition – automatische Texterkennung in gescannten Dokumenten.
DWH / UKLytics	Data Warehouse / Analytik-Plattform des UKL.
Averbis Health Discovery	KI-Plattform des Anbieters Averbis (Partner von Meierhofer) zur automatisierten Textanalyse medizinischer Dokumente.

0.2 PSP-Zuordnung Vorprojekt (PRINCE2-Produktstruktur) – FÜHRENDE STRUKTUR

Die PSP-Produktstruktur aus PSP_EMRGPT.md (ehem. PSP_Vorprojekt_v0.6.md) ist die **verbindliche, mit dem Kunden UKL abgestimmte Leitstruktur** dieses Vorprojekts. Jedes Ergebnis dieses Dokuments muss einem PSP-Produkt zuordenbar sein. Die folgende Tabelle bildet den Gesamtstatus aller PSP-Produkte ab.

PSP-Nr	PSP-Produkt	Abgedeckt in Kapitel
1. Projektorganisation		
1.1	Projektauftrag Vorprojekt	Kap. 0 (Management Summary)
1.2	Projektorganisation	Kap. 0.3, **Kap. 0.9 (RA**
1.3	Stakeholderanalyse	Kap. 0.4
1.4	Kommunikationsplan	Kap. 0.8
1.5	Vorläufiger Business Case	Kap. 0.5, **Kap. 20 (Kostengerüst)**
1.6	Risiko- und Chancenliste	Kap. 1.4 + QUALITAETSANALYS Risiko-Register
2. Fachlich		
2.1	Zielbild EMRgpt	Kap. 1, 2, 3.1
2.2	Use-Case-Beschreibung LLM-Anbindung	Kap. 3.3 + EMR GPT Case Beschreibung 02.docx
2.2.1	Kostenanalyse	Kap. 0.5 (Nutzen), **Ka (Kostengerüst)**
2.3	Fachliche Anforderungen	Kap. 3.2, 3.3, **Kap. 0.**
2.3.1	Daten aus UKLytics	

PSP-Nr	PSP-Produkt	Abgedeckt in Kapitel
		Kap. 8 (Datenzufluss DWH→GraphRAG)
3. Technisch		
3.1	Technische Zielarchitektur	Kap. 4-12, **Kap. 0.6**
3.1.1	Infrastruktur	Kap. 7, 8, **Kap. 0.7**
3.2	Marktanalyse	Kap. 16 (Variantenvergleich)
3.3	Analyse HYDMedia	Kap. 2.4, 3.2.6
3.3.1	Schnittstellenbeschreibung FHIR	Kap. 3.2.6.2, **Kap. 0.6 (Conformance Stater**
3.4	Analyse DMI Lösung	Kap. 2.1
~~3.5~~	~~Vergleich HYDMedia vs. DMI~~	–
3.6	Schnittstellenübersicht	Kap. 8, 10
3.7	Technisches Vorgehensmodell	Kap. 5, 6, 9
3.8	Zugriff auf ISILON	Kap. 3.2.6.1
3.9	KIS-Dokumentenzugriff (Meierhofer)	Kap. 16
4. Rechtebewertung und Sicherheit		
4.1	Datenschutzkonzept	Kap. 13.1
4.2	DSFA (Vorprüfung)	Kap. 13.1
4.3	Informationssicherheitsbewertung	Kap. 12, 13.4, **Kap. 19 (Incident-Response)**

PSP-Nr	PSP-Produkt	Abgedeckt in Kapitel
4.4	Berechtigungskonzept	Kap. 12, **Kap. 12.1.1**
4.5	SAP-Berechtigungsanalyse	Kap. 12.1, **Kap. 12.1.1 Kap. 0.6 (HYDMedia-Lücke)**
4.6	Logging-/ Nachvollziehbarkeitskonzept	Kap. 12.6
5. Management Summary		
5.1	Entscheidungsgrundlage Datenhaltung	Kap. 2
5.2	Entscheidungsgrundlage LLM-Anbindung	Kap. 5, 6, 7, 16
5.3	Management-Entscheidungsvorlage	Kap. 0
6. Abschlussdokumentation		
6.1	Empfehlung für Hauptprojekt	Kap. 0, 16.4
6.2	Grober Umsetzungsfahrplan	**Kap. 17**
6.3	Abnahmedokument Vorprojekt	–
Ergänzende Produkte (v1.4)		
–	Betriebskonzept (Gerüst)	**Kap. 18**
–	Incident-Response-Plan	**Kap. 19**

PSP-Nr	PSP-Produkt	Abgedeckt in Kapitel
–	Kostengerüst-Template	**Kap. 20**
–	Change-Management-Konzept	**Kap. 21**

Legende: ERGÄNZT = in dieser Version neu hinzugefügt | OFFEN = noch zu erstellen | ~~Durchgestrichen~~ = entfällt

---

0.3 Projektorganisation (PSP 1.2)

Rolle	Person/ Bereich	Verantwortung
Auftraggeber	[zu benennen]	Strategische Steuerung, Budgetfreigabe
Projektleiter (zu installieren)	[PM-Exzellenz erforderlich]	Operative Projektsteuerung, Meilensteincontrolling, Eskalation
IT-Architektur	Gert, Carina, Valentin	Technische Zielarchitektur, Schnittstellendesign
Fachbereich Medizin	Felix, Martin Neef, Niko v.D.	Use Cases, fachliche Anforderungen, Akzeptanztest
KIS-Verantwortlicher	Robert W.	M-KIS-Integration, Meierhofer-Abstimmung
Datenschutz	Hr. Sünkel	DSGVO-Konformität, DSFA
Informationssicherheit	S. Krause	Schutzbedarfsanalyse, Sicherheitsbewertung
Berechtigungen	Martin Schmeißer, Fr. Stallmach,	Berechtigungskonzept, SAP-Analyse

Rolle	Person/ Bereich	Verantwortung
	Fr. Schmidt-Morich	

Hinweis: Die Rolle des **Projektleiters für das Umsetzungsprojekt** ist derzeit unbesetzt und muss priorisiert installiert werden. PM-Exzellenz ist ein erklärtes Projektziel.

0.4 Stakeholder-Übersicht (PSP 1.3)

Stakeholder	Interesse	Einfluss	Strategie
Vorstand / Klinikumsleitung	Strategische Digitalisierung, Kostenkontrolle	Hoch	Regelmäßige Entscheidungsvorlag
Ärztlicher Direktor	Patientenversorgung, Effizienz	Hoch	Pilotierung, Einbindu Use-Case-Definition
IT-Leitung	Architekturkonformität, Betrieb, Security	Hoch	Technische Governar
Datenschutzbeauftragter	DSGVO-Konformität	Hoch	Frühzeitige Einbindu DSFA
ISB	Informationssicherheit, KRITIS	Hoch	Schutzbedarfsanalys
Pflegedienstleitung	Effizienz klinischer Prozesse	Mittel	Change-Managemen Schulung
Fachbereiche PSY/KJP	Besonderer Dokumentenschutz	Mittel	Sonderregelungen Berechtigungen
Meierhofer AG	KIS-Anbieter, Averbis-Partnerschaft	Mittel	Vertragsgestaltung, Schnittstellenabstimm
Dedalus (HYDMedia)	DMS-Anbieter, FHIR-Schnittstelle	Mittel	Vertragliche Zusage FHIR-Export
Klinisches Personal (Ärzte, Pflege)	Alltagstauglichkeit, Zeitersparnis	Hoch	Pilotierung, Feedback Schleifen, Champions
DMI GmbH & Co. KG	Dokumentenarchiv-Anbieter (AVP Infinity), Klassifizierungsinstanz	Mittel	Evaluierung als potenzielle HYDMedi Alternative
GreenBay Healthcare (hAIppokrates)	KI-Implementierungspartner, Sovereign-Cloud-Angebot	Mittel	Vertragliche Abstimm Preismodell und Betri

---

0.5 Vorläufiger Business Case (PSP 1.5 / PSP 2.2.1)

Quelle der Nutzendaten: „EMR GPT Use Case Beschreibung 02.docx" (Fachbereich-Erhebung)

Ausgangslage und Problemquantifizierung

Kennzahl	Ambulant	Stationär	Quelle
Suchanfragen pro Arzt/Tag	2–5	3–5	Fachbereich-Erhebung
Durchschnittliche Suchdauer	5–10 Min.	5–10 Min.	Fachbereich-Erhebung
Maximale Suchdauer (Einzelfälle)	bis 20 Min.	bis 20 Min.	Fachbereich-Erhebung
Gesamter Rechercheaufwand pro Arzt/Tag	15–50 Min.	15–50 Min.	Berechnung
Zusätzlicher Aufwand durch „Doppelfragen"	ca. 2–3 Min./ Doppelfrage	ca. 2–3 Min./ Doppelfrage	Fachbereich-Erhebung

[Annahme] Der theoretische Abrufbedarf liegt laut Fachbereich deutlich höher als die tatsächliche Nutzung, da HYDMedia aufgrund seiner Komplexität häufig nicht konsequent genutzt wird. Ärzte behelfen sich mit erneuter Befragung der Patienten.

Nutzenquantifizierung (Eigenlösung UKLGPT)

Nutzenkategorie	Quantifizierung	Annahme/ Herleitung
Zeitersparnis pro Suchanfrage	Von Ø 7,5 Min. auf ca. 10 Sek. (Zielwert)	GPT-gestützte Suche lt. Fachbereich-Erhebung
Zeitersparnis pro Arzt/Tag	15–50 Min. Recherchezeit → ca. 1–2 Min.	Bei 3–5 Anfragen × 10 Sek.
Hochrechnung UKL (Ärzte)	[Annahme: 500 Ärzte] × Ø 30 Min./Tag = **250 Arztstunden/Tag** eingesparte Recherchezeit	Muss mit Personaldaten validiert werden

Nutzenkategorie	Quantifizierung	Annahme/ Herleitung
Qualitative Verbesserung	Reduktion von Informationslücken, weniger Doppelfragen, höhere Patientensicherheit	Fachbereich-Erhebung (qualitativ)
Reduktion Frustrationslast	Signifikant – HYDMedia-Komplexität wird durch natürliche Sprache ersetzt	Fachbereich-Erhebung

Kostenabschätzung (Grobrahmen – noch zu detaillieren)

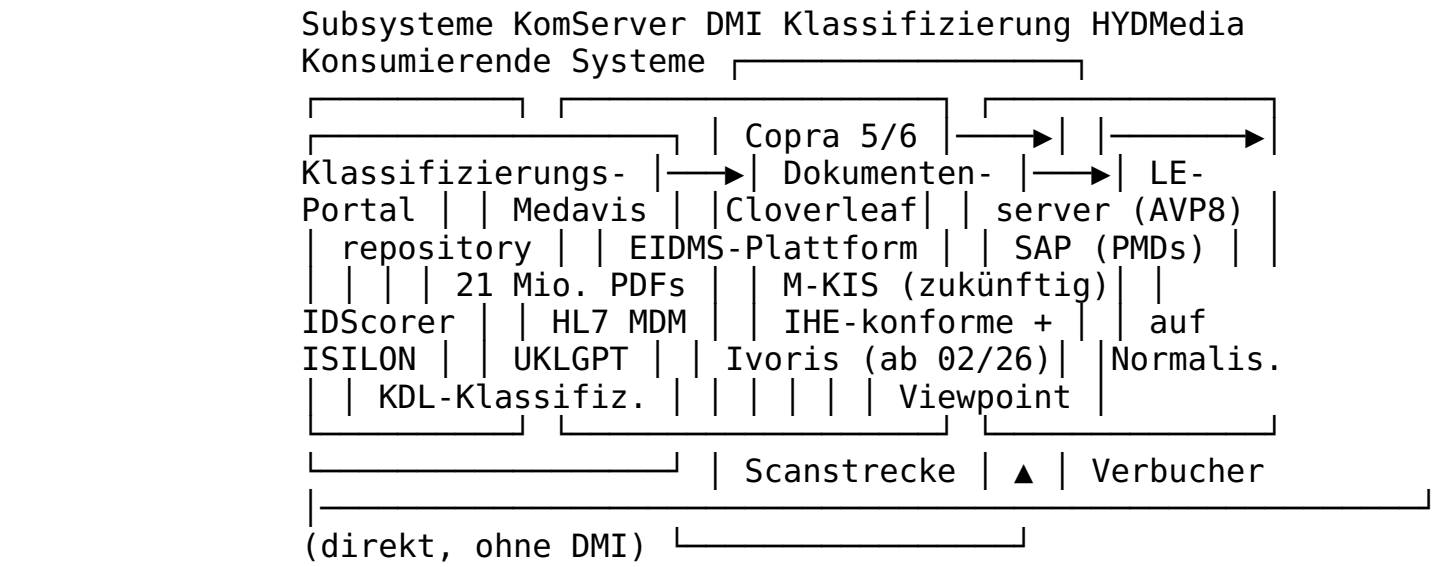
Kostenblock	Eigenlösung UKLGPT	Averbis/ Meierhofer	Anmerkung
Infrastruktur (GPU, Speicher, Netz)	Zu kalkulieren	Entfällt (Cloud/ SaaS)	On-Premise vs. Cloud
LLM-Lizenzen / API-Kosten	Zu kalkulieren	In Lizenz enthalten	Abhängig von Modellwahl
Personal (Entwicklung, Betrieb)	Zu kalkulieren	Gering (SaaS-Betrieb)	Kompetenzaufbau vs. Einkauf
OCR-Verarbeitung (21 Mio. PDFs)	Zu kalkulieren	Nicht enthalten	Beide Varianten benötigen OCR
Meierhofer/ Averbis-Lizenz	Entfällt	Angebot ausstehend	
Externe Beratung	Zu kalkulieren	Zu kalkulieren	

Handlungsbedarf: Die Kostenseite muss im weiteren Verlauf des Vorprojekts detailliert werden. Mindestens erforderlich: (1) Infrastruktur-Sizing und -Kosten durch IT, (2) Averbis-Angebot abwarten, (3) FTE-Planung für Entwicklung und Betrieb. *Detailliertes Kostengerüst-Template: siehe Kap. 20.*

0.6 Dokumentenpipeline und Datenqualität-Analyse (PSP 3.1 / 3.3)

Quellen: „Dokumentklassifizierung.pdf", „technische Konzeption_HydMedia und DMI Schnittstelle.docx", „HYDMedia_G6_6.10_-_FHIR_Conformance_Statement.docx"

Dokumentenpipeline im Ist-Zustand



Datenströme ins HYDMedia (Vollständigkeit)

Datenquelle	Weg ins HYDMedia	IHE/ KDL-konform	Status
SAP (PMDs → PDFs via x-tention/ MCI)	über DMI	Ja	Aktiv (aktuell 50%, nach DOK-Migration 100%)
Medavis (Radiologie)	MDM → DMI	Ja	Aktiv, inkl. Altdaten
Copra 5/6 (Intensivdoku)	Fileshare → DMI abholt	Ja	Aktiv, 900-Seiten-PDFs nach Abschluss
Viewpoint	MDM → DMI	Ja	Aktiv
IDScorer (PsyDoku)	MDM → DMI	Ja	Aktiv
Ivoris (Zahnmedizin)	MDM → DMI	Ja	Ab Ende Februar 2026

Datenquelle	Weg ins HYDMedia	IHE/ KDL-konform	Status
Scanstrecke (Druckdokumente)	DMI klassifiziert	Ja	Aktiv
Verbucher (diverse Systeme)	Direkt ins HYDMedia	**Nein**	Aktiv – **nicht IHE-konform**
SAP DOK-Migration	über DMI	Ja	**Geplant KW7/2026** – 20 Jahre SAP-Dokumente
Laborsystem	**Nicht angebunden!**	–	**LÜCKE** – Labordaten fehlen in HYDMedia, nur über UKLytics verfügbar

Datenqualitätsprobleme (Kritisch für UKLGPT)

Problem	Auswirkung auf UKLGPT	Lösungsansatz
Duplikate (3–5×)	Arztbriefe können durch Druck→Scan→Akte→Scan + Verbucherstrecke + MDM + SAP mehrfach im System landen	Dedup-Logik bei Indexierung, Hash-Vergleich, Timestamp-Gewichtung
Verbucher-Dokumente nicht IHE-konform	Ohne KDL-Klassifizierung schwer durchsuchbar, keine Registry-Registrierung	Nachträgliche Klassifizierung oder Ausschluss beim RAG-Embedding
OCR nur für computergenerierte Dokumente	Handschriftliche Dokumente und Bilder werden nicht erkannt	OCR-Installation geplant 17.02.26; Einschränkung dokumentieren
Fehlende Labordaten	Laborbefunde nicht über HYDMedia abrufbar	Separater FHIR-Connector zu UKLytics/Labor-LIMS erforderlich

Problem	Auswirkung auf UKLGPT	Lösungsansatz
SAP-DOK noch nicht migriert	Erst nach KW7/2026 werden 20 Jahre SAP-Dokumente in HYDMedia verfügbar sein	Abhängigkeit für UKLGPT-Vollbetrieb einplanen

HYDMedia FHIR-Schnittstelle (Conformance Statement G6 v6.10)

Quelle: Dedalus Healthcare, Conformance Statement HYDMedia G6

Unterstützte FHIR-Ressourcen:

FHIR-Ressource	HYDMedia-Objekt	Operationen
Patient	Patient	SEARCH, RECEIVE
Encounter	Case/Certificate	SEARCH, RECEIVE
DocumentReference	konfigurierbare DocumentClass	SEARCH, RECEIVE, PUSH
Binary	Content (PDF/ Dokument)	SEARCH, RECEIVE, PUSH
Account, Composition, Condition, Coverage	n/a	SEARCH, RECEIVE

Unterstützte Operationen: SEARCH (Abfrage), PUSH (Versand an Empfänger), RECEIVE (Empfang von Dokumenten), CONSUME/ PULL (Abruf externer Dokumente).

Unterstützte Profile: ISiK-Dokumentenaustausch Stufe 3 + ISiK-Basisstufe 3 (gematik).

Implikation für UKLGPT: Der FHIR-Connector muss primär `DocumentReference` (Metadaten) + `Binary` (Originaldokument) nutzen. Die Auflösung `DocumentReference → Binary` ist der zentrale technische Pfad für den Dokumentenzugriff. `Patient` und `Encounter` dienen der Kontext-Filterung.

HYDMedia-Berechtigungslücke (KRITISCH)

> **BEFUND:** Das aktuelle HYDMedia-Berechtigungskonzept setzt das **Need-to-Know/Need-to-Do-Prinzip NICHT um**.
> - Standardzugriff erfolgt aus SAP (Weiterleitung zu HYDMedia über Standardnutzer) – nur für aktive Akten. - Erweiterter Zugriff über Citrix (Windows-Anmeldung)

ermöglicht Zugriff auf **alle Akten aller Patienten** – obwohl AD-Gruppen existieren, sind diese in HYDMedia nicht für Zugriffssteuerung hinterlegt. - **Konsequenz für UKLGPT:** Das System darf NICHT das HYDMedia-Berechtigungskonzept übernehmen, sondern MUSS die Berechtigungshoheit aus M-KIS (bzw. aktuell SAP) ableiten. Die FHIR-Abfrage an HYDMedia darf nur nach erfolgter M-KIS-Berechtigungsprüfung erfolgen.

0.7 Preisreferenz hAIppokrates / GreenBay Healthcare (PSP 3.2)

Quelle: „hAippokrates_Pitchdeck – 27.01.26.pdf"

Modell	Monatliche Kosten	Lizenzen	Implementierung (einmalig)
S	1.000 €	50 inkl.	8.000 € zzgl. MwSt.
M	1.200 €	100 inkl.	8.000 € zzgl. MwSt.
L	1.400 €	200 inkl.	8.000 € zzgl. MwSt.
XL	2.500 €	Unbegrenzt	8.000 € zzgl. MwSt.

Preise bei 12 Monaten Mindestlaufzeit, freibleibend, zzgl. MwSt.

Portfolio-Übersicht GreenBay Healthcare: hAIppokrates (Dokumentensuche + Chat), DALLAS (KI-Anamnese-Avatar), KI-Strategie (Beratung), KI-Infrastruktur (Server-Setup), Server-Betreuung, Individuelle KI-Tools.

Betriebsmodelle: On-Premise (eigene Infrastruktur) oder Sovereign Cloud (Anbieter-betrieben, DSGVO-konform).

Ansprechpartner: Dr. med. David Baur (CAO), Elias Kohnen (Products & AI Solutions, Leipzig: +49 341 991985-53).

0.8 Kommunikationsplan (PSP 1.4)

Nr	Kommunikationsmaßnahme	Zielgruppe	Format
K-01	**Lenkungsausschuss-Bericht**	Vorstand, Klinikumsleitung	Entscheidungsvo (max. 5 Seiten) + Präsentation
K-02	**Projekt-Statusbericht**	IT-Leitung, Auftraggeber, Stakeholder	Statusbericht (Ampellogik, Risi Meilensteine)
K-03	**Technisches Jour fixe**		

Nr	Kommunikationsmaßnahme	Zielgruppe	Format
		IT-Architektur, Entwicklungsteam, Schnittstellenverantwortliche	Arbeitsbesprechu (Protokoll)
K-04	**Fachbereich-Abstimmung**	Ärztlicher Direktor, Fachbereich Medizin, Pflege	Workshop / Revi
K-05	**Datenschutz-/Sicherheits-Review**	DSB, ISB	Review-Meeting
K-06	**Externe Partner-Abstimmung**	Meierhofer, Dedalus, ggf. GreenBay	Jour fixe / Telko
K-07	**Change-Management / Klinik-Info**	Klinisches Personal (Ärzte, Pflege)	Newsletter, Info-Veranstaltung

Eskalationswege: - Stufe 1: Projektleiter → Auftraggeber - Stufe 2: Auftraggeber → Lenkungsausschuss - Stufe 3: Lenkungsausschuss → Vorstand

---

0.9 RACI-Matrix (PSP-Produkte)

Legende: R = Responsible (durchführend) | A = Accountable (verantwortlich) | C = Consulted (beratend) | I = Informed (informiert)

PSP-Produkt	Projektleiter	IT-Architektur	Fachbereich	DSB	ISB
1. Projektorganisation					
1.1 Projektauftrag	C	I	I	I	I
1.2 Projektorganisation	**A/R**	C	C	I	I
1.3 Stakeholderanalyse	**A/R**	C	C	C	C
1.4 Kommunikationsplan	**A/R**	I	I	I	I
1.5 Vorläufiger Business Case	R	C	C	I	I
1.6 Risiko- und Chancenliste	**A/R**	C	C	C	C
2. Fachlich					
2.1 Zielbild EMRgpt	**A**	R	R	C	I

PSP-Produkt	Projektleiter	IT-Architektur	Fachbereich	DSB	ISB
2.2 Use-Case-Beschreibung	C	C	**A/R**	I	I
2.2.1 Kostenanalyse	C	C	I	I	I
2.3 Fachliche Anforderungen	C	C	**A/R**	C	C
3. Technisch					
3.1 Technische Zielarchitektur	C	**A/R**	C	C	C
3.2 Marktanalyse	**A/R**	C	C	I	I
3.3 Analyse HYDMedia	C	C	I	I	I
3.3.1 Schnittstellenbeschr. FHIR	C	**A/R**	I	I	I
3.6 Schnittstellenübersicht	C	**A/R**	I	I	C
3.7 Techn. Vorgehensmodell	C	**A/R**	I	I	I
4. Rechtebewertung & Sicherheit					
4.1 Datenschutzkonzept	C	C	I	**A/R**	C
4.2 DSFA (Vorprüfung)	C	C	I	**A/R**	C
4.3 InfoSec-Bewertung	C	C	I	C	**A/R**
4.4 Berechtigungskonzept	C	C	I	C	**A**
4.5 SAP-Berechtigungsanalyse	C	I	I	C	**A**
4.6 Logging-Konzept	C	R	I	C	**A**
5. Management Summary					
5.3 Entscheidungsvorlage	**A/R**	C	C	C	C
6. Abschlussdokumentation					
6.1 Empfehlung Hauptprojekt	**A/R**	C	C	C	C
6.2 Umsetzungsfahrplan	**A/R**	C	C	I	I
6.3 Abnahmedokument	C	I	I	I	I

0.10 Anforderungsliste (PSP 2.3 – Entwurf)

Extrahiert aus dem Zielarchitektur-Dokument. Formale Priorisierung und Abnahmekriterien sind mit dem Fachbereich abzustimmen.

Funktionale Anforderungen

ID	Anforderung	Priorität	PSP-Bezug	Quelle (Kapitel)
FA-01	Semantische Suche in natürlicher Sprache über medizinische Altdaten (21 Mio. Dokumente)	MUSS	2.1, 3.1	Kap. 1.2, 3.1
FA-02	Konsolidierte, fallbezogene Antworten mit Quellenangabe (Dokument, Datum, Typ)	MUSS	2.1, 3.1	Kap. 1.2.2, 6.4
FA-03	Patientenkontextbezogene Recherche (nach Patient, Zeitraum, Dokumenttyp)	MUSS	2.2	Kap. 3.3.1
FA-04	Automatisierte Zusammenfassung von Patientenakten	SOLL	2.2	Kap. 3.3.2
FA-05	Unterstützung bei Visitenvorbereitung (Visitenlisten, offene Befunde)	SOLL	2.2	Kap. 2.7.1
FA-06	Leitlinienabgleich: Patientensituation gegen klinische Leitlinien prüfen	KANN	2.2	Kap. 2.7.1
FA-07	Strukturierte Übergabeberichte (Handover) generieren	SOLL	2.2	Kap. 2.7.1
FA-08	Integration in M-KIS-Oberfläche (Aufruf aus Patientenkontext)	MUSS	3.9	Kap. 4.1, 12.7.2
FA-09	GraphRAG-basierte Faktenabfrage (Diagnosen, Medikation, Labor aus DWH)	MUSS	2.3.1, 3.1	Kap. 7.3, 9.1
FA-10	Dokumenten-RAG: Volltextsuche über OCR-verarbeitete PDFs	MUSS	3.1, 3.3	Kap. 7.2.2, 10.1
FA-11		KANN	2.2	

ID	Anforderung	Priorität	PSP-Bezug	Quelle (Kapitel)
	Tumorboard-/ Konsilunterstützung (Fallzusammenstellung)			Kap. 2.7.1
FA-12	OCR-Verarbeitung der Dokumentenbasis (mind. computergenerierte Dokumente)	MUSS	3.1	Kap. 1.3.1, 0.6
FA-13	Automatisches SNOMED-CT-Tagging aller eingehenden Patientendokumente (NER-Pipeline)	SOLL	2.3.1, 3.1	Kap. 14.2.2
FA-14	Proaktives Evidence-Matching: Anzeige relevanter interner + externer Dokumente beim Fallaufruf	SOLL	2.2	Kap. 3.3.4, 7.2.1.2
FA-15	Europe PMC-Anbindung: Automatischer Abruf aktueller Publikationen zu Patientendiagnosen	SOLL	2.2	Kap. 7.2.1.1
FA-16	Facettierte Dokumentensuche über SNOMED-Tags (Diagnose, Prozedur, Substanz)	SOLL	2.2	Kap. 3.3.4, 14.2.2

Nicht-funktionale Anforderungen

ID	Anforderung	Priorität	PSP-Bezug	Quelle (Kapitel)
NFA-01	Antwortzeit < 30 Sekunden für Standard-Rechercheanfragen	MUSS	3.1	Kap. 0.5 (Zielwert 10 Sek.)
NFA-02	Verfügbarkeit > 99,5% während klinischer Kernzeiten (Mo-Fr 7-20 Uhr)	SOLL	3.1.1	– (noch zu definieren)
NFA-03	Berechtigungsprüfung vor jedem Datenzugriff (Gatekeeper-Prinzip)	MUSS	4.4, 4.5	Kap. 12.2, 12.7

ID	Anforderung	Priorität	PSP-Bezug	Quelle (Kapitel)
NFA-04	Revisionssicherer Audit-Trail aller Abfragen und Zugriffe	MUSS	4.6	Kap. 12.6
NFA-05	DSGVO-Konformität (Zweckbindung, Datenminimierung, Löschkonzept)	MUSS	4.1	Kap. 13.1
NFA-06	On-Premise-Betrieb (keine Patientendaten in externe Cloud)	MUSS	3.1.1	Kap. 0, 16.2
NFA-07	Episodische Datenverarbeitung mit TTL-Mechanismus	SOLL	4.1	Kap. 14.4
NFA-08	EU AI Act Konformität (Transparenz, Risikomanagement, Aufsicht)	MUSS	4.3	Kap. 13.3
NFA-09	Break-the-Glass Notfallzugriff mit dokumentierter Begründung	MUSS	4.4	Kap. 12.4
NFA-10	Skalierbarkeit: System muss für 500+ gleichzeitige Nutzer ausgelegt sein	SOLL	3.1.1	– (noch zu validieren)
NFA-11	Dokumentenschutz PSY/KJP (Sonderbehandlung sensibler Fachbereiche)	MUSS	4.4	Kap. 12.3.3
NFA-12	KRITIS-konforme Infrastruktur (BSI, IT-Sicherheitsgesetz)	MUSS	4.3	Kap. 13.4

Inhaltsverzeichnis

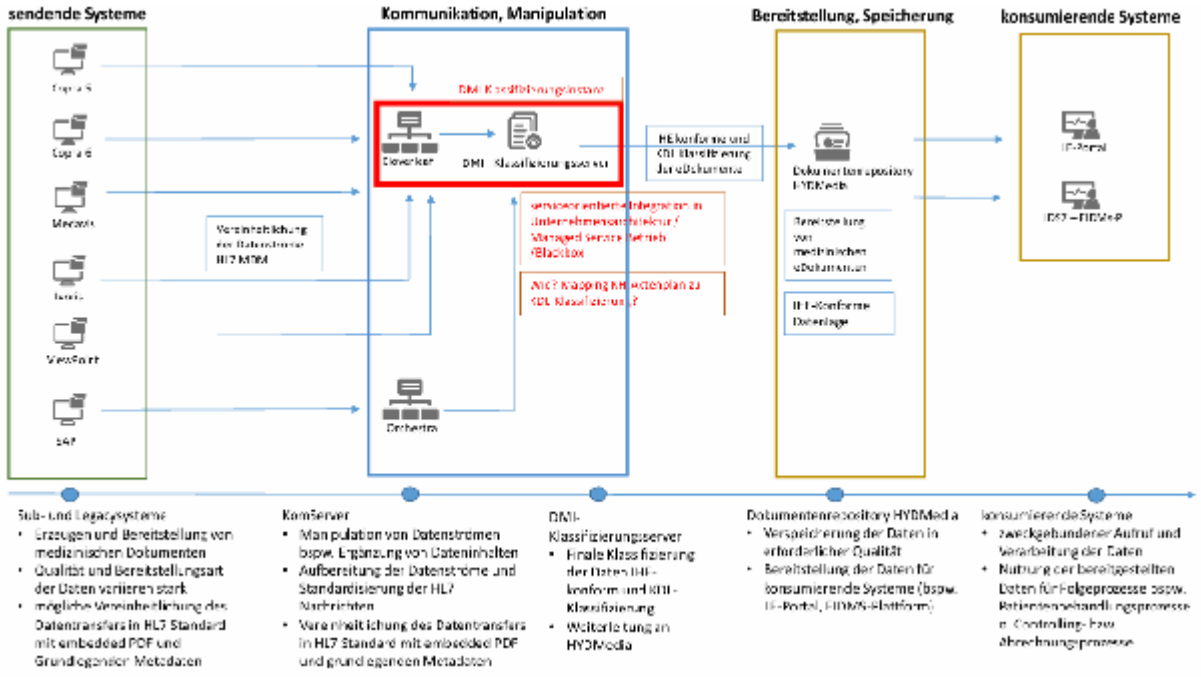
---

1. Zielbeschreibung: Technische Strategie zur Migration und Nutzung medizinischer Altdaten mittels UKLGPT {#1.-zielbeschreibung:-technische-strategie-zur-migration-und-nutzung-medizinischer-altdaten-mittels-uklgpt}

1.1 Einleitung und Hintergrund {#1.1-einleitung-und-hintergrund}

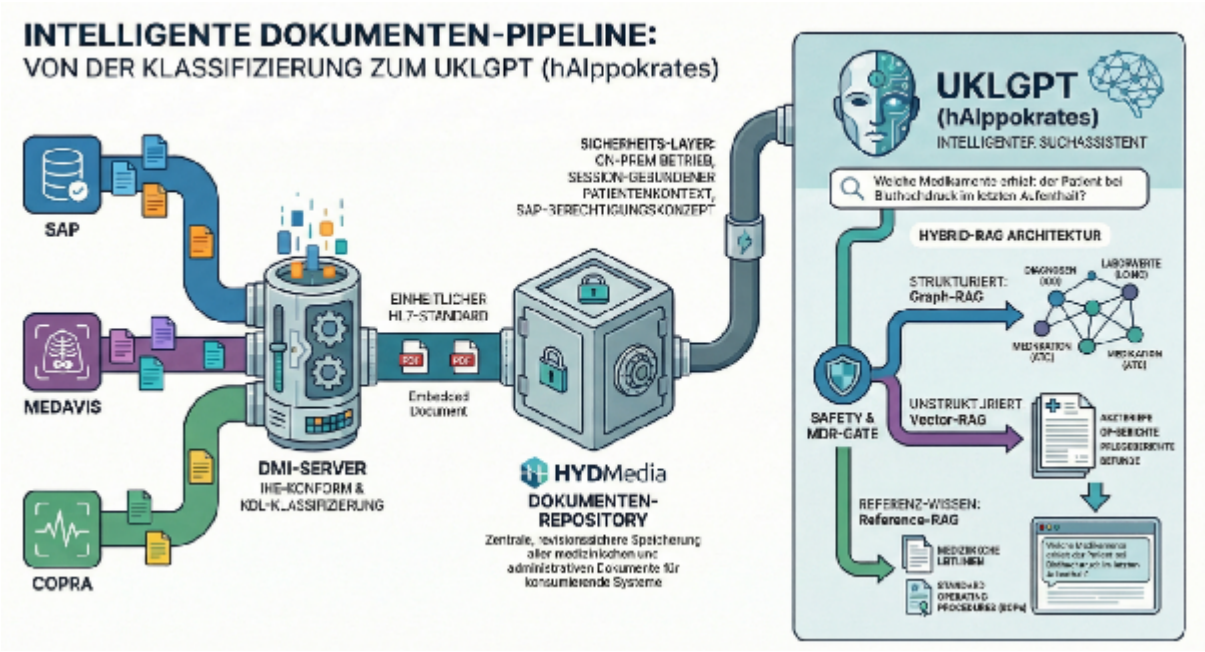
Das Universitätsklinikum Leipzig (UKL) steht vor der **Ablösung des bestehenden Krankenhausinformationssystems (KIS) SAP IS-H/ i.s.h.med** durch das neue KIS von Meierhofer (M-KIS). Im Rahmen dieses Systemwechsels werden aus regulatorischen und technischen Gründen **nur Rumpfdaten** (Leistungen, Falldaten usw.) in das neue M-KIS übernommen. Die klinisch hochrelevanten medizinischen Altdaten, die über Jahre im Altsystem und in diversen Subsystemen entstanden sind, müssen rechtssicher archiviert und für die weiterführende Patientenversorgung zugänglich bleiben.

Die Altdaten liegen aktuell fragmentiert in verschiedenen IT-Silos. Die Herausforderung besteht darin, diese heterogenen, zumeist unstrukturierten oder semi-strukturierten Daten (z. B. Arztbriefe, Befunde, OP-Berichte, Pflegeprotokolle) zu konsolidieren und sie dem klinischen Personal *kontextbezogen* und *effizient* zugänglich zu machen. Die bisherige manuelle Suche in Archivsystemen ist zeitaufwendig und ineffizient.



1.2 Zielbild der Strategie: Die Brücke der Altdaten (The Legacy Data Bridge) {#1.2-zielbild-der-strategie:-die-brücke-der-altdaten-(the-legacy-data-bridge)}

Die technische Strategie zielt darauf ab, die Integrität, Verfügbarkeit und semantische Erschließbarkeit der medizinischen Altdaten zu gewährleisten, indem eine leistungsstarke, KI-gestützte Recherche- und Assistenzebene über das Archivsystem gelegt wird.



Kernziel: Die Altdaten müssen für das klinische Personal in der M-KIS-Umgebung so schnell und präzise abrufbar sein, als wären sie Teil der aktiven elektronischen Patientenakte.

1.2.1 Architektonisches Ziel: Konsolidiertes Dokumentenarchiv

- **Konsolidierung:** Alle relevanten medizinischen Dokumente und Berichte aus SAP IS-H/i.s.h.med und assoziierten Altsystemen (z. B. Laborsysteme, Bildgebung) werden in einem zentralen, revisionssicheren Dokumentenarchiv (entweder **HYDMedia G6** oder **DMI AVP Infinity**) in einem standardisierten Format (**PDF/A**) abgelegt.

- **Rechtssicherheit:** Das gewählte Archivsystem muss die Anforderungen an ein revisionssicheres elektronisches Archiv (z. B. IDW PS 880, ISO 13485, BSI TR-03125) erfüllen, insbesondere hinsichtlich der Protokollierung, Unveränderbarkeit und Verfügbarkeit der Dokumente.

- **FHIR-Metadaten:** Die Archivierung muss die Erzeugung und Speicherung von strukturierten Metadaten (z. B. Patient-ID, Dokumententyp, Erstellungsdatum, Fachabteilung) im **HL7 FHIR DocumentReference**-Format gewährleisten, um die spätere Anbindung an den GraphRAG zu ermöglichen.

1.2.2 Anwendungsziel: UKLGPT als zentrale Recherche-Instanz

- **Intelligente Abfrage:** Der **UKLGPT Chatbot** wird als die primäre, intelligente Schnittstelle für die Recherche in diesen konsolidierten Altdaten etabliert.

- **Semantische Suche (Dokumenten-RAG):** Mittels **Retrieval-Augmented Generation (RAG)** werden die unstrukturierten Inhalte der archivierten PDF-Dokumente semantisch erschlossen. Dies ermöglicht es dem klinischen Personal, in **natürlicher Sprache** komplexe Fragen zu stellen, anstatt mit manuellen Suchmasken und Stichwörtern arbeiten zu müssen.

- **Kontextualisierte Antworten:** UKLGPT liefert die gefundenen Informationen nicht als rohe Dokumentenliste, sondern als **strukturierte, konsolidierte und fallbezogene Antworten**, basierend auf den Altdaten. Die Quellenangabe (welches Altdokument, wann erstellt) wird dabei zwingend mitgeliefert.

1.3 Archivierungsstrategie und die Rolle von UKLGPT {#1.3-archivierungsstrategie-und-die-rolle-von-uklgpt}

1.3.1 Dokumenten-Ingestion und Indexierung

Phase	Komponente	Beschreibung
1. Extraktion (PDF-Archiv)	Legacy-Export-Tool (SAP/ Subsysteme)	Vollständiger, einmaliger Export aller relevanten medizinischen Dokumente aus den Altsystemen in PDF/A-Format.
2. Archivierung	HYDMedia G6 / AVP Infinity	Revisionssichere Ablage der PDFs und Verwaltung der Basis-Metadaten.
3. Vorverarbeitung (OCR)	OCR-Engine (integriert oder extern)	Obligatorische Durchführung der **Optical Character Recognition (OCR)** für alle PDFs (insbesondere gescannte Dokumente), um Freitext durchsuchbar zu machen.
4. Semantische Indexierung	Embedding-Modell & Vektor-DB	Generierung von **Vektor-Embeddings** aus dem gesamten Textinhalt der Dokumente. Die Vektoren werden in einer **Vektor-Datenbank** (Teil des Patienten-Dokumenten-RAG) indexiert.

1.3.2 Die GraphRAG-Steuerung

Der GraphRAG spielt eine essenzielle Rolle bei der Steuerung der Altdaten-Recherche.

- **Metadaten-Hub:** Der Graph speichert die FHIR DocumentReference-Metadaten (wer, wann, welche Art von Dokument) zu den archivierten PDFs.
- **Patienten-Kontext:** Bei einer Anfrage (z. B. "Was war die Medikation des Patienten Max Mustermann im Jahr 2020?") nutzt der Graph die **Patienten-ID** und den **Zeitstempel** (2020) zur präzisen Identifizierung der **relevanten Altdokumente** (z. B. Entlassbriefe, Medikationspläne) im Archiv.
- **Gezieltes Retrieval:** Anstatt die gesamte Vektor-Datenbank zu durchsuchen, nutzt der Graph die ermittelten Metadaten, um

das anschließende Vektor-Retrieval auf die identifizierten, hochrelevanten Dokument-Chunks zu beschränken. Dies ist der **Schlüssel zur Effizienz**.

1.3.3 Sicherheits- und Berechtigungskonzept (Gatekeeper)

- **Mandantenfähigkeit:** Die Altdaten sind strikt Patient-gebunden und müssen über das im neuen KIS (M-KIS) geltende **Berechtigungsmanagement** geschützt werden.
- **Gatekeeper-Prinzip:** UKLGPT muss vor jeder Abfrage die **aktive Behandlungsbeziehung** oder die explizite **Need-to-Know-Berechtigung** des Nutzers im M-KIS oder SAP IS-H/ i.s.h.med-Ablöse-System verifizieren, bevor es die Metadaten-Abfrage über den GraphRAG startet.
- **Audit-Trail:** Jede Abfrage und der Zugriff auf Altdaten über UKLGPT muss revisionssicher protokolliert werden, um die Einhaltung der Datenschutz- und Compliance-Vorgaben zu gewährleisten.

1.4 Kritische Erfolgsfaktoren und Risikomanagement {#1.4-kritische-erfolgsfaktoren-und-risikomanagement}

Erfolgsfaktor	Details	Risiko bei Nichterfüllung
Vollständigkeit des Exports	Sicherstellung, dass *alle* klinisch relevanten Dokumente (insbesondere Vorbefunde, Tumorboard-Protokolle, OP-Dokumentation) aus SAP und Subsystemen migriert werden.	**Verlust klinisch essenzieller Informationen** und hohes juristisches Risiko.
Qualität der Metadaten (FHIR)	Generierung präziser, konsistenter und FHIR-konformer Metadaten während des Archivierungsprozesses zur Steuerung des GraphRAG.	**Unauffindbarkeit** von Dokumenten, da der Graph das Retrieval nicht korrekt steuern kann.
OCR-Qualität	Hohe Erkennungsrate für Text in gescannten und älteren Dokumenten.	Die semantische Suche schlägt bei fehlerhaft oder nicht erkannten Texten fehl;

Erfolgsfaktor	Details	Risiko bei Nichterfüllung
		"Verlust" von Informationen im Archiv.
Performance der Schnittstelle	Der Zugriff von UKLGPT auf die PDFs im Archiv (HYDMedia/AVP) muss hochperformant sein, um Echtzeit-Antworten zu ermöglichen.	Hohe Latenz im Chatbot; Ablehnung durch klinisches Personal aufgrund von **Zeitverlust**.

2. Nutzen, Anwendung und Innovation {#2.-nutzen,-anwendung-und-innovation}

2.1 Archivsystem für SAP i.s.h. med {#2.1-archivsystem-für-sap-i.s.h.-med}

Die Strategie, strukturierte medizinische Altdaten (z. B. Falldaten, Leistungsdaten, strukturierte Befunde aus dem abzulösenden KIS SAP IS-H/i.s.h.med) als PDF/A-Dokumente zu archivieren, ist primär durch **regulatorische und prozessuale Zwänge** (revisionssichere Archivierung, Systemwechsel) motiviert. Aus der Perspektive der KI-Architektur (UKLGPT/RAG) ist dies jedoch die **suboptimale Variante** und sollte kritisch gegen eine native, strukturierte Archivierung abgewogen werden.

Alternative: Native Archivierung in strukturierter und semistrukturierter Form

Die primäre Alternative besteht darin, die strukturierten Altdaten nicht in ein statisches, unstrukturiertes Format (PDF) zu "pressen", sondern sie in ihrer nativen, semantisch intakten Form zu archivieren, idealerweise in einem Format, das die spätere Nutzung im GraphRAG direkt ermöglicht.

2.2 Archivierung in FHIR-konformen Datenbanken/Repositories {#2.2-archivierung-in-fhir-konformen-datenbanken/repositories}

Ansatz: Die strukturierten Altdaten werden in ein FHIR-Repository (z.B. FHIR-Server) oder eine moderne analytische Datenbank migriert, die die FHIR-Semantik und -Ressourcen abbildet.

Aspekt	Vorteile	Nachteile
Datenqualität / Semantik	**Erhalt der Semantik:** Die Daten bleiben in ihrer ursprünglichen Struktur (z.B. Medikation, Laborwert mit Einheit, Diagnose mit Code) intakt und FHIR-konform. Kein Informationsverlust durch Transformation.	**Komplexität der Migration:** Erfordert eine hochkomplexe **FHIR-Mapping-Logik** für die Altdaten. Hoher Aufwand, um die oft inkonsistenten Altdaten zu bereinigen und zu standardisieren.
RAG-Nutzung (GraphRAG)	**Optimale Graph-Input:** Kann direkt in den GraphRAG oder eine andere graphbasierte Lösung importiert werden. Sofortige Nutzung von FHIR-konformen Knoten und Relationen.	**Datenmenge:** Für die initialen Big-Data-Migrationen können hohe Lizenz- und Betriebskosten für spezialisierte FHIR-Server anfallen.
Performance	**Hohe Performance:** Direkte Abfrage strukturierter Daten ist extrem schnell. Eliminierung der **OCR-Latenz** und der **Vektor-Retrieval-Latenz** für Fakten.	Nicht direkt relevant für die revisionssichere Langzeitarchivierung (dafür wäre ein spezialisiertes Archivsystem nötig).
Compliance/ Archiv	–	**Revisionssicherheit:** Erfordert eine zusätzliche Validierung und Zertifizierung (z.B. IDW PS 880) für die als Archiv genutzte FHIR-DB, was aufwendiger ist als bei dedizierten Archivsystemen (HYDMedia/AVP).

2.3 Archivierung in Vektor-Datenbanken (nativ für RAG) {#2.3-archivierung-in-vektor-datenbanken-(nativ-für-rag)}

Ansatz: Strukturierte Altdaten werden direkt in einem **Vektor-Archiv** gespeichert. Das heißt, anstatt den Umweg über PDF und OCR zu gehen, wird der Klartext (z.B. des Entlassbriefes) direkt in Vektoren überführt und mit Metadaten (Patient-ID, Datum) angereichert.

Aspekt	Vorteile	Nachteile
Datenqualität / Semantik	**Gezielte Indexierung:** Nur klinisch relevanter Text wird indexiert. Bessere Kontrollmöglichkeiten über die Chunking-Strategie.	**Verlust der FHIR-Struktur:** Die zugrundeliegende FHIR-Struktur wird für das RAG-Retrieval geopfert und nur der Text-Inhalt ist direkt nutzbar.
RAG-Nutzung	**Niedrigste Latenz:** Optimiert für Dokumenten-RAG-Abfragen (semantische Suche in Altdaten).	**Revisionssicherheit:** Ein Vektor-Archiv ist **kein revisionssicheres Archiv** im Sinne des deutschen Rechts (IDW PS 880). Das Originaldokument (PDF/A) muss parallel revisionssicher archiviert werden.
Kosten	Reduzierung der OCR-Kosten für die Altdaten (da kein PDF-zu-Text-Schritt notwendig ist, wenn Text nativ vorliegt).	Kosten für die Vektor-DB-Infrastruktur und das Embedding-Modell.

2.4 Bewertung des Ansatzes “Reine PDF-Archivierung” {#2.4-bewertung-des-ansatzes-“reine-pdf-archivierung”}

Der im Dokument beschriebene Ansatz – die Überführung *aller* klinischen Altdaten in das revisionssichere PDF/A-Archiv (HYDMedia/AVP) – **hat primär juristische Vorteile**, schafft aber erhebliche technische Hürden für die KI-Nutzung.

HYDMedia G6 ist kein „nativer" FHIR-Store, der beliebige FHIR-Ressourcen (wie Observation, Condition oder Medication) als JSON/XML speichert und unverändert wiedergibt. Es handelt sich vielmehr

um ein Enterprise Content Management (ECM) System mit einer **FHIR-Fassade**.

2.4.1 Technische Funktionsweise (Fassade vs. Store)

HYDMedia speichert keine FHIR-Ressourcen als solche. Stattdessen nimmt es FHIR-Anfragen entgegen und mappt diese auf seine internen Datenbankstrukturen (Patient, Fall, Dokument, Archiv).

- **Input:** Wenn eine FHIR-Nachricht an HYDMedia gesendet wird, wird der Dokumenteninhalt (Binary) im Dateisystem (z.B. Isilon) abgelegt und die Metadaten (DocumentReference) in die HYDMedia-SQL-Datenbank geschrieben.
- **Output:** Bei einer Abfrage baut HYDMedia die FHIR-Ressource „on-the-fly" aus seinen internen Daten zusammen und sendet sie zurück.

2.4.2 Unterstützte Ressourcen (Eingeschränkter Umfang)

HYDMedia unterstützt laut Conformance Statement nur eine begrenzte Menge an FHIR-Ressourcen, die für die Dokumentenverwaltung notwendig sind:

- DocumentReference (Metadaten zum Dokument)
- Binary (Der Dateiinhalt selbst, z.B. PDF)
- Patient (Stammdaten)
- Encounter (Fall/Aufenthalt)
- Organization, Practitioner

Nicht unterstützt werden typische strukturierte klinische Ressourcen, die für einen medizinischen Wissensgraphen benötigt würden, wie Observation (Laborwerte), MedicationStatement (Medikation) oder Condition (Diagnosen als Ressource). Diese Daten können in HYDMedia nur als Teil eines Dokuments (z.B. PDF-Arztbrief) liegen, aber nicht als einzeln abfragbare FHIR-Datenpunkte.

2.4.3 Herausforderungen in Vollständigkeit und Datenqualität

Aspekt	Vorteil des PDF-Ansatzes (Aktuell)	Nachteil aus KI/ RAG-Perspektive
Revisionssicherheit	**Höchste juristische Sicherheit:** PDF/A ist der De-facto-Standard für	**KI-Feindliches Format:** PDF ist ein Format für Druck und Anzeige,

Aspekt	Vorteil des PDF-Ansatzes (Aktuell)	Nachteil aus KI/ RAG-Perspektive
	revisionssichere Langzeitarchivierung. Dies ist **zwingend** für Altdaten erforderlich.	nicht für maschinelle Auswertung.
Einfachheit der Migration	**Prozessuale Einfachheit:** Der Export von SAP-Dokumenten als PDF ist ein etablierter Ablösungsprozess.	**Zusätzlicher Zwischenschritt:** Der gesamte Inhalt muss mittels **OCR** (Optical Character Recognition) erst wieder in maschinenlesbaren Text umgewandelt werden, um überhaupt durchsuchbar zu sein.
Datenqualität / Vollständigkeit	Erhält das Original-Layout des Dokuments (Visuelle Konsistenz).	**Qualitätsverlust durch OCR:** Fehler bei der OCR-Erkennung (insbesondere bei Scans, handschriftlichen Notizen oder schlechten Vorlagen) führen zu **Suchfehlern** im RAG.
Kosten / Performance	–	**Zusätzliche Kosten:** OCR-Lizenzen und Rechenzeit fallen für die **gesamte Altdatenmenge (21 Mio. Dokumente)** an. **Erhöhte Latenz:** OCR ist zeitintensiv.

2.4.4 Fazit und Empfehlung

Die Umwandlung strukturierter Altdaten in PDFs ist für die **revisionssichere Archivierung** zwingend erforderlich und muss beibehalten werden.

Für die **KI-Nutzung (UKLGPT)** sollte jedoch die **Alternative 1 (FHIR-Mapping für GraphRAG)** in Betracht gezogen werden, da der Weg über PDF + OCR + Vektor-DB zu einem **erheblichen Informationsverlust und einer Reduktion der Datenqualität** führen kann, insbesondere für Fakten, die im Altsystem strikt strukturiert waren (z.B. Medikationshistorie).

2.5 Hybride Strategie zur KI-Readiness und Compliance {#2.5-hybride-strategie-zur-ki-readiness-und-compliance}

Da HYDMedia nicht als Speicher für feingranulare, strukturierte FHIR-Daten dient, bestätigt dies die Entscheidung für die **Hybride Architektur-Strategie**. Dieser hybride Weg **erfüllt die Compliance-Anforderungen** (revisionssicheres PDF-Archiv) und **maximiert die Nutzbarkeit und Präzision der KI** (strukturierte Fakten im FHIR-GraphRAG).

Die empfohlene hybride Strategie kombiniert die zwingend notwendige revisionssichere Archivierung mit der KI-optimierten Aufbereitung von Altdaten, um maximale Compliance und KI-Effizienz zu gewährleisten:

- **Revisionssichere Archivierung (Pflicht):**
- Alle Altdokumente werden als **PDF/A** in den Systemen **HYDMedia/AVP** revisionssicher archiviert.
- **KI-Readiness durch Migration (Kür/Effizienz):**
- Alle **ursprünglich strukturierten Altdaten** (z.B. Diagnosen, Labor, Medikation) werden parallel in das **UKLytics-DWH migriert** und dort in **FHIR-Ressourcen** transformiert.
- **Intelligente GraphRAG-Fütterung:**
- Der GraphRAG nutzt primär die **FHIR-Daten** für Fakten und steuert den Dokumenten-RAG (für Freitext) über die Metadaten des PDF-Archivs (DocumentReference).

2.6 Nutzen der Einbindung von UKLGPT {#2.6-nutzen-der-einbindung-von-uklgpt}

UKLGPT stellt eine innovative Lösung für ein fundamentales Problem der modernen klinischen Versorgung dar: die Flut und die Fragmentierung medizinischer Information. Obwohl hochrelevantes Wissen – von aktuellen Forschungsergebnissen über interne Patientenakten bis hin zu nationalen und internationalen Leitlinien – prinzipiell vorhanden ist, ist es im zeitkritischen klinischen Alltag oft schwer, dieses konsistent, schnell und im richtigen Kontext nutzbar

zu machen. UKLGPT adressiert diese Lücke direkt durch die Bereitstellung eines kontextsensitiven, intelligenten Informationsassistenzsystems.

Die Kernvorteile der Anwendung manifestieren sich in folgenden Bereichen:

2.6.1 Steigerung der Effizienz und Reduktion der Kognitiven Last

- **Signifikanter Zeitgewinn im klinischen Alltag:** Durch die Fähigkeit, komplexe klinische Fragen schnell zu verarbeiten und unmittelbar kontextbezogene, quellenbasierte Antworten zu liefern, entfällt die zeitaufwendige manuelle Recherche in unterschiedlichen Dokumentensystemen (Krankenhausinformationssystem, Archiv, Leitlinien-Datenbanken). Dies ermöglicht es dem Fachpersonal, mehr Zeit direkt am Patienten zu verbringen.

- **Reduktion kognitiver Last:** UKLGPT agiert als intelligenter Aggregator. Es führt relevante Dokumente, Laborfakten, historische Patienteninformationen und aktuelle, evidenzbasierte Leitlinien passend zur gestellten klinischen Frage zusammen. Diese **konsolidierte Wissensbasis** minimiert die Notwendigkeit für den Arzt oder das Pflegepersonal, sich gleichzeitig eine Vielzahl fragmentierter Informationen im Gedächtnis zu behalten oder mühsam abzugleichen.

2.6.2 Erhöhung der Patientensicherheit und Qualitätssicherung

- **Verbesserte Patientensicherheit durch strukturierte, quellenbasierte Antworten:** Jede generierte Antwort basiert auf nachvollziehbaren, im System hinterlegten Quellen. Diese **Transparenz und Verifizierung** gewährleisten, dass die bereitgestellte Information aktuell, validiert und relevant für den spezifischen klinischen Fall ist. Dies senkt das Risiko von Fehlentscheidungen, die auf veralteten oder unvollständigen Informationen beruhen.

- **Verbesserte Nachvollziehbarkeit medizinischer Aussagen:** Da jede Empfehlung oder Faktenzusammenfassung quellenbasiert ist, wird die Argumentationskette hinter einer medizinischen Aussage transparent. Dies ist essenziell für die Dokumentation, die Fallbesprechung und die Qualitätssicherung.

2.6.3 Optimierung der Interdisziplinären Prozesse

- **Unterstützung interdisziplinärer Zusammenarbeit:** UKLGPT kann als zentrale Wissensplattform dienen, die sicherstellt, dass alle Mitglieder des Behandlungsteams – von der Pflege

über die unterschiedlichen Fachabteilungen bis hin zur Verwaltung – Zugriff auf die gleiche, konsistente und aktuelle Informationsbasis haben. Dies optimiert die Kommunikationswege und harmonisiert die Behandlungsstrategien.

- **Wissensmanagement und Schulung:** Die Anwendung kann auch zur schnellen Einarbeitung neuer Mitarbeiter oder zur Auffrischung von Fachwissen genutzt werden, indem sie komplexe klinische Szenarien und die dazugehörigen Standard Operating Procedures (SOPs) kompakt und verständlich zusammenfasst.

2.6.4 Definition des Anwendungsspektrums: Assistenz, nicht Autonomie

Es muss klar hervorgehoben werden, dass UKLGPT explizit als Informations- und Orientierungsassistenz konzipiert ist. Die Rolle des Systems ist strikt die eines Werkzeugs zur Informationsverbesserung.

Die Anwendung ist NICHT konzipiert als:

- Diagnose-System
- Therapie-System
- Autonomes Entscheidungs-System

Die **finale klinische Entscheidung** und die **Verantwortung** für die Behandlung verbleiben **uneingeschränkt beim behandelnden Fachpersonal**. UKLGPT liefert die bestmögliche Informationsgrundlage, interpretiert oder ersetzt jedoch nicht das klinische Urteilsvermögen.

2.7 Anwendungsszenarien und Abgrenzung {#2.7-anwendungsszenarien-und-abgrenzung}

Die vorliegende Lösung zielt darauf ab, klinisches Personal in wesentlichen Bereichen der Patientenversorgung durch den Einsatz intelligenter Informationsverarbeitung zu unterstützen. Dabei liegt der Fokus auf der Verbesserung der Informationslage, der Entscheidungsunterstützung und der Effizienz klinischer Prozesse.

2.7.1 Explizit unterstützte Use Cases (Ziel-Szenarien)

Die folgenden Anwendungsszenarien stellen die Kernbereiche dar, in denen das System einen direkten Mehrwert bietet:

- **Überblick über den aktuellen klinischen Status eines Patienten:**

- **Konsolidierte Patientenakte:** Schnelle, intuitive Zusammenfassung der relevantesten aktuellen und historischen Patientendaten (z.B. Vitalparameter, Laborwerte, Medikation, bildgebende Befunde) aus verschiedenen Quellsystemen.

- **Identifizierung von Auffälligkeiten/Trends:** Hervorhebung signifikanter Veränderungen oder kritischer Werte, um eine sofortige Fokussierung auf relevante klinische Entwicklungen zu ermöglichen.

- **Einordnung patientenbezogener Informationen im Lichte gültiger Leitlinien und evidenzbasierter Medizin:**

- **Leitlinienabgleich:** Automatische oder assistierte Prüfung der aktuellen Patientensituation gegen relevante nationale und internationale klinische Leitlinien (z.B. AWMF, ESC, NCCN) oder hausinterne Standards.

- **Evidenzbasierte Recherche:** Bereitstellung von kontextsensitiven, wissenschaftlichen Informationen zur Unterstützung diagnostischer und therapeutischer Überlegungen.

- **Vorbereitung auf klinische Besprechungen:**

- **Visiten:** Erstellung fokussierter Visitenlisten und patientenindividueller Vorlagen mit den aktuellsten Befunden und offenen Fragestellungen.

- **Übergaben (Handover):** Generierung strukturierter Übergabeberichte, die alle kritischen Informationen und anstehenden Aufgaben umfassen.

- **Tumorboards/Interdisziplinäre Konferenzen:** Zusammenstellung und Präsentation der gesamten onkologischen Historie, aller relevanten Bildgebungen und histopathologischen Berichte in einem standardisierten Format für eine effiziente Falldiskussion.

- **Konsile:** Bereitstellung eines prägnanten Konsilbefunds, der die Kernfrage und die relevanten Hintergrundinformationen zusammenfasst.

- **Strukturierte Recherche in der Patientenakte:**

- **Semantische Suche:** Ermöglichen einer intelligenten Suche, die nicht nur auf Schlüsselwörtern basiert, sondern auch klinische Konzepte und Zusammenhänge in Freitextfeldern erkennt.

- **Filterung und Aggregation:** Schnelle Filterung großer Datenmengen (z.B. alle Laborwerte eines bestimmten Typs über einen Zeitraum, alle verabreichten Antibiotika).

- **Unterstützung bei Dokumentation und Nachbereitung:**

- **Assisted Documentation:** Vorschläge für standardisierte Textbausteine oder Befunde auf Basis der erfassten Patientendaten, um die Dokumentationszeit zu reduzieren und die Vollständigkeit zu erhöhen.

- **Kodierunterstützung:** Vorbereitung von Datenstrukturen zur späteren Unterstützung der medizinischen Kodierung (z.B. ICD, OPS) durch Aggregation relevanter Diagnosen und Prozeduren.

2.7.2 Nicht-Ziel-Szenarien (Abgrenzung)

Die folgenden Szenarien liegen explizit außerhalb des Funktionsumfangs und der Verantwortung des Systems, da sie eine unmittelbare ärztliche oder pflegerische Entscheidung sowie die ethische und juristische Verantwortung erfordern:

- **Automatische Therapie- oder Diagnoseentscheidungen:** Das System fungiert ausschließlich als *Unterstützungsinstrument* und darf keine finalen medizinischen Entscheidungen treffen. Die Verantwortung für die Diagnosestellung und die Wahl der Therapie verbleibt vollständig beim behandelnden Arzt.

- **Individuelle ärztliche Empfehlungen ohne menschliche Plausibilitätsprüfung:** Obwohl das System evidenzbasierte Informationen bereitstellen kann, stellt es keine individuelle, auf den Einzelfall zugeschnittene Therapieempfehlung dar, die die einzigartigen Umstände des Patienten und das ärztliche Urteilsvermögen ersetzen könnte.

- **Bewertung oder Auswahl von Personen (z.B. Personaleinsatzplanung oder Eignungsfeststellung):** Der Fokus liegt rein auf der klinischen Patientenversorgung und nicht auf administrativen oder personalbezogenen Prozessen.

2.8 Innovation des Ansatzes

Die wahre Innovation und der signifikante Mehrwert unseres Ansatzes liegen nicht in der isolierten Verwendung eines Large Language Models (LLM), sondern in der **synergistischen Kombination streng definierter, domänenspezifischer Komponenten**. Diese Integration schafft ein robustes, kontrolliertes und klinisch sicheres KI-System:

2.8.1 FHIR-konformer klinischer Graph

Das Herzstück des Systems bildet ein **FHIR (Fast Healthcare Interoperability Resources)-konformer klinischer Graph**, implementiert in einer leistungsstarken Graphdatenbank wie **Neo4j**. Dies ermöglicht die Abbildung komplexer, hochgradig vernetzter

Patientenpfade, Behandlungszusammenhänge und klinischer Entitäten. Durch die Einhaltung des FHIR-Standards wird maximale Interoperabilität und die semantische Konsistenz der Daten über verschiedene Gesundheitseinrichtungen und Systeme hinweg gewährleistet. Die Graphenstruktur ist essenziell, um kausale oder zeitliche Zusammenhänge schnell und präzise abfragen zu können, was in linearen Datenbanken ineffizient wäre.

2.8.2 SNOMED-normalisierte Semantik

Alle klinischen Informationen, Diagnosen, Prozeduren und Medikationen werden mithilfe der **SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms)** normalisiert. SNOMED CT dient als weltweit anerkanntes, umfassendes klinisches Referenzvokabular. Diese Normalisierung ist entscheidend, um:

- **Terminologie-Inkonsistenzen** zu eliminieren (z. B. "Herzinfarkt" vs. "Myokardinfarkt").
- **Präzise Semantik** für das LLM bereitzustellen, wodurch die Gefahr von Halluzinationen, die auf missverstandener Terminologie basieren, minimiert wird.

2.8.3 Episodisches Patienten-RAG mit Zweckbindung

Wir nutzen einen spezialisierten **Retrieval-Augmented Generation (RAG)**-Ansatz, der nicht den gesamten Datenbestand abfragt, sondern sich auf **episodische Patientendaten mit klarer Zweckbindung** fokussiert. Das bedeutet:

- **Episodisch:** Die Abfrage beschränkt sich auf relevante Behandlungsabschnitte, Krankheitsverläufe oder spezifische Fragestellungen.
- **Zweckbindung:** Das Retrieval wird durch den konkreten klinischen Anwendungsfall (z.B. Medikationsprüfung, Differentialdiagnose) gesteuert.
- Dieses zielgerichtete RAG-System verbessert die **Antwortqualität**, reduziert die **Latenz** und erhöht die **Datensicherheit**, da nur die minimal notwendigen Informationen für die Generierung bereitgestellt werden.

2.8.4 Mehrstufige Prompt-Pipeline zur Qualitätssicherung

Um die Verlässlichkeit der KI-generierten Ergebnisse zu garantieren, kommt eine **mehrstufige Prompt-Pipeline** zum Einsatz. Diese Pipeline beinhaltet dedizierte Prüf- und Validierungsschritte, bevor die endgültige Antwort ausgegeben wird:

- **Stufe 1 (Generierung):** Das LLM generiert einen Rohvorschlag basierend auf dem RAG-Input.

- **Stufe 2 (Validierung/Korrektur):** Spezielle Prompt-Schritte überprüfen die generierte Antwort auf Konsistenz mit den Ursprungsdaten und der SNOMED-Semantik.

- **Stufe 3 (Formatierung/Erklärung):** Die Antwort wird in ein klinisch verwertbares Format gebracht und die Herkunft der Fakten (Zitierfähigkeit) wird sichergestellt.
  Diese Architektur dient als essenzieller Filtermechanismus gegen Halluzinationen und klinische Fehler.

2.8.5 Klare Trennung von Fakten, Dokumenten und Leitlinien

Ein fundamentaler Sicherheitsmechanismus ist die **strikt logische Trennung der Informationsquellen** im Retrieval-Prozess:

- **Fakten:** Strukturierte Patientendaten (z. B. Laborwerte, Vitalparameter).

- **Dokumente:** Unstrukturierte Patientenakten (z. B. Arztbriefe, OP-Berichte, Pflegeberichte).

- **Leitlinien:** Evidenzbasierte klinische Empfehlungen (z. B. AWMF-Leitlinien).

Durch diese klare Unterscheidung kann das KI-System präzise angeben, auf welcher Grundlage eine Aussage getroffen wird (z. B. "Laut Laborwert vom TT.MM.JJJJ ist der Wert X erhöht" vs. "Gemäß AWMF-Leitlinie Y sollte in diesem Fall Z erfolgen"). Dies erhöht die **Transparenz, Auditierbarkeit und die klinische Akzeptanz** des generierten Outputs.

2.8.6 Evidenz- und Studien-Domainservice – Die vierte Wissensschicht

Die vielleicht weitreichendste Innovation unserer Architektur liegt in der Einführung eines eigenständigen **Evidenz- und Studien-Domainservice**, der als vierte, dynamische Wissensschicht neben Fakten, Dokumenten und Leitlinien tritt. Im Gegensatz zu statischen Leitlinien-RAGs, die manuell befüllt und periodisch aktualisiert werden, handelt es sich hier um einen **lebenden, sich selbst aktualisierenden Wissensservice**, der drei fundamentale Fähigkeiten vereint:

1. Automatische Evidenz-Zuordnung (Evidence-Matching als Domainservice):

Der Service wird nicht als monolithischer Baustein implementiert, sondern als **entkoppelter Domainservice**, der über definierte Schnittstellen sowohl an die Patientendaten (GraphRAG) als auch an

die Wissensbasis (Globaler Wissens-RAG) angebunden werden kann. Konzeptionell bedeutet das:

- **Andockpunkt Patientendaten:** Bei jedem Patientenkontakt extrahiert der Service die SNOMED-kodierten Diagnosen aus dem GraphRAG und übersetzt sie automatisch in Suchanfragen gegen externe Evidenzquellen (Europe PMC, ClinicalTrials.gov). Das Ergebnis sind patientenspezifische Evidenz-Karten, die dem Kliniker proaktiv angezeigt werden – nicht als generische Literaturliste, sondern als **kontextualisierte, nach klinischer Relevanz gewichtete Empfehlungen**.

- **Andockpunkt Wissensbasis:** Derselbe Service speist gleichzeitig die globale Wissensbasis mit aktuellen Erkenntnissen. Neue hochzitierte Publikationen, aktualisierte Leitlinien und relevante Studienergebnisse werden automatisch in den Wissens-RAG-Index aufgenommen und stehen damit allen zukünftigen Patientenkontakten zur Verfügung.

2. LLM-gestützte Regelableitung aus Evidenz:

Ein wesentlicher konzeptioneller Unterschied zu reinen Literatur-Suchsystemen besteht darin, dass der Domainservice nicht nur Publikationen *findet*, sondern aus ihnen **klinisch anwendbare Regeln ableitet**. Durch den Einsatz von LLMs auf den abgerufenen Volltexten (Open Access via Europe PMC) können strukturierte Aussagen extrahiert werden:

- **Therapie-Regeln:** „Bei HFrEF mit LVEF < 35 % → SGLT2-Inhibitor erwägen (ESC Guidelines 2023, Klasse I, Evidenzgrad A)"
- **Kontraindikations-Checks:** „Metformin kontraindiziert bei eGFR < 30 ml/min (KDIGO 2024)"
- **Dosierungshinweise:** „Empagliflozin 10 mg/d – keine Dosisanpassung bei Niereninsuffizienz bis eGFR 20 ml/min (EMPEROR-Preserved, NEJM 2021)"

Diese aus der Literatur abgeleiteten Regeln werden nicht direkt als Handlungsanweisungen ausgegeben, sondern als **Evidenz-Annotationen** dem Kliniker zur Validierung vorgelegt. Der Kliniker kann bestätigen, anpassen oder verwerfen – bestätigte Regeln fließen in die hauseigene Regelbasis ein und stehen fortan als validierter Bestandteil der klinischen Entscheidungsunterstützung (CDS) zur Verfügung. So entsteht über die Zeit ein **lernender, evidenzbasierter Regelkatalog**, der die Brücke zwischen Literatur und Praxis schlägt.

3. Klinische Studienanbindung (Clinical Trial Matching):

Als dritte Säule integriert der Domainservice die automatische Suche nach **laufenden klinischen Studien**, die für den jeweiligen Patienten relevant sein könnten. Über die ClinicalTrials.gov API v2 (REST,

öffentlich, ohne Authentifizierung) werden SNOMED-Diagnosen in Studiensuchen übersetzt:

- **Automatische Studien-Vorschläge:** „Für Ihre Diagnose *Herzinsuffizienz* laufen aktuell 47 rekrutierende Studien, davon 3 an Zentren in Sachsen/Thüringen."
- **Eignungsprüfung:** Basierend auf dem Patientenprofil (Alter, Geschlecht, Diagnosen, Medikation) kann eine erste automatische Vorfilterung der Ein-/Ausschlusskriterien erfolgen.
- **Standortbezug:** Durch den geografischen Filter (`filter.geo`) werden Studien priorisiert, die am UKL selbst oder in erreichbarer Nähe durchgeführt werden.

Dies ist besonders für ein Universitätsklinikum von strategischer Bedeutung: Die Studienrekrutierung wird beschleunigt, Patienten erhalten frühzeitiger Zugang zu innovativen Therapien, und das Klinikum stärkt seine Position als Forschungsstandort.

Warum ist das eine wesentliche Innovation?

Die meisten existierenden klinischen KI-Systeme (einschließlich kommerzieller Anbieter wie hAIppokrates/GreenBay) beschränken sich auf die **Beantwortung von Fragen** auf Basis einer statischen Wissensbasis. Unser Ansatz geht fundamental weiter:

Merkmal	**Konventionelle Systeme**	**UKLGPT-Ansatz**
Evidenzbasis	Statisch, manuell kuratiert	Dynamisch, automatisch aktualisiert
Literaturanbindung	Keine oder manuell	Automatisches SNOMED → Europe PMC Matching
Regelableitung	Hardcodiert oder manuell	LLM-extrahiert, kliniker-validiert, lernend
Studienmatching	Nicht vorhanden	Automatisch via ClinicalTrials.gov + Patientenprofil
Architektur	Monolithisch	Domainservice, andockbar an Patient *und* Wissensbasis

Der Domainservice transformiert UKLGPT von einem *reaktiven Frage-Antwort-System* zu einem **proaktiven klinischen Wissenspartner**, der nicht nur antwortet, sondern aktiv relevante Evidenz, abgeleitete Regeln und Studienoptionen in den klinischen Workflow einbringt.

Zusammenfassend entsteht dadurch ein kontrolliertes, domänenspezifisches KI-System, das die Leistungsfähigkeit von

LLMs mit der Präzision, Struktur und Sicherheit klinischer Informationssysteme verbindet.

3. Fachliches Zielbild & Anwendung {#3.-fachliches-zielbild-&-anwendung}

3.1 Zielbild UKLGPT – Der Klinische Informationsassistent {#3.1-zielbild-uklgpt-–-der-klinische-informationsassistent}

UKLGPT – auch synonym als **EMR GPT** oder unter dem Arbeitstitel **hAIppokrates** bekannt – wird als hochspezialisierter, klinischer Informationsassistent konzipiert. Sein primäres Ziel ist die gezielte und kontextbezogene Abfrage medizinischer, administrativer und logistischer Informationen aus der komplexen IT-Systemlandschaft des Universitätsklinikums Leipzig.

3.1.1 Grundlage des Systems und Datenintegration:

Die Leistungsfähigkeit von UKLGPT basiert auf der tiefen Integration und Aggregation von Daten aus zentralen klinischen IT-Systemen:

1. **HYDMedia (Dokumentenmanagementsystem):** Dies bildet die Basis für den Zugriff auf archivierte und aktuelle Patientenakten, Befunde, Arztbriefe, Aufklärungsdokumente und alle weiteren strukturierten sowie unstrukturierten Dokumente.

2. **UKLytics (Echtzeit-DWH / Analytics- und Datenplattform):** Über diese Plattform erfolgt die Anbindung an primäre klinische und administrative Systeme (Krankenhausinformationssysteme, Subsysteme). UKLytics ermöglicht die Verarbeitung von Echtzeitdaten (z. B. Laborwerte, Vitalparameter) und die Bereitstellung vorprozessierter, validierter Daten für die KI-Anwendung.

3. **Optionale und zukünftige Integration:** Das System ist modular konzipiert, um perspektivisch weitere essenzielle Plattformen wie das **EIDMS** (Enterprise Imaging and Document Management System), spezialisierte **Laborsysteme** oder weitere Subsysteme (z. B. Medikationssysteme, OP-Planung) einzubeziehen und deren Daten zu harmonisieren.

3.1.2 Kernfunktionalität und Bereitstellung:

Der Assistent verarbeitet komplexe medizinische und administrative Anfragen und stellt die benötigten Informationen **fall-, zeit- und situationsbezogen** in **natürlicher Sprache** bereit. Dies ermöglicht es dem klinischen Personal, sich schnell einen umfassenden Überblick zu verschaffen, ohne manuelle Recherchen in verschiedenen Systemen durchführen zu müssen.

3.1.3 Positionierung und Abgrenzung – Fokus auf Assistenz:

Die klare strategische Positionierung von UKLGPT ist die einer **Informations- und Orientierungsassistenz**. Es ist von essenzieller Bedeutung, dass das System **ausdrücklich nicht** darauf abzielt, medizinische Entscheidungen zu automatisieren oder zu ersetzen. Die finale diagnostische und therapeutische Verantwortung verbleibt stets beim behandelnden Arzt.

3.1.4 Zentrale Mehrwerte und Unterstützung klinischer Routinen:

Die Implementierung von UKLGPT soll direkt zur **Unterstützung klinischer Routinen** und zur Optimierung der Patientenversorgung beitragen, indem es folgende Hauptziele adressiert:

- **Schnellerer und verlässlicher Informationszugang:** Sofortige und präzise Beantwortung von Fragen zur Patientenhistorie, Medikation oder aktuellen Befunden.
- **Bessere Übersicht über komplexe Patientenakten:** Strukturierung und Zusammenfassung umfangreicher und über Jahre gewachsener digitaler Akten zu verständlichen Überblicken.
- **Reduktion administrativer Tätigkeiten:** Minimierung des Zeitaufwands für die Informationssuche und Dokumentation, wodurch mehr Zeit für die unmittelbare Patientenbetreuung gewonnen wird.
- **Erhöhung von Effizienz, Sicherheit und Kontinuität:** Durch verbesserte Informationsflüsse werden Doppeluntersuchungen, Medikationsfehler und Verzögerungen in der Versorgung reduziert. Zudem wird die Übergabe von Patienten (z. B. Schichtwechsel) durch schnelle Kontextbereitstellung sicherer.

3.2 Ausgangssituation und Analyse des Ist-Zustands {#3.2-ausgangssituation-und-analyse-des-ist-zustands}

Obwohl das System HYDMedia im klinischen Alltag ein unverzichtbares Werkzeug darstellt und intensiv genutzt wird, zeigt

die aktuelle Analyse signifikante Effizienzdefizite bei Such- und Rechercheprozessen. Diese sind in verschiedenen kritischen klinischen Situationen besonders ausgeprägt und führen zu einer unnötigen Belastung des medizinischen Personals sowie zu potenziellen Verzögerungen in der Patientenversorgung.

3.2.1 Identifizierte Engpässe und zeitintensive Szenarien

Die folgenden klinischen Abläufe sind überproportional von den bestehenden Suchproblemen betroffen:

- **Stationärer Aufnahmeprozess:** Hier muss in kürzester Zeit eine umfassende Anamnese erstellt werden. Das Suchen von Vorbefunden, Medikationshistorien und Aufklärungsunterlagen bindet wertvolle Zeit und verzögert den eigentlichen Behandlungsbeginn.
- **Ambulante Wiedervorstellungen:** Bei Patienten, die bereits behandelt wurden, ist die schnelle Verfügbarkeit der vollständigen Behandlungshistorie (inklusive Therapieanpassungen und Verlaufsberichten) essenziell. Die aktuelle Situation erfordert oft langwieriges Durchforsten der Akten.
- **Spezialisierte Fallbesprechungen (Konsile, Übergaben, Tumorboards):** In diesen interdisziplinären Kontexten ist die unmittelbare, lückenlose Bereitstellung aller relevanten Informationen (etwa frühere Therapien, detaillierte Bildgebungsbefunde, histologische Ergebnisse) kritisch für fundierte Entscheidungen. Zeitverlust durch Recherche ist hier besonders kontraproduktiv.

3.2.2 Häufig gesuchte und schwer auffindbare Informationen

Die typischen Fragestellungen, die in klinischen Prozessen regelmäßig zu langwierigen und ineffizienten Suchprozessen in der Patientendokumentation führen, umfassen ein breites Spektrum an kritischen Patienteninformationen. Diese Suchvorgänge binden wertvolle Zeit des medizinischen Personals, die für die direkte Patientenversorgung effektiver genutzt werden könnte, und bergen zudem das Risiko, dass relevante Informationen übersehen werden.

Im Detail handelt es sich dabei insbesondere um die folgenden komplexen und zeitaufwendigen Informationsbedürfnisse:

- **Detaillierte Historie früherer Therapien und Behandlungen:** Eine lückenlose und granulare Aufschlüsselung aller in der Vergangenheit durchgeführten therapeutischen Maßnahmen, chirurgischen Eingriffe, und interventionellen Verfahren. Dies schließt auch die genauen Zeitpunkte, die verantwortlichen Abteilungen oder Leistungserbringer sowie die unmittelbaren Ergebnisse und Komplikationen dieser Behandlungen ein.

- **Aktuelle und frühere Medikationspläne sowie die Compliance des Patienten:** Die vollständige Übersicht über alle jemals verordneten Medikamente (einschließlich OTC-Präparaten und Nahrungsergänzungsmitteln, sofern bekannt), deren Dosierungen, Darreichungsformen, Indikationsgründe, Verordnungszeiträume und etwaige Dosisanpassungen oder Absetzungen. Ein essenzieller Aspekt ist hierbei die Dokumentation der *Patienten-Compliance* – also der tatsächlichen Einnahmetreue des Patienten – sowie Gründe für eine potenzielle Non-Compliance.

- **Vollständige Liste von Allergien und Unverträglichkeiten:** Eine hochkritische Sammlung von Informationen, die nicht nur bekannte Medikamentenallergien (mit genauer Angabe der manifestierten Reaktion) umfasst, sondern auch Allergien gegen Lebensmittel, Umweltstoffe (z.B. Pollen, Tierhaare) oder medizinische Hilfsmittel (z.B. Latex, Kontrastmittel). Die Unterscheidung zwischen echten immunologisch vermittelten Allergien und bloßen Unverträglichkeiten/Nebenwirkungen ist hierbei für die zukünftige Behandlungsplanung unerlässlich.

- **Externe Vorbefunde und Konsiliarberichte:** Die vollständige Akquise und Integration aller relevanten Dokumente, die außerhalb der aktuellen Behandlungseinrichtung erstellt wurden. Hierzu zählen Berichte von niedergelassenen Ärzten, Fachärzten, anderen Kliniken (z.B. Rehabilitationseinrichtungen, Spezialzentren), radiologische Befunde, Laborergebnisse sowie Epikrisen vorangegangener stationärer Aufenthalte. Die zeitnahe Verfügbarkeit dieser Informationen ist oft ein Flaschenhals in der Weiterbehandlung.

- **Ärztliche Aufklärungen und Einwilligungen (Informed Consent):** Die revisionssichere und leicht auffindbare Dokumentation aller durchgeführten Patientenaufklärungen bezüglich geplanter Diagnostik, therapeutischer Maßnahmen oder operativer Eingriffe. Dies umfasst die Bestätigung, dass der Patient die Risiken und Alternativen verstanden hat und seine *informierte* Zustimmung (Einwilligung) erteilt hat, was sowohl juristisch als auch ethisch von höchster Relevanz ist.

- **Der minutiöse bisherige Behandlungsverlauf (klinische Entscheidungen und Begründungen):** Die chronologische und detailreiche Abbildung der *Decision-Making-Kette* im Behandlungsverlauf. Dazu gehören die initialen Verdachtsdiagnosen, die im Verlauf gestellten Differentialdiagnosen, die Begründung für die Auswahl spezifischer diagnostischer Verfahren (z.B. warum eine bestimmte Bildgebung gewählt wurde), die Herleitung und Begründung von therapeutischen Strategiewechseln sowie die Dokumentation der interdisziplinären Konsultationen und Fallbesprechungen. Dies ist der Kern der klinischen Evidenz und des Qualitätsmanagements.

3.2.3 Das spezifische Problem externer Befunde

Ein zentrales und oft unterschätztes Problem stellen **externe Befunde** dar. Hierzu zählen beispielsweise Arztbriefe von niedergelassenen Kollegen, detaillierte CT- oder MRT-Befunde von externen Radiologien oder spezialisierte Laborergebnisse. Obwohl diese Dokumente zeitnah digitalisiert und in HYDMedia integriert werden, sind sie für die weiterbehandelnde Person oft schwer zugänglich oder ihre Existenz ist unklar. Sie gehen *zeitversetzt* ein und werden nicht immer unmittelbar den relevanten Hauptdokumenten zugeordnet, was ihre Auffindbarkeit massiv erschwert. Dies führt dazu, dass klinisch relevante Informationen in der digitalen Ablage "verloren gehen".

3.2.4 Quantifizierung des Suchaufwands und der Ineffizienz

Die Fachbereich-Erhebung zeigt erhebliche Effizienzverluste: Pro Arzt und Tag fallen **15–50 Minuten reine Recherchezeit** an (2–5 Anfragen × 5–10 Min. Suchdauer, Einzelfälle bis 20 Min.). Die detaillierte Aufschlüsselung findet sich im Business Case (Kap. 0.5).

3.2.5 Folgen der Ineffizienz

Der hohe Zeitaufwand hat weitreichende negative Konsequenzen:

- **Nicht-Nutzung vorhandener Informationen:** Aufgrund des hohen Aufwands werden vorhandene, aber schwer auffindbare Informationen nicht immer konsequent genutzt. Dies kann die Qualität der klinischen Entscheidungen mindern.
- **Redundante Befragungen:** Anstelle der Recherche kommt es häufig zu erneuten Befragungen der Patienten zu bereits bekannten Fakten („Doppelfragen“). Dies verursacht nicht nur **zusätzlichen Zeitaufwand** (ca. 2–3 Minuten pro Doppelfrage), sondern führt auch zu **Frustration** auf Seiten des Patienten und des Personals.
- **Potenzielle Informationslücken:** Die größten Risiken bestehen in der Entstehung potenzieller Informationslücken, die für die Patientensicherheit relevant sein können, insbesondere bei der Übernahme von Patienten durch neue Behandler.

3.2.6 HYDMedia – Schnittstelle zu UKLGPT: Technische Basis und Herausforderungen

Das Dokumentenmanagementsystem (DMS) **HYDMedia** ist als primäre Quelle für unstrukturierte und semi-strukturierte Patientendokumente (Dokumenten-RAG) in das UKLGPT-Konzept integriert. Die Anbindung und Nutzung dieser Quelle unterliegt spezifischen technischen Rahmenbedingungen und weist Herausforderungen in der Datenqualität auf.

3.2.6.1 Technische und Systemische Basis der Dokumentenverwaltung

Die technische und systemische Basis der Dokumentenverwaltung am UKL ist durch die zentrale Speicherung und Metadatenverwaltung von Dokumenten gekennzeichnet, die den Input für die geplanten RAG-Architekturen (Patienten-RAG, Dokumenten-RAG, GraphRAG) bilden.

Aktueller Bestand und Speichermedium:

- **Dokumentenumfang:** Im zentralen Dokumentenarchivsystem, **HYDMedia**, sind aktuell rund **21 Millionen PDFs** gespeichert. Diese immense Datenmenge stellt die Basis für die digitalen Patientenakten dar.
- **Speicherarchitektur:** Die eigentlichen PDF-Dateien liegen physisch auf einem dedizierten **Isilon-Speicher** (Scale-out NAS).
- **Metadatenverwaltung:** Die zugehörigen Metadaten, die essenziell für die Klassifizierung, Suche und Archivierung sind, werden separat in der **HYDMedia Datenbank** verwaltet.

Metadaten-Generierung und Klassifizierungslogik:

Die korrekte Verschlagwortung (Metadaten-Generierung) ist ein kritischer Prozess, um die Dokumente im Kontext der Patientenakte verorten und später über RAG-Systeme effizient abfragen zu können.

- **Quelle der Metadaten (IDX-Dateien):** Die Metadaten werden in Form von IDX-Dateien entweder direkt durch den **Scanprozess** (manuelle Dokumente) oder durch elektronische Dokumente, die die zentrale **DMI-Klassifizierungsinstanz** durchlaufen, erzeugt.
- **Anbindung der Quellsysteme:** Eine Vielzahl klinischer und administrativer Quellsysteme speist die DMI-Klassifizierung und damit HYDMedia. Zu diesen Systemen zählen:
- **Klinische Systeme:** Copra5/6, Medavis, Viewpoint, Cardworks, IDScorer.
- **Spezialsysteme:** Ivoris (dessen Anbindung und Dokumentenübertragung ab Ende Februar erwartet wird).
- **Administrativ/ERP-Systeme:** Aktuell werden bereits **50% der SAP-Dokumente** (insbesondere PMDs – Patientennahe Medizinische Dokumente) über diesen Kanal verarbeitet.
- **Registrierung:** Ein wichtiger Aspekt der Qualitätssicherung ist, dass alle Dokumente, die diesen zentralen Klassifizierungsinstanz-Prozess durchlaufen, in der zentralen

Registry registriert werden, was ihre Auffindbarkeit und Validität sicherstellt.

Ausblick und Zukünftige Entwicklungen (Einfluss auf den Scan-Input):

Die gesamte Architektur wird sich signifikant durch die bevorstehende Digitalisierungsoffensive verändern:

- **Ablösung analoger Prozesse:** Mit der geplanten Einführung der **elektronischen Patientenakte (ePA)**, die Hand in Hand mit der umfassenden **M-KIS-Umstellung** (Meierhofer Klinisches Informationssystem) geht, wird ein Großteil der bisherigen analogen Workflows abgelöst.
- **Reduktion des Scan-Volumens:** Es wird erwartet, dass Scans physischer Dokumente nach dieser Umstellung **nahezu vollständig entfallen** werden. Zukünftig werden Dokumente primär elektronisch entstehen und direkt in die Systeme eingespielt, was die Datenqualität und -aktualität für die RAG-Systeme deutlich verbessern wird.

3.2.6.2 Voraussetzungen und Schnittstellen

- **HYDMedia API:** Die prinzipielle Verfügbarkeit einer **FHIR-Schnittstelle** in HYDMedia wurde von Dedalus bestätigt (Dokumentation liegt vor). Zudem existiert eine HYDMedia WebApp (z.B. unter s050008132:6444/webapp).
- **Zugriffsrisiko (Dedalus-Position):** Direkte Datenbankzugriffe von Fremdsystemen werden aus datenschutzrechtlichen (DSGVO-Konformität) und sicherheitstechnischen Gründen (ISO 13485, BSI/TR-ESOR/IDW 880 PS W) abgelehnt, da die Protokollierung der Zugriffe außerhalb der HYDMedia-Applikation nicht rechtskonform gewährleistet werden kann. Es wird ein rechtskonformer Weg über einen **Export via FHIR** vorgeschlagen.
- **Anspruch zur Umsetzung:** Die Architektur zielt darauf ab, Patientendaten in Echtzeit zu suchen, diese im Hintergrund zu indexieren und die Originaldokumente in einem übersichtlichen Viewer anzuzeigen. Die Bedenken von Dedalus sind zwar nachvollziehbar, bergen jedoch erhebliche **Implikationen** für die Umsetzung der Zielarchitektur, die Echtzeit-Suchen und Indexierung vorsieht:

Potenzielle Verzögerung und Komplexität (Zeit- und Kostenfaktor):

- Der vorgeschlagene rechtskonforme Weg über einen **"Export via FHIR"** muss erst konzipiert, entwickelt und von Dedalus implementiert werden (oder zumindest die Schnittstelle dafür bereitgestellt und für den Export konfiguriert werden).

- Dies ist nicht trivial, da FHIR primär für den Austausch von strukturierten Daten gedacht ist. Das Abrufen *aller* benötigten Dokumente und Patientendaten in Echtzeit oder für eine initiale Voll-Indexierung über eine FHIR-Export-Schnittstelle kann, je nach Datenmenge und Performance der Schnittstelle, **sehr zeitaufwendig** sein und die angestrebte **Echtzeitfähigkeit gefährden**.

- Es muss geklärt werden, ob die bestehende, von Dedalus bestätigte FHIR-Schnittstelle bereits einen massenhaften, performanten Export der relevanten *Originaldokumente* und der zugehörigen Metadaten unterstützt. Falls nicht, muss diese Funktionalität **nachgerüstet** werden.

Architektonische Einschränkungen (Echtzeit vs. Export):

- Die Zielarchitektur erfordert die Suche und Indexierung **in Echtzeit**. Ein periodischer "Export" (auch via FHIR) ist typischerweise asynchron und zeitversetzt und steht damit im Widerspruch zu einer echten Near-Real-Time-Lösung, wie sie für die Indexierung von Dokumentenänderungen oder neuen Dokumenten benötigt wird.

- Die Architektur muss nun so umgestaltet werden, dass sie entweder mit der Latenz des FHIR-Exports leben kann oder eine andere, von Dedalus akzeptierte, ereignisgesteuerte Benachrichtigung über Datenänderungen ("Change Data Capture" oder ähnliches über FHIR Subscriptions) implementiert wird, was zusätzliche Komplexität schafft.

"Red Flag"-Potenzial (Reibungsverlust/Risiko):

- Die Ablehnung direkter Zugriffe in Kombination mit dem Verweis auf einen (noch zu spezifizierenden) FHIR-Export ist ein **Risiko**, weil es die Projektverantwortung für einen zentralen Aspekt der Datenversorgung (Performance, Umfang, Echtzeit) **aus der Hand des UKLGPT-Projektteams in die Hand des Dedalus-Teams verlagert**.

- Es entsteht eine kritische **Abhängigkeit** von Dedalus für die Bereitstellung einer hochperformanten, exportfähigen und rechtskonformen Schnittstelle. Wenn diese Schnittstelle nicht schnell und in der benötigten Qualität geliefert werden kann, gerät das gesamte UKLGPT-Projekt in Verzug.

- Die Begründung mit der fehlenden Protokollierung bei direkten Datenbankzugriffen ist ein formaljuristisches Argument, das jedoch über die Notwendigkeit hinwegtäuscht, eine *gleichwertig performante* Alternative zu schaffen.

Fazit:

Das Projekt muss **sofort** die Spezifikation des "Exports via FHIR" mit Dedalus klären. Es muss sichergestellt werden, dass:

- Der Export sowohl die benötigten strukturierten Daten (für GraphRAG) als auch die Originaldokumente (für Dokumenten-RAG) in **Masse** und **Echtzeit** liefern kann.
- Die Performance der Schnittstelle die Anforderungen der Indexierung (insbesondere des initialen Ladevorgangs) erfüllt.
- Eine klare Zusage und ein Zeitplan von Dedalus für die Bereitstellung dieser exportfähigen FHIR-Schnittstelle vorliegen.

Die aktuelle Formulierung ist ein **direkter Weg in eine kritische Abhängigkeit und potenzielle Zeitverzögerung**, wenn die FHIR-Schnittstelle nicht bereits die notwendige Export-Performance bietet.

3.2.6.3 Herausforderungen in Vollständigkeit und Datenqualität

Problem	Details
Duplikate	Dokumente können durch unterschiedliche Zuwege (elektronisch aus Subsystemen, Scanstrecke, Verbucher) mehrfach in HYDMedia vorhanden sein (teilweise 3–5-fach) – oft mit unterschiedlichen Zeitstempeln. Eine nachträgliche Zuordnung von Duplikaten für nicht in der Registry registrierte Dokumente findet nicht statt.
Unvollständigkeit	**Labordaten** sind aktuell nicht an HYDMedia angebunden und müssen separat aus **UKLytics** bezogen werden. Die **SAP DOK Migration** (Dokumente der letzten 20 Jahre) ist zwar geplant, erfordert aber einen mehrmonatigen Prozess.
Mangelnde Strukturierung	Dokumente, die nicht über die DMI-Schnittstelle eingebunden wurden, sind **nicht IHE-konform** und **nicht KDL-klassifiziert** (Registry). Dies erschwert eine semantische Durchsuchbarkeit. Fehler bei der händischen Eingabe in der Nomenklatur beim Verbucher können ebenfalls zu Suchfehlern führen.
Fehlende OCR-Basis	Für eine semantische Suche in den 21 Mio. PDFs ist eine **OCR-Erkennung** (Optical Character Recognition) zwingend erforderlich, da handschriftliche oder

Problem	Details
	bildbasierte Inhalte aktuell nicht durchsuchbar sind.

3.2.6.4 Strategische Entscheidungen und ToDos

Bereich	ToDo / Maßnahme
OCR-Integration	Installation der OCR-Funktionalität für HYDMedia (geplant für 17.02.2026).
UKLGPT API-Anbindung	Prüfung, wie über die **FHIR Connectoren** eine *DocumentReference* zur Originaldokument-Datei (Binary) aufgelöst werden kann. Ziel ist es, die Originaldokumente aus dem DMS in den RAG-Prozess zu integrieren. Ggf. Anbindung einer weiteren HYDMedia API (z.B. REST-API für die OCR-Datenbank).
Archiv-Strategie (DMI/ AVP)	Es ist zu prüfen, ob zur vollen KI-Readiness eine erweiterte AVP-Version (**AVP Infinity**) notwendig ist. Dies würde einen Release-Wechsel der DMI-Klassifizierungsinstanz bedeuten und potenziell HYDMedia als Dokumentenarchiv ablösen, was jedoch ein eigenes, komplexes Projekt (inkl. Infrastruktur und Migration) nach sich ziehen würde. **Prüfung der WebAPI von DMI (x-tention).**
RAG-Strategie	Umsetzung konkreter, fallbezogener Abfragen über RAG (ggf. auch Long Context RAG), um die Inhalte der Dokumente semantisch zu erschließen und bereitzustellen.
SNOMED-Autotagging	Aufbau der NER-Pipeline (MedCAT/cTAKES/ SciSpacy) zur automatischen SNOMED-CT-Annotation aller Patientendokumente (vgl. Kap. 14.2.2).
Europe PMC-Anbindung	Integration der Europe PMC REST-API für automatisches Evidence-Matching basierend auf SNOMED-Diagnose-Codes (vgl. Kap. 7.2.1.1).
Snowstorm Terminologie-Server	Evaluierung On-Premise-Snowstorm (Phase 2) vs. öffentlicher IHTSDO-Endpunkt (Phase 1) für SNOMED-CT $lookup/$expand (vgl. Kap. 14.2.1).

3.3 Use Cases von UKLGPT {#3.3-use-cases-von-uklgpt}

Das vorgeschlagene UKLGPT-System ist ein spezialisierter, KI-gestützter klinischer Informationsassistent, der darauf ausgelegt ist, die klinischen Arbeitsabläufe signifikant zu optimieren und die Effizienz in der medizinischen Dokumentation und Recherche zu steigern. Es fokussiert sich auf die Verarbeitung und Bereitstellung von medizinischen Daten aus der elektronischen Patientenakte (ePA).

3.3.1 Use Case 1: Intelligente Dokumenten- und Befundrecherche

3.3.1.1 Problemstellung und Herausforderung:

Die klinische Recherche ist derzeit ein mühsamer und zeitaufwendiger Prozess. Medizinische Informationen sind in einer Vielzahl von Systemen und Formaten fragmentiert:

- **HYDMedia:** Archivsystem für gescannte Dokumente und ältere Berichte.
- **Laborsysteme (z. B. LabCentre):** Spezifische Datenbanken für Laborwerte und Mikrobiologie.
- **Patientenorganizer/Klinische Arbeitsplätze:** Ablageorte für Arztbriefe, Entlassungsberichte und bildgebende Befunde.

Die Notwendigkeit, Informationen aus diesen verschachtelten Strukturen und heterogenen Ablagen manuell zusammenzuführen, führt zu einem hohen Zeitaufwand und erhöht das Risiko von Suchfehlern oder dem Übersehen relevanter Informationen. Dies bindet wertvolle klinische Zeit, die für die Patientenversorgung benötigt wird.

3.3.1.2 Innovative Lösung durch UKLGPT:

UKLGPT fungiert als zentrale, intelligente Suchschnittstelle, die es dem klinischen Personal ermöglicht, Abfragen in **natürlicher, umgangssprachlicher Sprache** zu stellen. Anstatt komplexe Suchparameter oder Systemmasken bedienen zu müssen, können Ärzte und Pflegekräfte direkte klinische Fragen formulieren.

3.3.1.3 Beispiele für erweiterte Abfragen in natürlicher Sprache:

- „Welche Medikamente, insbesondere Antihypertensiva, erhielt der Patient während seines letzten stationären Aufenthalts vor dem aktuellen?"

- „Wann genau wurde die letzte Magenspiegelung (Ösophago-Gastro-Duodenoskopie) durchgeführt und war der Befund auffällig (z.B. Hinweise auf Barrett-Ösophagus oder Ulcera)?"
- „Waren die Leberwerte (insbesondere GOT, GPT, Gamma-GT) in den letzten sechs Monaten schon einmal so hoch wie aktuell, und welche Maßnahmen wurden damals ergriffen?"
- „Zeige mir den letzten CT-Befund aus der Radiologie, der die Abdominalregion betrifft, und vergleiche die Größe der Läsion X mit dem Vorbefund."
- „Was war der detaillierte letzte Tumorboard-Beschluss vom [Datum], einschließlich der empfohlenen Chemotherapie-Linie und des geplanten Follow-up?"
- „Liste alle dokumentierten Allergien und Unverträglichkeiten des Patienten auf und zeige die Quelle der Dokumentation."

Der Assistent gewährleistet dabei stets die strikte Einhaltung der Datenschutz- und Berechtigungsrichtlinien, indem er **ausschließlich** die Dokumente, Befunde und strukturierten Daten durchsucht, für die der jeweilige Nutzer **explizit berechtigt** ist. Die Ergebnisse werden nicht nur als rohe Daten, sondern **kontextualisiert und übersichtlich aufbereitet** zurückgeliefert.

3.3.1.4 Erzielter Nutzen und Mehrwert:

- **Massive Zeitersparnis:** Die Recherchezeit wird von potenziell mehreren Minuten auf Sekunden reduziert. Dies ermöglicht eine schnellere Entscheidungsfindung am Patientenbett.
- **Signifikante Reduktion von Fehlern:** Such- und Übertragungsfehler (z. B. Abschreiben falscher Werte) werden minimiert, da die Informationen direkt aus der Quelle generiert werden.
- **Optimierte Entscheidungsgrundlagen:** Durch den schnellen, vollständigen und kontextualisierten Zugriff auf die Gesamtheit der Patientenakte werden die diagnostischen und therapeutischen Entscheidungen im klinischen Alltag fundierter.

3.3.2 Use Case 2: Automatisierte Zusammenfassung von Patientenakten

Angesichts der zunehmenden Komplexität und des Umfangs elektronischer Patientenakten wird eine schnelle Einarbeitung in den Fall eines Patienten kritisch.

3.3.2.1 Funktionalität:

UKLGPT nutzt seine Fähigkeit zur Sprachverarbeitung, um umfangreiche Akten (z.B. mehrere hundert Seiten) zu analysieren

und eine strukturierte, prägnante Zusammenfassung zu erstellen. Diese Zusammenfassung hebt die **wesentlichen klinischen Inhalte** hervor, kategorisiert nach:

- **Anamnese:** Wichtige Vorerkrankungen, Risikofaktoren, Familienanamnese.
- **Aktuelle und frühere Diagnosen:** Chronologische Auflistung der relevantesten Diagnosen.
- **Therapien:** Durchgeführte Behandlungen, Operationen, Medikationshistorie.
- **Krankheitsverlauf:** Wichtige Ereignisse und Entwicklung des Zustands über die Zeit.

3.3.2.2 Erzielter Nutzen und Mehrwert:

- **Schneller, umfassender Überblick:** Dies ist essenziell bei der **Aufnahme** neuer Patienten, bei der **Schichtübergabe** im interdisziplinären Team oder bei der Durchführung eines **Konsils** durch einen Facharzt.
- **Entlastung des klinischen Personals:** Die manuelle Erstellung von Zusammenfassungen oder das langwierige Einarbeiten in die Akte entfällt oder wird stark verkürzt.
- **Verbesserte Kontinuität:** Der Informationsverlust bei Personalwechsel oder Verlegung wird minimiert, was die Qualität und Kontinuität der Behandlung verbessert.

3.3.3 Use Case 3: Unterstützung bei der Qualitätssicherung und Dokumentations-Compliance (Optional)

Als optionales Modul kann UKLGPT zur Überprüfung der formalen Qualität der klinischen Dokumentation eingesetzt werden.

3.3.3.1 Funktionalität:

Das System kann Dokumente (z.B. OP-Berichte, Anamnesebögen) automatisch scannen und Hinweise auf **formale Unvollständigkeiten oder Inkonsistenzen** geben. Beispiele hierfür sind:

- **Fehlende Pflichtangaben:** Prüfung, ob alle gesetzlich oder intern vorgeschriebenen Felder (z.B. Unterschrift, Aufklärungsvermerk, ICD-Codes) vorhanden sind.
- **Inkonsistenzen:** Erkennung von Widersprüchen (z.B. unterschiedliche Angaben zu Allergien in verschiedenen Dokumenten).

3.3.3.2 Erzielter Nutzen und Mehrwert:

- **Unterstützung bei der Nachbereitung:** Ärzte können unmittelbar nach der Dokumentation auf fehlende Angaben hingewiesen und zur Korrektur aufgefordert werden.
- **Vorbereitung auf Audits:** Das System unterstützt bei der Einhaltung von Dokumentationsrichtlinien und der Vorbereitung auf interne und externe Audits.

3.3.4 Use Case 4: Automatisches Evidence-Matching und SNOMED-Tagging

3.3.4.1 Problemstellung und Herausforderung:

Kliniker stehen vor zwei fundamentalen Herausforderungen:

- **Fehlende Evidenz-Anbindung:** Aktuelle wissenschaftliche Erkenntnisse (Leitlinien, Studien) sind vom klinischen Arbeitsplatz aus schwer zugänglich. Die manuelle Recherche in PubMed/Europe PMC dauert pro Fragestellung 10–30 Minuten und findet daher im Alltag kaum statt.
- **Fehlende semantische Erschließung:** Patientendokumente (Arztbriefe, Befunde, OP-Berichte) sind als unstrukturierter Freitext gespeichert und weder standardisiert kodiert noch mit internationalen Terminologien verknüpft. Eine gezielte Suche nach „allen Patienten mit Herzinsuffizienz" erfordert derzeit manuelles Durchlesen statt einer Facetten-Suche.

3.3.4.2 Innovative Lösung durch UKLGPT:

A) Proaktives Evidence-Matching:

Beim Öffnen eines Patientenfalls analysiert UKLGPT automatisch die vorliegenden Diagnosen, Medikationen und Prozeduren und zeigt dem Kliniker drei Informationsebenen an:

1. **Relevanteste interne Dokumente** – die 5 wichtigsten Patientendokumente, gerankt nach semantischer Ähnlichkeit und Aktualität.
2. **Passende interne Leitlinien/SOPs** – hauseigene Behandlungspfade und Protokolle, die zu den aktiven Diagnosen passen.
3. **Aktuelle externe Evidenz (Europe PMC)** – die neuesten und meistzitierten Publikationen zu den Hauptdiagnosen des Patienten, abgerufen über die Europe PMC REST-API (vgl. Kap. 7.2.1.1).

Beispiel-Szenario: Ein Patient mit der Diagnose „Herzinsuffizienz (NYHA III)" wird aufgerufen. UKLGPT zeigt automatisch: * Interne Dokumente: letzter Echokardiographie-Befund, Entlassungsbrief Kardiologie, Medikamentenplan * Interne Leitlinie: UKL-SOP „Management der Herzinsuffizienz" * Externe Evidenz: 3 aktuelle

Studien zu HFrEF-Therapie aus Europe PMC (mit DOI-Link, Journal, Zitationen)

B) Automatisches SNOMED-Tagging:

Jedes in das System eingehende Patientendokument wird automatisch durch eine NER-Pipeline (Named Entity Recognition) analysiert und mit SNOMED-CT-Codes annotiert (vgl. Kap. 14.2.2). Der Kliniker profitiert davon durch:

- **Facettierte Suche:** „Zeige mir alle Dokumente dieses Patienten, die sich auf `Diabetes mellitus Typ 2` beziehen" – unabhängig von der im Freitext verwendeten Formulierung.
- **Automatische Diagnose-Verlinkung:** Dokumente werden im GraphRAG mit den entsprechenden SNOMED-Konzepten verknüpft und erscheinen als Kontext bei verwandten Abfragen.
- **Cross-Patient-Analyse (anonymisiert, Forschungsmodus):** Identifikation ähnlicher Fälle über SNOMED-basierte Kohortenbildung.

3.3.4.3 Erzielter Nutzen und Mehrwert:

- **Evidenzbasierte Entscheidungen ohne Zusatzaufwand:** Der Kliniker erhält aktuelle Evidenz proaktiv, ohne selbst recherchieren zu müssen.
- **Zeitersparnis Evidenzrecherche:** Von Ø 15 Min. manueller PubMed-Suche auf 0 Min. (automatische Bereitstellung).
- **Verbesserte Dokumentenqualität:** Durch SNOMED-Tagging werden alle Dokumente maschinenlesbar und international standardisiert kodiert.
- **Grundlage für klinische Forschung:** Die SNOMED-annotierten Dokumente bilden eine strukturierte Datenbasis für retrospektive Studien und Qualitätssicherung.

3.4 Wichtiger Hinweis, Abgrenzung und Haftungsausschluss zur Nutzung des UKLGPT-Systems (UKLGPT) {#3.4-wichtiger-hinweis,-abgrenzung-und-haftungsausschluss-zur-nutzung-des-uklgpt-systems-(uklgpt)}

Das **UKLGPT-System** (im Folgenden auch als **UKLGPT** bezeichnet) ist ein hochentwickeltes KI-gestütztes Tool, das **ausschließlich zur Unterstützung** des medizinischen Fachpersonals in der täglichen klinischen Praxis konzipiert wurde. Seine Funktion ist strikt auf die **Bereitstellung von Informationen und assistierenden Textvorschlägen** begrenzt.

3.4.1 Abgrenzung der Verantwortlichkeiten:

Es muss unmissverständlich klargestellt werden, dass **UKLGPT keine eigenständige fachliche Bewertung** medizinischer Sachverhalte vornimmt. Das System ist **nicht befugt**, Diagnosen zu stellen, verbindliche Therapieempfehlungen auszusprechen oder medizinische Entscheidungen zu treffen. Die durch das System generierten oder bereitgestellten Informationen, Analysen oder Dokumentationsvorschläge dienen lediglich als **Hilfsmittel** und müssen stets einer **gründlichen kritischen Prüfung** durch den behandelnden Arzt oder das klinische Fachpersonal unterzogen werden.

3.4.2 Haftungsausschluss und alleinige Verantwortung des Anwenders:

Die **alleinige und vollständige Verantwortung** für die Interpretation der vom System gelieferten Informationen, die daraus abgeleiteten klinischen Entscheidungen, die daraus resultierenden medizinischen Maßnahmen sowie die gesamte medizinische Dokumentation **verbleibt uneingeschränkt beim behandelnden Arzt oder dem medizinischen Fachpersonal.**

Die Nutzung von UKLGPT entbindet das medizinische Personal **in keiner Weise** von seiner Sorgfaltspflicht, seinem Fachwissen und seiner klinischen Urteilsfähigkeit. Im Falle von Fehlern, Auslassungen oder nicht sachgerechten Interpretationen der Systeminformationen liegt die Haftung allein beim Anwender.

3.4.3 Nutzungsbedingungen:

Die Inanspruchnahme und Verwendung des UKLGPT-Systems hat **zwingend und ausnahmslos** im Rahmen der geltenden **internen Richtlinien** der Einrichtung sowie der strengen **Datenschutz- und Sicherheitsvorgaben** zu erfolgen. Jegliche missbräuchliche oder nicht autorisierte Nutzung ist untersagt und kann zu internen Konsequenzen führen.

4. Frontend / Benutzeroberfläche (UI) {#4.-frontend-/-benutzeroberfläche-(ui)}

4.1 Rolle und Funktion des Frontends {#4.1-rolle-und-funktion-des-frontends}

Das Frontend agiert nicht als ein **generisches, offenes Chat-Interface**, sondern ist primär als ein **klinisches Interaktions- und**

Steuerungsinstrument konzipiert und implementiert. Es ist die zentrale Schnittstelle, die eine sichere, kontextbezogene und zweckgebundene Nutzung des Systems gewährleistet. Seine Hauptaufgabe ist es, sicherzustellen, dass jede Nutzeranfrage und die daraus resultierende Interaktion in einem **klaren fachlichen, organisatorischen und rechtlichen Kontext** der Patientenversorgung eingebettet sind. Dies verhindert den missbräuchlichen oder unkontrollierten Einsatz außerhalb definierter klinischer Prozesse. Die Oberfläche dient somit als **"Guardrail"** für die Anwendung.

4.2 Zentrale Funktionen der Benutzeroberfläche {#4.2-zentrale-funktionen-der-benutzeroberfläche}

Die UI muss essenzielle Funktionen bereitstellen, die für einen nahtlosen und sicheren klinischen Workflow unerlässlich sind:

- **Sichere Authentifizierung und Autorisierung:** Die Anmeldung erfolgt obligatorisch über **Single Sign-on (SSO)**, typischerweise integriert in das **Active Directory** oder ein vergleichbares Identity- und Access-Management-System der klinischen Einrichtung.

- **Patienten- und Encounter-Kontext:**

- Die UI muss die **Auswahl des aktiven Patienten** (Patienten-ID, Name) und des spezifischen **Begegnungs- oder Fall-Kontexts (Encounter)** deutlich anzeigen und verwalten.

- Alle Interaktionen sind fest an diesen Kontext gebunden.

- **Sichtbare Zweckbindung:** Die primäre Intention der Nutzung muss durchgehend sichtbar sein (**klinische Versorgung, Entscheidungsassistenz**), um die Abweichung von der medizinischen Nutzung zu verhindern.

- **Strukturierte und evidenzbasierte Darstellung der Antwort:** Die Ergebnisse der Assistenzsysteme dürfen nicht als monolithischer Textblock präsentiert werden, sondern müssen in klar definierte, nachvollziehbare Komponenten zerlegt werden:

- **Patientenfakten und -historie:** Übersichtliche, **strukturierte und zeitlich geordnete** Anzeige der relevanten Basisdaten (Diagnosen, Medikation, Allergien etc.), die in die Antwort eingeflossen sind.

- **Dokumentenausschnitte und Quellennachweis:** Direkte Einbettung der spezifischen **Textausschnitte** aus den Patientenakten oder sonstigen Quelldokumenten, auf denen die Antwort basiert, inklusive **klarer Kennzeichnung der Quelle** (z.B. Name des Dokuments, Datum).

- **Leitlinien und Evidenz:** Explizite Nennung und Verlinkung der zugrundeliegenden medizinischen **Leitlinien, Fachliteratur oder Evidenz-Quellen**, die zur Generierung der Empfehlung herangezogen wurden.

- **Proaktives Evidence-Panel (Kap. 7.2.1.2):** Beim Öffnen eines Patientenfalls wird automatisch ein **Evidence-Panel** angezeigt, das drei Bereiche umfasst:

- **Interne Dokumente:** Top-5 relevanteste Patientendokumente (semantisch gerankt).

- **Interne Leitlinien:** Passende SOPs und Hausprotokolle.

- **Externe Evidenz (Europe PMC):** Top-5 aktuelle Publikationen pro Hauptdiagnose (Titel, Journal, Jahr, Zitationen, DOI-Link). Die Evidenz-Karten sind klickbar und öffnen den Volltext bei Open-Access-Artikeln.

- **SNOMED-Tag-Anzeige:** Patientendokumente zeigen ihre automatisch vergebenen **SNOMED-CT-Tags** als klickbare Badges an, die eine facettierte Navigation ermöglichen (z. B. „Zeige alle Dokumente mit Tag *Herzinsuffizienz*").

4.3 Essenzielle Sicherheits- und Transparenzelemente {#4.3-essenzielle-sicherheits--und-transparenzelemente}

Um die klinische Verantwortung klar beim Arzt zu belassen und Compliance zu gewährleisten, muss das Frontend folgende permanente Sicherheitselemente aufweisen:

- **Haftungsausschluss und Hinweis zur Entscheidungsfindung:** Ein **permanenter und unübersehbarer Hinweis** muss klarstellen, dass es sich um eine **Informationsassistenz** handelt und **keine automatische Entscheidungsgewalt** oder Handlungsanweisung abgeleitet werden darf. Die **klinische Letztverantwortung** verbleibt beim behandelnden Personal.

- **Kennzeichnung von Unsicherheiten und Datenlücken:** Das System muss **explizit visualisieren**, wenn die Antwort auf **unsicheren oder widersprüchlichen Daten** beruht, wenn **Daten fehlen** oder wenn die zugrundeliegende Evidenz **gering** ist. Dies erhöht die Sorgfaltspflicht des Nutzers.

- **Explizite Anzeige bei Notfallzugriffen (Break-Glass):** Wenn ein Zugriff auf Daten oder Funktionalitäten außerhalb der normalen Autorisierung (z.B. im Notfall) unter Verwendung eines **"Break-Glass"-Mechanismus** erfolgt, muss dieser Zustand **unmissverständlich** in der UI kenntlich gemacht werden. Dies dient der nachträglichen revisionssicheren Protokollierung und Überwachung dieser kritischen Zugriffe.

5. Applikations- und Orchestrierungsschicht {#5.-applikations--und-orchestrierungsschicht}

Die Applikations- und Orchestrierungsschicht bildet das zentrale Steuerungs- und Logikzentrum des Systems. Sie fungiert als die kritische Middleware, die die Rohdaten und Fähigkeiten der darunterliegenden Schichten (Datenzugriff und Basis-KI-Modelle) in strukturierte, kontextsensitive und nützliche Anwendungen für den Endnutzer umsetzt. Ihre Hauptaufgabe ist es, die Komplexität des Backends zu abstrahieren und eine nahtlose, intelligente Interaktion zu ermöglichen.

5.1 Verantwortung {#5.1-verantwortung}

Diese Schicht trägt die Verantwortung für die gesamte Interaktion und die Qualität der ausgegebenen Ergebnisse. Ihre Entscheidungsmechanismen sind entscheidend für die Nutzererfahrung und die Korrektheit der medizinischen Informationen:

- **Analyse und Interpretation der Anfrage:** Sie entscheidet, wie eine eingehende Nutzeranfrage (natürliche Sprache, strukturierter Input, oder API-Aufruf) semantisch interpretiert, in Absichten (Intents) zerlegt und auf die verfügbaren Domänen abgebildet wird.

- **Strategische Datenquellenauswahl:** Sie wählt dynamisch und kontextabhängig aus, welche spezifischen Datenquellen (z. B. FHIR-Server, Dokumenten-Repository, Leitlinien-Datenbank) für die Beantwortung der Anfrage genutzt werden müssen. Dies optimiert die Latenz und die Relevanz des Abfrageprozesses.

- **Ergebnisaggregation und -veredelung:** Sie steuert, wie die von den verschiedenen Subsystemen generierten Ergebnisse kombiniert, gewichtet, normalisiert und in ein kohärentes, benutzerfreundliches Format gebracht werden. Sie stellt sicher, dass die Präsentation den definierten Qualitäts- und Sicherheitsstandards entspricht.

5.2 Kernkomponenten {#5.2-kernkomponenten}

Die Leistungsfähigkeit dieser Schicht beruht auf dem Zusammenspiel mehrerer spezialisierter Module:

- **Domain- und Intent-Klassifikator (DIC):** Dieses Modul analysiert die eingehende Anfrage und identifiziert die thematische Domäne (z. B. Diagnostik, Therapieempfehlung, Medikationsmanagement) sowie die konkrete Absicht des Nutzers (z. B. "Medikament X suchen", "Wechselwirkungen prüfen"). Es ist die erste Weiche im Verarbeitungsprozess.

- **Kontext-Assembler:** Dieses essenzielle Modul sammelt und hält den *state* der aktuellen Interaktion. Es assembliert alle notwendigen Kontextparameter, die die Suchergebnisse personalisieren und präzisieren müssen:

- **Patientenkontext:** Demografische Daten, Vorerkrankungen, Allergien, aktuelle Medikation (aus FHIR).

- **Encounter-Kontext:** Aktuelle Behandlungssituation, Standort, behandelndes Team.

- **Zweck-Kontext:** Der spezifische Grund der Abfrage (z. B. Entscheidungshilfe, Dokumentationsprüfung, Schulung).

- **Tool- und Pipeline-Router:** Basierend auf dem **Intent** und dem **Kontext** entscheidet dieser Router, welche der verfügbaren KI- oder Datenpipelines (Tools) in welcher Reihenfolge aktiviert werden müssen, um die Anfrage optimal zu beantworten (z. B. RAG-Pipeline vs. Berechnungs-Engine vs. LLM-Generierung).

- **Regel- und Guardrail-Engine:** Dieses Modul implementiert die kritischen Sicherheits- und Qualitätsregeln. Es stellt sicher, dass generierte Antworten medizinisch ethisch, gesetzlich konform und institutionsspezifisch korrekt sind. Es fungiert als ein finaler Filter, der potenziell unsichere oder nicht autorisierte Ausgaben korrigiert oder blockiert. Es umfasst Business-Logik und harte Stopps (Guardrails) basierend auf Richtlinien.

5.3 Orchestrierter Ablauf (vereinfacht) {#5.3-orchestrierter-ablauf-(vereinfacht)}

Der Prozess einer Anfrage durchläuft eine stringente, hochgradig orchestrierte Kette von Schritten, um Relevanz, Sicherheit und Qualität zu gewährleisten:

- **Eingang Nutzeranfrage:** Erfassung der initialen Anfrage über die UI oder API.

- **Domain- und Intent-Klassifikation:** Das DIC-Modul identifiziert *was* der Nutzer will und *worum* es geht.
- **Sicherheits- und Kontextprüfung:** Die Guardrail-Engine und der Kontext-Assembler prüfen die Berechtigungen des Nutzers und reichern die Anfrage mit relevanten Patientendaten an (z. B. "Prüfe Medikation für Patient Max Mustermann").
- **GraphRAG-Abfragen (FHIR-Fakten):** Die erste datenintensive Abfrage wird gestartet. Sie nutzt die Graph-Struktur von FHIR-Daten, um präzise, faktische Patientendaten und Beziehungen abzurufen (z. B. Labordaten, Diagnosen).
- **Dokumenten-RAG (HYDMedia):** Parallel oder sequenziell erfolgt die Abfrage der unstrukturierten oder semistrukturierten Hausdokumentation (z. B. Entlassbriefe, OP-Berichte in HYDMedia), um tiefere klinische Einsichten zu gewinnen.
- **Leitlinien-RAG:** Abfrage der externen oder internen medizinischen Leitlinien, um die Faktenchecks und Empfehlungen auf Basis des aktuellen medizinischen Standards zu untermauern.
- **Ergebnisaggregation und Generierung:** Die Ergebnisse der RAG-Pipelines werden aggregiert und dem LLM zur finalen Generierung einer kohärenten Antwort vorgelegt.
- **Übergabe an Qualitätssicherung und Final Guardrail:** Die generierte Antwort wird abschließend durch die Regel-Engine auf Sicherheits- und Konformitätsverstöße geprüft, bevor die Ausgabe an den Nutzer erfolgt.

6. Qualitätssicherung der KI-Antworten: Die Prompt-Pipeline als Kontrollinstanz {#6.-qualitätssicherung-der-ki-antworten:-die-prompt-pipeline-als-kontrollinstanz}

Die Gewährleistung einer hohen und verlässlichen Antwortqualität im Kontext sensitiver Anfragen, insbesondere im medizinischen Bereich, ist ein zentrales Anliegen. Dies erfordert eine strikte Kontrolle, die über die inhärente Unsicherheit und statistische Natur von Large Language Models (LLMs) hinausgeht.

6.1 Grundprinzip: Determinismus durch Orchestrierung {#6.1-grundprinzip:-determinismus-durch-orchestrierung}

Die Antwortqualität wird bewusst nicht allein dem Sprachmodell (LLM) überlassen. Stattdessen wird sie durch eine **mehrstufige, deterministische Prompt-Orchestrierung** abgesichert. Dieses Grundprinzip transformiert den Prozess von einem unsicheren, statistischen Generierungsprozess in eine kontrollierte Abfolge von validierten Schritten, die durch spezifische, aufgabenspezifische Prompts gesteuert werden. Jeder Schritt dient der Verfeinerung, Validierung oder Kontextanreicherung, um die Wahrscheinlichkeit von Fehlern zu minimieren.

6.2 Domain-aware Prompt Orchestration: Fachliche Weichenstellung {#6.2-domain-aware-prompt-orchestration:-fachliche-weichenstellung}

Um eine unnötige Belastung spezialisierter Validierungsschritte zu vermeiden und die Sicherheit zu erhöhen, wird jede eingehende Anfrage einer verpflichtenden **fachlichen Domäne** zugeordnet.

- **Domänen-Filter:** Die Architektur sieht vor, dass nur Anfragen, die eindeutig als **medizinisch** klassifiziert werden, die nachfolgenden, hochspezialisierten und sicherheitskritischen Stufen der medizinischen Pipeline durchlaufen.
- **Sicherheitsaspekt:** Nicht-medizinische Anfragen werden über eine separate, weniger restriktive Pipeline bearbeitet. Dies stellt sicher, dass medizinische Validierungslogik nur dort angewendet wird, wo sie zwingend erforderlich ist, und erlaubt gleichzeitig eine angepasste Bearbeitung anderer Anfragetypen (z.B. organisatorischer oder technischer Natur).

6.3 Detaillierte Medizinische Prompt-Pipeline {#6.3-detaillierte-medizinische-prompt-pipeline}

Die medizinische Pipeline ist eine kritische Kontrollstruktur, die sequenziell Abarbeitung und Validierung gewährleistet (vgl. Kap. 6.3). Sie besteht aus den folgenden obligatorischen Schritten:

- **Domain Classification:** Klassifizierung der Anfrage, um die Zugehörigkeit zur medizinischen Domäne zu bestätigen und die Notwendigkeit der Aktivierung dieser Pipeline festzustellen.
- **Intent- & Safety-Triage:** Bestimmung der Absicht (z.B. Informationssuche, Entscheidungsunterstützung, Diagnose)

und erste Sicherheitsüberprüfung, um gefährliche oder potenziell missbräuchliche Anfragen frühzeitig zu erkennen und abzuweisen.

- **Strukturierte Kontext-Erhebung:** Das System führt einen strukturierten Dialog, um alle notwendigen Informationen (Symptome, Dauer, Vorerkrankungen etc.) in einem vorab definierten Format zu erfassen, um die Grundlage für eine präzise Antwort zu schaffen.

- **Kontext-Normalisierung (FHIR / SNOMED):** Die erhobenen und textuellen Kontextdaten werden in standardisierte medizinische Terminologien (wie FHIR für Datenstruktur und SNOMED CT für Konzepte) überführt. Dies eliminiert Unklarheiten in der natürlichen Sprache und ermöglicht eine exakte Wissenssuche.

- **Medizinische Kategorisierung:** Die normalisierten Daten werden zur präzisen fachlichen Einordnung verwendet, um die nachfolgenden Such- und Validierungsschritte auf relevante Fachgebiete zu beschränken.

- **Wissens-Retrieval (Graph + RAG):** Über ein Retrieval-Augmented Generation (RAG)-System erfolgt die Suche nach Evidenz. Hierbei wird sowohl ein strukturierter medizinischer Wissens-Graph als auch eine Datenbank von validierten Fachtexten durchsucht, um eine faktische Grundlage zu schaffen.

- **Reranking:** Die Ergebnisse des Wissens-Retrievals werden nach Relevanz, Aktualität und Evidenzgrad neu bewertet, um sicherzustellen, dass nur die qualitativ hochwertigsten Quellen in die Synthese einfließen.

- **Medizinische Validierung:** Ein spezifisches LLM-Modul oder ein Regelwerk überprüft die gesammelten Informationen und die geplante Antwort auf Konsistenz mit etablierten medizinischen Leitlinien und Standards.

- **Antwort-Synthese:** Erst in diesem letzten Schritt generiert das Sprachmodell die finale, strukturierte und verständliche Antwort, basierend *ausschließlich* auf den zuvor validierten und angereicherten Informationen.

6.4 Ergebnis und Wirkung auf die Antwortqualität {#6.4-ergebnis-und-wirkung-auf-die-antwortqualität}

Die Implementierung dieser Prompt-Pipeline führt zu fundamentalen Verbesserungen der Systemleistung und der Verlässlichkeit:

- **Signifikante Reduktion von Halluzinationen:** Durch die strenge Bindung der Antwort-Synthese an das durch Retrieval

gesicherte und normalisierte Wissen wird die Tendenz des LLM, ungesicherte Informationen zu generieren, stark reduziert.

- **Klare Trennung von Information und Entscheidung:** Die Architektur erzwingt eine Differenzierung zwischen dem Bereitstellen von neutralen, evidenzbasierten **Informationen** und dem Treffen von **Entscheidungen** (was außerhalb der Zuständigkeit des Systems liegt). Dies ist essenziell für die Einhaltung ethischer und rechtlicher Standards.

- **Reproduzierbare, auditierbare Antworten:** Jeder Schritt der Pipeline ist protokollierbar und nachvollziehbar. Die Antwort ist nicht das Ergebnis eines "Black Box"-Prozesses, sondern die deterministische Aggregation validierter Zwischenergebnisse, was eine lückenlose Auditierbarkeit (Nachprüfbarkeit) ermöglicht.

7. Datenschicht – Überblick und detaillierte Architektur {#7.-datenschicht-–-überblick-und-detaillierte-architektur}

Die Architektur der Datenschicht ist ein fundamentaler Bestandteil des Systems und gewährleistet die effiziente und zielgerichtete Verarbeitung von Informationen. Sie ist strikt in vier primäre Wissensdomänen unterteilt, um eine klare Trennung der Datenquellen und der jeweiligen Abfrage-Mechanismen (Retrieval-Augmented Generation, RAG) zu gewährleisten.

7.1 Die vier Wissensdomänen {#7.1-die-vier-wissensdomänen}

Die Trennung in Domänen ermöglicht es, für jede Art von Information die optimale Speicher- und Abrufstrategie zu wählen, was die Präzision und Relevanz der generierten Antworten signifikant erhöht.

Domäne	Technologie	Primärer Dateninhalt	Zweck und Fokus
Strukturierte Fakten (Struktur-RAG)	GraphRAG	DWH-Daten, FHIR/ SNOMED-Klassifikationen	Bereitstellung eines präzisen, klinischen Kontextes und Steuerung der Abfragepfade (Faktenbasiert).
Unstrukturierte Dokumente (Patienten-		HYDMedia G6 Dokumente, klinische	Erfassung des Patientenverlaufs, detaillierte Befunde,

Domäne	Technologie	Primärer Dateninhalt	Zweck und Fokus
Dokumenten-RAG)	Vektor-DB (episodisch, zeitbasiert)	Berichte, Arztbriefe	individuelle Historie (Kontext).
Globales Wissen (Globaler Wissens-RAG)	Vektor-DB + Europe PMC API	Leitlinien, Europe PMC / PubMed-Artikel, Medizinische Klassifikationen	Bereitstellung evidenzbasierter Informationen, medizinischer Evidenz und fachlicher Einordnung (Leitlinien). Automatisches Evidence-Matching via SNOMED → Europe PMC (Kap. 7.2.1.1).
Evidenz- und Studien-Domainservice (Kap. 2.8.6)	Europe PMC + ClinicalTrials.gov API + LLM-Regelableitung	Aktuelle Publikationen, laufende klinische Studien, LLM-extrahierte Therapieregeln	Dynamische, sich selbst aktualisierende Wissensschicht: patientenspezifische Evidenz-Karten, automatisches Trial-Matching, LLM-gestützte Ableitung klinischer Regeln mit Human-in-the-Loop-Validierung. Andockbar an Patientendaten *und* Wissensbasis (Kap. 7.2.1.3, 7.2.1.4).

7.2 Detaillierte RAG-Architekturkomponenten {#7.2-detaillierte-rag-architekturkomponenten}

Die Umsetzung dieser Domänen erfolgt über spezialisierte RAG-Komponenten:

7.2.1 Globaler Wissens-RAG

- **Technologie:** Vektor-Datenbank (z. B. Weaviate, Qdrant – On-Premise-fähig, vgl. NFA-06)
- **Datenquellen:** Aktuelle medizinische Leitlinien (national und international), relevante Artikel aus der wissenschaftlichen

Datenbank PubMed, Standard-Klassifikationssysteme (z. B. ICD-10, OPS).

- **Funktion:** Dieser RAG dient als Quelle für allgemeines, medizinisches Expertenwissen. Er liefert die *Evidenz* und die *Einordnung* klinischer Sachverhalte in den breiteren medizinischen Kontext. Er gewährleistet, dass die generierten Empfehlungen auf dem aktuellen Stand der Forschung und den gültigen Standards basieren.

7.2.1.1 Europe PMC als primäre externe Evidenzquelle

Für die Anbindung externer wissenschaftlicher Literatur wird **Europe PMC** (https://europepmc.org/) als primäre Datenquelle eingesetzt. Europe PMC bietet Zugang zu über **43 Millionen Publikationen** (PubMed, PMC, Preprints, Patents) und stellt eine offene, DSGVO-konforme REST-API bereit, die ohne Authentifizierung nutzbar ist.

API-Endpunkte und Nutzung:

Endpunkt	Funktion	Einsatz in UKLGPT
`/search`	Volltextsuche mit strukturierten Feldern (TITLE, ABSTRACT, DISEASE, CHEM, GENE_PROTEIN, PUB_YEAR)	Abruf relevanter Publikationen zu aktuellen Patientendiagnosen
`/{source}/{id}/citations`	Zitationsnetzwerk eines Artikels	Identifikation der einflussreichsten Arbeiten zu einem Thema
`/{source}/{id}/references`	Referenzliste eines Artikels	Rückverfolgung der Evidenzkette
`/{source}/{id}/textMinedTerms`	Text-Mining-Annotationen (Genes, Chemicals, Diseases)	Automatische Anreicherung mit strukturierten Entitäten
`/{id}/fullTextXML`	Volltextzugriff (Open Access)	Tiefere Analyse bei relevanten Treffern

Suchstrategie – SNOMED-zu-Europe-PMC-Mapping:

Die Verknüpfung zwischen Patientendaten und externer Evidenz erfolgt über die SNOMED-CT-Kodierung der Patientendiagnosen. Der Ablauf:

1. **SNOMED-Extraktion:** Die SNOMED-CT-Codes der aktuellen Patientendiagnosen werden aus dem GraphRAG extrahiert (z. B. `84114007` = Herzinsuffizienz).
2. **Term-Expansion:** Über die Snowstorm FHIR-API (`$lookup`) werden die bevorzugten Bezeichnungen und Synonyme des SNOMED-Konzepts abgerufen.
3. **Europe PMC Query:** Die Terme werden als strukturierte Suchanfrage übersetzt: `DISEASE:"heart failure" AND PUB_YEAR:[2024 TO 2026] AND (LANG:"eng" OR LANG:"ger") AND IN_EPMC:y`
4. **Ranking:** Ergebnisse werden nach `sort_cited:y` (Zitationshäufigkeit) und `sort_date:y` (Aktualität) sortiert; die Top-N werden in den Wissens-RAG-Index aufgenommen.
5. **Caching:** Bereits abgerufene Artikel werden im lokalen Wissens-RAG-Index zwischengespeichert (TTL: 30 Tage), um API-Last zu minimieren.

Antwortformate: JSON (primär) oder XML; Metadaten umfassen `pmid`, `doi`, `title`, `authorString`, `journalTitle`, `pubYear`, `citedByCount`, `isOpenAccess`.

7.2.1.2 Automatisches Evidence-Matching (Internes + Externes Wissen)

Das **Evidence-Matching** ist der Kernmechanismus, der bei jedem Patientenkontakt automatisch die relevantesten internen und externen Dokumente identifiziert und dem Kliniker proaktiv anzeigt. Der Prozess läuft in drei Schichten:

Schicht 1 – Patientenkontext aus GraphRAG: * Beim Öffnen eines Patientenfalls werden die aktiven Diagnosen (SNOMED-kodiert), aktuelle Medikation (ATC/SNOMED), laufende Prozeduren und das Alter/Geschlecht aus dem GraphRAG extrahiert. * Diese Kontextinformationen bilden das **Patienten-Profil** (klinischer Fingerabdruck).

Schicht 2 – Internes Dokumenten-Matching (Patienten-Dokumenten-RAG + Wissens-RAG): * Das Patienten-Profil wird als Embedding-Query gegen den **Patienten-Dokumenten-RAG** ausgeführt → relevanteste eigene Dokumente (Arztbriefe, Befunde, OP-Berichte) nach semantischer Ähnlichkeit. * Parallel dazu wird das Patienten-Profil gegen den **internen Wissens-RAG** (vorgehaltene Leitlinien, SOPs, Hausprotokolle) gematcht → relevanteste interne Leitlinien und Protokolle.

Schicht 3 – Externes Evidence-Matching (Europe PMC): * Die Top-3-Diagnosen des Patienten werden über das SNOMED-zu-Europe-PMC-Mapping (s. 7.2.1.1) in Suchanfragen übersetzt. * Besondere Berücksichtigung erfahren: * **Aktualität:** Publikationen der letzten 2 Jahre werden priorisiert. * **Relevanz:** Klinische Leitlinien

(`SRC:MED`) und systematische Reviews werden höher gewichtet. * **Zitationshäufigkeit:** Hochzitierte Arbeiten erhalten Priorität. * Die Ergebnisse werden als kompakte Evidenz-Karten im UI angezeigt (Titel, Journal, Jahr, Zitationen, Abstract-Snippet, DOI-Link).

Ergebnis-Darstellung im Frontend (Kap. 4):

Bereich	Inhalt	Aktualisierung
Interne Dokumente	Top-5 relevanteste Patientendokumente (nach Ähnlichkeit + Aktualität)	Bei jedem Fallaufruf
Interne Leitlinien	Passende SOPs, Hausprotokolle, klinikspezifische Pfade	Bei jedem Fallaufruf
Externe Evidenz	Top-5 Europe-PMC-Treffer pro Hauptdiagnose (Titel, Journal, Jahr, Zitationen)	Tägliches Refresh, bei Diagnoseänderung sofort

Datenschutz-Konformität: Es werden keine Patientendaten an Europe PMC übermittelt. Die Suchanfragen enthalten ausschließlich medizinische Fachbegriffe (SNOMED-Terme), keine personenbezogenen Daten.

Referenzimplementierung: `Codebeispiele/europe-pmc-api/europe_pmc_examples.py` (Suche, Leitlinien, Text-Mining, Evidence-Matching) und `Codebeispiele/snomed-fhir-api/snomed_to_europepmc_bridge.py` (End-to-End: SNOMED → Terme → Europe PMC).

7.2.1.3 Klinische Studiendatenbanken als Datenquelle

Neben der wissenschaftlichen Literatur (Europe PMC) bindet der Evidenz-Domainservice (vgl. Kap. 2.8.6) **klinische Studienregister** als eigenständige Datenquelle an. Ziel ist die automatische Identifikation laufender Studien, die für die Diagnosen eines Patienten relevant sein könnten – insbesondere für ein Universitätsklinikum ein strategischer Vorteil bei der Studienrekrutierung.

Primäre Quelle: ClinicalTrials.gov API v2

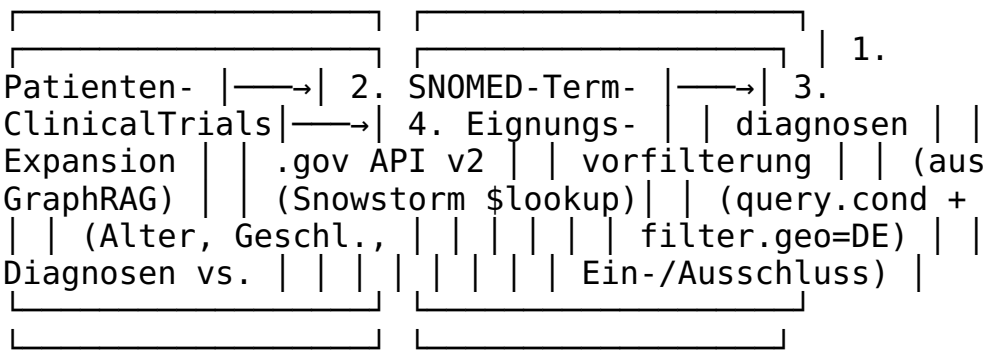
ClinicalTrials.gov ist das weltweit größte Studienregister (> 500.000 Studien) und bietet seit 2024 eine moderne REST-API (v2, OpenAPI 3.0):

Parameter	Funktion	Beispiel
`query.cond`		`query.cond=heart+failure`

Parameter	Funktion	Beispiel
	Suche nach Krankheitsbegriff (Condition)	
`query.intr`	Suche nach Intervention (Medikament, Prozedur)	`query.intr=sacubitril`
`filter.overallStatus`	Filter auf Studienstatus	`RECRUITING`, `NOT_YET_RECRUITING`
`filter.geo`	Geografischer Filter (Standort)	Filter auf Sachsen/Deutschland
`pageSize`	Ergebnisse pro Seite (max. 1.000)	`pageSize=20`
`format`	Antwortformat	`json` oder `csv`

Beispiel-Abfrage: `https://clinicaltrials.gov/api/v2/studies?query.cond=heart+failure&filter.overallStatus=RECRUITING&pageSize=1`

Integration in den UKLGPT-Workflow:



1. **SNOMED → Suchterm:** Die Patientendiagnosen (SNOMED-kodiert) werden über Snowstorm `$lookup` in englische MeSH-kompatible Terme übersetzt.
2. **Studiensuche:** Die Terme werden als `query.cond` an ClinicalTrials.gov gesendet, gefiltert auf `RECRUITING` + geografische Nähe.
3. **Eignungsvorfilterung:** Das Patientenprofil (Alter, Geschlecht, aktive Diagnosen, Medikation) wird gegen die Eligibility-Criteria der gefundenen Studien abgeglichen. Dies erfolgt LLM-gestützt, da die Eligibility-Kriterien als Freitext vorliegen.
4. **Ergebnisdarstellung:** Relevante Studien werden als Karten im UI angezeigt (NCT-Nummer, Titel, Phase, Rekrutierungsstatus, Standort, Kontakt).

Sekundäre Quellen:

Register	Status	Zugriff
EU CTIS (Clinical Trials Information System)	Seit 01/2025 verpflichtend für EU-Studien. Kein offizielles REST-API, Webportal unter euclinicaltrials.eu	Semi-automatisch über R-Paket `ctrdata` oder Web-Scraping
DRKS (Deutsches Register Klinischer Studien, BfArM)	WHO-Primärregister für Deutschland. API geplant, aktuell JSON-Export möglich	JSON-Export über drks.de; perspektivisch automatisiert über geplante API
WHO ICTRP (International Clinical Trials Registry Platform)	Meta-Register, enthält DRKS-Daten	Batch-Download, keine REST-API

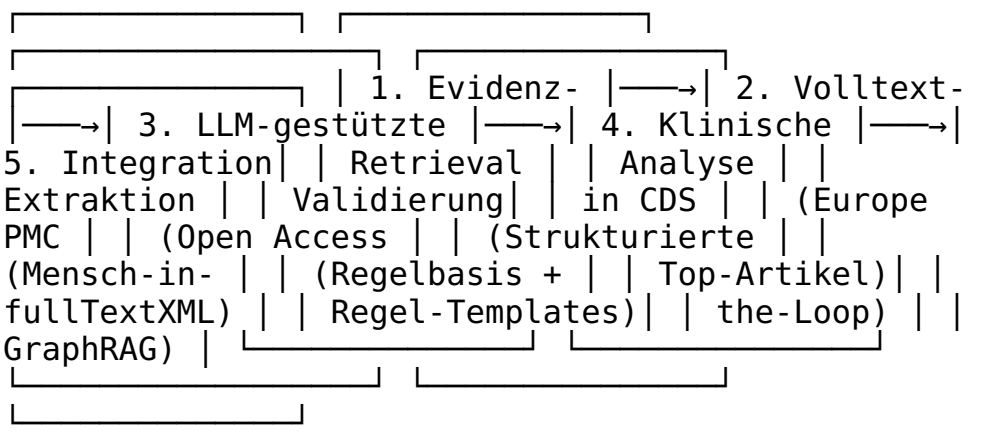
Datenschutz: Analog zu Europe PMC werden ausschließlich medizinische Fachbegriffe an externe Studienregister übermittelt. Es fließen keine Patientendaten (Name, Geburtsdatum, Fallnummer) nach außen.

7.2.1.4 LLM-gestützte Regelableitung aus Evidenz

Ein konzeptioneller Kernbaustein des Evidenz-Domainservice (vgl. Kap. 2.8.6) ist die Fähigkeit, aus der abgerufenen Literatur nicht nur *Informationen* zu liefern, sondern **klinisch anwendbare Regeln** zu extrahieren, die in die hauseigene Entscheidungsunterstützung (CDS) einfließen können.

Abgrenzung: Dieser Mechanismus ersetzt keine klinischen Leitlinien und keine ärztliche Urteilsbildung. Er dient als *Vorschlagssystem*, das neue Evidenz identifiziert und strukturiert aufbereitet. Jede abgeleitete Regel durchläuft einen **humanen Validierungsprozess**, bevor sie in die aktive Regelbasis aufgenommen wird.

Workflow der Regelableitung:



Schritt 1 – Evidenz-Retrieval: Der Europe-PMC-Service identifiziert hochzitierte, aktuelle Publikationen zu den relevanten Diagnosen (vgl. Kap. 7.2.1.1). Priorisiert werden: Systematische Reviews, Leitlinien-Updates, RCTs mit > 100 Zitationen.

Schritt 2 – Volltextanalyse: Für Open-Access-Artikel wird der Volltext über die Europe PMC `fullTextXML`-API abgerufen. Bei Nicht-OA-Artikeln werden Abstract und Conclusion ausgewertet.

Schritt 3 – LLM-Extraktion: Ein spezialisierter Prompt extrahiert aus dem Volltext strukturierte Regel-Kandidaten nach einem definierten Template:

Feld	Beschreibung	Beispiel
`condition`	SNOMED-kodierte Bedingung	`84114007` (Heart failure) + `LVEF < 35%`
`action`	Empfohlene Maßnahme	„SGLT2-Inhibitor initiieren"
`evidence_grade`	Evidenzgrad (wenn angegeben)	Klasse I, Evidenzgrad A
`source`	Quellennachweis (DOI, PMID)	DOI: 10.1093/eurheartj/ehab368
`contraindications`	Gegenanzeigen (wenn extrahiert)	„eGFR < 20 ml/min"
`confidence`	LLM-Konfidenz der Extraktion	0.92

Schritt 4 – Klinische Validierung (Human-in-the-Loop):

Extrahierte Regel-Kandidaten werden **nicht** automatisch aktiviert. Stattdessen werden sie in eine **Review-Queue** eingestellt, die von klinischem Fachpersonal bearbeitet wird:

- **Bestätigt:** Die Regel wird in die aktive CDS-Regelbasis aufgenommen, mit Quellennachweis und Gültigkeitsdatum.
- **Modifiziert:** Der Kliniker passt Bedingungen, Dosierungen oder Kontraindikationen an lokale Gegebenheiten an.
- **Verworfen:** Die Regel wird als nicht relevant oder nicht korrekt markiert; das Feedback fließt in die Prompt-Optimierung zurück.

Schritt 5 – Integration in CDS:

Validierte Regeln werden als strukturierte Objekte im GraphRAG gespeichert und stehen der Prompt-Pipeline (Kap. 6.3, Schritt „Medizinische Validierung") zur Verfügung. Bei zukünftigen

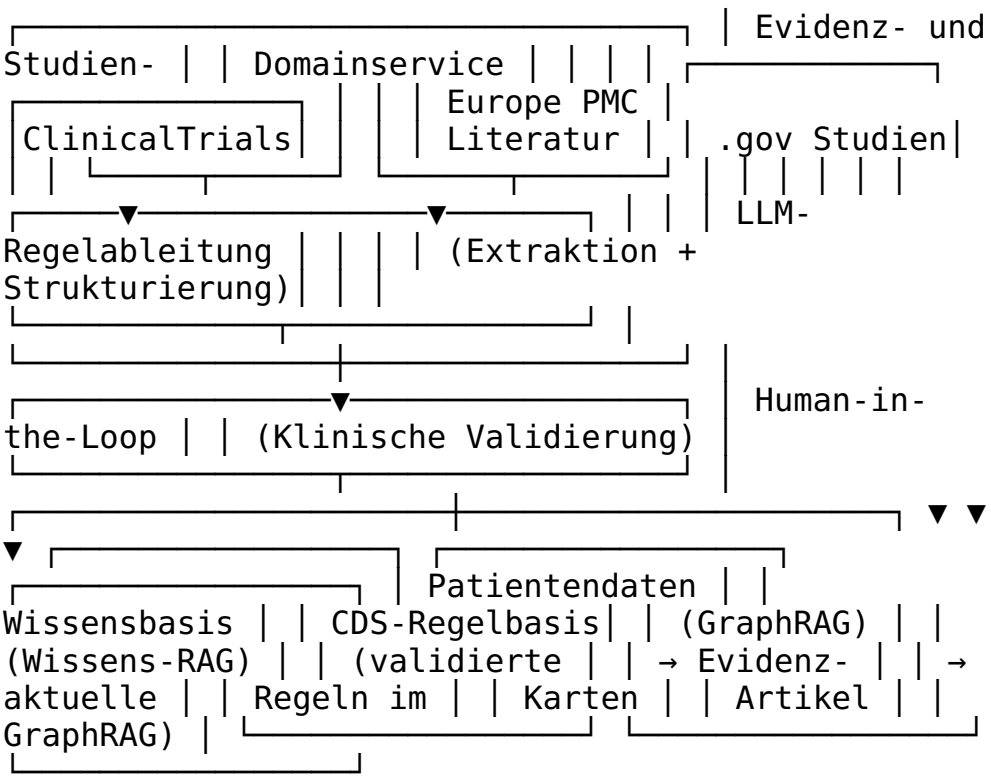
Patientenkontakten werden die Regeln automatisch gegen das Patientenprofil gematcht:

- Patient hat SNOMED 84114007 (Herzinsuffizienz) + LVEF 30% → Regel wird ausgelöst → Hinweis an Kliniker: „Gemäß ESC 2023 (validiert am TT.MM.JJJJ): SGLT2-Inhibitor erwägen."

Governance und Qualitätssicherung:

- Jede Regel hat eine **Gültigkeitsdauer** (TTL, default: 24 Monate). Abgelaufene Regeln werden automatisch zur Re-Validierung vorgelegt.
- Jede Regelanwendung wird im **Audit-Trail** protokolliert (Regel-ID, Patientenkontext, Kliniker-Reaktion).
- Regeln können nach Fachgebiet, Evidenzgrad und Validierungsstatus gefiltert werden.
- Eine **Dashboard-Ansicht** zeigt den aktuellen Stand der Regelbasis: Anzahl aktiver Regeln, offene Review-Queue, Regelanwendungen pro Woche, Ablehnungsquote.

Zusammenspiel der Domainservice-Komponenten:



Der Domainservice wirkt damit auf **drei Ebenen** gleichzeitig: Er liefert dem Kliniker patientenspezifische Evidenz (links), hält die Wissensbasis aktuell (Mitte) und baut über die Zeit eine validierte, evidenzbasierte Regelbasis auf (rechts). Diese dreifache Wirkung – Information, Wissen, Regeln – ist das, was UKLGPT von einem reinen Frage-Antwort-System zu einem **lernenden klinischen Wissenspartner** transformiert.

7.2.2 Patienten-Dokumenten-RAG

- **Technologie:** Vektor-Datenbank, mit einem Fokus auf episodischer und zeitbasierter Indexierung. Die Daten sind mandantenfähig und streng auf den jeweiligen Patienten beschränkt.
- **Datenquellen:** Alle unstrukturierten oder semistrukturierten klinischen Dokumente aus dem HYDMedia G6 System (z. B. OP-Berichte, Entlassungsbriefe, Laborbefunde als Freitext, Pflegeprotokolle).
- **Funktion:** Er liefert den individuellen *Kontext* und den *Verlauf* des Patienten. Durch die Vektorisierung dieser Dokumente können semantische Ähnlichkeiten schnell abgerufen werden, um beispielsweise frühere, relevante Details oder ähnliche Symptomverläufe zu identifizieren.

7.2.3 Struktur-RAG (Graph-RAG)

- **Technologie: GraphRAG** (Integration von Wissensgraphen und LLMs).
- **Datenquellen:** Strukturierte Daten aus dem Data Warehouse (DWH), die bereits in ein standardisiertes, klinisches Datenmodell (z. B. FHIR-Ressourcen oder SNOMED CT-Konzepte) transformiert wurden.
- **Funktion:** Dies ist die Quelle für harte, **strukturierte Fakten**. Der Graph-RAG ermöglicht es, komplexe Beziehungen zwischen klinischen Entitäten (Patienten, Diagnosen, Medikationen, Behandlungen) zu navigieren. Er dient zur *Steuerung* des gesamten RAG-Prozesses, indem er präzise Abfragen liefert (z. B. "Welche Medikation erhielt der Patient *nach* der Diagnose X?") und den initialen, klinischen Kontext festlegt. Die Graphenstruktur verhindert "Halluzinationen" bei Faktenfragen, da die Antworten direkt aus den strukturierten DWH-Daten abgeleitet werden.

7.3 Rolle des GraphRAG: Der Graph als zentrale Wissens- und Steuerungsebene {#7.3-rolle-des-graphrag:-der-graph-als-zentrale-wissens--und-steuerungsebene}

Der in der Architektur verankerte Graph spielt eine fundamentale und vielschichtige Rolle in der gesamten Retrieval-Augmented Generation (RAG)-Kette. Er fungiert nicht nur als bloße Datenablage, sondern als aktiver Steuerungsknotenpunkt und primäre Wissensautorität.

Die Hauptfunktionen des Graphen im GraphRAG-System sind:

7.3.1 Klinische Kontextautorität (wer, wann, was, wo):

- Der Graph bildet die zentralen Entitäten des klinischen Umfelds – Patienten, Diagnosen, Behandlungen, Fachbereiche, Zeitpunkte und Standorte – in einem semantischen Netz ab.

- Diese reichhaltige Struktur ermöglicht es, eine Abfrage präzise zu kontextualisieren und die komplexen Beziehungen zwischen medizinischen Fakten und Ereignissen zu navigieren. Er beantwortet die fundamentalen W-Fragen und liefert somit die notwendige Orientierung für nachfolgende RAG-Schritte.

7.3.2 Steuerungsinstanz für Dokumenten-RAG (welche Dokumente sind relevant):

- Anstatt sämtliche Dokumente nach relevanten Informationen zu durchsuchen, nutzt der Graph seine strukturelle Intelligenz zur Vorfilterung.

- Basierend auf den extrahierten Entitäten und den Beziehungen zum klinischen Fall steuert der Graph gezielt, welche Dokument-Chunks, Textabschnitte oder spezifischen Dateien (z. B. Behandlungspläne, Laborberichte, Fachartikel) für die Beantwortung einer Nutzeranfrage im nachgeschalteten Vektor-RAG-Prozess als relevant markiert werden. Dies erhöht die Präzision und reduziert die Latenz erheblich.

7.3.3 Primäre Quelle für strukturierte Fakten:

- Im Gegensatz zu unstrukturiertem Text liefert der Graph klar definierte, ontologisch verknüpfte Fakten (z. B. "Medikament X ist kontraindiziert bei Zustand Y").

- Diese strukturierten Informationen können direkt zur Beantwortung von Faktenfragen verwendet oder zur Anreicherung der generierten Antwort aus den Dokumenten genutzt werden, wodurch die faktische Korrektheit und Zuverlässigkeit der Ausgabe signifikant gesteigert wird.

7.3.4 Brücke zwischen Dokumenten, Leitlinien und Klassifikationen:

- Der GraphRAG-Ansatz ermöglicht es, lose Textdokumente (Berichte), standardisierte medizinische Leitlinien (Protokolle) und formale Klassifikationssysteme (z. B. ICD-10, OPS-Kataloge) semantisch miteinander zu verknüpfen.

- Diese Verknüpfungsfunktion schafft eine kohärente Wissensbasis, die es ermöglicht, eine Information aus einem Patientenbericht mit der entsprechenden Empfehlung aus einer nationalen Leitlinie zu validieren und mit den relevanten Klassifikationscodes zu versehen. Er transformiert die

Datenlandschaft von isolierten Silos zu einem integrierten, *vernetzten klinischen Wissensökosystem*.

7.4 GraphRAG – Fachliches Datenmodell: Eine detaillierte Betrachtung {#7.4-graphrag-–-fachliches-datenmodell:-eine-detaillierte-betrachtung}

Dieses Kapitel beleuchtet das zugrundeliegende fachliche Datenmodell für die GraphRAG-Implementierung, welches auf dem HL7 FHIR (Fast Healthcare Interoperability Resources) Standard basiert und durch die Integration von Terminologien erweitert wird. Ziel ist es, eine semantisch reichhaltige Graphenstruktur zu schaffen, die komplexe klinische Zusammenhänge effizient abbilden und für erweiterte Abfragen im Rahmen von Retrieval Augmented Generation (RAG) nutzen kann.

7.4.1 FHIR-basierte Knoten (Labels) und ihre Bedeutung

Die Kernstruktur des Graphenmodells wird durch eine Reihe von Knoten (Labels) definiert, die direkt auf wesentlichen FHIR-Ressourcen basieren. Jeder dieser Knoten repräsentiert eine klinisch oder administrativ relevante Einheit:

Graph Label	FHIR Ressource	Fachliche Bedeutung im Graphen
Patient	Patient	Der zentrale Akteur im Gesundheitssystem. Enthält demografische Daten und dient als Ausgangspunkt für alle patientenbezogenen klinischen Informationen.
Encounter	Encounter	Repräsentiert eine Interaktion zwischen Patient und Leistungserbringer (z.B. Arztbesuch, Krankenhausaufenthalt). Dient als wichtiger zeitlicher und kontextueller Ankerpunkt.
EpisodeOfCare	EpisodeOfCare	Fasst zusammenhängende Begegnungen und Behandlungen für ein spezifisches Gesundheitsproblem oder

Graph Label	FHIR Ressource	Fachliche Bedeutung im Graphen
		über einen bestimmten Zeitraum zusammen (z.B. Behandlung einer chronischen Krankheit).
Observation	Observation	Stellt Messungen, Testergebnisse oder klinische Feststellungen dar (z.B. Blutdruckwerte, Laborergebnisse). Sind häufig mittels LOINC codiert.
Condition	Condition	Beschreibt festgestellte Gesundheitsprobleme, Diagnosen, oder Allergien/ Unverträglichkeiten. Diese werden primär über SNOMED CT und sekundär über ICD codiert.
Procedure	Procedure	Erfasste Eingriffe, Operationen oder therapeutische Maßnahmen, die am Patienten durchgeführt wurden.
MedicationStatement / MedicationRequest	MedicationStatement / MedicationRequest	**MedicationRequest** beschreibt die ärztliche Anordnung eines Medikaments; **MedicationStatement** dokumentiert die tatsächliche Einnahme des Patienten. Beide sind für die Medikationshistorie kritisch.
AllergyIntolerance	AllergyIntolerance	Dokumentiert bekannte Überempfindlichkeiten oder Allergien des Patienten gegen Substanzen.
Practitioner	Practitioner	Die individuelle Person, die eine Gesundheitsdienstleistung erbringt (z.B. Arzt, Pflegekraft).
PractitionerRole	PractitionerRole	

Graph Label	FHIR Ressource	Fachliche Bedeutung im Graphen
		Beschreibt die spezifische Rolle eines Practitioner in einer Organization (z.B. "Hausarzt in Praxis-Musterstadt"). Dient der Kontextualisierung der Leistungserbringung.
CareTeam	CareTeam	Eine Gruppe von Leistungserbringern, die gemeinsam an der Versorgung eines Patienten beteiligt sind.
Organization	Organization	Eine Institution, die Gesundheitsleistungen erbringt (z.B. Krankenhaus, Arztpraxis).
DocumentReference	DocumentReference	Verweist auf externe klinische Dokumente (z.B. Entlassbriefe, Befunde, Arztberichte). Über diesen Knoten können die unstrukturierten Textinhalte in den RAG-Prozess eingebunden werden.

7.4.2 Terminologie-Knoten zur semantischen Anreicherung

Um die klinischen Daten semantisch abfragbar zu machen und eine effektive Abstraktion von spezifischen Werten zu ermöglichen, werden separate Knoten für standardisierte medizinische Terminologien verwendet. Diese entkoppeln die klinischen Instanzen (wie Condition oder Observation) von ihren Codierungen:

Label	Standard	Zweck im Graphen
SNOMEDConcept	SNOMED CT	Dient der feingranularen, mehrdimensionalen Codierung von Diagnosen, Prozeduren, Körperstrukturen etc. Ermöglicht semantische Abfragen über Konzeptbeziehungen innerhalb von SNOMED CT.
ICDCode	ICD-10 / ICD-11	Wird für die administrative und abrechnungsrelevante

Label	Standard	Zweck im Graphen
		Klassifikation von Diagnosen verwendet (z.B. im Kontext von Condition).
HPOConcept	Human Phenotype Ontology	Speziell für die Codierung von Phänotypen und Symptomen, insbesondere im Bereich seltener oder genetischer Erkrankungen.
LOINCCode	LOINC	Standard zur Codierung von Laboruntersuchungen, Messungen und Beobachtungen (z.B. im Kontext von Observation).

7.4.3 Zentrale Beziehungen und die Abbildung klinischer Prozesse

Die Verbindungen (Kanten) zwischen den Knoten definieren die klinischen und administrativen Zusammenhänge. Sie sind das Herzstück des Graphenmodells und ermöglichen die Navigation und Kontextualisierung von Informationen.

Beziehung (Muster: (Quelle)-[RELATION]->(Ziel))	Erklärung des Zusammenhangs
(Patient)-[:HAS_ENCOUNTER]->(Encounter)	Der Patient hatte diese spezifische klinische Encounter (Begegnung).
(Encounter)-[:PART_OF]->(EpisodeOfCare)	Die Encounter ist Teil einer umfassenderen, zusammenhängenden EpisodeOfCare.
(Encounter)-[:HAS_CONDITION]->(Condition)	Während dieser Encounter wurde diese spezifische Condition (Diagnose/ Gesundheitsproblem) festgestellt oder behandelt.
(Condition)-[:CODED_AS]->(SNOMEDConcept)	Die klinische Condition ist primär mit diesem SNOMEDConcept codiert.
(Condition)-[:ALSO_CODED_AS]->(ICDCode)	Die klinische Condition trägt zusätzlich diesen administrativen ICDCode.
(Encounter)-[:HAS_PROCEDURE]->(Procedure)	Während der Encounter wurde diese Procedure (Prozedur/ Eingriff) durchgeführt.

Beziehung (Muster: (Quelle)-[RELATION]->(Ziel))	Erklärung des Zusammenhangs
(Encounter)-[:HAS_MEDICATION]->(MedicationStatement/Request)	Im Rahmen der Encounter wurde dieses Medikament verordnet oder dokumentiert.
(Encounter)-[:HAS_OBSERVATION]->(Observation)	Diese Observation (Messung/Befund) wurde während der Encounter erhoben.
(Observation)-[:CODED_AS]->(LOINCCode)	Die Art der Observation (z.B. "Systolischer Blutdruck") ist durch diesen LOINCCode standardisiert.
(Encounter)-[:HAS_DOCUMENT]->(DocumentReference)	Die Encounter führte zur Erstellung oder zum Verweis auf dieses klinische DocumentReference.
(CareTeam)-[:HAS_MEMBER]->(PractitionerRole)	Das CareTeam umfasst diese spezifischen PractitionerRole-Mitglieder.
(PractitionerRole)-[:ASSIGNED_TO]->(Practitioner)	Die spezifische PractitionerRole ist der physischen Person Practitioner zugeordnet.
(Encounter)-[:MANAGED_BY]->(CareTeam)	Die gesamte Encounter wurde von diesem CareTeam betreut und koordiniert.

7.4.4 Metadaten zur Abbildung von Verlauf und Aktualität

Ein entscheidendes Merkmal dieses Datenmodells ist die **Berücksichtigung zeitlicher Gültigkeiten**. Nahezu alle klinisch relevanten Knoten (wie Condition, Observation, Procedure, etc.) enthalten obligatorisch **zeitliche Gültigkeiten** (dargestellt als Properties, z.B. effectivePeriodStart und effectivePeriodEnd). Diese Metadaten sind essenziell, um:

- **Verlaufsinformationen** korrekt abzubilden (z.B. wann eine Diagnose gestellt wurde oder eine Medikation aktiv war).
- Die **Aktualität** der im RAG-Prozess verwendeten Informationen zu gewährleisten.

Diese zeitliche Dimension ermöglicht es, Fragen zu beantworten wie: "Welche Diagnosen waren *aktiv* in der EpisodeOfCare des letzten Jahres?" und erhöht damit die Präzision der aus dem Graphen abgeleiteten Antworten.

8. Datenzufluss: Echtzeit-DWH → GraphRAG {#8.-datenzufluss:-echtzeit-dwh-→-graphrag}

Dieser Abschnitt beschreibt den geplanten Datenfluss aus der zentralen *Echtzeit-Data-Warehouse (DWH)*-Umgebung in die Graphdatenbank, die als analytische Schicht für komplexe Beziehungsanalysen und 360-Grad-Sichten dient.

8.1 Datenquellen (Quellsysteme des DWH) {#8.1-datenquellen-(quellsysteme-des-dwh)}

Die Datenbasis für das Graph-Modell wird primär aus den harmonisierten und integrierten Daten des zentralen DWH bezogen. Dieses DWH aggregiert heterogene Daten aus den folgenden primären operativen Quellsystemen:

Quellsystem	Beschreibung und Relevanz
SAP ISH / ISHmed	Das zentrale Krankenhausinformationssystem (KIS) für administrative und medizinische Prozesse. Liefert Falldaten, Patientenstammdaten, Leistungsdokumentation und Abrechnungsinformationen, die essenziell für die Abbildung des Patientenpfads sind.
KIS Meierhofer	Das neue Krankenhausinformationssystem, das zukünftig als zentrales KIS genutzt wird.
PDMS Copra	Das Patientendaten-Managementsystem, typischerweise eingesetzt auf Intensivstationen (ITS) und in der Anästhesie. Liefert hochfrequente, kritische Vitalparameter, Medikation und Verlaufsdaten.
Polypoint	Ein System zur Personal- und Einsatzplanung. Liefert Informationen über die Verfügbarkeit, Qualifikation und Schichtpläne des medizinischen und pflegerischen Personals. Relevant für die Verknüpfung von Behandlungen mit den verantwortlichen Ärzten/Pflegeteams.
SAP HR, MARA, FICO	**SAP HR (Human Resources)** liefert strukturierte Daten zu Mitarbeitern (Rollen, Organisationseinheiten). **MARA (Material Management)** liefert Daten zu Medikamenten und medizinischem Material. **FICO (Finance and**

Quellsystem	Beschreibung und Relevanz
	Controlling) liefert Kosten- und Leistungsdaten. Diese Systeme ergänzen die klinischen Daten um administrative und logistische Kontextinformationen.

8.2 Integrationsmuster und ETL-Logik {#8.2-integrationsmuster-und-etl-logik}

Die Integration zwischen dem Echtzeit-DWH und GraphRAG folgt einem spezifischen Muster, um die Aktualität, Konsistenz und Historisierbarkeit der Graphenstruktur zu gewährleisten.

8.2.1 Übertragungsfrequenz und Technologie

- **Near-Real-Time (NRT) / Event-basiert (CDC):** Dies ist das bevorzugte Muster. Änderungen in den Quelldaten des DWH (z. B. eine neue Diagnose, eine geänderte Medikation, ein verschobener Termin) werden mittels *Change Data Capture (CDC)* fast unmittelbar erkannt und als Events an die Integrationsstrecke übermittelt. Dies minimiert die Latenz und ermöglicht zeitnahe Analysen.

- **Micro-Batch:** Für weniger zeitkritische oder voluminöse Daten, bei denen CDC technisch aufwändig wäre, können sehr kleine, frequente Batches (z. B. alle 5 bis 15 Minuten) verwendet werden.

8.2.2 Transformation und Mapping-Pipeline

Die Daten durchlaufen einen mehrstufigen Transformationsprozess, um von einem relationalen/tabellarischen Schema in die flexible Graphenstruktur überführt zu werden:

1. **DWH-Schema → FHIR-Ressourcen:** Die strukturierten DWH-Daten werden zunächst auf den **FHIR (Fast Healthcare Interoperability Resources)**-Standard gemappt. Dies dient als Zwischenschicht und Brücke zu einem international anerkannten, interoperablen Gesundheitsdatenstandard (z.B. Patient, Encounter, Condition, Procedure, MedicationRequest).

2. **FHIR-Ressourcen → Graphschema:** Die FHIR-Ressourcen werden anschließend in die spezifischen *Nodes* (Knoten) und *Relationships* (Beziehungen) des Graphschemas übersetzt. **Beispiel:** Ein FHIR-Patient wird zum (Patient)-Knoten; eine FHIR-Encounter wird zum (Encounter)-Knoten, verbunden über die Beziehung [:HAS_ENCOUNTER].

8.2.3 Datenaktualisierung (Upserts)

- **Upserts mit fachlicher Schlüsseldefinition:** Statt einfacher Inserts oder Deletes werden *Upsert*-Operationen (Update or Insert) verwendet. Dies ist kritisch für Graphdatenbanken, um Redundanzen zu vermeiden und die Konsistenz von Knoten über die Zeit zu gewährleisten.

- **Technisch:** Es wird der MERGE-Befehl in Cypher (Graph-Abfragesprache) genutzt.

- **Fachlich:** Die **fachlichen Schlüssel** (z.B. Patienten-ID, Fall-ID, Prozeduren-Code) dienen als eindeutige Identifikatoren, um zu prüfen, ob ein Knoten oder eine Beziehung bereits existiert. Nur wenn die Kombination des fachlichen Schlüssels *und* der übermittelten Daten abweicht, erfolgt ein Update oder ein neuer Eintrag (bei Historisierung).

8.2.4 Historisierung

- **Historisierung medizinisch relevanter Änderungen:** Nicht jede Datenänderung führt zu einem einfachen Update (Überschreiben) eines Knoten-Properties. Für klinisch oder forensisch relevante Daten (z.B. Diagnosen, Medikationspläne, Befunde) muss die zeitliche Abfolge der Änderungen dokumentiert werden.

- **Implementierung im Graph:** Dies wird erreicht, indem die Beziehungen (Relationships) zwischen Knoten mit **Zeitstempel-Properties** (z.B. gültig_von, gültig_bis) versehen werden, oder indem neue, zeitgestempelte Knoten erstellt werden, die über eine Beziehung auf den Ursprungsknoten verweisen (z.B. (Patient)-[:HAT_DIAGNOSE_ZU_ZEITPUNKT {datum: '2026-01-10'}]->(Diagnose)).

9. GraphRAG – Retrieval-Strategie im Detail {#9.-graphrag-–-retrieval-strategie-im-detail}

Die Retrieval Augmented Generation (RAG) im Kontext von Graphen (GraphRAG) stellt eine hochentwickelte Strategie dar, um die Leistungsfähigkeit von Large Language Models (LLMs) durch präzise, faktenbasierte Kontextinformationen aus einem klinischen Wissensgraphen zu erweitern. Diese Strategie gliedert sich in mehrere, aufeinander aufbauende oder komplementäre Schritte.

9.1 Deterministische Abfragen (Primary Path) – Die Faktenbasis {#9.1-deterministische-abfragen-(primary-path)-–-die-faktenbasis}

Der "Primary Path" dient der obligatorischen Extraktion **harter Fakten** und unverzichtbarer, strukturierter Informationen direkt aus dem Graphen. Diese Abfragen basieren auf vordefinierten, robusten Templates und werden *immer* vor jeder nachfolgenden semantischen Suche ausgeführt, um eine korrekte und sichere LLM-Antwort zu gewährleisten. Sie nutzen die Stärken der Graphentechnologie – die schnelle, exakte Navigation und Abfrage von Beziehungen – mittels der Cypher-Abfragesprache.

9.1.1 Zentrale Anwendungsbeispiele und ihre Bedeutung:

Abfrageziel	Details / Kontextuelle Bedeutung
Aktiver Encounter eines Patienten	Die Feststellung des aktuellen Behandlungsfalls (z. B. Krankenhausaufenthalt, Ambulanzbesuch) ist die primäre Verankerung für alle weiteren Informationen.
Aktuelle Diagnosen (SNOMED normalisiert)	Die Abfrage von Diagnosen, die nach einem standardisierten Vokabular (z. B. SNOMED CT) kodiert sind, liefert die Grundlage für die medizinische Interpretation und Risikobewertung.
Aktuelle Medikation inkl. Allergien	Ein kritischer Pfad zur Sicherstellung der Patientensicherheit. Die genaue Liste der verabreichten Medikamente und bekannter Unverträglichkeiten muss *immer* als Kontext dienen.
Letzte relevante Laborwerte	Abruf von zeitlich aktuellen und klinisch bedeutsamen Ergebnissen (z. B. Nierenfunktionsparameter, Entzündungszeichen) zur Beurteilung der aktuellen physiologischen Lage.
Zuständiges CareTeam / Station	Organisatorische Details, die für die Weiterleitung, Eskalation oder die Kontaktaufnahme relevant sind.

9.1.2 Funktion

Diese Abfragen stellen sicher, dass das LLM nicht halluziniert und die Antwort auf einer unbestreitbaren, strukturierten Datenbasis aufbaut.

Der Output sind aggregierte Datensätze, die als prägnante Eingabe in den Prompt des LLM integriert werden.

9.2 Graph-Neighborhood Retrieval – Der fokussierte Kontext {#9.2-graph-neighborhood-retrieval-–-der-fokussierte-kontext}

Das Neighborhood Retrieval geht über die reinen Fakten hinaus und zielt darauf ab, den unmittelbar relevanten Kontext für komplexere oder weniger strikt definierte Fragen zu liefern.

9.2.1 Mechanismus

Es wird ein **begrenzter Subgraph** um eine oder mehrere zentrale Entitäten (z. B. den aktuellen Encounter, eine spezifische Diagnose) extrahiert. Die Begrenzung erfolgt typischerweise durch eine maximale Anzahl von **Hops** (z. B. 1 bis 2 Kantenentfernungen), um die Informationsmenge kontrollierbar zu halten und die Relevanz zu maximieren.

9.2.2 Verwendung für komplexe Kontextfragen

Diese Methode kommt zum Einsatz, wenn die reine Faktenbasis nicht ausreicht. Beispiele hierfür sind:

- **"Welche Prozeduren wurden im Zusammenhang mit der Herzinsuffizienz in den letzten 48 Stunden durchgeführt?"** (Erfordert die Navigation über Diagnose → Encounter → Prozeduren mit Zeitstempel.)
- **"Welche Laborwerte zeigten in den letzten 7 Tagen einen signifikanten Trend?"** (Erfordert die Einbeziehung benachbarter Zeitreihen- und Trend-Knoten.)

9.3 Graph-to-Text + Embeddings – Semantische Ähnlichkeit und erweiterte Suche (optional) {#9.3-graph-to-text-+-embeddings-–-semantische-ähnlichkeit-und-erweiterte-suche-(optional)}

Diese optionale, aber mächtige Erweiterung erlaubt es, die Strukturinformationen des Graphen für die **semantische Suche** nutzbar zu machen, insbesondere bei Fragestellungen, die Ähnlichkeiten oder Mustererkennung erfordern.

9.3.1 Prozessschritte:

1. **Serialisierung des Subgraphen in kontrollierte Textform (Graph-to-Text):** Komplexe Subgraphen (z. B. der gesamte Behandlungsverlauf eines Patienten oder das Profil einer Patientengruppe) werden mithilfe spezifischer Regeln oder kleiner, spezialisierter LLMs in eine kohärente, kontrollierte Textdarstellung überführt. Dies kann beispielsweise die Form "Patient A hat Diagnose X, wurde mit Medikament Y behandelt und hatte Laborwert Z" annehmen.

2. **Embedding:** Dieser Text wird anschließend in einen hochdimensionalen Vektor (Embedding) umgewandelt.

3. **Speicherung und Retrieval:** Diese Embeddings werden in einem **Vektorindex** abgelegt. Dies kann ein separater Vektorindex (z. B. Weaviate, Qdrant – On-Premise) oder der integrierte Vector Index der Graphen-DB sein.

9.3.2 Einsatzgebiet:

Diese Methode ist ideal für Fragestellungen des Typs **„Patient mit ähnlichem Profil“** oder **„Zeige Behandlungsverläufe, die diesem stark ähneln“**. Das LLM kann so auf Basis der semantischen Ähnlichkeit (Vector Search) im hochdimensionalen Raum auf historisches Wissen zugreifen, das sonst nur durch sehr komplexe, langsame Cypher-Abfragen zu finden wäre. Es ermöglicht die Überführung des strukturierten Graphwissens in das Domänenwissen des LLM.

10. Verzahnung GraphRAG ↔ Dokumenten-RAG (HYDMedia) – Eine integrierte Architektur zur Wissensgewinnung {#10.-verzahnung-graphrag-↔-dokumenten-rag-(hydmedia)-–-eine-integrierte-architektur-zur-wissensgewinnung}

Die effektive Integration von strukturierten Graphen-Informationen (GraphRAG) und unstrukturierten Dokumentendaten (Dokumenten-RAG, hier basierend auf HYDMedia-Speicher) ist ein Kernelement für umfassendes Retrieval und präzise Antworten. Diese Sektion beschreibt, wie die beiden RAG-Systeme verzahnt werden, wobei der **DocumentReference**-Knoten die zentrale Brücke bildet.

10.1 Die Rolle des DocumentReference-Knotens als Metadaten-Hub {#10.1-die-rolle-des-documentreference-knotens-als-metadaten-hub}

Der DocumentReference-Knoten im Wissensgraphen dient primär als leichtgewichtiger Verweis und **enthält ausschließlich Metadaten** zum extern gespeicherten Dokument, nicht das Dokument selbst. Diese Metadaten sind essenziell für die Steuerung der Ingestion und das Retrieval:

- **Dokumenttyp (z.B. Entlassbrief, Laborbefund, Konsilbericht):** Dient der initialen fachlichen Filterung und Priorisierung bei der Abfrage.
- **Datum/Zeitraum:** Ermöglicht die zeitliche Einschränkung des Suchraumes, relevant für episodische oder zeitkritische Anfragen.
- **Encounter-Bezug (Behandlungsfall-ID):** Stellt den direkten Kontextbezug im Graphen her und ist der Schlüssel zur Autorisierung und zum Zugriff auf patientenbezogene Informationen.
- **Storage-Pointer (HYDMedia/ISILON):** Der eigentliche Speicherort-Verweis (URI/URL), der das Dokument im zugehörigen Archivierungssystem (HYDMedia für die Primärablage oder ISILON für Backups/sekundäre Archivierung) eindeutig identifiziert.

Zweck: Durch diese Trennung wird der Graph schlank gehalten und bleibt hoch performant. Die komplexen, speicherintensiven Dokumenteninhalte werden erst bei Bedarf über den Storage-Pointer abgerufen.

10.2 Steuerung der Dokumenten-Ingestion und des Zugriffs {#10.2-steuerung-der-dokumenten-ingestion-und-des-zugriffs}

Die Dokumenten-Ingestion in das RAG-System, d.h. das tatsächliche Laden, Parsen und Embedding des unstrukturierten Textes, wird strikt durch den Wissensgraphen gesteuert:

- **Autorisierter Subgraph als Zugriffsfilter:** Es werden **nur DocumentReferences geladen**, die Teil des **autorisierten Subgraphen** sind. Dies stellt sicher, dass nur Dokumente in den RAG-Index gelangen, zu denen der aktuelle Nutzer oder die aktuelle Anfrage Kontext und Zugriffsrechte besitzt (z.B. über den Encounter-Bezug oder die Organisationszugehörigkeit). Der Graph übernimmt somit die Funktion eines primären Sicherheits- und Relevanzfilters.

- **Ressourcenschonende und aktuelle Embedding-Strategie:**

- **Nur episodisch embedded:** Dokumente werden nicht dauerhaft oder sofort nach Erstellung in den RAG-Vektorindex überführt. Das Embedding erfolgt "episodisch", d.h. bedarfsgesteuert im Kontext eines aktuellen Behandlungsfalls oder einer spezifischen Abfrage. Dies reduziert die Größe und Wartungsintensität des Vektorindexes drastisch.

- **Mit TTL (Time-To-Live) gespeichert:** Die generierten Embeddings werden im Vektorspeicher nur für einen begrenzten Zeitraum (TTL) vorgehalten. Nach Ablauf der TTL werden sie automatisch gelöscht. Dies gewährleistet, dass der Vektorindex stets aktuelle Daten enthält und Speicherkapazitäten effizient genutzt werden, insbesondere da sich der medizinische Kontext eines Falls (der Encounter) mit der Zeit ändert.

- **Fachliche Vorfilterung:** Bereits vor dem eigentlichen Embedding erfolgt eine **fachliche Vorselektion** basierend auf den Metadaten des DocumentReference-Knotens:

- **Typ:** Irrelevante Dokumenttypen für die aktuelle Fragestellung (z.B. Administratives, das keinen direkten medizinischen Inhalt liefert) werden ausgeschlossen.

- **Zeitraum:** Dokumente außerhalb eines relevanten Zeitfensters (z.B. Befunde, die älter als fünf Jahre sind und für die Akutversorgung nicht mehr relevant) werden ignoriert.

Fazit zur Steuerung: Die Kopplung von GraphRAG (für Metadaten, Beziehungen und Autorisierung) und Dokumenten-RAG (für den inhaltsbasierten Abruf und das Embedding) schafft ein hochgradig kontrolliertes, sicheres und performantes System zur Wissensgewinnung aus großen, heterogenen Datenbeständen.

11. Chat-Orchestrierung (Tool-Routing und Antwortgenerierung) {#11.-chat-orchestrierung-(tool-routing-und-antwortgenerierung)}

Die Chat-Orchestrierung stellt den zentralen Mechanismus dar, um Benutzeranfragen effizient und präzise durch die verfügbaren Wissensdomänen zu routen und eine umfassende, nachvollziehbare Antwort zu generieren.

11.1 Phasen der Verarbeitung {#11.1-phasen-der-verarbeitung}

11.1.1 Intent-Erkennung und Kontextualisierung

Zunächst wird die Benutzeranfrage durch ein spezialisiertes Large Language Model (LLM) analysiert, um den **Intent** (die Absicht, z.B. "Diagnose-Validierung", "Therapie-Empfehlung", "Faktenabfrage zum Patienten") zu bestimmen. Parallel erfolgt eine **Kontextualisierung**, d.h., die Extraktion relevanter medizinischer Entitäten (z.B. Patient, Krankheit, Medikamente, Laborwerte) zur Vorbereitung des Tool-Routings.

11.1.2 Sequenzielles Tool-Routing (Wissensquellen-Abfrage)

Die ermittelten Entitäten steuern die Abfrage der dedizierten Retrieval-Augmented Generation (RAG) Systeme. Die Reihenfolge dieser Abfragen ist **sequenziell und priorisiert**, um die Relevanz und Qualität der Fakten zu maximieren:

- **a) GraphRAG (Strukturierte FHIR-Fakten): Höchste Priorität** hat die Abfrage des Patienten-Graphen, welcher strukturierte, interoperable klinische Daten im FHIR-Format (Fast Healthcare Interoperability Resources) enthält. Hier werden primär **patientenspezifische Fakten** (Diagnosen, Labor, Medikation, Anamnese) abgerufen. Die Ergebnisse sind hochstrukturiert und direkt zitierbar.

- **b) Dokumenten-RAG (HYDMedia):** Im nächsten Schritt werden **patientenbezogene Freitext-Informationen** aus dem Dokumenten-Archiv (HYDMedia-System) abgefragt. Dies umfasst Arztbriefe, Befunde, OP-Berichte oder Pflegeprotokolle, die nicht im strukturierten Graphen abgebildet sind. Die Ergebnisse sind Text-Chunks (Zitate).

- **c) Leitlinien-RAG (Globale Evidenz):** Zuletzt erfolgt die Abfrage der globalen Evidenzbasis. Hierzu zählen medizinische Leitlinien (z.B. AWMF, nationale / internationale Fachgesellschaften), Studien und Standard-Referenzwerke. Diese dienen zur **Validierung** der patientenspezifischen Fakten und zur Ableitung generischer Empfehlungen.

11.1.3 Synthese und Antwortstruktur

Das LLM aggregiert die Ergebnisse aller RAG-Abfragen und synthetisiert sie zu einer einzigen, mehrteiligen Antwort. Die **Antwortstruktur** folgt einem klaren, fest definierten Format:

- **Patientenfakten (Graph, strukturiert):** Direkte, zusammenfassende Aussagen, die aus dem FHIR-Graphen stammen (z.B. "Der Patient hat die Diagnose F32.9

(Depressive Episode, nicht näher bezeichnet) seit dem 01.01.2025. Aktuelle Medikation: Citalopram 20mg 1-0-0.").

- **Dokumentenzitate (HYDMedia):** Direkte, belegbare Text-Ausschnitte aus den Quell-Dokumenten, die die strukturierten Fakten ergänzen oder kontextualisieren (z.B. *Zitat aus dem Entlassbrief: "Die Stimmungslage des Patienten ist im Vergleich zur Aufnahme deutlich gebessert, jedoch persistieren ruminierende Gedanken."*).

- **Leitlinien / Evidenz:** Allgemeine medizinische Empfehlungen oder Evidenzgrade, die sich auf die erkannten Entitäten beziehen und die vorgeschlagenen Schlussfolgerungen untermauern.

11.2 Klinische Notwendigkeit und Risikomanagement {#11.2-klinische-notwendigkeit-und-risikomanagement}

Die strikte Trennung und die strukturierte Ausweisung der Quellen sind nicht nur ein technologisches, sondern vor allem ein **klinisches Erfordernis**.

Diese Methodik ist bewusst gewählt zur **klinischen Nachvollziehbarkeit und Haftungsreduktion**. Sie gewährleistet, dass jede Aussage des Chatbots auf einer identifizierbaren, **primären Datenquelle** beruht. Ein klinischer Anwender (Arzt/ Pflegekraft) muss in der Lage sein, die Herkunft jeder Information (Patientenakte vs. Standardwissen) sofort zu erkennen und zu überprüfen. Dies ist essenziell für die Übernahme der Verantwortung im klinischen Entscheidungsprozess. Durch die klare Zitierung wird das Risiko einer "halluzinierten" oder falsch zugeordneten Information minimiert.

12 Berechtigungsmanagement & Sicherheit (SAP-geführt) – Detaillierte Ausgestaltung {#12-berechtigungsmanagement-&-sicherheit-(sap-geführt)-–-detaillierte-ausgestaltung}

12.1 Grundsatz: SAP IS-H als „Master of Permission" und die Architektur der Datenhoheit {#12.1-grundsatz:-sap-is-h-als-„master-of-permission"-und-die-architektur-der-datenhoheit}

Das Berechtigungs- und Sicherheitskonzept von UKLGPT / hAIppokrates basiert auf dem unumstößlichen Prinzip der **Datenhoheit des führenden klinischen Systems**. Der Chatbot ist konzeptionell darauf ausgelegt, **keine eigene fachliche Berechtigungslogik** zu implementieren, zu interpretieren oder gar zu erweitern.

SAP IS-H (SAP Healthcare), als zentrales Krankenhausinformationssystem (KIS), ist der **alleinige „Master of Permission"**. Die Berechtigungsdefinitionen (inkl. Behandlungsauftrag, Organisationseinheiten-Zuordnung, funktionale Rechte, Dienst-/Planstellen) sind dort hinterlegt und werden in Echtzeit oder nahezu in Echtzeit repliziert.

Implikationen dieses Architekturprinzips:

- **Kein Patientenkontext ohne gültige SAP-Berechtigung:** Die Initialisierung jeglicher KI-Aktivität ist zwingend an eine positive Berechtigungsprüfung im führenden System geknüpft.
- **Kein Retrieval-Augmented Generation (RAG)-Aufbau ohne vorherige Freigabe:** Die Zusammenstellung der patientenspezifischen Wissensbasis (Embeddings, Dokumentauszüge) ist ein nachgeordneter Schritt, der erst nach erfolgreicher Autorisierung erfolgt.
- **Datenschutz durch Architektur („Privacy by Design"):** Durch die strikte Trennung von Berechtigungsdefinition (SAP) und technischer Durchsetzung (UKLGPT) wird sichergestellt, dass die KI-Komponente per Design nicht über die notwendigen Informationen verfügt, um Berechtigungen zu umgehen oder eigenständig zu vergeben.

KRITISCHER HINWEIS – Migration der Berechtigungshoheit (QA-Befund C-01):

SAP IS-H wird im **Oktober 2026 abgeschaltet**. Das hier beschriebene Prinzip der SAP-geführten Berechtigung muss daher auf den Nachfolger **M-KIS (Meierhofer)** migriert werden. Die Architektur ist so konzipiert, dass das Prinzip „führendes KIS = Master of Permission" systemunabhängig gilt. Die konkrete Migration erfordert:

1. **Abstimmung mit Meierhofer:** Klärung, welche Berechtigungs-APIs und Behandlungsauftragsdaten M-KIS bereitstellt.
2. **Übergangsphase:** Während des Parallelbetriebs (SAP + M-KIS) muss ein Routing definiert werden, welches System für welchen Fall die Berechtigungshoheit hält.
3. **DWH-Migration:** Die SAP-Replikation im Echtzeit-DWH muss auf M-KIS-Datenquellen umgestellt werden.
4. **Testkonzept:** Vollständiger Regressionstest des Berechtigungsflusses nach M-KIS-Umstellung.

Im weiteren Text wird „SAP IS-H" als aktueller Master beschrieben. Nach Migration gilt identische Logik für M-KIS.

12.1.1 SAP-Berechtigungsmodell im Detail (Ist-Stand am UKL)

Quelle: „Projektdokumentation Berechtigungen SAP klinische Module.docx", „Produkt Berechtigungskonzept.docx"

Das produktive SAP-Berechtigungsmodell am UKL operiert auf **zwei Steuerungsebenen**, die beide in UKLGPT abgebildet werden müssen:

Ebene 1 – Behandlungsauftrag (Datenraum): Der Behandlungsauftrag steuert, **wann** ein Nutzer auf Patientendaten zugreifen darf. Er definiert den zugänglichen Datenraum auf Basis der OE-Zuordnung von Benutzer und Patient.

Behandlungsauftrag-Typ	Auslöser	Geltungsdauer	UKLGPT-Mapping
Dynamisch (automatisch)	Termine, KLAU, Bewegungen im SAP erzeugen sofort BA	Ambulant: 200 Tage, Stationär: 14 Tage nach Entlassung	`ACCESS_GRANTE` bei aktiver BA-Relation im Graph
Temporär (Selbstbedienung)	Benutzer beantragt Zugriff manuell	Bis 24:00 Uhr des Beantragungstages	Zeitlich befristetes Token mit Loggin
Notfallnutzer			

Behandlungsauftrag-Typ	Auslöser	Geltungsdauer	UKLGPT-Mapping
	Medizinischer Notfall (Patientensicherheit)	Zeitlich begrenzt, vollständig geloggt	Break-the-Glass gemäß Kap. 12.4

Ebene 2 – Funktionale Berechtigungen (Sammelberechtigungsgruppen): Die SAP-Berechtigungen steuern, **was** ein Nutzer tun darf. Aufbau: Transaktionen → Berechtigungsgruppen (BG) → Sammelberechtigungsgruppen (SBG) → Benutzergruppen.

Benutzergruppe	SBG-Typ	Besonderheit
Ärzte	Standard-SBG + OE-spezifische SBG	Vollzugriff auf Patientendaten der eigenen OE
Pflege	Standard-SBG + OE-spezifische SBG	Lesend + dokumentierend
Verwaltung	Standard-SBG (eingeschränkt)	Abrechnungsbezogener Zugriff
Forschung	Über Antragsverfahren (DIZ)	Separater Genehmigungsworkflow

OE-spezifischer Dokumentenschutz (PSY/KJP/PST):

> Besonders schützenswerte Dokumente der Psychiatrie, Psychosomatik und Kinder-/Jugendpsychiatrie unterliegen einem **zusätzlichen dokumentenspezifischen Zugriffsschutz**. Fachfremde Einrichtungen dürfen auf diese Dokumente NICHT zugreifen, selbst wenn ein Behandlungsauftrag vorliegt.
>
> Betroffene Dokumenttypen: `PSY_VER_01`, `ÄRZ_VERL_P`, `PST_VER_01`, `PST_BEF_01` (OEs: PST-1, PST-T, PSTA1, PSTA2, E03-1, E03-2, J00-3, KJPA1, KJPA2)
>
> **UKLGPT-Implikation:** Diese Dokumenttypen müssen bei der FHIR-Abfrage über `DocumentReference`-Metadaten gefiltert werden, bevor sie in den RAG-Prozess eingespeist werden (vgl. Kap. 12.3.3 Ebene 3: Dokumentenschutz).

IGA-System-Integration (Zielzustand): Die automatische Berechtigungsverwaltung erfolgt künftig über das **IGA (Identity Governance and Administration)**-System, das HR-Daten (Planstelle, Stelle, Dienstart, Teilbereich „10" = Patientenversorgung) automatisch in SAP-Rollen überführt. UKLGPT muss diese Automatik

berücksichtigen und bei Personalwechseln die Berechtigungen in Echtzeit reflektieren.

VIP-Behandlung: Keine gesonderte Einschränkung vorgesehen.
Inaktive Patienten: Eigene BG erforderlich – standardmäßig KEIN automatischer Zugriff.

12.2 Gatekeeper-Prinzip: Der Patient-Scoped RAG-Ansatz {#12.2-gatekeeper-prinzip:-der-patient-scoped-rag-ansatz}

UKLGPT agiert als **Patient-Scoped RAG-System**. Dies bedeutet, dass die gesamte Verarbeitung des Large Language Models (LLM) auf den Daten eines **eindeutig identifizierten, autorisierten Patienten** beschränkt ist. Die Berechtigungsprüfung fungiert hier als strikter **Gatekeeper** und erfolgt **zwingend vor** dem Aufbau des patientenspezifischen Wissensraums.

Der Autorisierungsprozess (Logik):

- **Initialfrage:** Darf der aktuell authentifizierte **User X** den konkreten **Patienten Y** *zum jetzigen Zeitpunkt* basierend auf den definierten SAP IS-H-Regeln (Behandlungsauftrag, OE-Zugehörigkeit etc.) sehen?
- **Ergebnis A (Negativ):** ❌ **Nein** → Die KI-Anwendung verweigert den Zugriff. Es findet **kein RAG-Aufbau** statt, **keine Daten** werden geladen, und es kann **keine Antwort** generiert werden. Die Sitzung wird protokolliert (versuchter Zugriff).
- **Ergebnis B (Positiv):** ✅ **Ja** → Der **RAG-Wissensraum** (Indexierung der relevanten Dokumente und Daten) wird aufgebaut. Dieser kann gegebenenfalls durch eine nachgelagerte **Sub-Filterung sensibler Inhalte** (siehe 12.3.3, Ebene 3) weiter eingeschränkt werden, um das „Need-to-Know"-Prinzip maximal durchzusetzen.

12.3 Die Drei Ebenen der gestaffelten Zugriffskontrolle {#12.3-die-drei-ebenen-der-gestaffelten-zugriffskontrolle}

Der Zugriff wird durch eine gestaffelte, hierarchische Prüfung sichergestellt, die die Identität, den Behandlungsauftrag und den Dokumentenschutz berücksichtigt.

12.3.1 Ebene 1: Kontext-Check (Identitäts- und Rollenprüfung)

Diese Ebene stellt die **eindeutige und unveränderbare Identifikation** des Anwenders und des klinischen Kontexts sicher.

- **Initialisierung:** Der Aufruf des Chatbots erfolgt ausschließlich aus dem M-KIS (medizinisches klinisches Informationssystem) heraus. Dabei werden mindestens die **User-ID** (aus dem M-KIS-Kontext) und die **Fallnummer** (Patientenkontext) sicher an den Chatbot übergeben.

- **Authentifizierung:** Abgleich der übergebenen User-ID mit dem **Windows Single Sign-On (SSO)** oder dem **Active Directory (AD)**, um die Authentizität des Benutzers zu verifizieren.

- **IAM-Mapping:** Die erfolgreiche Identität wird auf die relevanten SAP-Daten abgebildet: **SAP-Personalnummer**, **Planstelle** und die zugeordnete **Organisationseinheit (OE)** (Abteilung, Station).

Ziel: Eine revisionssichere, eindeutige Zuordnung des KI-Zugriffs zu einer realen Person mit einer definierten klinischen Rolle.

12.3.2 Ebene 2: Behandlungsauftrag (Überprüfung des Datenraums)

Der Behandlungsauftrag ist das **zentrale DSGVO-Instrument** zur Einhaltung des **„Need-to-Know“-Prinzips**. Die Prüfung erfolgt in Echtzeit gegen das **Echtzeit-DWH (SAP-Replikation)** oder einen dedizierten **Graph-Datenbank-Layer**, der die Berechtigungen performant abbildet.

- **Szenario A – Dynamisch (Automatisch/Regulär):** Die primäre und bevorzugte Zugriffsform.

- Der User ist im SAP der aktuell **behandelnden Organisationseinheit (OE)** zugeordnet (z. B. der internistischen Station 2).

- Der Patient ist der entsprechenden OE zugeordnet (liegt auf dieser Station / hat einen Termin in dieser Ambulanz).

- **Ergebnis:** Zugriff wird **automatisch und sofort erlaubt**, da ein aktiver Behandlungsauftrag durch die klinische Zugehörigkeit impliziert ist.

- **Szenario B – Temporär (Manuell/Ad-hoc):** Für Konsile, Vertretungen oder nicht-klinische, aber berechtigte Zugriffe (z. B. Kodierung, MDK-Vorbereitung).

- **Kein aktiver Behandlungsauftrag** nach Szenario A vorhanden.

- **Chatbot-Interaktion:** Der Chatbot fordert explizit eine **Bestätigung** des Zugriffs und eine **Begründung** (z. B. "Konsil Anästhesie", "Vertretung").

- **Ergebnis:** Der Zugriff wird **zeitlich begrenzt** gewährt (z. B. bis zum Ende des Arbeitstages oder der Schicht). Dies wird mit **vollständiger Auditierung** des manuellen Eingriffs dokumentiert.

- **Szenario C – Erweitert (Interdisziplinär/Generell):** Für zentrale, fallübergreifende Einheiten.

- Einheiten wie die **Zentrale Notaufnahme (ZNA)**, **Anästhesie-Pool** oder der **Springer-Pool** müssen auf eine breitere Palette von Fällen zugreifen können.

- **Abbildung:** Dies erfolgt entweder über übergeordnete **OE-Beziehungen** im SAP (hierarchische Berechtigung) oder über dedizierte **Flags** im DWH/Graph-DB-Layer, die eine temporär erweiterte Berechtigung signalisieren.

12.3.3 Ebene 3: Dokumentenschutz (Sub-Filterung Sensibler Inhalte)

Selbst bei einem gültigen Behandlungsauftrag (Ebene 2) können spezielle Dokumente zusätzlichen, **dokumentenspezifischen Zugriffsbeschränkungen** unterliegen.

- **Betroffene Bereiche:** Hochsensible Fachbereiche wie **Psychiatrie (PSY), Psychosomatik (PST)** oder die **Kinder- und Jugendpsychiatrie (KJP)**.

- **Mechanismus:** Vor dem Embedding und der Indexierung der Dokumente erfolgt eine **Prüfung der Dokument-Metadaten** (z. B. aus HYDMedia über FHIR DocumentReference).

- **Regel-Engine:**

- IF Dokument.Fachabteilung $\in$ {PSY, PST, KJP}

- AND User.Fachabteilung $\notin$ {PSY, PST, KJP}

- THEN Dokument wird nicht in den RAG-Index geladen (Sub-Filterung).

Konsequenz: Der Chatbot hat technisch keinen Zugriff auf diese Inhalte und kann **keine Aussagen** dazu treffen, was eine maximale Sicherheit des Fachgeheimnisses gewährleistet.

12.4 Sonderfall: Notfallmodus („Break-the-Glass") {#12.4-sonderfall:-notfallmodus-(„break-the-glass")}

Zur Gewährleistung der Patientensicherheit in kritischen Situationen muss ein kontrollierter Ausnahmezugang möglich sein, der die restriktiven Berechtigungen temporär übersteuert.

- **Aktivierung:** Der Notfallmodus muss **explizit im User Interface (UI)** durch den Anwender ausgelöst und bestätigt werden.
- **Umfang:** Er gewährt einen **zeitlich begrenzten Vollzugriff** auf alle Daten des aktuellen Patientenfalls.
- **Sicherheitsmaßnahmen:**
- **Zwingende, lückenlose Protokollierung** des Zugriffs und der Begründung.
- **Automatisches Alerting** (z. B. an den Datenschutzbeauftragten (DSB) oder die Vorgesetzten), das den „Break-the-Glass"-Vorgang in Echtzeit meldet.

12.5 Technische Umsetzung {#12.5-technische-umsetzung}

Um die komplexen hierarchischen und temporären Berechtigungsbeziehungen performant zu überprüfen, wird ein Graph-Datenbank-Layer eingesetzt.

- **Graph-Modellierung:** Die Beziehungen werden als Kanten und Knoten abgebildet:
- User ↔ OE ↔ Patient (Permanente/Statische Berechtigungen)
- **Temporäre Zugriffe:** Abbildung als **zeitlich begrenzte Kanten** (z. B. [:HAS_TEMP_ACCESS {until: '2026-02-10T23:59:59'}]).
- **Notfallstatus:** Explizites **Flag** ({is_emergency: true}) an der User-Patient-Kante.
- **Rolle des Graphen:** Der Graph dient ausschließlich der **Durchsetzung (Enforcement)** der Berechtigungen durch hochperformante Queries. Die **Definition (Master Data)** der Rechte bleibt weiterhin in SAP IS-H.

12.6 Audit- und Logging-Konzept: Revisionssichere Dokumentation {#12.6-audit--und-logging-konzept:-revisionssichere-dokumentation}

Jeder Interaktion des Users mit der KI, die potenziell auf Patientendaten zugreift, muss revisionssicher und gerichtsfest protokolliert werden, um die Transparenz und Nachvollziehbarkeit zu gewährleisten.

12.6.1 Protokollierte Parameter (Core-Audit Trail):

- **Wer?** (Eindeutige User-ID, Rolle, Planstelle)
- **Wann?** (ISO-Zeitstempel der Anfrage)
- **Worauf?** (Patienten-ID, Fallnummer)
- **Warum?** (Zugriffsgrund: dynamisch, temporär, Notfall)
- **Was?** (Die **initiale User-Anfrage** – die KI-generierte Antwort selbst wird **nicht** im Audit-Trail gespeichert, um Trainingsdaten zu vermeiden, aber die Tatsache der Anfrage wird dokumentiert).

12.6.2 Löschkonzept für RAG-Artefakte:

Die hochsensiblen, patientenspezifischen Datenstrukturen, die für den RAG-Prozess generiert werden (Embeddings und ggf. temporäre Dokumentkopien), werden unmittelbar nach Beendigung des Prozesses gelöscht, um das Risiko der Persistenz zu minimieren:

- **Sofortige Löschung bei:**
- Wechsel des Patienten durch den Anwender
- Logout aus dem M-KIS/der Anwendung
- Automatischem Timeout der Sitzung (Inaktivität)

12.7 Ende-zu-Ende-Berechtigungsfluss (Detailliertes textuelles Ablaufmodell) {#12.7-ende-zu-ende-berechtigungsfluss-(detailliertes-textuelles-ablaufmodell)}

Der gesamte Berechtigungsfluss zur Gewährleistung des datenschutzkonformen Zugriffs auf Patienteninformationen ist **vollständig deterministisch, nicht bypassbar** und **durchgängig auditierbar**. Dies stellt sicher, dass zu keinem Zeitpunkt eine

Verarbeitung von Patientendaten ohne explizit validierte Zugriffsberechtigung stattfindet.

Der Ablauf gliedert sich in folgende zwingende Schritte:

12.7.1 Initialisierung & Authentifizierung (User-Login)

Der **User** meldet sich über die standardisierten und gehärteten Klinik-IT-Systeme per Windows-Anmeldung (Active Directory) oder Single Sign-On (SSO) an. Diese Schritte etablieren die eindeutige **digitale Identität** des Nutzers im Kliniknetzwerk.

12.7.2 Anforderungsinitiierung (M-KIS Aufruf)

Das Medizinisches Klinik-Informations-System (**M-KIS**) sendet einen API-Aufruf an das UKLGPT-System. Dieser Aufruf beinhaltet zwingend die eindeutige **User-ID** (aus Schritt 1) und die konkrete **Fallnummer** (Encounter-ID) des Patienten, dessen Daten verarbeitet werden sollen.

12.7.3 Vorsperre (UKLGPT Gate)

UKLGPT implementiert eine strikte „Fail-Safe“-Architektur: Es wird **keine** Datenverarbeitung, keine API-Antwort und kein weiterer Verarbeitungsschritt initiiert, solange die **finale ACCESS_GRANTED-Entscheidung** aussteht. Der Berechtigungs-Check ist der erste logische Schritt.

12.7.4 Identitäts- und Rollenzuordnung (IAM/AD)

Das **Identity and Access Management (IAM)** System, basierend auf dem **Active Directory (AD)** der Klinik, führt eine Tiefenprüfung der User-ID durch. Es werden die für die Berechtigungsprüfung kritischen organisatorischen Attribute des Users ermittelt:

- Zugehörige **SAP-Personalnummer**.
- Aktuelle **Planstelle** (Funktionszuordnung).
- Zugehörige **Organisationseinheit (OE)** (Ort der Leistungserbringung).

12.7.5 Medizinisch-Funktionale Rechte (SAP IS-H Validierung)

Unter Verwendung der in Schritt 4 ermittelten Attribute fragt UKLGPT das Krankenhaus-Informationssystem (**SAP IS-H**) ab. SAP IS-H liefert die **Behandlungsauftragsrelation** (Ist der User am Fall beteiligt?) und die **funktionalen Rechte** (Darf der User *prinzipiell* in seiner Rolle *diese* Art von Daten einsehen?). Dies stellt die Einhaltung der medizinischen Notwendigkeit ("Need-to-Know") sicher.

12.7.6 Strukturelle Berechtigung

Parallel oder nachfolgend wird das **Echtzeit-Data-Warehouse (DWH)**, primär in Form eines **Graphdatenbanksystems**, konsultiert. Hier wird die **dynamische, strukturelle Beziehung** in Echtzeit geprüft:

- Beziehung User ↔ Organisationseinheit (OE) ↔ Patient.
- Es wird geprüft, ob die OE, der der User zugeordnet ist, auch diejenige ist, die aktuell den Behandlungsauftrag für den Patienten hält.
- Diese Prüfung verhindert den Zugriff von Personal, das zwar im Haus, aber nicht am aktuellen Behandlungsort oder in der zuständigen Abteilung tätig ist.

12.7.7 Finale Gatekeeper-Entscheidung:

Das System aggregiert die Ergebnisse der Schritte 3, 4, 5 und 6. Die Entscheidung ist binär:

- **ACCESS_GRANTED:** Nur wenn alle Prüfschritte (Identität, Rolle, Behandlungsauftrag, funktionale Rechte, strukturelle Zuordnung) positiv durchlaufen wurden. In diesem Fall wird der **Patient-Scoped Retrieval-Augmented Generation (RAG)** Prozess für diesen spezifischen Fall gestartet.
- **ACCESS_DENIED:** Bei Nichterfüllung auch nur einer einzigen Bedingung. Der Prozess wird **sofort und irreversibel abgebrochen**, ohne dass der User Zugriff auf Patientendaten erhält. Eine detaillierte Fehlermeldung wird protokolliert.

12.7.8 Zwingende Sicherheitshypothese:

Es existiert **kein technischer oder logischer Pfad** innerhalb des UKLGPT-Systems, der die Verarbeitung oder Verfügbarmachung von Patientendaten ermöglicht, ohne dass eine **SAP-validierte und strukturell bestätigte Berechtigung** vorliegt.

12.8 Verknüpfung der Berechtigungsprüfung mit der Prompt-Pipeline (Orchestrierungslogik) {#12.8-verknüpfung-der-berechtigungsprüfung-mit-der-prompt-pipeline-(orchestrierungslogik)}

Die umfassende Berechtigungsprüfung nach Kapitel 12.7 ist **kein vorgeschaltetes, externes Modul**, sondern ein **integraler und zwingender Bestandteil der Prompt-Orchestrierung** durch das

UKLGPT-System. Sie steuert den gesamten Datenfluss in der nachgeschalteten Retrieval-Augmented Generation (RAG) Kette.

Die Zwingende Regel der Orchestrierung:Keine Prompt-Pipeline, kein Daten-Retrieval und keine LLM-Inferenz darf ohne den vorangegangenen Status *ACCESS_GRANTED* initiiert werden.

Diese Verknüpfung wird durch folgende technische und logische Design-Entscheidungen gewährleistet:

12.8.1 Vorklassifikation ohne Daten (Domain Classification):

Die initiale **Domain Classification** (z.B. "Ist die Anfrage eine Laborwertanfrage oder eine Anamnese-Zusammenfassung?") erfolgt **ausschließlich auf Basis des reinen User-Prompts** und *ohne* jeglichen Zugriff auf Patientendaten. Dies ist der einzige Verarbeitungsschritt, der vor dem ACCESS_GRANTED Status ausgeführt werden darf.

12.8.2 Zwingender Encounter (Kontext-Erhebung):

Ab dem Moment, in dem die **Kontext-Erhebung** (Retrieval von Daten) beginnen müsste, ist ein **gültiger, aktiver und berechtigter Encounter** (Fallnummer mit ACCESS_GRANTED) **zwingend erforderlich**.

12.8.3 Daten-Filterung am Ursprung (Kontext-Assembler):

Der **Kontext-Assembler**, die Komponente, die Daten aus den Quellsystemen sammelt, erhält **ausschließlich** die IDs von **freigegebenen Graph-Fakten** und **freigegebenen Dokument-IDs**. Diese IDs werden direkt aus dem Ergebnis der Berechtigungsprüfung (Schritt 6 und 7) abgeleitet. Der Assembler kann physisch keine Daten abrufen, deren ID nicht explizit als berechtigt markiert wurde.

12.8.4 Technisches Nicht-Existieren (Data Obfuscation):

Alle **nicht freigegebenen Inhalte** (z.B. Dokumente, zu denen der User keinen Behandlungsauftrag hat) werden für die nachfolgenden Prozesse (Embedding-Suche, Vektordatenbank-Retrieval) als **technisch nicht existent** behandelt. Sie werden weder im Index gesucht, noch als Ergebnis zurückgegeben. Das System agiert, als wären diese Daten für diesen User-Kontext physisch nicht vorhanden.

12.8.5 Systematischer Sicherheitsgewinn:

Durch diese Architektur wird nicht nur der unzulässige Datenzugriff *verhindert*, sondern auch die Qualität der Ausgabe **systematisch gesteuert**:

- **Verhinderung von Halluzinationen:** Da nur autorisierte und damit kontextuell relevante Fakten in den Prompt gelangen, wird die Grundlage für fehlerhafte, "halluzinierte" Antworten signifikant reduziert.
- **Ausschluss unzulässiger Datenreferenzen:** Es ist technisch unmöglich, dass das LLM auf Inhalte referenziert, die aus datenschutzrechtlichen Gründen nicht freigegeben wurden. Die gesamte Antwort des LLM basiert ausschließlich auf dem **Patient-Scoped Kontext**.

13. Compliance {#13.-compliance}

Die Entwicklung und der Betrieb des UKLGPT-Systems als spezialisiertes KI-Assistenzsystem im klinischen Umfeld unterliegt strengen regulatorischen Anforderungen, die über die reine DSGVO-Konformität hinausgehen.

13.1 Datenschutz & Regulierung (im Kontext der DSGVO) {#13.1-datenschutz-&-regulierung-(im-kontext-der-dsgvo)}

Die Implementierung dieses Systems erfolgt unter strenger Beachtung der Vorgaben der Datenschutz-Grundverordnung (DSGVO), insbesondere im Hinblick auf die Verarbeitung besonderer Kategorien personenbezogener Daten (Gesundheitsdaten, Art. 9 DSGVO).

13.1.1 Datenschutzkonzept

Das Datenschutzkonzept des UKLGPT-Systems ist ein grundlegendes, rechtsverbindliches Dokument, das den rechtskonformen Umgang mit den besonders schützenswerten Patientendaten (Art. 9 DSGVO) detailliert beschreibt. Es berücksichtigt die Kernprinzipien der DSGVO, insbesondere die **Zweckbindung** und die **Datenminimierung**.

Das Konzept identifiziert und bewertet die datenschutzrechtlichen Risiken, die aus dem Einsatz eines KI-gestützten RAG-Systems resultieren. Basierend auf dieser Analyse werden die notwendigen technischen und organisatorischen Schutzmaßnahmen (TOMs) verbindlich definiert. Die Erstellung dieses Konzepts, einschließlich der obligatorischen Datenschutz-Folgenabschätzung (DSFA) gemäß

Art. 35 DSGVO, ist die zwingende Voraussetzung für die spätere Freigabe des Systems. Das Gesamtkonzept setzt sich aus der initialen Datenschutzanalyse und dem daraus abgeleiteten Maßnahmenkatalog zusammen.

Die Einhaltung der DSGVO-Konformität dient als primäres Qualitätskriterium. Die finale Freigabe des Konzepts durch den **Datenschutzbeauftragten (Hr. Sünkel)** ist das maßgebliche Abnahmekriterium.

13.1.2 Kernprinzipien der DSGVO-Konformität:

- **Zweckbindung (Art. 5 Abs. 1 lit. b DSGVO):** Die Nutzung der Daten ist ausschließlich auf den Behandlungs- und Versorgungskontext beschränkt (**Default-Einstellung**). Eine sekundäre Nutzung, etwa für Forschungszwecke oder kommerzielle Analysen, ist durch technische und organisatorische Maßnahmen ausgeschlossen, es sei denn, der Patient hat hierzu eine explizite, informierte Einwilligung erteilt oder es besteht eine gesetzliche Grundlage.

- **Datenminimierung (Art. 5 Abs. 1 lit. c DSGVO):** Es werden nur die für die unmittelbare Beantwortung der medizinischen Fragestellung erforderlichen Daten verarbeitet. Dies wird durch zwei Mechanismen sichergestellt:

- **Episodische Subgraphen:** Der RAG-Aufbau (Retrieval-Augmented Generation) greift nur auf einen minimalen, **kontextspezifischen Ausschnitt** der Patientenakte (den Subgraphen) zu, der für den aktuellen Encounter relevant ist.

- **Time-to-Live (TTL):** Die im RAG-Prozess temporär generierten Datenstrukturen werden mit einer sehr kurzen Lebensdauer (TTL) versehen und sofort nach Abschluss der Abfrage gelöscht (**Sekundärdatenhaltung ist ausgeschlossen**).

- **Rechtmäßigkeit der Verarbeitung (Art. 6 & 9 DSGVO):** Die Verarbeitung von Gesundheitsdaten erfolgt i.d.R. auf Basis von Art. 9 Abs. 2 lit. h (Gesundheitsvorsorge oder medizinische Diagnose) in Verbindung mit nationalen Bestimmungen (§ 22 Abs. 1 Nr. 1 lit. b BDSG).

- **Rechenschaftspflicht und Dokumentation (Art. 5 Abs. 2, Art. 30 DSGVO):**

- **Datenschutz-Folgenabschätzung (DSFA):** Aufgrund der Verarbeitung besonderer Kategorien personenbezogener Daten (Art. 9) und der Nutzung neuer Technologien (KI-gestütztes RAG-System) ist die Durchführung einer **Datenschutz-Folgenabschätzung (DSFA) gemäß Art. 35 DSGVO zwingend erforderlich**. Die DSFA dient zur Bewertung der Risiken und zur Definition geeigneter Schutzmaßnahmen.

13.1.3 Anforderungen an das KI-Audit-Log (Art. 32 Abs. 1 lit. d, Art. 5 Abs. 1 lit. f DSGVO – Integrität und Vertraulichkeit):

Ein revisionssicheres Logging-System ist implementiert, um die Nachvollziehbarkeit jeder Datenabfrage zu gewährleisten und Missbrauchsfälle zu identifizieren:

- **Identifikation des Zugreifenden: Wer** hat die Abfrage initiiert (Arzt/Personal, mit Zeitstempel)?
- **Bezug zum Betroffenen: Welcher Patient / Encounter** war Gegenstand der Abfrage?
- **Art der Anfrage: Welcher Query-Typ (Template-ID)** oder welche Struktur der Abfrage wurde verwendet?
- **Keine Speicherung von Volltexten:** Die sensiblen Inhalte der Patientenakte oder die generierten Antworten werden **nicht** im Audit-Log gespeichert, um die Datenminimierung zu wahren.

13.1.4 Verbot der automatisierten Entscheidungsfindung (Art. 22 DSGVO):

- Das System ist explizit als **Informationsassistenz** konzipiert. Es darf **keine automatische Entscheidungsfindung** im Sinne des Art. 22 DSGVO treffen, die rechtliche Wirkung entfaltet oder den Betroffenen in ähnlicher Weise erheblich beeinträchtigt.
- Der Mensch (der behandelnde Arzt) bleibt stets in der Verantwortung und trifft die abschließende medizinische Entscheidung (Prinzip **"Human in the Loop"**). Das System dient lediglich der strukturierten Informationsaufbereitung (kein Clinical Decision Support **(CDS-Automatismus)**).

13.1.5 Ableitung für die Datenschutz-Folgenabschätzung (DSFA)

Die Architektur und die technischen Schutzmechanismen wurden entwickelt, um die in der DSFA zentral identifizierten Risiken proaktiv und explizit zu adressieren und das datenschutzrechtliche Restrisiko signifikant zu reduzieren:

Zweck: Diese Vorprüfung bewertet, ob eine formale Datenschutz-Folgenabschätzung erforderlich ist. Sie analysiert Art, Umfang und Risiko der Datenverarbeitung. Dadurch werden regulatorische Anforderungen frühzeitig erkannt. Die Vorprüfung reduziert rechtliche Unsicherheiten. Sie unterstützt eine rechtssichere Projektplanung.

Zusammensetzung: Prüfdokument

Qualitätskriterien: Vollständig

Abnahmekriterien: Zustimmung Datenschutzbeauftragter

Verantwortlich: Datenschutzbeauftragter (Hr. Sünkel)

DSFA-Risiko	Adressierung durch Architekturelement	Schutzmaßnahme
Unbefugter Zugriff (Art. 32)	Externe Autorisierungsebene	**SAP-geführter Gatekeeper** vor dem RAG-Aufbau: Gewährleistung, dass nur authentifiziertes und autorisiertes Personal auf die Daten zugreifen kann.
Zweckentfremdung (Art. 5 Abs. 1 lit. b)	Technischer Kontext-Zwang	**Zwang zum Encounter-Kontext:** Der Zugriff auf Daten ist technisch an einen aktiven Behandlungsfall (Encounter) gebunden, was eine zweckfremde Nutzung stark erschwert.
Sekundärdatenhaltung (Art. 5 Abs. 1 lit. c)	Löschkonzept	**TTL und sofortige Löschung:** Temporär erstellte Subgraphen und Indexstrukturen werden unmittelbar nach Abschluss der Abfrage gelöscht (Datenminimierung).
Fehlinformationen / Halluzination (Art. 5 Abs. 1 lit. d)	Qualitätssicherung der Ausgabe	**Mehrstufige Prompt-Pipeline mit Validierung:** Die generierte Antwort wird technisch auf Plausibilität und Konsistenz mit den Quelldokumenten geprüft.
	Auditierbarkeit	**Revisionssicheres KI-Audit-Log:**

DSFA-Risiko	Adressierung durch Architekturelement	Schutzmaßnahme
Mangelnde Nachvollziehbarkeit (Art. 5 Abs. 1 lit. a/f)		Lückenlose Protokollierung von *Wer, Wann, Auf Wen, Mit Welcher Anfrage*, um Integrität und Rechenschaftspflicht zu gewährleisten.

13.1.6 Landesspezifische Cloud-Restriktionen Sachsen (Besprechung 27.02.2026)

Im Rahmen der Besprechung vom 27.02.2026 wurde ein bisher nicht dokumentiertes Compliance-Risiko identifiziert: In Sachsen bestehen landesweite Restriktionen zur Cloud-Nutzung bei Universitätskliniken. Die genaue Rechtsgrundlage (Landesverordnung vs. Vereinbarung mit dem Landesdatenschützer) ist noch zu klären.

Relevanz für EMRGPT: - Die Eigenlösung EMRGPT ist davon **nicht betroffen** (vollständig On-Premise). - Die Averbis/Meierhofer-Variante nutzt **Azure OpenAI in Schweden** (EU Datazone) für die LLM-Verarbeitung und ist damit potenziell betroffen. - Falls Azure in Sachsen nicht genehmigt wird, wäre ein Wechsel auf deutsche Anbieter (Stackit, Arvato, Telekom) erforderlich, was **höhere Kosten** verursacht. - Ein **externes Datenschutzgutachten** (z. B. KPMG, ~100.000 €) könnte zusätzlich erforderlich werden.

Nächste Schritte: 1. Termin mit dem sächsischen Landesdatenschützer (März 2026) – AVVs und Architekturskizze vorab bereitstellen. 2. Ergebnis dokumentieren und ggf. in Kostengerüst (Kap. 20) einpreisen. 3. Referenzen aus anderen Bundesländern (z. B. Klinikum Rheine/Azure) als Vergleich heranziehen.

Quellen: Besprechungsprotokoll 27.02.2026 (Fachablage), Datenschutz-AVVs (ausstehend).

13.2 EU-Verordnung über Medizinprodukte (EU MDR – Verordnung (EU) 2017/745) {#13.2-eu-verordnung-über-medizinprodukte-(eu-mdr-–-verordnung-(eu)-2017/745)}

Das UKLGPT-System ist darauf ausgelegt, Informationen bereitzustellen und Vorschläge zu generieren (Informationsassistenz, siehe 3.4). Die Abgrenzung zur **medizinischen Zweckbestimmung** ist kritisch:

13.2.1 Abgrenzung zur Medizinprodukte-Software:

Solange das System **ausschließlich zur Informationsaggregation und -aufbereitung** dient und die **Letztentscheidung** unzweifelhaft beim klinischen Fachpersonal liegt, ist eine Einstufung als Medizinprodukt nach MDR unwahrscheinlich.

13.2.2 Risiko der Zweckbestimmung:

Falls Funktionen implementiert werden, die eine **direkte diagnostische oder therapeutische Empfehlung** beinhalten oder eine **automatisierte Auswertung von Patientendaten** mit klinischer Schlussfolgerung (z.B. "Wahrscheinlichkeit für Diagnose X ist hoch") ermöglichen, würde das System als **Medizinprodukte-Software** eingestuft.

Konsequenz bei Einstufung als Medizinprodukt: Es müssten strenge Anforderungen an das Qualitätsmanagementsystem (QMS), Risikomanagement, klinische Bewertung und Konformitätsbewertung (oftmals Klasse IIa oder höher) erfüllt werden, was den Entwicklungsaufwand signifikant erhöht.

13.2.3 Architektonische Maßnahme (Guardrails):

Die **Regel- und Guardrail-Engine** (siehe 5.2) muss aktiv verhindern, dass das System medizinische Entscheidungen oder verbindliche Handlungsanweisungen ausgibt.

13.3 EU AI Act (Verordnung zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz) {#13.3-eu-ai-act-(verordnung-zur-festlegung-harmonisierter-vorschriften-für-künstliche-intelligenz)}

Der EU AI Act (Verordnung (EU) 2024/1689), der am 1. August 2024 in Kraft getreten ist, betrifft den Einsatz von UKLGPT direkt.

- **Einstufung als Hochrisiko-KI-System:** KI-Systeme, die im Gesundheitswesen eingesetzt werden und Entscheidungen *beeinflussen* oder zur *Patiententriage* beitragen, werden im Regelfall als **Hochrisiko-KI** eingestuft (Annex III).
- **Auswirkung auf UKLGPT:** Da UKLGPT zur Unterstützung der klinischen Entscheidungsfindung eingesetzt wird, muss es die Anforderungen an Hochrisiko-KI erfüllen, selbst wenn es kein Medizinprodukt ist.
- **Anforderungen an Hochrisiko-KI:**
- **Daten-Governance und Bias-Prävention:** Es muss sichergestellt werden, dass die für das Training und den Betrieb

verwendeten Datensätze (insbesondere für das LLM und die RAG-Komponenten) repräsentativ und frei von systematischen Diskriminierungen oder Bias sind, die die Patientensicherheit oder die Gleichbehandlung gefährden könnten.

- **Robustheit, Genauigkeit und Cybersicherheit:** Notwendigkeit der systematischen Überwachung und Überprüfung der Genauigkeit der KI-Antworten (siehe 6.3 Prompt-Pipeline) sowie des Schutzes vor externen Manipulationen (Adversarial Attacks).

- **Transparenz und Protokollierung:** Die Ausgabe muss so strukturiert sein, dass sie die Herkunft der Information transparent macht (Quellenzitierung, siehe 11.2). Die Protokollierung (Audit-Log, siehe 12.6) muss detailliert nachweisen, welche Information zu welchem Zeitpunkt abgerufen wurde.

- **Menschliche Aufsicht (Human Oversight):** Das System muss so konzipiert sein, dass der Nutzer die Ergebnisse leicht interpretieren, validieren und ggf. ignorieren kann (permanenter Haftungsausschluss im UI, siehe 4.3).

13.4 IT-Sicherheitsgesetz und KRITIS-Regulierung (Deutschland) {#13.4-it-sicherheitsgesetz-und-kritis-regulierung-(deutschland)}

Als Krankenhausinformationssystem (KIS) oder als kritische Komponente desselben fällt die KI-Plattform unter die Regulierung Kritischer Infrastrukturen (KRITIS), sofern die Klinik die entsprechende Schwellenwert-Größe erreicht.

- **BSI-KritisV:** Das System muss die **Mindestanforderungen an die IT-Sicherheit** gemäß den BSI-KritisV erfüllen. Dies erfordert die Implementierung eines robusten **Informationssicherheits-Managementsystems (ISMS)** und die Durchführung regelmäßiger Audits (z.B. nach ISO 27001).

- **Meldepflichten:** Sicherheitsvorfälle, die die Funktion des Systems oder die Verfügbarkeit von Patientendaten kritisch beeinträchtigen, müssen unverzüglich an das Bundesamt für Sicherheit in der Informationstechnik (BSI) gemeldet werden.

Die Schutzbedarfsanalyse (SBA) ist ein fundamentaler Bestandteil des Informationssicherheits-Managementsystems (ISMS) und dient der Bewertung des kritischen Wertes der im UKLGPT-System verarbeiteten Informationen. Sie erfolgt nach der Basismethodik, wobei die Schutzziele **Vertraulichkeit (V)**, **Integrität (I)** und **Verfügbarkeit (A)** betrachtet werden.

Da es sich bei UKLGPT um ein System zur **Informationsassistenz** handelt und die **klinische Letztentscheidung stets beim Anwender verbleibt**, wird die Sicherheitsbewertung unter Berücksichtigung dieses architektonischen **Guardrails** durchgeführt.

13.4.1 Schutzbedarfsfeststellung

Die Bewertung basiert auf der Tatsache, dass das System **Patientendaten (Gesundheitsdaten gem. Art. 9 DSGVO)** und **klinisches Wissen** verarbeitet, dessen Fehlanwendung oder Verlust eine hohe Gefährdung für die Patientenversorgung darstellen kann.

Kriterium	Definition	Einstufung (UKLGPT)	Begründung
Vertraulichkeit (V)	Schutz vor unbefugter Preisgabe.	**Sehr Hoch**	Verarbeitung hochsensibler Patientendaten (Diagnosen, Befunde, Medikation, Verlaufsdaten). Bei unbefugtem Zugriff entstehen gravierende Schäden für die Betroffenen (Datenschutzverletzung, Verstoß gegen ärztliche Schweigepflicht).
Integrität (I)	Schutz vor unbefugter oder unbeabsichtigter Veränderung oder Verlust.	**Hoch**	Das System ist eine **Informationsassistenz** und *speichert* keine Primärdaten. **ABER**: Die Korrektheit der **abgerufenen** und **generierten** Information ist kritisch. Eine Veränderung der Fakten oder Leitlinien könnte zu falschen klinischen Entscheidungen des Arztes führen (Fehlinformation).
Verfügbarkeit (A)	Schutz vor Ausfall oder Nichtnutzbarkeit des Systems/ der Daten.	**Mittel**	Das System ist ein Assistenzsystem (kein primäres KIS). Ein Ausfall verhindert die KI-gestützte Wissensgewinnung, was zu einem erhöhten manuellen Suchaufwand führt, die

Kriterium	Definition	Einstufung (UKLGPT)	Begründung
			primäre klinische Versorgung bleibt jedoch über das KIS/M-KIS funktionsfähig.

13.4.2 Risikoreduktion durch Architektonisches Design (Guardrail-Effekt)

Die Architekturentscheidung, das System explizit auf die **reine Informationsassistenz** zu beschränken (siehe 13.1.4), hat eine direkte Auswirkung auf die Risikobewertung, insbesondere für die **Integrität**:

Schutzbedarfs-ziel	Risikofaktor ohne Guardrail	Architektonische Gegenmaßnahme	Risikoreduzierte Einstufung
Integrität (I)	Das System trifft eine **automatisierte Entscheidung** (z.B. "Verabreiche Medikament X"). Fehlerhafte Information führt direkt zu Patientenschaden.	**"Human in the Loop" (Art. 22 DSGVO-Konformität)**: Der Arzt trifft die Letztentscheidung. Die KI liefert nur *Informationen*. **Zusätzlich:** Mehrstufige Prompt-Pipeline (6.3) prüft Konsistenz.	**Hoch** (statt Sehr Hoch) – Das Risiko liegt nicht in der Automatisierung, sondern in der fehlerhaften *Bereitstellung* von Informationen, die ein menschlicher Prüfschritt abmildert.
Vertraulichkeit (V)	Unkontrollierter Zugriff auf alle Datenbestände.	**SAP-geführter Gatekeeper (12.2)** und **Patient-Scoped RAG** (12.3): Der Zugriff wird auf den Behandlungsauftrag begrenzt.	**Sehr Hoch** – Die Gefahr bleibt, aber die technische Angriffsfläche ist maximal reduziert (durch SSO und Gatekeeper).
Verfügbarkeit (A)	Systemausfall verhindert Notfallversorgung.	**Assistenzsystem**: Primäre Dokumentation und Notfallfunktionalität verbleiben im KIS.	**Mittel** – Der Ausfall verursacht Verzögerungen, aber keinen direkten Stillstand der Primärversorgung.

13.4.3 Fazit der Schutzbedarfsanalyse

Auf Basis der festgestellten Schutzbedarfe ergibt sich der folgende Gesamt-Mindestschutzbedarf für das UKLGPT-System:

- **Vertraulichkeit (V): Sehr Hoch**
- **Integrität (I): Hoch**
- **Verfügbarkeit (A): Mittel**

Die Einstufung als **Hochrisiko-KI** (gemäß EU AI Act) und die Verarbeitung von Gesundheitsdaten (**Art. 9 DSGVO**) erzwingen dennoch die Implementierung von Sicherheitsmaßnahmen, die dem **SEHR HOHEN** Schutzbedarf der Vertraulichkeit und der daraus abgeleiteten hohen Anforderungen an die Integrität (Fehlinformationsvermeidung) gerecht werden müssen. Die Architektur liefert durch die explizite Rolle als *Assistenz* die notwendigen **organisatorischen und technischen Guardrails**, um die Risiken kontrolliert zu halten und die Einhaltung der Rechenschaftspflicht (Art. 5 Abs. 2 DSGVO) zu gewährleisten.

13.5 Nationale Berufsordnungen und Haftungsrecht {#13.5-nationale-berufsordnungen-und-haftungsrecht}

- **Ärztliche Verantwortung:** Die Architektur muss die **Letztverantwortung des Arztes** für die klinische Entscheidung juristisch unangreifbar machen. Die Haftungsausschlussklauseln (3.4) müssen technisch und prozessual gestützt werden, indem das System keine Entscheidungsautonomie beansprucht.
- **Dokumentationspflicht:** Da die KI-Assistenz in den klinischen Workflow eingreift, müssen die Ergebnisse und die ihnen zugrundeliegenden Quellen lückenlos dokumentierbar sein (Audit-Trail).

Regulatorische Anforderung	Betroffenes Architekturelement	Erfüllungsstrategie im UKLGPT
DSGVO Art. 9 (Gesundheitsdaten)	Berechtigungs-Layer, RAG-Architektur	**SAP-geführter Gatekeeper** (12.1), **TTL & Löschkonzept** temporärer Embeddings (12.6.2), **Zwang zur Zweckbindung** (13.1).
EU MDR (Medizinprodukt)	Guardrail-Engine, Funktionsumfang	**Strikte Limitierung** der Funktionen auf **reine Informationsassistenz**, um die Einstufung als

Regulatorische Anforderung	Betroffenes Architekturelement	Erfüllungsstrategie im UKLGPT
		Medizinprodukt zu vermeiden (3.4).
EU AI Act (Hochrisiko-KI)	Prompt-Pipeline, Audit-Log	**Transparenz** durch Quellennachweis (11.2), **Genauigkeit** durch Prompt-Validierung (6.3), **Menschliche Aufsicht** durch UI-Hinweise (4.3).
KRITIS/BSI-KritisV	Schnittstellen, Betriebsumgebung	Implementierung eines **ISMS** und Einhaltung technischer Mindestanforderungen zur Gewährleistung der Systemverfügbarkeit und Integrität.

14. Leitprinzipien der Datenarchitektur und -verarbeitung {#14.-leitprinzipien-der-datenarchitektur-und--verarbeitung}

Die folgenden Leitprinzipien definieren den Rahmen für die Architektur, die semantische Modellierung, die Datenverarbeitung und die Sicherheitsstrategie unserer digitalen Versorgungsplattform. Sie gewährleisten Interoperabilität, klinische Relevanz, Nachvollziehbarkeit und Compliance.

14.1 FHIR-zentrierte Semantik (Fast Healthcare Interoperability Resources) {#14.1-fhir-zentrierte-semantik-(fast-healthcare-interoperability-resources)}

- **Primäre Abbildungsgrundlage**: Strukturierte klinische und administrative Daten werden vorrangig und konsistent entlang des offiziellen FHIR-Ressourcenmodells (aktuell R4 oder R5) abgebildet. FHIR dient als zentraler Standard für den Datenaustausch und die interne Datenrepräsentation, um die

Interoperabilität mit nationalen und internationalen Gesundheitssystemen zu maximieren.

- **Nutzung von FHIR-Profilen**: Wo nötig, werden spezifische FHIR-Profile (z.B. aus der Medizininformatik-Initiative, MII) angewendet, um die nationale Spezifität und die Einhaltung lokaler Anforderungen sicherzustellen.

14.2 SNOMED CT als klinische Referenzterminologie {#14.2-snomed-ct-als-klinische-referenzterminologie}

- **Zentrale Klassifikation**: SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) wird als obligatorische, hierarchische Referenzterminologie für die detaillierte Kodierung von Diagnosen, Prozeduren, Symptomen und klinischen Beobachtungen eingesetzt.

- **Ergänzende Terminologien**: Zur Abdeckung spezifischer Anwendungsfälle und nationaler/internationaler Reporting-Anforderungen wird SNOMED CT ergänzt durch:

- **ICD-10/11**: Für die Abrechnung und offizielle statistische Berichterstattung.

- **HPO (Human Phenotype Ontology)**: Zur tiefen Phänotypisierung, insbesondere in der Genetik und bei seltenen Erkrankungen.

- **LOINC (Logical Observation Identifiers Names and Codes)**: Für die standardisierte Kodierung von Laborergebnissen und klinischen Messungen.

14.2.1 SNOMED CT Terminologie-Server (Snowstorm / FHIR-API)

Für die technische Anbindung an SNOMED CT wird ein **FHIR-konformer Terminologie-Server** eingesetzt. Die Referenzimplementierung ist **Snowstorm** (SNOMED International, Open Source), die sowohl als öffentlicher Service als auch On-Premise betrieben werden kann.

FHIR-Operationen für UKLGPT:

Operation	Endpunkt	Einsatz
`$lookup`	`CodeSystem/$lookup?system=http://snomed.info/sct&code={SCTID}`	Abruf von Bezeichnungen, Definitionen und Eigenschaften eines Konzepts (z. B. für Anzeige im UI oder

Operation	Endpunkt	Einsatz
		Prompt-Anreicherung)
`$expand` (ECL)	`ValueSet/$expand?url=http://snomed.info/sct?fhir_vs=ecl/{ECL}&filter={text}`	Hierarchische Suche mit Expression Constraint Language (z. B. `< 84114007` = alle Untertypen von Herzinsuffizienz)
`$translate`	`ConceptMap/$translate`	Mapping zwischen SNOMED CT und ICD-10/LOINC
`$subsumes`	`CodeSystem/$subsumes`	Prüfung hierarchischer Beziehungen (z. B. „Ist Diagnose X ein Untertyp von Y?")

Betriebsmodell:

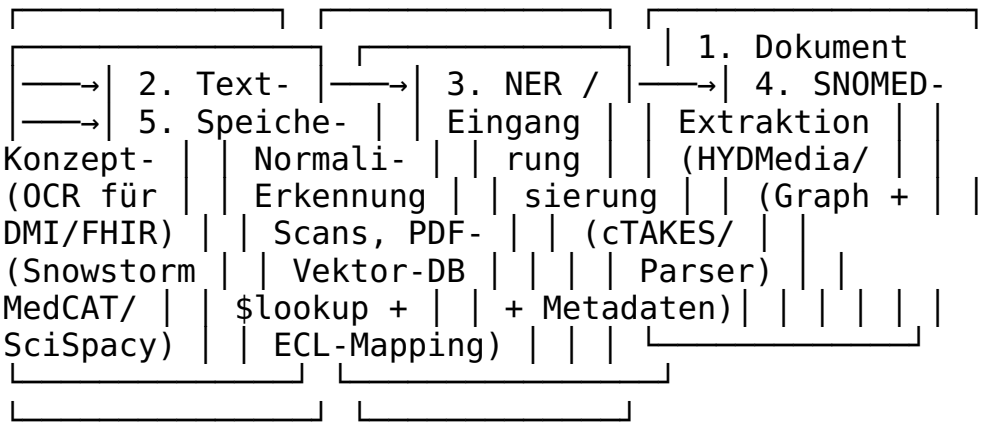
- **Phase 1 (Pilotierung):** Nutzung des öffentlichen IHTSDO-Servers (`https://snowstorm.ihtsdotools.org/fhir`) mit Fallback auf lokalen Cache.
- **Phase 2 (Produktion):** On-Premise Snowstorm-Instanz im UKL-Rechenzentrum (Elasticsearch-Backend, Docker-basiert), um Latenzanforderungen (< 50 ms pro Lookup) und Datenschutzanforderungen zu erfüllen.
- **Referenzimplementierung:** IHTSDO/SNOMED-in-5-minutes (GitHub) – stellt Beispielcode in Python, Java, JavaScript u. a. bereit.

14.2.2 SNOMED-Autotagging-Pipeline für Patientendokumente

Alle Patientendokumente, die in das UKLGPT-System gelangen (aus HYDMedia, DMI oder direkt aus Quellsystemen), werden automatisch mit **SNOMED-CT-Codes** annotiert. Dieses Autotagging ist ein zentrales Element, das drei Ziele verfolgt:

1. **Semantische Durchsuchbarkeit:** Dokumente werden nicht nur per Volltextsuche, sondern über standardisierte medizinische Konzepte auffindbar.
2. **Evidence-Matching:** Die SNOMED-Tags eines Dokuments ermöglichen das automatische Verknüpfen mit externer Evidenz (Europe PMC, vgl. Kap. 7.2.1.1).
3. **GraphRAG-Anbindung:** Die extrahierten SNOMED-Konzepte werden als Kanten im Wissensgraphen mit dem jeweiligen `DocumentReference`-Knoten verknüpft.

Pipeline-Architektur (5 Stufen):



Stufe 1 – Dokumenteingang: * Trigger: Neues Dokument in HYDMedia (via FHIR DocumentReference-Event) oder DMI-Klassifizierungsinstanz. * Metadaten (Patient, Fall, Dokumenttyp, Zeitstempel) werden aus der FHIR-Nachricht extrahiert.

Stufe 2 – Textextraktion: * **PDF-Dokumente:** OCR-Verarbeitung (HYDMedia OCR-Modul, ab 02/2026 verfügbar) oder Apache Tika. * **Strukturierte Nachrichten (HL7/FHIR):** Direkter Textauszug aus den relevanten Feldern. * **Scans:** OCR mit medizinisch trainiertem Modell (höhere Genauigkeit bei Fachbegriffen).

Stufe 3 – Named Entity Recognition (NER) / Konzepterkennung: * Einsatz eines medizinischen NER-Modells (Optionen: **MedCAT**, **cTAKES**, **SciSpacy** mit `en_ner_bc5cdr_md`-Modell). * Erkannte Entitätstypen:

Entitätstyp	Beispiel	SNOMED-Hierarchie
Diagnosen / Befunde	„Herzinsuffizienz", „Barrett-Ösophagus"	`< 404684003` (Clinical Finding)
Prozeduren	„Gastroskopie", „CT Abdomen"	`< 71388002` (Procedure)
Substanzen / Medikamente	„Metoprolol", „Ramipril"	`< 105590001` (Substance)
Körperstrukturen	„linker Ventrikel", „Leber"	`< 123037004` (Body Structure)
Organismen	„Helicobacter pylori"	`< 410607006` (Organism)

Stufe 4 – SNOMED-Normalisierung: * Jede erkannte Entität wird über die Snowstorm FHIR-API normalisiert: * `$expand` mit ECL-Filter zur Auflösung in den korrekten SNOMED-CT-Code. * `$lookup` zum Abruf der kanonischen Bezeichnung und des Definitionsstatus. * Konfidenz-Score der NER-Erkennung wird mitgeführt (Schwellwert: ≥ 0.85 für automatische Übernahme, < 0.85 → Review-Queue). * **ICD-10-Crosswalk:** Über `$translate` (ConceptMap) werden

ergänzend ICD-10-Codes abgeleitet, sofern ein valides Mapping existiert.

Stufe 5 – Speicherung: * Die extrahierten SNOMED-Tags werden als `Coding`-Elemente in `DocumentReference.category` (CodeableConcept) oder einer dedizierten FHIR-Extension gespeichert. * Im **GraphRAG** wird für jedes getaggte Konzept eine Kante `DocumentReference → hasSnomedTag → Concept` angelegt. * Im **Vektor-DB**-Index werden die SNOMED-Codes als zusätzliche Filterattribute (Facetten) gespeichert, um hybride Suche (Semantik + Facette) zu ermöglichen.

Durchsatz und Skalierung:

Szenario	Geschätzter Durchsatz	Anmerkung
Neudokumente (laufender Betrieb)	~500–1.000 Dokumente/Tag	Echtzeit-Verarbeitung (< 5 Sek./Dokument)
Altdaten-Migration (21 Mio. PDFs)	Batch-Verarbeitung über 6–12 Monate	Priorisierung nach Aktualität (jüngste Fälle zuerst)

Qualitätssicherung: * **Stichproben-Audit:** 1 % der automatisch getaggten Dokumente werden monatlich durch klinisches Fachpersonal validiert. * **Feedback-Loop:** Korrekturen fließen als Trainingsdaten zurück in das NER-Modell (kontinuierliche Verbesserung).

Referenzimplementierungen (siehe `Codebeispiele/`): * `snomed-fhir-api/snomed_fhir_examples.py` – $lookup, $expand (ECL), $translate, NER-Normalisierung * `snomed-fhir-api/snomed_to_europepmc_bridge.py` – Integrierter Workflow: SNOMED → Europe PMC * `europe-pmc-api/europe_pmc_examples.py` – Evidence-Matching, Leitliniensuche, Text-Mining * `snomed-snowstorm-api/` – Original IHTSDO-Beispiele in Python, JavaScript, Go, Ruby, PHP, curl

14.3 Trennung der Wissensdomänen und Datenhaltungsschichten {#14.3-trennung-der-wissensdomänen-und-datenhaltungsschichten}

- **Modulare Architektur**: Die Datenhaltung wird in logisch getrennte Domänen unterteilt, um Skalierbarkeit, gezielte Aktualisierung und spezifische Sicherheitsanforderungen zu gewährleisten:

- **Globales Wissen (Globaler Wissens-RAG)**: Umfasst externe, allgemein gültige Informationsquellen wie klinische Leitlinien, wissenschaftliche Literatur (primär über **Europe PMC** REST-

API, vgl. Kap. 7.2.1.1) und Ontologien. Diese Domäne dient als Grundlage für Entscheidungsunterstützungssysteme (Clinical Decision Support, CDS). Die Verknüpfung mit Patientendaten erfolgt über **SNOMED-CT-basiertes Evidence-Matching** (Kap. 7.2.1.2).

- **Patientenbezogene Dokumente (HYDMedia)**: Unstrukturierte oder semi-strukturierte Dokumente wie Arztbriefe, Befunde, Aufklärungsbögen und Bilder, die im Rahmen der Versorgung entstehen. Diese werden in einem hochsicheren Dokumentenarchiv verwaltet.

- **Strukturierte Versorgungsdaten (GraphRAG)**: Die im FHIR-Format abgebildeten, strukturierten Patientendaten (z.B. Medikation, Vitalparameter, Diagnosen) werden in einer graph-basierten oder Retrieval-Augmented Generation (RAG)-fähigen Datenbank gespeichert, um komplexe Abfragen und KI-gestützte Analysen zu ermöglichen.

14.4 Episodische Datenverarbeitung (Privacy by Design) {#14.4-episodische-datenverarbeitung-(privacy-by-design)}

- **Zweckbindung und Datenminimierung**: Patientenbezogene Daten werden primär nur für die Dauer der *aktiven Versorgungsepisode* verarbeitet. Dieses Prinzip stellt sicher, dass Daten nicht unnötig lange im aktiven Verarbeitungsspeicher verbleiben.

- **Time-to-Live (TTL)**: Zur strikten Einhaltung der Speicherbegrenzung wird ein TTL-Mechanismus implementiert. Außerhalb einer aktiven Episode sind die Daten nur über einen manuell initiierten, expliziten *Abruf* (z.B. durch eine autorisierte Fachkraft für eine nachfolgende Episode) zugänglich, was die Verarbeitung auf das absolut Notwendige beschränkt.

14.5 Least Privilege & Need-to-Know (Zugriffskontrolle) {#14.5-least-privilege-&-need-to-know-(zugriffskontrolle)}

- **Basis**: Die Zugriffskontrolle basiert auf den Prinzipien des **Least Privilege** (geringste notwendige Berechtigung) und **Need-to-Know** (Zugriff nur auf die Daten, die zur Erfüllung der aktuellen Aufgabe notwendig sind).

- **Implementierung**: Dies wird über **AD-basierte (Active Directory) RBAC (Role-Based Access Control)** und dynamische **ABAC (Attribute-Based Access Control)**-Policies durchgesetzt. ABAC ermöglicht eine feingranulare, kontextabhängige Steuerung des Zugriffs (z.B. "Darf nur auf

Daten von Patienten zugreifen, die aktuell auf dieser Station behandelt werden").

14.6 Nachvollziehbarkeit und klinische Sicherheit (Governance) {#14.6-nachvollziehbarkeit-und-klinische-sicherheit-(governance)}

- **Transparenz (Quellen)**: Jede Information, die in der Plattform verarbeitet wird und für klinische Entscheidungen relevant ist, muss eine klare, verifizierbare Quelle aufweisen (z.B. FHIR Provenance Ressourcen).
- **Audit-Trail**: Ein umfassendes, manipulationssicheres **Audit-Protokoll** (Audit-Log) zeichnet jeden Zugriff, jede Verarbeitung und jede Änderung an patientenbezogenen Daten auf, um die forensische Nachvollziehbarkeit zu gewährleisten und datenschutzrechtlichen Anforderungen zu genügen.
- **Guardrails**: Es werden technische und prozessuale **Guardrails** implementiert (z.B. Plausibilitätsprüfungen, Warnmeldungen bei potenziell gefährlichen Medikationskombinationen), um die klinische Sicherheit zu erhöhen und menschliche Fehler zu minimieren.

15. Gesamtarchitektur – Überblick {#15.-gesamtarchitektur-–-überblick}

Die Zielarchitektur ist konsequent nach dem Leitprinzip **"vom klinischen Nutzer zur gesicherten Datenquelle"** ausgerichtet. Dieses Prinzip stellt sicher, dass die Anwendung stets die Kontrolle über die dargebotenen Informationen behält und eine hohe Qualität sowie eine nachvollziehbare Herkunft der Antworten gewährleistet wird.

Zentrale Idee und ein fundamentaler architektonischer Pfeiler ist die Festlegung, dass **nicht das Large Language Model (LLM)** die alleinige Kontrolle über Inhalt, Kontext und Antwortlogik ausübt. Stattdessen übernehmen die **Applikations- und Qualitätssicherungsschicht** die Steuerung und Orchestrierung. Das LLM dient hierbei primär als leistungsstarker Generator und Interpretationswerkzeug, dessen Output durch die nachgeschalteten Schichten validiert und kontextualisiert wird.

Die Gesamtarchitektur ist in klar definierte, voneinander getrennte Ebenen (Layered Architecture) gegliedert, um Modularität, Skalierbarkeit, einfache Wartbarkeit und eine strikte Trennung der Verantwortlichkeiten (Separation of Concerns) zu gewährleisten:

15.1 Frontend / UI (Präsentationsschicht): {#15.1-frontend-/-ui-(präsentationsschicht):}

- Verantwortlich für die Interaktion mit dem klinischen Nutzer (z.B. Eingabe von Prompts, Anzeige der Antworten).
- Stellt eine intuitive und an klinische Workflows angepasste Benutzeroberfläche (User Interface, UI) bereit.
- Hier erfolgt die Erfassung und initialen Validierung der Nutzereingaben.

15.2 Applikations- und Orchestrierungsschicht (Business/Service-Schicht): {#15.2-applikations--und-orchestrierungsschicht-(business/service-schicht):}

- Das Herzstück der Logik und Steuerung.
- Verantwortlich für die Verarbeitung der Nutzeranfragen und die Koordination des gesamten Prozesses.
- Definiert die Abfolge der Schritte (z.B. Kontextanreicherung, Prompt-Generierung, Aufruf des LLMs, Post-Processing).
- Implementiert die Geschäftslogik und stellt sicher, dass alle Schritte gemäß den klinischen Anforderungen ausgeführt werden.

15.3 Qualitätssicherung über Prompt-Pipeline (RAG-Orchestrierung): {#15.3-qualitätssicherung-über-prompt-pipeline-(rag-orchestrierung):}

- Eine kritische Zwischenschicht zur Gewährleistung von Halluzinationsfreiheit und Faktenbasierung (Groundedness).
- Hier werden die Nutzereingaben mit relevanten Kontextdaten (Retrieval-Augmented Generation, RAG) angereichert.
- Die Pipeline steuert die Auswahl der Abrufstrategien (z.B. welcher RAG-Speicher wird konsultiert?).
- Stellt die finale, strukturierte Prompt-Anweisung für das LLM zusammen.

- Führt ggf. eine Validierung des LLM-Outputs gegen die abgerufenen Quelldaten durch.

15.4 Datenschicht (Retrieval Augmented Generation - RAG): {#15.4-datenschicht-(retrieval-augmented-generation---rag):}

- Enthält alle gesicherten, validierten und strukturierten Wissensquellen.
- Gewährleistet den Zugriff auf unterschiedliche Arten von Wissensbasen, die jeweils spezifische Anwendungsfälle abdecken:
- **GraphRAG:** Für komplexe Zusammenhänge, Kausalitäten und Relationen (z.B. klinische Pfade, Krankheitsmodelle).
- **Dokumenten-RAG:** Für unstrukturierte und semi-strukturierte Dokumente (z.B. Leitlinien, SOPs, Studienberichte).
- **Wissens-RAG:** Für strukturiertes, kuratiertes Wissen (z.B. Terminologien, Medikamenteninformationen).

15.5 Schnittstellen (Integrationsschicht): {#15.5-schnittstellen-(integrationsschicht):}

- Stellt die Anbindung an externe und interne Drittsysteme sicher (z.B. KIS, ePA, Laborinformationssysteme).
- Definiert standardisierte APIs für den sicheren und interoperablen Datenaustausch.
- Diese Schicht ist essenziell für die Echtzeit-Kontextualisierung mit patientenspezifischen Daten.
- **Externe Evidenz-Schnittstellen:**
- **Europe PMC REST-API** (https://europepmc.org/RestfulWebService): Abruf wissenschaftlicher Literatur über strukturierte Suchanfragen (DISEASE, CHEM, GENE_PROTEIN). Kein personenbezogener Datenfluss nach extern. (Kap. 7.2.1.1)
- **Snowstorm FHIR-API** (SNOMED CT Terminologie-Server): FHIR R4-konforme Operationen ($lookup, $expand, $translate, $subsumes) für SNOMED-CT-Normalisierung und hierarchische Abfragen. Phase 1 extern, Phase 2 On-Premise. (Kap. 14.2.1)

15.6 Berechtigungs- und Sicherheitskonzept (Querschnittsfunktion): {#15.6-berechtigungs--und-sicherheitskonzept-(querschnittsfunktion):}

- Eine übergreifende, nicht-funktionale Anforderung, die alle Schichten durchdringt.
- Umfasst **Authentifizierung** (Wer ist der Nutzer?) und **Autorisierung** (Was darf der Nutzer sehen/tun?).
- Gewährleistet die Einhaltung aller relevanten Datenschutzbestimmungen (z.B. DSGVO, klinische Schweigepflicht).
- Implementiert Logging und Auditing zur Nachvollziehbarkeit jeder Interaktion.

16. Variantenvergleich: UKLGPT (Eigenlösung) vs. Averbis/ Meierhofer (Marktlösung) {#16.-variantenvergleich}

16.1 Hintergrund und Einordnung (PSP 3.2 Marktanalyse, PSP 3.9 KIS-Dokumentenzugriff)

Im Rahmen des Vorprojekts ist eine neutrale Bewertung der am Markt verfügbaren Alternativen durchzuführen. Der MKIS-Hersteller **Meierhofer AG** bietet in Kooperation mit **Averbis** (einem Spezialisten für NLP und medizinische Textanalyse) eine integrierte KI-Lösung an, die direkt im M-KIS verfügbar sein wird.

Quelle: E-Mail von Samira Grass (Meierhofer AG) vom 09.02.2026 an UKL-Projektteam, Betreff: „Beantwortung Rückfragen Averbis Präsentation" (siehe Fachablage/Averbis.txt).

16.1.1 Beschreibung der Averbis/Meierhofer-Lösung

Die Averbis-Lösung umfasst zwei Kernmodule:

1. **Medical Summary:** Automatisierte Zusammenfassung von Patienteninformationen aus verschiedenen Quellen, inkl. Anbindung an HYDMedia-Archiv.

2. **Gesprächsdokumentation:** KI-gestützte Dokumentation ärztlicher Gespräche mit automatischer Zusammenfassung.

Bestätigte Fähigkeiten (Stand 09.02.2026): - Anbindung an HYDMedia über den Kommunikationsserver der Klinik bestätigt - Averbis Health Discovery Technologie kann Informationen aus dem Archiv für Medical Summary nutzen - Fallbezug (nicht nur patientenübergreifend) ab März 2026 geplant - Kennzeichnung KI- vs. Arzt-Eingaben in M-KIS Version 15.1.0 - Feedback-Mechanismus (Daumen hoch/runter) in Medical Summary vorhanden - Medikamentenübernahme derzeit nicht automatisch (MDR-Zertifizierung erforderlich) - Angebot in Erstellung (Dienstleistungsaufwände von Averbis ausstehend)

Technische Architektur (Besprechung 27.02.2026)

Am 27.02.2026 fand ein technischer Klärungstermin mit Averbis, Meierhofer, 4K Analytics und dem UKL-Projektteam statt. Averbis präsentierte eine Architekturskizze (siehe Fachablage/ Averbis_Integration Architekturbild.jpg) mit folgender Komponentenstruktur:

Komponenten im Krankenhaus-Netzwerk (On-Premise):

Komponente	Funktion	Datenfluss
Medical Summary User Interface	Webanwendung für Endanwender (Ärzte)	Zugang über Reverse Proxy (HTTPS)
Medical Summary Backend	Zentrale Datenhaltung – alle Daten werden lokal persistiert	Empfängt Patientendaten (PID, Name, Geb.Datum), Falldaten (Diagnosen, Laborwerte, Medikation) von M-KIS via HL7
Health Discovery (NLP)	On-Premise-NLP-Engine von Averbis zur Informationsextraktion	Kommuniziert via HTTPS mit Azure OpenAI
Arztbriefschreibung	Generiert Epikrisen auf Basis extrahierter Daten	Empfängt Epikrise-Daten vom Backend
Vektordatenbank	Lokale Speicherung der Embeddings	Embeddings werden in der Cloud berechnet, aber

Komponente	Funktion	Datenfluss
		lokal gespeichert

Cloud-Komponente: - **Azure OpenAI** (EU Datazone, Schweden): LLM wird ausschließlich temporär zur Verarbeitung genutzt. Keine persistente Datenhaltung in der Cloud. Alternative Hosting-Optionen: Stackit (DE), Arvato, Telekom – Wahl abhängig von Datenschutzvorgaben des jeweiligen Bundeslandes.

Anbindung Archiv/CMS (HYDMedia): - **PULL:** Historische Daten werden bei Fallanlage (ADT-Ereignis) über REST-API oder FHIR-Schnittstelle aus HYDMedia abgerufen. - **PUSH:** Neue Dokumente, die nicht über M-KIS laufen, können via FHIR/HL7-Nachrichten direkt an das Medical Summary Backend gesendet werden. - **Schnittstellenwahl:** UKL bevorzugt REST-API (FHIR biete in diesem Kontext keinen Mehrwert); Averbis kann beide Varianten integrieren. Entscheidung steht aus.

ADT-basierte Archivstrategie (bestätigt): - Kein initialer Massenimport aller 20+ Mio. Dokumente. Stattdessen ereignisgesteuerter Abruf: Beim Anlegen eines neuen Patientenfalls in M-KIS (ADT-Meldung) werden die historischen Dokumente dieses Patienten aus HYDMedia geladen und in die Medical Summary integriert. - Fokus auf aktuelle/jüngere Akten; Tiefenrecherche im Archiv nur für Spezialfälle. - Optionale Vor-Klassifikation via Erkennungs-/Validierungsprompts (z. B. Arztbrief, OP-Bericht) mit Nutzerkorrekturmöglichkeit.

Berechtigungskonzept: - SSO über Identity Provider mit OIDC-Protokoll. M-KIS-Berechtigungen werden durchgeschleust. Identity Provider ist Teil der NEXT-Betriebsumgebung und zum 1.10. kommunikationsbereit. Keine Zusatzkosten für Identity-Integration.

Quellen: Besprechungsprotokoll vom 27.02.2026 (Fachablage), Architekturskizze (Fachablage/Averbis_Integration Architekturbild.jpg), Mail Dr. Cundius (Fachablage/ Mail_'Carina_Averbis.pdf).

Piloterfahrungen und Referenzen (Stand 27.02.2026)

Referenzkunde	Status	Anmerkung
Asklepios Kliniken	Testphase (bis 3 Monate)	Nach Testphase Lizenzierung und Live-Betrieb
UMG Greifswald	Verzögert	IT-Leiterwechsel; Testphase geplant
Klinikum Rheine	Produktiv (Azure)	Nicht direkt vergleichbar für Sachsen (Cloud-Restriktionen)

Kostenindikatoren (Stand 27.02.2026)

Position	Betrag	Quelle
Teststellung (aktueller Funktionsumfang, Dienstleistung)	**~90.000 €**	Mail Dr. Cundius
Averbis-Lizenz + Meierhofer-Lizenz (p.a.)	**~200.000 €/Jahr**	Mail Dr. Cundius (Meierhofer-Lizenz noch nicht beziffert)
Archiv-/HYDMedia-Anbindung (Entwicklung)	**10–20 Personentage** zusätzlich	Besprechung 27.02.2026 (Averbis-Schätzung)
Identity-Integration (OIDC/SSO)	Im Standardumfang enthalten	Keine Zusatzkosten

Hinweis: Die exakte Aufwandsabschätzung für die HYDMedia-Anbindung ist erst nach Detailprüfung der Datenlage (Dokumententypen, Qualität, KDL-Labelstatus) möglich.

Offene Risiken und Qualitätsprobleme (Besprechung 27.02.2026)

Risiko	Beschreibung	Auswirkung
OCR-Rückwirkung unklar	OCR in HYDMedia seit Feb. 2026 implementiert; ob rückwirkend für alle Bestandsdokumente angewendet wird, ist ungeklärt	Nicht maschinenlesbare PDFs können nicht von der NLP-Engine verarbeitet werden
KDL-Klassifikation lückenhaft	Systematische KDL-Klassifizierung erst seit ~2 Jahren; rückwirkende Klassifizierung vermutlich nicht erfolgt	Filterlogik in Medical Summary kann Dokumententypen (Arztbrief, Laborbefund etc.) nicht zuverlässig unterscheiden
Heterogene Dokumentenqualität	Mehrfach-Scans, Handyfotos, inkonsistente Zeitstempel im Archiv	Erhöhter Dienstleistungsaufwand über die geschätzten 10–20 PT hinaus

Risiko	Beschreibung	Auswirkung
Cloud-Freigabe Sachsen	Landesweite Cloud-Restriktionen in Sachsen; Termin mit Landesdatenschützer angesetzt (März 2026)	Ggf. Stackit (DE) statt Azure erforderlich → höhere Kosten; ggf. Datenschutzgutachten (~100.000 €)
Zeitplan-Konflikt	Projektteam priorisiert KIS-Einführung; parallele Systemeinführung zum 1.10. kritisch	Eskalation an Lenkungsausschuss/ Vorstand erforderlich (Dr. Vasipki)

16.2 Strukturierter Variantenvergleich

Kriterium	UKLGPT (Eigenlösung)	Averbis/Meierhofer
Datenhoheit & Kontrolle	**Vollständig beim UKL.** Alle Daten, Modelle und Pipelines unter eigener Kontrolle. On-Premise.	Teilweise beim UKL. Medical Summary Backend + Vektordatenbank on-prem; LLM-Verarbeitung temporär in Azure Cloud (EU, Schweden). Keine persistente Cloud-Datenhaltung, aber Verarbeitungshoheit liegt bei Azure/Averbis. (Stand: Besprechung 27.02.2026)
Archivdaten-Zugriff (21 Mio. PDFs)	**Vollzugriff** über GraphRAG + Dokumenten-RAG. Semantische Suche, kontextualisierte Antworten über gesamten Datenbestand.	ADT-basierter Abruf aus HYDMedia bestätigt (REST-API oder FHIR). Keine Massenindexierung – Dokumente werden erst bei Fallanlage patientenbezogen geladen. Archivzugriff technisch machbar (10–20 PT Aufwand), aber abhängig von OCR-/KDL-Qualität der Altdaten. (Stand: Besprechung 27.02.2026)
Archivierungs-Strategie	Hybride Strategie: PDF/A-Archiv (Compliance) + FHIR-GraphRAG (KI-Readiness). Maximale Nutzung der Altdaten.	Kein eigenes Archivierungskonzept. Setzt auf vorhandene HYDMedia-Infrastruktur.

Kriterium	UKLGPT (Eigenlösung)	Averbis/Meierhofer
		Keine FHIR-Graph-Anreicherung.
Funktionsumfang	Umfassend: Intelligente Recherche, Patientenakten-Zusammenfassung, Leitlinienabgleich, Kodierunterstützung, interdisziplinäre Workflows. Erweiterbar.	Fokussiert: Medical Summary + Gesprächsdokumentation. Weitere Module anbieterspezifisch und kostenpflichtig.
Berechtigungskonzept	**Detailliertes 3-Ebenen-Modell** (Kontext, Behandlungsauftrag, Dokumentenschutz) mit Break-the-Glass. SAP/ M-KIS-geführt.	M-KIS-Berechtigungen werden via SSO/OIDC (Identity Provider, NEXT-Umgebung) durchgeschleust – bestätigt (27.02.2026). Feinsteuerung PSY/KJP-Schutz und dokumentenspezifischer Schutz weiterhin unklar.
Unabhängigkeit / Vendor Lock-in	**Technologisch unabhängig.** Offene Standards (FHIR, SNOMED). LLM austauschbar. Keine Bindung an einzelnen Anbieter.	**Hoher Vendor Lock-in.** Abhängigkeit von Meierhofer + Averbis. Funktionserweiterungen nur über Anbieter. Preisgestaltung nicht kontrollierbar.
Innovationstiefe	**Hoch.** GraphRAG mit FHIR/SNOMED, mehrstufige Prompt-Pipeline, episodisches RAG mit Zweckbindung, deterministische Qualitätssicherung.	Mittel. NLP-basierte Textanalyse (Averbis Health Discovery). Kein GraphRAG, kein FHIR-basiertes Wissensnetz.
Qualitätssicherung KI	**Mehrstufige Prompt-Pipeline** mit Domain-Classification, SNOMED-Normalisierung, Validierung, Quellentrennung (Fakten/Dokumente/ Leitlinien).	Feedback-Mechanismus (Daumen hoch/runter). Interne Validierungslogik von Averbis (intransparent für UKL).
Time-to-Market	Länger: Eigenentwicklung erfordert Aufbau von	Kürzer: M-KIS 15.1.0 mit Averbis-Funktionen könnte parallel zur KIS-

Kriterium	UKLGPT (Eigenlösung)	Averbis/Meierhofer
	Team, Infrastruktur, Pilotierung. MVP realistisch ab Q4/2026 (vgl. Kap. 17.2).	Einführung verfügbar sein. Testphase bis 3 Monate. Aber: Zeitplan-Konflikt – Projektteam priorisiert KIS-Einführung (Besprechung 27.02.2026).
Kosten (Einschätzung)	Höhere Initialkosten (Infrastruktur, Personal, LLM-Lizenzen). Langfristig niedrigere laufende Kosten durch Eigenhoheit.	**Teststellung ~90.000 €**, laufend **~200.000 €/Jahr** (Averbis- + Meierhofer-Lizenz, Meierhofer-Anteil noch nicht beziffert) + 10–20 PT für HYDMedia-Anbindung. Langfristig steigende Lizenzkosten möglich. (Quelle: Mail Dr. Cundius, 27.02.2026)
Compliance (EU AI Act, DSGVO)	Volle Kontrolle über Compliance-Maßnahmen. Transparenz der KI-Pipeline gewährleistet Hochrisiko-KI-Anforderungen.	Compliance-Verantwortung teilweise beim Anbieter. Transparenz der Averbis-Algorithmen für UKL nicht vollständig gegeben.
Strategischer Wert	**Hoch.** Aufbau interner KI-Kompetenz. Universitätsklinikum als Innovationsführer. Nachnutzungspotenzial, Forschungsanbindung.	Mittel. Standardprodukt. Kein Kompetenzaufbau am UKL. Kein Alleinstellungsmerkmal.

16.3 Bewertungsmatrix (gewichtet)

Kriterium	Gewicht	UKLGPT	Averbis/ Meierhofer
Datenhoheit & Kontrolle	20%	5 (1,0)	2 (0,4)
Funktionsumfang	15%	5 (0,75)	3 (0,45)
Berechtigungskonzept	15%	5 (0,75)	3 (0,45)
Unabhängigkeit	15%	5 (0,75)	1 (0,15)
Innovationstiefe	10%	5 (0,5)	3 (0,3)
Time-to-Market	10%	2 (0,2)	4 (0,4)

Kriterium	Gewicht	UKLGPT	Averbis/ Meierhofer
Kosten (initial)	10%	2 (0,2)	3 (0,3)
Strategischer Wert	5%	5 (0,25)	2 (0,1)
Gesamtscore	**100%**	**4,4**	**2,55**

(Skala: 1 = schlecht, 5 = sehr gut. Gewichteter Score in Klammern.)

16.4 Empfehlung

Die **Eigenlösung UKLGPT** wird empfohlen. Die Argumente:

1. **Maximale Datenhoheit** und Unabhängigkeit von externen Anbietern – strategisch essenziell für ein Universitätsklinikum.
2. **Vollständige Nutzung der Altdaten** (21 Mio. Dokumente) durch hybride Architektur (GraphRAG + Dokumenten-RAG), die Averbis nicht bietet.
3. **Tiefgreifende Qualitätssicherung** durch mehrstufige Prompt-Pipeline, die für den klinischen Einsatz mit Patientendaten unverzichtbar ist.
4. **Strategischer Kompetenzaufbau** in KI und Datenarchitektur am UKL – Basis für Forschungskooperationen und weitere Digitalisierungsvorhaben.
5. **Kein Vendor Lock-in** – alle Komponenten sind auf offenen Standards aufgebaut und austauschbar.

Die **Averbis/Meierhofer-Lösung** hat Vorteile bei Time-to-Market und könnte als **Übergangslösung** für die Medical Summary (Gesprächsdokumentation) während der UKLGPT-Entwicklung dienen. Dies sollte geprüft werden, sofern die Kosten vertretbar sind und kein langfristiger Lock-in entsteht.

16.4.1 Risiken der Averbis-Variante

Risiko	Bewertung	Anmerkung (v2.1)
Cloud-Datenverarbeitung (DSGVO-kritisch bei Patientendaten)	**Hoch**	Bestätigt: LLM-Verarbeitung via Azure OpenAI (Schweden). Cloud-Freigabe in Sachsen unklar – Termin Landesdatenschützer März 2026. Ggf. Stackit (DE) erforderlich (höhere Kosten).
Fehlende Transparenz der KI-Pipeline (EU AI Act Hochrisiko)	Hoch	Unverändert

Risiko	Bewertung	Anmerkung (v2.1)
Kein Zugriff auf Altdaten-Gesamtbestand über GraphRAG	Hoch	Präzisiert: ADT-basierter Einzelabruf bestätigt – kein semantischer Gesamtzugriff, kein GraphRAG.
Langfristig steigende Lizenzkosten ohne Exit-Strategie	**Hoch**	Neu bewertet: ~200.000 €/Jahr Lizenz + 90.000 € Teststellung. TCO 5 Jahre: >1 Mio. € (ohne Infrastruktur).
Abhängigkeit von Produktroadmap eines Drittanbieters	Hoch	Bestätigt: Exklusivvertrieb Meierhofer–Averbis.
OCR-/KDL-Qualitätsrisiko im Archiv	**Hoch**	Neu: Rückwirkende OCR/KDL unklar. Kann Dienstleistungsaufwand erheblich über 10–20 PT treiben.
Datenschutzgutachten Sachsen	**Mittel**	Neu: Ggf. externes Gutachten (KPMG o. ä.) erforderlich (~100.000 €).

16.4.2 Offene Punkte für Finalisierung der Bewertung

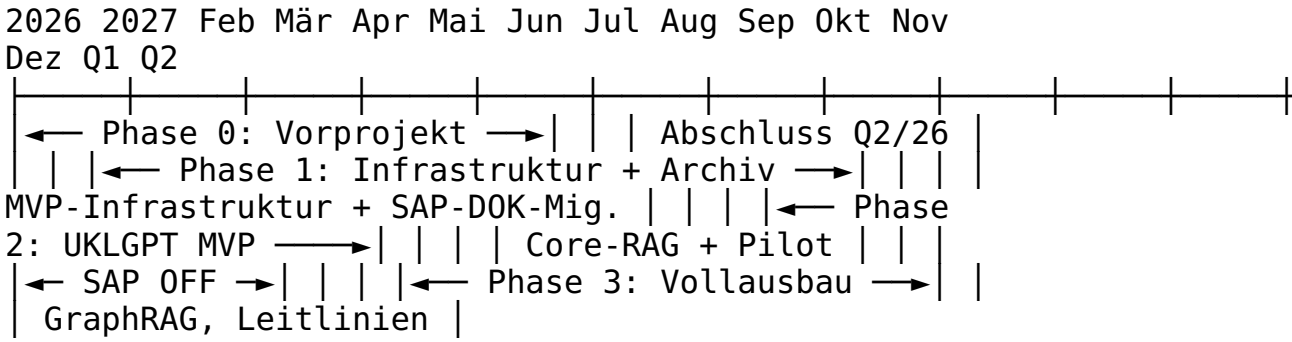
1. ~~Angebot von Meierhofer/Averbis liegt noch nicht vor – Kosten müssen verglichen werden.~~ **TEILWEISE GEKLÄRT (27.02.2026):** Kostenindikatoren vorhanden (Teststellung ~90.000 €, Lizenz ~200.000 €/a), Meierhofer-Lizenzanteil noch nicht beziffert. TCO-Vergleich ausstehend.
2. ~~Technischer Klärungstermin mit Meierhofer ist anzusetzen.~~ **ERLEDIGT (27.02.2026):** Besprechung mit Averbis, Meierhofer, 4K Analytics durchgeführt. Architektur dokumentiert (Kap. 16.1.1).
3. Prüfung, ob Averbis als Übergangslösung (Medical Summary only) parallel zur EMRGPT-Entwicklung einsetzbar ist, ohne Lock-in zu erzeugen. **Weiterhin offen.**
4. **NEU:** Schnittstellenspezifikation HYDMedia (REST vs. FHIR) klären und mit Dedalus abstimmen.
5. **NEU:** Cloud-/Datenschutzklärung Sachsen – Ergebnis des Datenschützer-Termins (März 2026) abwarten. AVVs und Architekturskizze vorab bereitstellen.
6. **NEU:** Detailanalyse Dokumentenqualität im HYDMedia-Archiv: OCR-Status, KDL-Label-Qualität, repräsentative Stichprobe zur Risikobewertung.
7. **NEU:** Eskalation Zeitplan/Übersichtsfunktion bis 1.10. an Lenkungsausschuss (Abstimmung mit Dr. Vasipki).

---

17. Grober Umsetzungsfahrplan (PSP 6.2) {#17.-grober-umsetzungsfahrplan}

Harter Constraint: SAP i.s.h.med wird im Oktober 2026 abgeschaltet. Alle archivrelevanten Maßnahmen müssen bis dahin produktiv sein.

17.1 Phasenübersicht



17.2 Detaillierter Phasenplan

Phase 0: Vorprojekt-Abschluss (bis Ende Q2/2026)

Meilenstein	Zieldatum	Deliverable	Abhängigkeit	Go/No-Go
M0.1	Mär 2026	Vorprojekt-Ergebnisse konsolidiert, QA abgeschlossen	–	–
M0.2	Mär 2026	Variantenvergleich finalisiert (Averbis-Angebot liegt vor)	Meierhofer-Angebot	–
M0.3	Apr 2026	Kostenseite Business Case fertiggestellt	IT-Sizing, Averbis-Angebot	–
M0.4	Apr 2026	**LA-Entscheidung: Umsetzungsvariante + Ressourcenfreigabe**	M0.1–M0.3	**Go/No-Go**
M0.5	Mai 2026	Projektleiter installiert, Team aufgestellt	M0.4	–
M0.6	Jun 2026			–

Meilenstein	Zieldatum	Deliverable	Abhängigkeit	Go/ No-Go
		Vorprojekt formal abgenommen (PSP 6.3)	Alle PSP-Produkte	

Phase 1: Infrastruktur und Archiv-Migration (Mai – Sep 2026)

Meilenstein	Zieldatum	Deliverable	Abhängigkeit	Go/ No-Go
M1.1	Mai 2026	GPU-Infrastruktur bestellt / bereitgestellt	Budgetfreigabe (M0.4)	–
M1.2	Jun 2026	SAP-DOK-Migration nach HYDMedia abgeschlossen	DMI, IT (gestartet KW7/2026)	–
M1.3	Jul 2026	OCR-Pilottest abgeschlossen (1000 Dokumente, Qualitätsrate dokumentiert)	OCR-Installation (Feb 2026)	–
M1.4	Jul 2026	Dedalus FHIR-Schnittstelle validiert (Performance-Test, Scope bestätigt)	Dedalus-Zusage	**Go/ No-Go**
M1.5	Aug 2026	M-KIS-Berechtigungskonzept fertig, Meierhofer-Abstimmung abgeschlossen	M-KIS Go-Live-Planung	–
M1.6	Sep 2026	Basisinfrastruktur produktionsreif (Vektor-DB, Graph-DB, LLM-Instanz)	M1.1	–

Phase 2: UKLGPT MVP (Jul – Dez 2026)

Meilenstein	Zieldatum	Deliverable	Abhängigkeit	Go/ No-Go
M2.1	Aug 2026	Dokumenten-RAG Prototyp (Suche	M1.3, M1.4	–

Meilenstein	Zieldatum	Deliverable	Abhängigkeit	Go/ No-Go
		über OCR-PDFs, Pilotstichprobe)		
M2.2	Sep 2026	Berechtigungs-Integration M-KIS → UKLGPT validiert	M1.5	–
M2.3	Sep 2026	Audit-Trail und Logging implementiert	–	–
M2.4	**Okt 2026**	**MVP Go-Live: Dokumentensuche über HYDMedia (1 Pilotstation)**	M1.6, M2.1–M2.3, **SAP-Abschaltung**	**Go/ No-Go**
M2.5	Nov 2026	Pilotfeedback ausgewertet, Iterationsplanung	M2.4	–
M2.6	Dez 2026	MVP auf 3 weitere Stationen ausgerollt	M2.5	–

Phase 3: Vollausbau (Q1–Q2/2027)

Meilenstein	Zieldatum	Deliverable	Abhängigkeit	Go/ No-Go
M3.1	Jan 2027	GraphRAG-Integration (strukturierte Fakten aus UKLytics/DWH)	DWH-Anbindung	–
M3.2	Feb 2027	Leitlinien-RAG (Globales Wissen) integriert	–	–
M3.2a	Feb 2027	**Europe PMC-Anbindung** produktiv (Evidence-Matching, Kap. 7.2.1.1)	M3.2	–
M3.3	Mär 2027	Vollständige OCR-Verarbeitung des Dokumentenbestands abgeschlossen	Rechenkapazität	–
M3.3a	Mär 2027	**SNOMED-Autotagging-Pipeline** produktiv	M3.3, OCR	–

Meilenstein	Zieldatum	Deliverable	Abhängigkeit	Go/ No-Go
		(NER + Snowstorm, Kap. 14.2.2)		
M3.3b	Mär 2027	**Snowstorm On-Premise** bereitgestellt und validiert (Kap. 14.2.1)	Infrastruktur	–
M3.4	Apr 2027	Labor-FHIR-Connector (UKLytics) produktiv	DWH-Team	–
M3.5	Mai 2027	**Klinikweiter Rollout UKLGPT** (inkl. Evidence-Panel, SNOMED-Tags)	M3.1–M3.4	**Go/ No-Go**
M3.6	Jun 2027	Projektabschluss, Übergabe an Linienbetrieb	Betriebskonzept	–

17.3 Kritischer Pfad

```
SAP-DOK-Migration (KW7–Jun 2026) │ ▼ Dedalus FHIR validiert (Jul 2026) ──► Dokumenten-RAG Prototyp (Aug 2026) │ │ ▼ ▼ M-KIS-Berechtigungen (Aug 2026) ──────► MVP Go-Live (Okt 2026) ◄── SAP OFF │ ▼ Vollausbau (Q1-Q2/2027)
```

Kritische Abhängigkeiten: 1. **Dedalus FHIR-Zusage** – ohne validierte Schnittstelle kein Dokumentenzugriff 2. **SAP-DOK-Migration** – ohne migrierte Altdaten kein vollständiger Dokumentenbestand 3. **M-KIS-Berechtigungshoheit** – ohne M-KIS als Permission-Master kein Sicherheitskonzept nach SAP-Abschaltung 4. **GPU-Infrastruktur** – Vorlauf für Beschaffung ca. 8–12 Wochen

17.4 Rollback-Strategie

Szenario	Fallback	Trigger
FHIR-Schnittstelle nicht performant	Direkter ISILON-Zugriff mit eigenem Audit-Layer	M1.4 scheitert
MVP nicht rechtzeitig zu SAP-Abschaltung	HYDMedia Viewer als manuelle Recherche-Alternative (Ist-Zustand)	M2.4 scheitert

Szenario	Fallback	Trigger
Berechtigungs-Migration verzögert	Temporäre Whitelist + verschärftes Logging	M1.5 verzögert
OCR-Qualität unzureichend	Beschränkung auf computergenerierte Dokumente, Handschrift nachgelagert	M1.3 zeigt < 90%

Hinweis: Dieser Fahrplan ist ein Grobentwurf auf Basis der aktuellen Informationslage. Die detaillierte Planung erfolgt durch den installierten Projektleiter im Rahmen des Hauptprojekts.

---

18. Betriebskonzept – Gerüst (PSP-Ergänzung) {#18.-betriebskonzept}

Dieses Kapitel beschreibt die Mindestanforderungen für den Produktivbetrieb von UKLGPT. Es dient als Gerüst, das durch IT-Betrieb und Infrastruktur-Team mit konkreten Werten befüllt werden muss.

18.1 Service-Level-Agreements (SLA)

SLA-Parameter	Zielwert (Vorschlag)	Messverfahren	Verantwortlich
Verfügbarkeit	≥ 99,5% (Mo–Fr 7:00–20:00 Uhr)	Uptime-Monitoring (z.B. Prometheus/ Grafana)	IT-Betrieb
Geplante Wartungsfenster	Sa 22:00–So 06:00 Uhr	Vorabkommunikation 5 Werktage	IT-Betrieb
Antwortzeit	≤ 30 Sek. (P95) für Standard-Rechercheanfragen	APM-Monitoring (End-to-End)	IT-Betrieb
Wiederherstellungszeit (RTO)	≤ 4 Stunden (Prio 1), ≤ 24 Stunden (Prio 2)	Incident-Ticket-System	IT-Betrieb
Datenverlust-Toleranz (RPO)	≤ 1 Stunde (Graph-DB), 0 (Audit-Log)	Backup-Monitoring	IT-Betrieb
Max. gleichzeitige Nutzer	≥ 500	Load-Testing (vor Go-Live)	IT-Infrastruktur

18.2 Monitoring-Architektur

18.2.1 Infrastruktur-Monitoring

Komponente	Metriken	Schwellwert (Alarm)	Tool (Vorschlag)
GPU-Server (LLM-Inferenz)	GPU-Auslastung, VRAM, Temperatur	> 90% / > 85°C	Prometheus + DCGM Exporter
Vektor-Datenbank	Abfrage-Latenz, Index-Größe, Speicher	Latenz > 5 Sek.	Prometheus
Graph-Datenbank (Neo4j)	Heap-Nutzung, Query-Latenz, Transaktionen/Sek.	Heap > 85%, Latenz > 3 Sek.	Neo4j Ops Manager
LLM-Instanz	Inferenz-Latenz, Token/Sek., Queue-Tiefe	Queue > 50, Latenz > 20 Sek.	Custom Exporter
Netzwerk/Speicher (ISILON)	I/O-Durchsatz, Latenz, Kapazität	Kapazität > 90%	Grafana

18.2.2 Applikations-Monitoring

Metrik	Beschreibung	Alarm-Schwellwert
Anfragen/Minute	Gesamtlast auf UKLGPT-API	> 1000/Min. (Kapazitätswarnung)
Fehlerrate (5xx)	Anteil fehlgeschlagener Anfragen	> 1%
Berechtigungsprüfungen/Ablehnungen	Anzahl abgelehnter Zugriffe	Spike > 3σ (Anomalieerkennung)
RAG-Retrieval-Qualität	Anteil leerer Suchergebnisse	> 20% (Hinweis auf Indexierungsproblem)
Audit-Log-Integrität	Hash-Ketten-Validierung	Jede Inkonsistenz = Prio 1

18.2.3 Klinisches Qualitäts-Monitoring

Metrik	Beschreibung	Erhebung
Nutzerzufriedenheit	Thumbs-Up/Down pro Antwort	In-App-Feedback
Eskalationsrate	Anteil Anfragen, die an manuellen Support eskaliert werden	Ticket-System
Halluzinations-Meldungen	Vom Nutzer gemeldete faktische Fehler	In-App-Meldefunktion

18.3 Support-Modell

Support-Level	Beschreibung	Reaktionszeit	Verantwortlich
Level 1	Anwender-Support: Bedienungsfragen, Passwort-Resets, Zugriffsprobleme	≤ 4 Stunden (Kernzeit)	IT-Service-Desk (UKL)
Level 2	Technischer Support: Applikationsfehler, Performance-Probleme, Datenqualität	≤ 8 Stunden	UKLGPT-Betriebsteam (2–3 FTE)
Level 3	Architektur/ Entwicklung: Infrastruktur-Ausfälle, LLM-Anpassungen, komplexe Fehleranalysen	≤ 24 Stunden	IT-Architektur + ggf. externer Partner

Hinweis: Sicherheitsvorfälle (SEC-1/SEC-2) unterliegen den kürzeren Reaktionszeiten des Incident-Response-Plans (Kap. 19.2: ≤ 30 Min. bei Kritisch) und werden nicht über den regulären Support-Prozess gesteuert.

Bereitschaft: Kernzeit Mo–Fr 7:00–20:00. Außerhalb: Rufbereitschaft für Prio-1-Vorfälle (Ausfall, Sicherheitsvorfall).

18.4 Backup und Disaster Recovery

Komponente	Backup-Strategie	Frequenz	Aufbewahrung	Wiederh
Graph-Datenbank (Neo4j)	Inkrementelles Backup + wöchentliches Full-Backup	Täglich / wöchentlich	30 Tage	Quartals
Vektor-Datenbank	Snapshot + Re-Indexierung aus Quelldaten möglich	Wöchentlich	14 Tage	Bei Re-I implizit
Audit-Log	Append-Only, Replikat auf separatem Storage	Echtzeit-Replikation	≥ 10 Jahre (Aufbewahrungspflicht)	Halbjähr
Konfiguration/ Code	Git-Repository + Infrastructure-as-Code	Bei jeder Änderung	Unbegrenzt	Bei jede
LLM-Modell-Artefakte	Versioniertes Model-Registry	Bei jeder Modelländerung	Alle Versionen	Bei Roll

Disaster-Recovery-Szenario: 1. Totalausfall GPU-Server → Failover auf Backup-Instanz (Ziel: < 4h RTO) 2. Datenbankkorruption → Restore aus letztem konsistentem Backup + Re-Indexierung 3. Sicherheitsvorfall → Isolierung, Forensik, Restore aus Clean-Backup (→ siehe Kap. 19)

18.5 Kapazitätsplanung

Ressource	MVP (Okt 2026)	Vollausbau (2027)	Skalierungstrigger
GPU (LLM-Inferenz)	1× NVIDIA A100 80GB (o.ä.)	2–4× A100 (Load-Balancing)	Queue-Tiefe > 30 sustained
RAM (Graph-DB)	128 GB	256–512 GB	Heap > 80%
	~5 TB	~20 TB	Kapazität > 75%

Ressource	MVP (Okt 2026)	Vollausbau (2027)	Skalierungstrigger
Storage (Vektor-DB + Embeddings)			
Storage (Audit-Log)	500 GB	2 TB/Jahr (wachsend)	Kapazität > 80%

Anmerkung: Konkrete Sizing-Werte müssen durch IT-Infrastruktur auf Basis von Lasttests validiert werden.

18.6 Release- und Änderungsmanagement

Prozess	Beschreibung	Frequenz
Minor Releases (Bugfixes, Prompt-Tuning)	Staging → Test → Produktion, kein Wartungsfenster nötig	Nach Bedarf (Rolling Update)
Major Releases (Modellwechsel, Architekturänderung)	Staging → Abnahmetest → Wartungsfenster → Produktion + Rollback-Plan	Quartalsweise
Notfall-Patches (Sicherheit)	Sofort-Deployment nach 4-Augen-Prinzip	Bei Bedarf
Modell-Updates (LLM-Versionswechsel)	Parallelbetrieb Alt/Neu → A/B-Vergleich → Cutover	Nach Evaluierung

---

19. Incident-Response-Plan (PSP 4.3 Ergänzung) {#19.-incident-response-plan}

Dieser Plan definiert die Reaktion auf Sicherheitsvorfälle, KI-Fehlfunktionen und Datenschutzverletzungen im UKLGPT-Betrieb.

19.1 Vorfallkategorien

Kategorie	Beschreibung	Schweregrad	Beispiele
SEC-1		**Kritisch**	

Kategorie	Beschreibung	Schweregrad	Beispiele
	Datenschutzverstoß / Unbefugter Zugriff auf Patientendaten		Berechtigungsfehler → Patient A sieht Daten von Patient B; Audit-Log-Manipulation
SEC-2	Sicherheitslücke in Infrastruktur	**Hoch**	Unautorisierter Netzwerkzugriff auf GPU-Server; kompromittierte API-Keys
AI-1	Systematische KI-Fehlfunktion (Halluzination mit klinischer Relevanz)	**Hoch**	LLM erfindet Medikation, die nicht in Akte steht; falsche Laborwerte zitiert
AI-2	Qualitätsverschlechterung der KI-Antworten	**Mittel**	Zunehmend leere Ergebnisse; sinkende Nutzerzufriedenheit; Index-Drift
OPS-1	Systemausfall (UKLGPT nicht erreichbar)	**Hoch**	GPU-Server-Ausfall; Datenbank-Crash; Netzwerkproblem
OPS-2	Performance-Degradation	**Mittel**	Antwortzeiten > 60 Sek.; Queue-Überlauf

19.2 Eskalationsmatrix

Schweregrad	Reaktionszeit	Eskalationsstufe	Sofortmaßnahme
Kritisch (SEC-1)	≤ 30 Min.	ISB + DSB + IT-Leitung + Vorstand	System isolieren, Zugriff sperren, Forensik starten
Hoch (SEC-2, AI-1, OPS-1)	≤ 2 Stunden	ISB + IT-Betriebsteam + Projektleiter	Betroffene Komponente isolieren / deaktivieren
Mittel (AI-2, OPS-2)	≤ 8 Stunden	IT-Betriebsteam + IT-Architektur	Monitoring intensivieren, Workaround kommunizieren

19.3 Meldepflichten (DSGVO Art. 33/34 + BSI-KRITIS)

DSGVO-Meldepflicht (Art. 33/34)

Schritt	Frist	Verantwortlich	Aktion
1. Erkennung	Sofort	Monitoring / Nutzer-Meldung	Vorfall im Incident-Ticket dokumentieren
2. Erstbewertung	≤ 4 Stunden	DSB + ISB	Personenbezogene Daten betroffen? Risiko für Betroffene?
3. Meldung an Aufsichtsbehörde	**≤ 72 Stunden** nach Erkennung	DSB	Falls Risiko für Rechte/Freiheiten → Meldung an Sächsischen DSB
4. Benachrichtigung Betroffener	Unverzüglich	DSB + Klinikumsleitung	Falls hohes Risiko → individuelle Benachrichtigung
5. Forensische Analyse	≤ 5 Werktage	ISB + IT-Sicherheit	Root-Cause-Analyse, Audit-Log-Auswertung
6. Maßnahmen + Abschlussbericht	≤ 10 Werktage	Projektleiter + ISB + DSB	Technische Korrektur, Prozessanpassung, Dokumentation

BSI-KRITIS-Meldepflicht (§ 8b BSI-Gesetz, vgl. Kap. 13.4)

Das UKL unterliegt als Krankenhaus der kritischen Infrastruktur (KRITIS-Sektor Gesundheit) einer eigenständigen Meldepflicht gegenüber dem BSI:

Schritt	Frist	Verantwortlich	Aktion
1. Feststellung KRITIS-Relevanz	≤ 4 Stunden	ISB	Ist die Verfügbarkeit/ Integrität einer kritischen Dienstleistung (Patientenversorgung) betroffen?
2. BSI-Meldung (Erstmeldung)	**Unverzüglich** nach Erkennung	ISB	Meldung über BSI-Meldeportal

Schritt	Frist	Verantwortlich	Aktion
			(Kontaktstelle nach § 8b Abs. 3)
3. BSI-Folgemeldung	Nach Analyse	ISB	Ursache, Auswirkungen, ergriffene Maßnahmen

Hinweis: Die DSGVO- und BSI-Meldepflichten bestehen parallel und unabhängig voneinander. Bei SEC-1-Vorfällen sind in der Regel beide Meldewege auszulösen.

19.4 KI-spezifische Incident-Response

Bei Halluzination mit klinischer Relevanz (AI-1):

1. **Sofort:** Betroffene Anfrage + Antwort im Audit-Log identifizieren
2. **≤ 2h:** Reproduzierbarkeit prüfen (gleiche Eingabe → gleicher Fehler?)
3. **≤ 4h:** Falls systematisch → Betroffenen Dokumentenbereich/ RAG-Index isolieren
4. **≤ 24h:** Root-Cause-Analyse (Indexierungsfehler? Prompt-Drift? Modellfehler?)
5. **≤ 48h:** Korrektur deployen, betroffene Nutzer informieren
6. **Danach:** Testfall in automatisierte Qualitätsprüfung aufnehmen

Bei Berechtigungsfehler (SEC-1):

1. **Sofort:** UKLGPT-Zugriff für betroffenen Nutzerkreis sperren
2. **≤ 30 Min.:** Audit-Log auswerten: Wer hat was gesehen?
3. **≤ 4h:** Scope des Verstoßes bewerten (Anzahl betroffener Patienten/Datensätze)
4. **≤ 72h:** DSGVO-Meldung (falls erforderlich, siehe 19.3)
5. **≤ 5 Tage:** Technische Ursache beheben, Berechtigungslogik patchen
6. **Danach:** Penetrationstest auf Berechtigungsebene wiederholen

19.5 Post-Incident-Review

Jeder Vorfall ab Schweregrad "Hoch" wird innerhalb von 10 Werktagen in einem **Post-Incident-Review** aufgearbeitet:

Element	Beschreibung
Timeline	Wann wurde der Vorfall erkannt, eskaliert, behoben?
Root Cause	Was war die technische/organisatorische Ursache?

Element	Beschreibung
Impact	Welche Daten/Patienten/Nutzer waren betroffen?
Lessons Learned	Was muss sich ändern (Prozess, Technik, Monitoring)?
Action Items	Konkrete Maßnahmen mit Verantwortlichen und Fristen

---

20. Kostengerüst-Template (PSP 2.2.1 / 1.5 Ergänzung) {#20.-kostengerüst-template}

Dieses Template strukturiert alle relevanten Kostenblöcke. Die konkreten Werte müssen durch IT-Infrastruktur, Controlling und ggf. externe Anbieter befüllt werden.

20.1 Einmalige Investitionskosten (CAPEX)

Kostenblock	Beschreibung	Schätzung	Quelle/ Verantwortlich	S
GPU-Hardware	NVIDIA A100/H100 oder vergleichbar (1–4 Karten)	[ ] €	IT-Infrastruktur	C
Server-Hardware	Rack-Server für LLM-Inferenz, Graph-DB, Vektor-DB	[ ] €	IT-Infrastruktur	C
Storage-Erweiterung	Zusätzlicher ISILON-/SAN-Speicher für Embeddings + Audit	[ ] €	IT-Infrastruktur	C
Netzwerk	Ggf. 10/25 GbE Anbindung GPU-Cluster	[ ] €	IT-Infrastruktur	C
Softwarelizenzen (einmalig)	Neo4j Enterprise, ggf. OCR-Lizenzen	[ ] €	IT + Einkauf	C
Externe Beratung / Implementierung	KI-Architektur, Implementierungsunterstützung	[ ] €	PMO + Einkauf	C
Implementierung hAIppokrates (Alternative)	8.000 € zzgl. MwSt. (lt. Pitchdeck)	9.520 € brutto	Kap. 0.7	F V
Dedalus FHIR-Schnittstelle	Konfiguration/Entwicklung FHIR-Fassade (Kap. 0.6)	[ ] €	Dedalus + IT	C

Kostenblock	Beschreibung	Schätzung	Quelle/ Verantwortlich	S
DWH/ETL-Pipeline	CDC/Micro-Batch Pipeline DWH→GraphRAG (Kap. 8)	[ ] €	IT-Architektur	O
M-KIS-Integration	API-Anbindung Berechtigungsprüfung (Kap. 12.7.2)	[ ] €	Meierhofer + IT	O
Schulung / Change-Management	Trainer, Materialien, Pilotbegleitung (Kap. 21)	[ ] €	PMO	O
Snowstorm On-Premise	Elasticsearch-Backend für SNOMED CT Terminologie-Server (Kap. 14.2.1)	[ ] €	IT-Infrastruktur	O
NER-Pipeline (SNOMED-Autotagging)	Setup MedCAT/cTAKES, Training auf UKL-Dokumenten (Kap. 14.2.2)	[ ] €	IT + KI-Team	O
SUMME CAPEX		**[ ] €**		

20.2 Laufende Betriebskosten (OPEX, p.a.)

Kostenblock	Beschreibung	Schätzung p.a.	Quelle/ Verantwortlich	Statu
Personal: Betriebsteam	2–3 FTE (DevOps/ML-Ops, Applikationsbetreuung)	[ ] €	PMO/ Controlling	OFFE
Personal: Fachliche Betreuung	0,5 FTE (Prompt-Pflege, QA, Fachbereichskoordination)	[ ] €	PMO/ Controlling	OFFE
Stromkosten GPU	Ca. 300–700W pro GPU × 24/7	[ ] €	IT-Infrastruktur	OFFE
Softwarelizenzen (laufend)	Neo4j, ggf. LLM-API-Kosten (bei Cloud-Modell)	[ ] €	IT + Einkauf	OFFE
Wartung/ Support Hardware	Herstellerwartung Server + GPU	[ ] €	IT-Infrastruktur	OFFE
OCR-Verarbeitung	Laufende OCR für neue Dokumente	[ ] €	IT	OFFE
Averbis/ Meierhofer-Lizenz (Alternative)	Averbis-Lizenz + Meierhofer-Lizenz (Meierhofer-Anteil noch nicht beziffert)	**~200.000 €**	Mail Dr. Cundius (27.02.2026)	**KONK**

Kostenblock	Beschreibung	Schätzung p.a.	Quelle/ Verantwortlich	Statu
Externer Support / Beratung	Optional: Architektur-Reviews, Security-Audits	[ ] €	PMO	OFFE
SUMME OPEX p.a.		**[ ] €**		

20.3 Vergleichsübersicht Eigenlösung vs. Averbis/Meierhofer

Kostenart	Eigenlösung UKLGPT	Averbis/Meierhofer	Anmerkung
CAPEX (einmalig)	[ ] € (Hardware + Implementierung)	**~90.000 €** (Teststellung) + 10–20 PT HYDMedia-Anbindung	Eigenlösung: h HW-Investition Averbis: Testste allein 90 k€ (Qu Mail Dr. Cundiu 27.02.2026)
OPEX p.a. (Betrieb)	[ ] € (Personal + Infra + Lizenzen)	**~200.000 €/Jahr** (Averbis- + Meierhofer-Lizenz, Meierhofer-Anteil noch offen) + geringer Personalaufwand	Averbis: **deutli höhere laufen Kosten** als ursprünglich geschätzt. Ven Lock-in.
TCO 5 Jahre	**[ ] €**	**≥ 1.090.000 €** (90 k€ + 5 × 200 k€, ohne Anbindungsaufwand und Datenschutzgutachten)	Entscheidungs – Averbis-TCO bezifferbar
Personalaufbau	2,5–3,5 FTE (Kompetenzaufbau)	0,5–1 FTE (Administration)	Eigenlösung: K how bleibt im H
Strategischer Wert	Hoch (Datenhoheit, Erweiterbarkeit)	Niedrig (Abhängigkeit von Anbieter-Roadmap)	Nicht in € bezif
Datenschutzgutachten (ggf.)	Im internen DSB abgedeckt	Ggf. **~100.000 €** (externes Gutachten bei Cloud-Restriktionen Sachsen)	Risiko nur bei A Cloud-Variante

Handlungsbedarf: 1. IT-Infrastruktur: GPU-Sizing und Hardware-Kosten ermitteln 2. Controlling: FTE-Kosten kalkulieren (Entgeltgruppen TV-L) 3. ~~Einkauf: Averbis-Angebot von Meierhofer anfordern (Samira Grass)~~ **TEILWEISE ERLEDIGT (27.02.2026):**

Kostenindikatoren liegen vor (90 k€ Teststellung, ~200 k€/a Lizenz). Detailliertes Angebot mit Meierhofer-Lizenzanteil noch ausstehend. 4. PMO: TCO-Vergleich erstellen – **Averbis-Seite nun weitgehend bezifferbar (≥ 1,09 Mio. € / 5 Jahre)**. Eigenlösung-Kosten noch offen. 5. **NEU:** Prüfung Zusatzkosten Datenschutzgutachten (~100 k€) bei Cloud-Variante in Sachsen.

---

21. Change-Management-Konzept (Klinische Einführung) {#21.-change-management-konzept}

Dieses Konzept adressiert das Akzeptanz-Risiko (R-05 im Risiko-Register) und beschreibt die schrittweise Einführung von UKLGPT im klinischen Alltag.

21.1 Pilotierungsstrategie

Phase A: Pilotstation (Okt–Nov 2026, parallel zu MVP)

Aspekt	Beschreibung
Pilotstation	1 Station mit hoher Motivation und breitem Dokumentenbedarf (Vorschlag: Innere Medizin oder Chirurgie, Abstimmung mit Fachbereich)
Pilotgruppe	10–15 Ärzte + 5 Pflegekräfte (freiwillige Teilnahme)
Dauer	4–6 Wochen
Begleitung	1 Fach-Champion + 1 IT-Ansprechpartner vor Ort
Feedback-Erhebung	Wöchentliches kurzes Online-Feedback (5 Min.), Exit-Interview am Ende
Erfolgskriterien	≥ 70% der Pilotnutzer bewerten System als "hilfreich" oder "sehr hilfreich"; keine SEC-1-Vorfälle

Phase B: Erweiterter Pilot (Dez 2026–Feb 2027)

Aspekt	Beschreibung
Erweiterung	3–4 weitere Stationen (inkl. 1 chirurgische, 1 interdisziplinäre)

Aspekt	Beschreibung
Anpassungen	Basierend auf Phase-A-Feedback: Prompt-Optimierung, UI-Verbesserungen, Schulungsmaterial
Erfolgskriterien	Nutzung ≥ 3×/Tag/Arzt im Durchschnitt; Fehlerrate AI-1 < 1%

Phase C: Klinikweiter Rollout (ab Q2/2027)

Aspekt	Beschreibung
Rollout-Wellen	Je 5–10 Stationen pro Monat
Voraussetzung	Phase-B-Erfolgskriterien erreicht; Betriebskonzept (Kap. 18) vollständig umgesetzt
Begleitung	Champions-Netzwerk (siehe 21.2)

21.2 Champions-Netzwerk

Rolle	Anzahl	Aufgabe	Profil
Ärztliche Champions	1 pro Klinik/ Abteilung (Ziel: 15–20)	Multiplikator, Erstansprechpartner, Feedback-Kanal	Technologieaffine Ärzte mit klinischer Autorität
Pflege-Champions	1 pro großer Station (Ziel: 10–15)	Pflegespezifische Anwendungsfälle, Schulungsunterstützung	Erfahrene Pflegekräfte mit Digitalisierungsinteresse
IT-Liaison	1–2 gesamt	Technische Fragen vor Ort, First-Level-Triage	IT-Betriebsteam-Mitglied

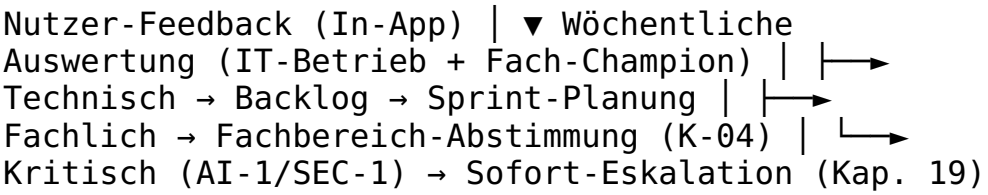
Anreize: Fortbildungspunkte (CME), Namentliche Nennung im Projekt, ggf. Freistellung für Champion-Aktivitäten.

21.3 Schulungskonzept

Schulungsformat	Zielgruppe	Dauer	Inhalt	Zeitpunkt
Kickoff-Präsentation	Alle Mitarbeiter der Pilotstation	30 Min.	Was ist UKLGPT? Was kann es? Was nicht? Datenschutz.	Vor Pilotstart

Schulungsformat	Zielgruppe	Dauer	Inhalt	Zeitpunkt
Hands-on-Workshop	Pilotnutzer (Ärzte + Pflege)	60 Min.	Live-Demo, eigene Anfragen stellen, Feedback geben	Pilotstart-Woche
Quick-Reference-Card	Alle Nutzer	–	1-Seiter: Zugang, Beispielanfragen, Dos/Don'ts, Support-Kontakt	Bei Freischaltung
E-Learning-Modul	Alle klinischen Mitarbeiter (Rollout)	20 Min.	Selbstlernkurs mit Praxisbeispielen, Quiz	Ab Rollout-Phase
Champions-Training	Champions	90 Min.	Vertiefung: Prompt-Strategien, Troubleshooting, Feedback-Prozess	Vor Pilotstart
Auffrischung	Alle Nutzer	15 Min.	Updates, neue Features, Erfahrungsberichte	Quartalsweise

21.4 Feedback- und Verbesserungsprozess



21.5 Erfolgsmessung Change-Management

KPI	Zielwert	Messzeitpunkt
Nutzungsrate (Anfragen/Tag/Arzt)	≥ 3	Ab Pilotstart, monatlich
Nutzerzufriedenheit (Skala 1–5)	≥ 3,5	Monatliche Kurzumfrage
Anteil geschulter Nutzer	≥ 90% (vor Freischaltung)	Pro Rollout-Welle
Champion-Abdeckung	≥ 80% der Abteilungen	Ab Rollout-Phase
Reduktion Recherchezeit (validiert)	≥ 50% ggü. Baseline	3 Monate nach Rollout

---

22. SAP IS-H/i.s.h.med Connector – Architektur, Schnittstellen und KI-Integration {#22.-sap-ish-connector}

Dieses Kapitel analysiert die technischen Möglichkeiten und Herausforderungen der Anbindung von UKLGPT an das bestehende KIS SAP IS-H/i.s.h.med. Es beschreibt die verfügbaren Schnittstellen, Integrationspfade und die Strategie für die Berechtigungsmigration auf M-KIS.

22.1 SAP IS-H/i.s.h.med als KIS – Systemübersicht {#22.1-systemübersicht}

SAP IS-H (Industry Solution for Healthcare) und i.s.h.med bilden zusammen das derzeit am UKL eingesetzte Krankenhausinformationssystem (KIS), das sowohl administrative als auch klinische Prozesse abdeckt:

- **SAP IS-H** (Branchenkomponente Krankenhaus): Deckt den gesamten administrativen Patientenpfad ab – von der Aufnahme über Verlegung und Entlassung bis zur Abrechnung. IS-H läuft als Branchenlösung auf SAP ERP Central Component (ECC) und ist eng mit den SAP-Standardmodulen FI/CO, MM und HR integriert.

- **i.s.h.med** (Klinisches System): Spezialisiert auf die klinische Dokumentation und medizinische Prozesse. Bietet Echtzeit-Zugriff auf die vollständige Patienteninformation und ist das einzige KIS, das vollständig in die SAP-Plattform integriert ist.

22.1.1 Zentrales Datenmodell

Das Datenmodell von SAP IS-H basiert auf einer hierarchischen Struktur transparenter SAP-Tabellen:

``` NPAT (Patientenstammdaten) ├── NADR (Adressdaten) ├── NGPA (Geschäftspartner / Personen) ├── NFPZ (Fall-Personen-Zuordnung, z.B. einweisender Arzt) └── NFAL (Fälle / Cases) – Schlüssel: PATID (1:n) ├── NBEW (Bewegungen zum Fall) – Schlüssel: FALNR (1:n) │ Aufnahme, Verlegung, Entlassung, ambulante Besuche ├── NDIA (Diagnosen) – zum Fall zugeordnet ├── NLEI (Erbrachte Leistungen / Services Performed) │ Beziehungen zu NBEW, NORG, NTPK └── NDOC (Dokumentzuordnung) Administrationsinformationen zu Dokumenten
```

NORG (Organisatorische Einheiten / OE) Abbildung der Klinikstruktur (Stationen, Ambulanzen, Fachabteilungen) ```

Zentrale ABAP Views: - **N_VBEWPAT** (View über NBEW, NFAL, NPAT): Zusammengeführte Sicht auf Bewegungen, Fälle und Patientenstammdaten.

22.1.2 Kernmodule und Relevanz für UKLGPT

Modul / Bereich	Funktion	Relevanz für UKLGPT
Patientenverwaltung	Aufnahme, Entlassung, Verlegung (ADT)	Falldaten, Bewegungen für Berechtigungsprüfung
Fallmanagement	Stationäre, teilstationäre, ambulante Fälle	Grundlage für Behandlungsauftrag
Leistungserfassung	Medizinische und pflegerische Leistungen	Leistungsdaten für GraphRAG
Diagnosen-/ Prozedurenverwaltung	ICD/OPS-Kodierung	Diagnosen für FHIR-Transformation
Klinische Dokumentation (i.s.h.med)	PMDs (Patientennahe Medizinische Dokumente), Arztbriefe, Befunde	Dokumentenzugriff für Dokumenten-RAG
Berechtigungsverwaltung	Behandlungsauftrag, OE-Zuordnung, Dokumentenschutz	**Kritisch** – Master of Permission (vgl. Kap. 12.1)
SAP HR	Mitarbeiterdaten, Planstellen, Organisationseinheiten	IAM-Mapping für Berechtigungen

---

22.2 Verfügbare Connectoren und Schnittstellen {#22.2-connectoren-und-schnittstellen}

22.2.1 BAPI (Business Application Programming Interface)

SAP stellt für IS-H spezifische BAPIs bereit, die über das RFC-Protokoll aufgerufen werden können:

BAPI	Funktion	Relevanz
BAPI_PATCASE_GETINPATADMISS	Anzeige stationärer Falldaten inkl. Aufnahmeinformationen	Abfrage von Falldaten für Berechtigungsprüfung
BAPI_BUSPARTNER_CREATE/ CHANGE	Anlage/Änderung von Geschäftspartnern (Patienten)	Patientenstammdaten
BAPI_HCSRVCAT_ADDITEM/ CHANGEITEM	Leistungskatalog-Positionen (GHS-Leistungen, PEPP)	Leistungsdaten
BAPI_MEDDOC_CREATE_LINK	Verknüpfung medizinischer Dokumente mit Archivdaten	Dokumentenzuordnung für i.s.h.med

Technischer Zugriff: BAPIs werden über das SAP RFC-Protokoll aufgerufen. Für die externe Anbindung stehen folgende Connectoren zur Verfügung: - **SAP JCo** (Java Connector) – für Java-basierte Anwendungen - **SAP NCo** (.NET Connector) – für .NET-basierte Anwendungen - **SAP Cloud Connector** – für Cloud-basierte Integration via SAP BTP

22.2.2 HL7-Schnittstellen

SAP IS-H unterstützt standardmäßig **HL7 v2.x** (insbesondere ADT-Nachrichten für Aufnahme/Entlassung/Verlegung) und **HL7 MDM** (Medical Document Management) für den Dokumentenaustausch.

Am UKL ist der Kommunikationsserver **Cloverleaf** als zentrale Nachrichtenplattform im Einsatz (vgl. Dokumentenpipeline in Kap. 0.6), der HL7-Nachrichten normalisiert und an nachgelagerte Systeme (DMI-Klassifizierung, HYDMedia) weiterleitet.

22.2.3 FHIR-Fähigkeiten

SAP hat als erster Hersteller die **ISiK-konforme FHIR-Zertifizierung** für Krankenhaussysteme erhalten. Die FHIR-Schnittstelle für SAP Patient Management unterstützt:

- **ISiK Basismodul**: Patient, Encounter, Condition, Procedure, Observation (gemäß gematik-Spezifikation)
- **FHIR Messaging**: Lose Kopplung von Informationssystemen, getriggert durch SAP-Systemveränderungen
- **ISiK Terminplanung**: Austausch von Patienteninformationen und Terminmanagement via FHIR

HYDMedia FHIR-Fassade (vgl. Kap. 0.6, 2.4): HYDMedia ist **kein nativer FHIR-Store**, sondern ein ECM-System mit FHIR-Fassade. Strukturierte klinische Ressourcen wie Observation (Laborwerte),

MedicationStatement (Medikation) oder Condition (Diagnosen) werden **nicht** als einzeln abfragbare FHIR-Datenpunkte unterstützt.

22.2.4 IHE-Profile

Am UKL werden folgende IHE-Integrationsprofile eingesetzt:

- **IHE XDS.b** (Cross-Enterprise Document Sharing): Dokumentenaustausch über Registry/Repository-Architektur – realisiert über DMI-Klassifizierungsserver und HYDMedia
- **IHE PIX/PDQ** (Patient Identifier Cross-Referencing / Patient Demographics Query): Patienten-Identifikation
- **IHE-konforme Dokumentenklassifizierung**: Über DMI-Klassifizierungsserver (AVP8) mit KDL-Klassifizierung

---

22.3 KI-Integrationspfade {#22.3-ki-integrationspfade}

22.3.1 Primärer Datenpfad: UKLytics-DWH (indirekt)

Gemäß der UKLGPT-Architektur (Kap. 8) erfolgt der Datenzugriff auf SAP-IS-H-Daten **nicht direkt**, sondern über das Echtzeit-DWH (UKLytics):

```
SAP IS-H/i.s.h.med ⟶ Echtzeit-DWH (UKLytics) ⟶ FHIR-Transformation ⟶ GraphRAG (Neo4j)
```

Integrationsmuster (Kap. 8.2): - **Near-Real-Time (NRT) / Event-basiert (CDC):** Änderungen in den SAP-Quelldaten werden mittels Change Data Capture erkannt und als Events übermittelt. - **Micro-Batch:** Für weniger zeitkritische Daten alle 5–15 Minuten.

22.3.2 Vergleich der Integrationsmethoden

Methode	Vorteile	Nachteile	Empfehlung
DWH-Replikation (CDC)	Entkoppelt, performant, skalierbar	Latenz (Sekunden bis Minuten)	**Empfohlener Hauptpfad**
BAPI/RFC über JCo/ NCo	Echtzeit, strukturierter Zugriff	Enge SAP-Kopplung, Komplexität	Nur für Berechtigungsprüfung
FHIR-Schnittstelle (ISiK)	Standardisiert, REST-basiert	Eingeschränkter Ressourcenumfang	Ergänzend möglich
OData/REST via SAP Gateway	Modern, webfreundlich	Begrenzte IS-H-spezifische Services	Langfristig relevant (S/4HANA)

22.3.3 Echtzeit-Berechtigungsprüfung über SAP IS-H

Dies ist der **kritischste Integrationspunkt** für UKLGPT. Der technische Ablauf (vgl. Kap. 12.7):

```
1. User-Login (AD/SSO) 2. M-KIS sendet API-Aufruf an UKLGPT (User-ID + Fallnummer) 3. UKLGPT: Fail-Safe-Vorsperre (kein Zugriff ohne ACCESS_GRANTED) 4. IAM/AD: Auflösung der SAP-Personalnummer, Planstelle, OE 5. SAP IS-H Validierung: Behandlungsauftrag + funktionale Rechte 6. Echtzeit-DWH/Graph: Strukturelle Berechtigung (User ↔ OE ↔ Patient) 7. Finale Gatekeeper-Entscheidung: ACCESS_GRANTED oder ACCESS_DENIED
```

Abfragbare Berechtigungsdaten aus SAP IS-H: - Behandlungsauftragsrelation (dynamisch, temporär, Notfall) - OE-Zuordnung des Benutzers - Sammelberechtigungsgruppen (SBG) - Funktionale Rechte (Transaktionen, Berechtigungsgruppen) - Dokumentenschutz-Flags (PSY/KJP/PST, vgl. Kap. 12.1.1, 12.3.3)

22.3.4 Dokumentenzugriff für UKLGPT

SAP i.s.h.med verwaltet Patientennahe Medizinische Dokumente (PMDs). Der Zugriffspfad für UKLGPT:

```
HYDMedia FHIR-API ⟶ DocumentReference (Metadaten) ⟶ Binary (PDF-Original) ⟶ Patient/Encounter (Kontextfilterung)
```

Status am UKL: - ca. 50% der SAP-Dokumente bereits über den DMI-Klassifizierungsweg (IHE/KDL-konform) in HYDMedia - **SAP-DOK-Migration** (geplant ab KW7/2026): 20 Jahre SAP-Dokumente werden in HYDMedia migriert – danach 100% Abdeckung

22.3.5 RAG-Anbindung pro Schicht

RAG-Schicht	Datenquelle	SAP-Bezug	Update-Frequenz
Patient-RAG (GraphRAG)	Strukturierte Fakten (ICD-Diagnosen, LOINC-Labor, ATC-Medikation)	SAP IS-H via DWH → FHIR → Neo4j	Near-Real-Time (CDC)
Dokumenten-RAG (Vector-RAG)	Arztbriefe, Befunde, OP-Berichte als PDF	SAP PMDs → HYDMedia → FHIR Binary	Bei Neueingang (Event-basiert)
	Leitlinien, SOPs		Periodisch

RAG-Schicht	Datenquelle	SAP-Bezug	Update-Frequenz
Reference-RAG		Kein SAP-Bezug	
Evidenz-RAG	Europe PMC, ClinicalTrials.gov	Kein SAP-Bezug	Automatisches Matching (Kap. 7.2.1)

22.4 Datenextraktion und FHIR-Transformation {#22.4-datenextraktion}

22.4.1 Extraktionspfade aus SAP IS-H

Extraktionsmethode	Datentyp	Ziel	Werkzeug
SAP BW DataSources	Falldaten (NFAL), Bewegungen (NBEW), Diagnosen (NDIA), Leistungen (NLEI)	UKLytics-DWH	SAP BW-Extraktoren
BAPI-basierter Export	Patientenstamm, Falldaten	Migrationstool	JCo/NCo
HL7 MDM-Replay	Dokumente (PMDs)	HYDMedia	Cloverleaf
Direkte Tabellenextraktion	NPAT, NFAL, NBEW, NDIA, NLEI, NDOC	DWH / FHIR-Mapping	SAP Open SQL / ABAP

22.4.2 FHIR-Mapping (SAP → FHIR R4)

Das FHIR-Mapping der SAP-IS-H-Daten erfolgt gemäß MII-Profilen (Medizininformatik-Initiative) und ISiK-Spezifikation:

SAP IS-H Daten	SAP-Tabelle	FHIR-Ressource	Besonderheit
Patientenstamm	NPAT, NADR	Patient	MPI-Abgleich erforderlich
Fall	NFAL	Encounter	Stationär/ Ambulant-Differenzierung
Bewegung	NBEW		Verlegungshistorie

SAP IS-H Daten	SAP-Tabelle	FHIR-Ressource	Besonderheit
		Encounter (Location)	
Diagnose	NDIA	Condition	ICD-10-GM → SNOMED-CT-Mapping
Leistung	NLEI	Procedure	OPS → SNOMED-CT-Mapping
OE-Zuordnung	NORG	Organization / Location	Hierarchische Struktur
Dokument-Verweis	NDOC	DocumentReference	Link zum PDF in HYDMedia
Behandlungsauftrag	SAP-spezifisch	(kein Standard-FHIR-Mapping)	Custom Extension erforderlich

22.4.3 Graph-Transformation (FHIR → Neo4j)

```
SAP-Tabellen (NFAL, NBEW, NDIA, NLEI, NPAT) ⟶ DWH-Schema (UKLytics) ⟶ FHIR-Ressourcen (Patient, Encounter, Condition, Procedure, MedicationRequest) ⟶ Graph-Knoten und Relationships (Neo4j/Cypher)
```

Upsert-Strategie (Kap. 8.2.3): Verwendung von `MERGE`-Befehlen in Cypher mit fachlichen Schlüsseln (Patienten-ID, Fall-ID, Prozeduren-Code) als eindeutige Identifikatoren.

22.5 Migration der Berechtigungshoheit auf M-KIS {#22.5-berechtigungsmigration}

SAP IS-H wird im **Oktober 2026** abgeschaltet. Die Berechtigungshoheit muss auf M-KIS (Meierhofer) migriert werden. Das Architekturprinzip „führendes KIS = Master of Permission" gilt systemunabhängig (vgl. Kap. 12.1).

22.5.1 Migrationsprogramm

Schritt	Maßnahme	Zeitraum	Verantwortlich
1	Abstimmung Berechtigungs-APIs mit Meierhofer	Q2/2026	IT-Architektur, Robert W.
2	Prototypisierung M-KIS-Berechtigungsconnector	Q2–Q3/2026	IT-Architektur

Schritt	Maßnahme	Zeitraum	Verantwortlich
3	Routing-Definition Parallelbetrieb (SAP + M-KIS)	Q3/2026	IT-Betrieb
4	DWH-Umstellung SAP → M-KIS-Datenquellen	Q3/2026	UKLytics-Team
5	Vollständiger Regressionstest Berechtigungsfluss	Q3/2026	QA, IT-Sicherheit
6	Cutover auf M-KIS als alleiniger Master	Okt 2026	Projektleitung

22.5.2 Graph-Modellierung der Berechtigungen

```
(User)-[:GEHOERT_ZU]->(OE) (Patient)-
[:AUFENTHALT_IN]->(OE) (User)-
[:HAS_BEHANDLUNGSAUFTRAG {typ: 'dynamisch',
gueltig_bis: timestamp}]->(Patient) (User)-
[:HAS_TEMP_ACCESS {until: '2026-02-27T23:59:59'}]-
>(Patient) (User)-[:HAS_EMERGENCY_ACCESS
{is_emergency: true, logged_at: timestamp}]-
>(Patient)
```

22.6 Herausforderungen und Risiken {#22.6-herausforderungen}

22.6.1 Architekturelle Herausforderungen

Herausforderung	Beschreibung	Mitigationsstrategie
SAP-Abschaltung Okt 2026	IS-H steht danach nicht mehr als Echtzeit-Datenquelle zur Verfügung	Frühzeitige Abstimmung mit Meierhofer, Parallelbetriebs-Routing
HYDMedia-Berechtigungslücke	HYDMedia setzt Need-to-Know **nicht** um – Standardzugriff über Standardnutzer	Berechtigungshoheit aus M-KIS/SAP ableiten, FHIR-Abfrage erst nach KIS-Prüfung
SAP-DOK-Migration ausstehend	20 Jahre Altdokumente erst nach	

Herausforderung	Beschreibung	Mitigationsstrategie
	KW7/2026 in HYDMedia verfügbar	Abhängigkeit in Projektplanung einplanen (Kap. 17)
Fehlende Labordaten in HYDMedia	Laborbefunde nicht über HYDMedia abrufbar	Separater FHIR-Connector zu UKLytics/Labor-LIMS
Verbucher-Dokumente nicht IHE-konform	Ohne KDL-Klassifizierung schwer durchsuchbar	Nachträgliche Klassifizierung oder Ausschluss
Duplikate in Dokumentenpipeline	Arztbriefe können 3–5× im System landen	Dedup-Logik bei Indexierung (Hash-Vergleich, Timestamp-Gewichtung)

22.6.2 Technische Limitierungen

Limitierung	Auswirkung auf KI-Integration
PDF als Archivformat	Informationsverlust bei OCR-Rücktransformation; „KI-feindliches Format" (vgl. Kap. 2.1, 2.4)
Kein nativer FHIR-Store in HYDMedia	Kein Zugriff auf feingranulare klinische Ressourcen (Observation, MedicationStatement, Condition)
RFC-Protokoll-Abhängigkeit	Keine native REST/HTTP-Integration möglich; JCo/NCo-Connector erforderlich
Behandlungsauftrag ohne FHIR-Mapping	SAP-Behandlungsauftrag hat kein Standard-FHIR-Äquivalent – Custom Extension erforderlich
Dedalus-Abhängigkeit	Kritische Abhängigkeit vom Dedalus-Team für performante FHIR-Exportschnittstelle

22.6.3 Organisatorische Herausforderungen

Herausforderung	Beschreibung
Parallelbetrieb SAP/M-KIS	Routing-Definition erforderlich: Welches System hält für welchen Fall die Berechtigungshoheit?

Herausforderung	Beschreibung
Springer-Problematik	Mitarbeiter mit wechselnden Stationszuordnungen erfordern dynamische Berechtigungszuweisung
VIP-Patienten	Aktuell keine gesonderte Einschränkung vorgesehen – erhöhtes Missbrauchsrisiko
Inaktive Patienten	Eigene BG erforderlich – standardmäßig kein automatischer Zugriff. Muss in UKLGPT abgebildet werden
Zeitdruck	SAP-Abschaltung Okt 2026 bei gleichzeitigem M-KIS-Rollout erzeugt extremen Zeitdruck für Berechtigungsmigration

---

22.7 Empfehlungen für die Connector-Strategie {#22.7-empfehlungen}

1. **Primärer Datenpfad:** DWH-Replikation (CDC) über UKLytics – entkoppelt, performant, skalierbar
2. **Berechtigungsprüfung:** Echtzeit-Graph-Layer (Neo4j) mit SAP-IS-H-Replikation bis Okt 2026, danach M-KIS
3. **Dokumentenzugriff:** HYDMedia FHIR-API (DocumentReference → Binary) mit vorgelagerter Berechtigungsprüfung
4. **FHIR-Transformation:** DWH-Schema → FHIR R4 (MII-Profile) → Graph-Schema als dreistufige Pipeline
5. **Berechtigungsmigration:** Frühzeitige Prototypisierung des M-KIS-Berechtigungsconnectors parallel zum SAP-Betrieb
6. **Sofortige Klärung mit Dedalus:** Performance-Zusage und Zeitplan für die FHIR-Exportschnittstelle

---

99 Feedback und ToDos {#99-feedback-und-todos}

Gerne prüfe ich das vorliegende Konzept der **Zielarchitektur UKLGPT** auf Konsistenz und identifiziere potenzielle Lücken.

Das Konzept ist in seinen Kernprinzipien sehr **konsistent** und weist eine hohe logische Kohärenz auf, insbesondere im Hinblick auf Sicherheit und klinische Verantwortung. Die identifizierten **Lücken** betreffen hauptsächlich die Detaillierung spezifischer technischer Mechanismen sowie organisatorische und metrische Aspekte.

I. Konsistenz (Stärken des Konzepts)

Das Gesamtkonzept ist durch eine klare, disziplinierte Architektur gekennzeichnet. Die Konsistenz zeigt sich in folgenden Punkten:

Bereich	Konsistenz (Stärke)
Klinische Verantwortung	Die Positionierung als reine **Informations- und Orientierungsassistenz** ist explizit und wird durchgehend betont (Kap. 2.6.4, 3.1.3). Dies stellt die **finale klinische Entscheidung und Verantwortung des Fachpersonals** juristisch und prozessual unangreifbar.
Datengovernance & Sicherheit	Es gibt eine klare und konsistente Strategie, um Sicherheit (Privacy by Design) und Nachvollziehbarkeit zu gewährleisten (Abschnitt 2). Die Prinzipien **Least Privilege & Need-to-Know** (Kap. 14.5), die **Episodische Datenverarbeitung/TTL** (Kap. 14.4) und der umfassende **Audit-Trail** (Kap. 12.6) stützen sich gegenseitig.
Terminologie und Semantik	Die Festlegung auf **FHIR** (Interoperabilität) und **SNOMED CT** (klinische Referenzterminologie) als zentrale Standards (Kap. 14.1, 14.2, 2.1, 2.2) ist fundamental für die Präzision und minimiert die Gefahr von KI-Halluzinationen aufgrund terminologischer Inkonsistenzen.
Qualitätssicherung (Guardrails)	Die mehrstufige **Prompt-Pipeline** (Kap. 6) und die **klare Trennung der Wissensdomänen** (Fakten, Dokumente, Leitlinien; Kap. 14.3) sind konsistente und essenzielle Mechanismen, um die Verlässlichkeit des KI-Outputs systematisch zu filtern und zu auditieren.
Problem-Lösungs-Fit	Das System ist darauf ausgelegt, die im Ist-Zustand (3.2) identifizierte **Fragmentierung medizinischer Information** und den **hohen manuellen Suchaufwand** (15–50 Minuten pro Arzt/Tag) direkt durch eine konsolidierte Wissensbasis und natürliche Sprachanfragen zu lösen.

II. Lücken (Potenzielle Ergänzungsbereiche)

Obwohl die Architektur robust ist, gibt es Bereiche, deren Detaillierung das Konzept weiter stärken würde:

Bereich	Lücke / Fehlende Detaillierung	Empfohlene Ergänzung/ Konkretisierung
Lösung "Externe Befunde"	Das Problem der **zeitversetzten und unklaren Zuordnung** externer Befunde in HYDMedia (3.2.3) wird zwar präzise beschrieben, die spezifische Lösung des UKLGPT-Systems zur **Behebung dieser Lücke** ist jedoch unkonkret ("zentrale, intelligente...").	Detaillierung des technischen Prozesses: Wie genau wird UKLGPT/RAG die Dokumente beim **Eingang** semantisch taggen und den korrekten Patientenepisoden zuordnen, bevor eine Suche initiiert wird?
Management von Leitlinien	Die Domäne **"Globales Wissen (Evidence-Base)"** (2.3) umfasst klinische Leitlinien und wissenschaftliche Literatur. Es fehlt die Spezifizierung, **wie deren Aktualität, Vollständigkeit und Einbindung in den RAG-Prozess technisch sichergestellt wird.**	Klarer Prozess für: **(1) Data Ingestion** (Schnittstelle zu AWMF, PubMed, etc.), **(2) Versionierung** (welche Leitlinienversion ist gerade gültig?) und **(3) Plausibilitätsprüfung** (Verzahnung mit der Prompt-Pipeline).
"Zweckbindung" RAG-System	Das Prinzip des **"Episodischen Patienten-RAG mit Zweckbindung"** (1.3.3) ist strategisch wichtig für die Datensicherheit,	Detaillierung der **ABAC (Attribute-Based Access Control) Policies** (2.5), die die Zweckbindung im Retrieval-Prozess steuern. Wie wird der klinische Anwendungsfall

Bereich	Lücke / Fehlende Detaillierung	Empfohlene Ergänzung/ Konkretisierung
	aber die **technische Logik zur Definition und Durchsetzung der Zweckbindung** (z.B. ein Regelwerk) ist nicht beschrieben.	("Medikationsprüfung") technisch in ein Retrieval-Limit übersetzt?
Quantifizierung und ROI	Der hohe Suchaufwand (3.2.4) ist quantifiziert, es fehlen jedoch die **Prognosen und Messgrößen für den Erfolg (ROI)** der Lösung.	~~Ergänzung um Erfolgsmetriken~~ **TEILWEISE BEHOBEN (v1.4):** SLA-Zielwerte in Kap. 18.1, Change-Management-KPIs in Kap. 21.5 (Nutzungsrate, Zufriedenheit, Recherchezeit-Reduktion). Detaillierte ROI-Berechnung erfordert befülltes Kostengerüst (Kap. 20).
~~**Organisation und Adoption**~~	~~Fehlt ein Abschnitt zur strategischen Umsetzung und den notwendigen organisatorischen Schritten.~~	**BEHOBEN (v1.4):** Change-Management-Konzept in Kap. 21 mit 3-Phasen-Pilotierung, Champions-Netzwerk, Schulungskonzept (6 Formate), Feedback-Prozess, Erfolgsmessung.

Das Dokument „Zielarchitektur UKLGPT mit Patienten-RAG, Dokumenten-RAG und GraphRAG (FHIR/SNOMED)“ ist inhaltlich sehr detailliert und bietet eine klare Begründung für die Strategie. Die Konzentration auf die Nutzung von FHIR-Metadaten und GraphRAG zur Steuerung der Recherche in Altdaten ist ein architektonischer Schlüssel.

Hier ist eine Analyse hinsichtlich **Konsistenz** und **inhaltlicher Verbesserungsmöglichkeiten**:I. Konsistenz und Strukturelle Verbesserungsmöglichkeiten

Aspekt	Inkonsistenz / Verbesserung	Vorschlag zur Behebung
Gliederungsstruktur	Der Abschnitt "Gesamtarchitektur – Überblick" (mit Unterpunkten 3.1, 3.2, 3.3, etc.) beginnt nach dem Abschnitt "Fachliches Zielbild & Anwendung" (mit Unterpunkten 3.1, 3.2, 3.3, etc.) erneut mit einer "3.". Dies ist irreführend und bricht die fortlaufende Nummerierung.	Die Überschrift **"3. Gesamtarchitektur – Überblick"** sollte in **"4. Gesamtarchitektur – Überblick"** umbenannt werden (und die nachfolgenden Hauptabschnitte 4, 5, 6, etc. entsprechend hochgezählt werden).
Terminologie	Es wird mehrfach der Begriff "UKLGPT Chatbot" (z.B. in 1.2.2) verwendet, während der Fokus der technischen Architektur auf dem "UKLGPT-RAG-System" liegt.	Für die Backend-Architektur sollte der konsistente Begriff **UKLGPT-RAG-System** oder **UKLGPT** verwendet werden. Der Begriff *Chatbot* sollte auf die Frontend-/UI-Schicht beschränkt werden.

II. Inhaltliche Verbesserungsmöglichkeiten

Bereich	Herausforderung / Verbesserungsmöglichkeit	Detaillierung für den Inhalt
Strategische Abwägung (PDF/A vs. FHIR-Nativ)	Es besteht eine starke inhaltliche Spannung zwischen der gewählten Strategie (Archivierung als PDF/A aus regulatorischen Gründen) und der als "suboptimal" beschriebenen Nachteile dieser Wahl (Abschnitt 2.1).	Fügen Sie in Abschnitt **1.2.1/ Fazit** eine präzisere Rechtfertigung ein. Begründen Sie, warum der regulatorische Zwang (Revisionssicherheit nach IDW PS 880, etc.) die in Kauf genommene *Semantik-Einbuße* und den *OCR*-

Bereich	Herausforderung / Verbesserungsmöglichkeit	Detaillierung für den Inhalt
		Aufwand überwiegt. Verdeutlichen Sie: **PDF/A ist die regulatorische Basis, GraphRAG die semantische Ergänzung.**
Auswahl des Archivsystems	Es werden zwei spezifische Archivsysteme genannt ("entweder **HYDMedia G6** oder **DMI AVP Infinity**" in 1.2.1), was auf eine noch **ausstehende technische Entscheidung** hindeutet.	Präzisieren Sie die Auswahlkriterien (z.B. Kosten, API-Performance, Skalierbarkeit für den Export) oder verschieben Sie die Nennung beider Namen in einen **"To-Do- oder Entscheidungs"-Abschnitt** (z.B. 99) mit einem klaren Meilenstein.
Gatekeeper-Performance	Das *Gatekeeper-Prinzip* (1.3.3) zur Prüfung der aktiven Behandlungsbeziehung ist kritisch für die klinische Akzeptanz. Die Verifizierung vor **jeder** Abfrage impliziert eine extrem niedrige Latenz.	**Erläutern Sie in 1.3.3 oder 12.7** kurz die **Performance-Anforderung** an die Gatekeeper-Schnittstelle (z.B. "Ziel-Latenz unter 50 ms") und welche technische Lösung (z.B. direkter Datenbank-Lookup statt API-Aufruf) diese Anforderung gewährleisten soll.
GraphRAG-Steuerung des Retrievals	Die Funktion des GraphRAG als "Schlüssel zur Effizienz" (1.3.2) durch die Beschränkung des Vektor-Retrievals ist klar, aber der *technische Ablauf* könnte detaillierter sein.	Beschreiben Sie explizit, **wie die Metadaten des Graphen die Vektor-Datenbank filtern/steuern** (z.B. "Der Graph liefert eine präzise Liste von *Chunk-IDs* oder *Document-Hashes*, die das Vektor-Retrieval gezielt

Bereich	Herausforderung / Verbesserungsmöglichkeit	Detaillierung für den Inhalt
		einschränken (Pre-Filter-Strategie).'').
Datenzufluss ETL-Logik (Kapitel 8)	Der Abschnitt **8.2 Integrationsmuster und ETL-Logik** ist ein kritischer Punkt für die Datenqualität und Konsistenz des Graphen, aber der Inhalt ist im vorliegenden Ausschnitt nicht vorhanden.	Fügen Sie einen **detaillierten Sub-Abschnitt** ein, der die **Transformations- und Validierungsregeln** beim Übergang vom DWH zur Graphen-DB beschreibt (z.B. Mapping von SAP-Codes auf SNOMED CT/FHIR, Umgang mit inkonsistenten oder fehlenden Daten).

[image3]: