arc42-Architekturdokumentation: Leistungsabrechnung LEA

Fallbeispiel Krankenversicherung · Nov 16, 2026 · @Bastian

Über dieses Dokument

LEA ist die Plattform, mit der die fiktive LimmatVita Krankenversicherung AG Leistungsrechnungen automatisiert prüft und abrechnet. Dieses Dokument beschreibt ihre Architektur nach dem arc42-Template (Version 8) in allen zwölf Kapiteln.

Aspekt	Inhalt
Charakter	Fallbeispiel. Unternehmen, Kennzahlen, Teams und Entscheide sind erfunden. Tarife, Standards, Gesetze und Azure-Dienste sind real und unter «Quellen» belegt.
Systemumfang	Release 1: ambulante Rechnungen nach TARDOC und ambulanten Pauschalen, elektronisch und als Dokument. Stationäre SwissDRG-Rechnungen, Medikamente und Pflege folgen in späteren Releases.
Architekturstand	Baseline für Release 1. Der Pilot läuft im Schattenbetrieb parallel zum Altsystem.
Zielpublikum	Architektur-Board, Entwicklungsteams, Plattformbetrieb, Informationssicherheit, Datenschutzberatung, interne Revision
Notation	Kästen und Pfeile, angelehnt an C4 (Kontext, Container, Komponenten). Prozessmodelle in BPMN liegen im Repository, nicht in diesem Dokument.
Flüchtige Details	Versionsnummern von Bibliotheken, Sizing und Konfiguration stehen im Repository (Version Catalog, Kustomize-Overlays). Dieses Dokument hält fest, was länger als ein Release gilt.

Rechtliche Aussagen sind Entwurfsannahmen des Fallbeispiels und keine Rechtsberatung.

1 Einführung und Ziele

LEA soll 85 % der ambulanten Leistungsrechnungen ohne manuellen Eingriff prüfen und abrechnen und dabei jeden Entscheid nachvollziehbar halten.

1.1 Aufgabenstellung

LimmatVita versichert rund 620′000 Personen in der obligatorischen Krankenpflegeversicherung (OKP) und in Zusatzversicherungen. Pro Jahr treffen etwa 8 Mio. Leistungsbelege ein: im Mittel 32′000 pro Arbeitstag, von Januar bis März bis 60′000. 82 % kommen elektronisch als XML, 18 % als Papier, PDF oder Foto.

Seit dem 1. Januar 2026 werden ambulante ärztliche Leistungen nach TARDOC und ambulanten Pauschalen abgerechnet; TARMED ist abgelöst (BAG). Das Altsystem wurde dafür mit hart codierten Regeln nachgezogen. Seither ist die Dunkelverarbeitungsquote elektronischer Rechnungen von 71 % auf 58 % gefallen. Dokumentbelege werden vollständig von Hand erfasst.

Kennzahl	Heute (Altsystem)	Ziel mit LEA
Dunkelverarbeitung elektronische Rechnungen	58 %	90 %
Dunkelverarbeitung Dokumentbelege	0 %	60 %
Dunkelverarbeitung über alle Kanäle	48 %	85 %
Durchlaufzeit dunkel verarbeiteter Fälle (95. Perzentil)	1 Arbeitstag (Nachtlauf)	5 Minuten
Regeländerung bis Produktion	6 bis 10 Wochen	höchstens 10 Arbeitstage

Wesentliche Funktionen:

ID	Funktion	Beschreibung
F1	Eingang	XML-Rechnungen nach generalInvoiceRequest 5.0 (Forum Datenaustausch) sowie Dokumentbelege aus Kundenportal, App und Scan-Strasse annehmen
F2	Dokumentenverständnis	Belege klassifizieren und in ein kanonisches Rechnungsmodell überführen, mit Konfidenz je Feld

ID	Funktion	Beschreibung
F3	Prüfung	Formal-, Deckungs- und Tarifprüfung sowie Auffälligkeitsprüfung gegen einen versionierten Regelstand
F4	Fallsteuerung	Dunkel verarbeiten oder in einen Arbeitskorb leiten, mit Fristen, Eskalation und Vier-Augen-Prinzip
F5	Abrechnung	Kostenbeteiligung berechnen, Leistungsabrechnung und Zahlungsauftrag erzeugen, Rechnungsantwort an Leistungserbringer senden
F6	Nachweis	Entscheidhistorie lückenlos führen, Entscheide zum Stichtag reproduzieren, Belege archivieren

Nicht im Umfang: Kostengutsprachen, Regress, Prämieninkasso, Tarifverhandlung und Vertragsmanagement.

1.2 Qualitätsziele

Die fünf Ziele sind nach Priorität geordnet. Bei Konflikten gewinnt das höhere Ziel. Messbare Szenarien dazu stehen in Kapitel 10.

Prio	Qualitätsziel	Was das konkret heisst
1	Nachvollziehbarkeit und Korrektheit	Jeder Entscheid lässt sich mit Eingangsdaten, Regelstand und Modellversion reproduzieren. Fehlentscheide in der Dunkelverarbeitung liegen unter 0.5 % (Stichprobe).
2	Vertraulichkeit	Gesundheitsdaten werden nur in der Schweiz gespeichert und bearbeitet. Zugriff nach Need-to-know, jeder Zugriff protokolliert.
3	Automatisierungsgrad	Mindestens 85 % Dunkelverarbeitung über alle Kanäle, 95 % dieser Fälle in unter 5 Minuten entschieden.
4	Änderbarkeit	Neue Tarifversion oder Regeländerung in höchstens 10 Arbeitstagen produktiv. Reine Regeldaten ohne Software-Deployment.

Prio	Qualitätsziel	Was das konkret heisst
5	Robustheit	Kein Beleg geht verloren. Fällt ein AI-Dienst aus, läuft die XML-Verarbeitung unverändert weiter.

1.3 Stakeholder

Rolle	Erwartung an die Architektur
Leiterin Leistungen (Auftraggeberin)	Automatisierungsgrad, Durchlaufzeit, Kosten je Beleg
Sachbearbeitende Leistungen (rund 140 Vollzeitstellen)	Ein Cockpit, verständliche Prüfhinweise, keine Doppelerfassung
Tarifspezialistinnen und -spezialisten	Regeln selbst pflegen, gegen historische Fälle testen und freigeben
Vertrauensärztlicher Dienst	Abgeschirmter Zugriff auf medizinische Angaben
Datenschutzberater und CISO	Datenhaltung Schweiz, Datenschutz-Folgenabschätzung, Zugriffskonzept
Interne Revision, Aufsicht (BAG für OKP, FINMA für Zusatzversicherung)	Nachvollziehbare Entscheide, kontrollierte Auslagerung, AI-Governance
Entwicklungsteams (3 Teams, 21 Personen)	Klare Modulgrenzen, schnelle Rückmeldung aus Tests
Plattformteam	Einheitliches Deployment, einheitliche Observability
Versicherte	Schnelle, verständliche Rückerstattung; menschliche Überprüfung auf Wunsch
Leistungserbringer	Zügige Zahlung, strukturierte und begründete Rückweisungen

2 Randbedingungen

Drei Vorgaben prägen die Architektur am stärksten: Gesundheitsdaten bleiben in der Schweiz, jeder Entscheid muss begründbar sein, und der Betrieb läuft auf der bestehenden Azure-Plattform.

2.1 Technische Randbedingungen

ID	Randbedingung	Hintergrund und Folge
T1	Azure ist die strategische Cloud. Produktion in Switzerland North, Switzerland West nur für Backup und Notfall.	Gruppenvorgabe; die Landing Zone des Plattformteams besteht bereits.
T2	Laufzeitplattform ist Kubernetes (AKS), ausgeliefert wird per GitOps mit Argo CD.	Plattformstandard. Teams haben keinen schreibenden Clusterzugriff.
T3	Backend in Java mit Spring Boot, relationale Daten in PostgreSQL.	Unternehmensstandard und Know-how der Teams.
T4	Geschäftsprozesse werden in BPMN auf Flowable modelliert.	Flowable ist seit 2021 für Vertragsmutationen im Einsatz; Betriebserfahrung vorhanden.
T5	Das Kernsystem «Bestand» (Partner, Policen, Deckungen) bleibt führend und on-premises. Zugriff nur über dessen REST-API.	Ablösung nicht vor 2029. LEA darf dessen Daten puffern, aber nicht ändern.
T6	Elektronische Rechnungen folgen den Standards des Forum Datenaustausch: generalInvoiceRequest und generalInvoiceResponse 5.0. Version 4.5 nimmt LimmatVita nur noch für Storni und Mahnungen zu Altfällen an.	Branchenstandard (Forum Datenaustausch).
T7	Stationäre SwissDRG-Rechnungen mit medizinischem Datensatz laufen über die zertifizierte Datenannahmestelle nach Art. 59a KVV.	Nicht Teil von Release 1. LEA erhält daraus später nur die freigegebenen Angaben.
T8	AI-Dienste werden nur genutzt, wenn die Verarbeitung in der Schweiz erfolgt und der Anbieter die Daten nicht zum Training verwendet.	Folgt aus R3. Schränkt die Modellauswahl ein (siehe ADR-004, Risiko RK1).

2.2 Organisatorische Randbedingungen

ID	Randbedingung	Hintergrund und Folge
O1	Drei Teams entwickeln LEA: «Eingang und Dokumente», «Prüfung und Regeln», «Abrechnung und Cockpit». Das Plattformteam stellt Cluster, Pipelines und Observability.	Bausteingrenzen folgen den Teamgrenzen.
O2	Die Regelpflege liegt beim Tarifteam im Fachbereich, nicht in der IT.	Regeln sind Daten mit eigenem Freigabeprozess.
O3	Software-Releases alle zwei Wochen, Regeldaten bei Bedarf täglich.	Regeldaten und Code haben getrennte Lieferwege.
O4	Vier-Augen-Prinzip für manuelle Zahlungsfreigaben über CHF 5'000 und für jede Regelfreigabe.	Internes Kontrollsystem.
O5	Schattenbetrieb parallel zum Altsystem, bis die Abweichungsquote vier Wochen in Folge unter 0.5 % liegt.	Abnahmekriterium für die Umschaltung je Rechnungstyp.
O6	Architekturdokumentation, ADRs und BPMN-Modelle liegen im Repository und werden per Pull Request geändert.	Docs as Code; dieses Dokument wird daraus erzeugt.

2.3 Regulatorische Randbedingungen

ID	Randbedingung	Folge für die Architektur
R1	KVG und KVV: Leistungen müssen wirksam, zweckmässig und wirtschaftlich sein. Der Versicherer prüft die Rechnungen darauf.	Prüfkern mit Tarif- und Auffälligkeitsregeln; begründete Rückweisungen.
R2	Der vertrauensärztliche Dienst gibt nur jene medizinischen Angaben weiter, die für den Leistungsentscheid nötig sind (Art. 57 KVG).	Abgeschirmter Bereich mit eigenen Rollen und eigener Datenablage.

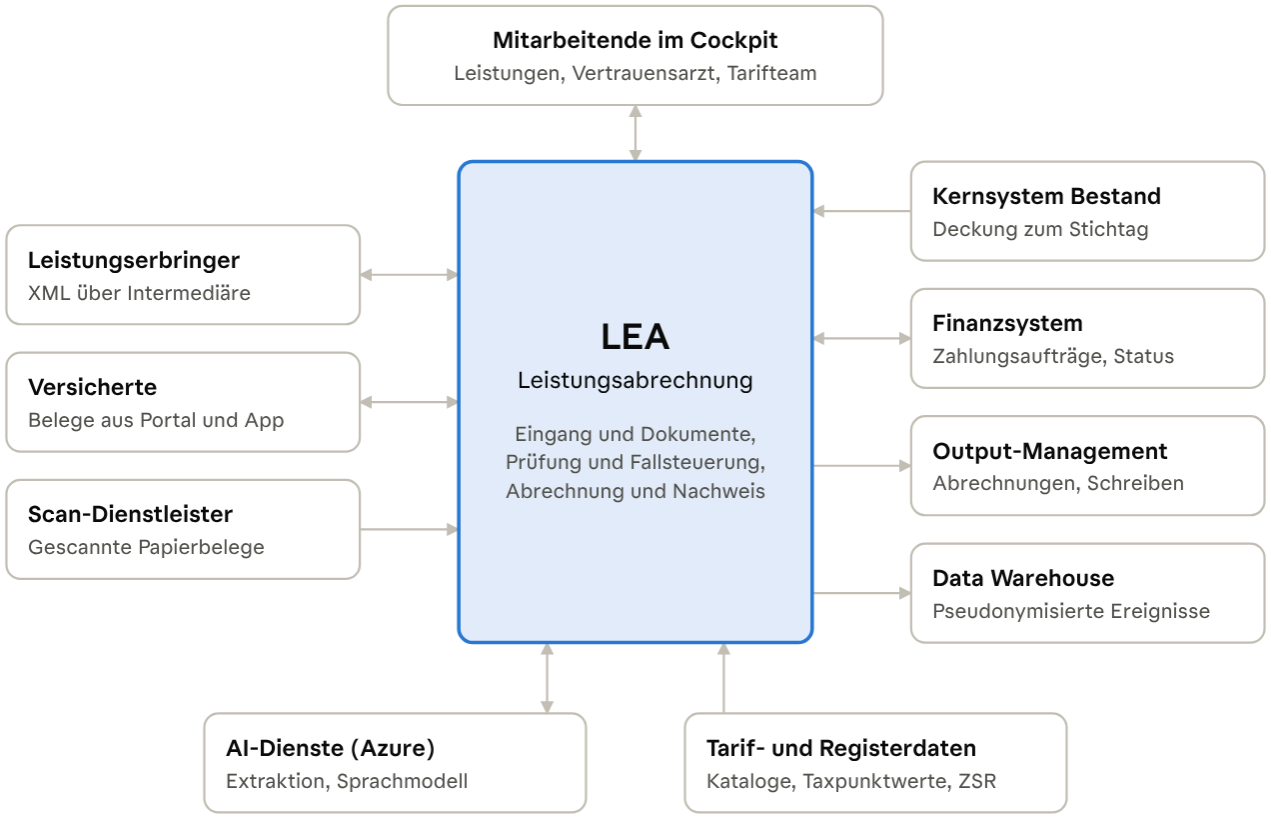
ID	Randbedingung	Folge für die Architektur
R3	Datenschutzgesetz (DSG): Gesundheitsdaten sind besonders schützenswert. Im OKP-Bereich handelt der Versicherer als Bundesorgan, in der Zusatzversicherung als Privater.	Datenhaltung Schweiz, Datenschutz-Folgenabschätzung, Mandantentrennung OKP und Zusatz.
R4	Automatisierte Einzelentscheide (Art. 21 DSG): Die betroffene Person wird informiert und kann eine Überprüfung durch einen Menschen verlangen. Bundesorgane kennzeichnen solche Entscheide.	Kennzeichnung auf der Leistungsabrechnung, Einsprachekanal, der einen manuellen Fall eröffnet.
R5	Aufsicht durch das BAG (OKP) und die FINMA (Zusatzversicherung). Das FINMA-Rundschreiben 2018/3 regelt Auslagerungen; LimmatVita wendet es gruppenweit an.	Prüfrechte, dokumentierte Exit-Strategie, portable Technologiewahl.
R6	FINMA-Aufsichtsmitteilung 08/2024 zu Künstlicher Intelligenz: Inventar, Tests, laufende Überwachung, Dokumentation, Erklärbarkeit, unabhängige Überprüfung.	AI-Governance in Kapitel 8.4.
R7	Belege und Entscheide werden 10 Jahre unveränderbar aufbewahrt (Art. 958f OR).	Archiv mit Schreibschutz (WORM), getrennt von den operativen Datenbanken.

3 Kontextabgrenzung

LEA verantwortet die Rechnung vom Eingang bis zum Zahlungsauftrag und zur Leistungsabrechnung. Vertragsdaten, Zahlungsverkehr und Druck bleiben in den Nachbarsystemen.

3.1 Fachlicher Kontext

LEA verbindet drei Eingangskanäle mit vier Nachbarsystemen



Fachlicher Kontext · 3 Eingangskanäle, 4 Nachbarsysteme

Pfeile zeigen die Richtung des Datenflusses. Doppelpfeile stehen für Rückmeldungen wie Rechnungsantwort oder Zahlungsstatus.

Nachbar	Liefert an LEA	Erhält von LEA
Leistungserbringer (Praxen, Spitalambulatorien, Therapeuten), angebunden über Intermediäre	Elektronische Rechnungen, Storni, Antworten auf Rückfragen	Rechnungsantwort: akzeptiert, Rückfrage oder begründete Rückweisung
Versicherte, über Kundenportal und App	Rückforderungsbelege als PDF oder Foto, Einsprachen	Status des Belegs, Leistungsabrechnung
Scan-Dienstleister	Gescannte Papierbelege mit Metadaten	Quittung je Lieferung

Nachbar	Liefert an LEA	Erhält von LEA
Kernsystem Bestand	Versicherte Person, Deckungen und Produkte zum Behandlungsdatum	Nichts; LEA liest nur
Tarif- und Registerdaten	Tarifkataloge (TARDOC, ambulante Pauschalen), Taxpunktwerte, Leistungserbringer-Register (ZSR, GLN)	Nichts
Finanzsystem	Zahlungsstatus, Rückläufer	Zahlungsaufträge an Leistungserbringer und Versicherte
Output-Management	Versandstatus	Aufträge für Leistungsabrechnungen und Schreiben
Sachbearbeitung, vertrauensärztlicher Dienst, Tarifteam	Entscheide, Korrekturen, Regeländerungen	Arbeitskörbe, Prüfhinweise, Testergebnisse
AI-Dienste der Cloud-Plattform	Extrahierte Felder, Textvorschläge	Belegbilder, minimierte Textausschnitte
Data Warehouse	Nichts	Pseudonymisierte Ereignisse und Kennzahlen
Altsystem Leistungen (nur im Schattenbetrieb)	Dessen Entscheid je Rechnung	Nichts

3.2 Technischer Kontext

Partner	Kanal	Format	Hinweise
Intermediäre	HTTPS mit gegenseitiger TLS-Authentisierung, Abholung alle 2 Minuten	generalInvoiceRequest und generalInvoiceResponse 5.0 (XML)	Schema-Validierung am Eingang; Quittung erst nach dauerhafter Ablage

Partner	Kanal	Format	Hinweise
Kundenportal und App	REST, OAuth 2.0	PDF, JPEG, HEIC bis 20 MB	Virenprüfung vor Ablage; Upload ist idempotent über eine Beleg-ID des Portals
Scan-Dienstleister	SFTP, vier Lieferungen pro Arbeitstag	PDF/A mit JSON-Begleitdatei	Stapelweise Quittung, fehlende Seiten werden zurückgemeldet
Kernsystem Bestand	REST über ExpressRoute, synchron	JSON	Zeitlimit 800 ms; Antworten 15 Minuten gepuffert; stichtagsbezogene Abfrage
Tarifkataloge, Taxpunktwerte	Dateiimport durch das Tarifteam	Kataloge der Tariforganisation, CSV	Import erzeugt einen neuen Regelstand im Entwurf, nie direkt produktiv
Register ZSR und GLN	Täglicher Abgleich	CSV	Historisiert; Prüfung gegen den Stand am Behandlungsdatum
Finanzsystem	Kafka-Ereignis, Adapter ruft REST-API des Finanzsystems	JSON	Zahlungslauf und Bankanbindung liegen im Finanzsystem
Output-Management	REST, asynchrone Rückmeldung	JSON mit Dokumentdaten	Vorlagen liegen im Output-Management
Azure AI Document Intelligence	REST über privaten Endpunkt	Belegbild hin, JSON zurück	Nur Switzerland North
Azure OpenAI	REST über privaten Endpunkt	JSON	Regionale Bereitstellung; nur minimierte Texte ohne Namen und Versichertennummer
Data Warehouse	Kafka-Topics	Avro	Pseudonymisierung vor der Veröffentlichung

Partner	Kanal	Format	Hinweise
Microsoft Entra ID	OpenID Connect	Tokens	Anmeldung der Mitarbeitenden, Workload Identity für Dienste

4 Lösungsstrategie

LEA trennt, was sich unterschiedlich schnell ändert: Regeln sind versionierte Daten, Abläufe sind BPMN-Modelle, AI-Dienste sitzen hinter austauschbaren Schnittstellen. Der Entscheid selbst fällt immer in deterministischem Code oder durch einen Menschen.

4.1 Vom Qualitätsziel zum Ansatz

Qualitätsziel	Ansatz	Vertieft in
1 Nachvollziehbarkeit	Der Prüfkern ist eine reine Funktion: Rechnung, Kontext und Regelstand hinein, Prüfergebnis heraus. Jeder Entscheid speichert Eingaben, Regelstand-ID und Modellversionen.	5.2, 8.2, ADR-002
1 Nachvollziehbarkeit	AI liefert Daten und Hinweise, nie den Entscheid. Kürzungen und Rückweisungen stammen aus Regeln oder von Menschen.	8.4, ADR-008
2 Vertraulichkeit	Alle Daten und alle Verarbeitung in Switzerland North, private Endpunkte, eigene Schlüssel. AI-Dienste erhalten nur, was sie für die Aufgabe brauchen.	7, 8.3
3 Automatisierungsgrad	Zwei Eingangspfade münden in ein kanonisches Rechnungsmodell. Konfidenzschwellen je Feld bestimmen, was dunkel läuft.	5.3, 6.2
3 Automatisierungsgrad	Die Fallsteuerung behandelt Dunkelverarbeitung als Normalpfad. Menschen kommen nur über definierte Aufgaben ins Spiel.	6.1, ADR-001

Qualitätsziel	Ansatz	Vertieft in
4 Änderbarkeit	Regeln sind zeitlich versionierte Daten, die das Tarifteam pflegt, gegen historische Fälle testet und freigibt.	8.2
4 Änderbarkeit	Acht fachlich geschnittene Services mit hexagonalem Aufbau. Nachbarsysteme sind über Adapter entkoppelt.	5.1, ADR-007
5 Robustheit	Asynchrone Kopplung über Kafka mit Transactional Outbox und idempotenten Verbrauchern. Fällt ein AI-Dienst aus, gehen Dokumentbelege in die manuelle Erfassung.	6.3, 8.6, ADR-005, ADR-006

4.2 Technologieentscheide im Überblick

Bereich	Wahl	Kurzbegründung
Backend	Java, Spring Boot	Unternehmensstandard (T3)
Prozesssteuerung	Flowable mit BPMN, DMN für Routingregeln	Vorgabe T4; Aufgaben, Fristen und Eskalation ohne Eigenbau
Messaging	Apache Kafka, betrieben mit Strimzi auf AKS	Portabel, volle Kafka-Semantik (ADR-005)
Operative Daten	Azure Database for PostgreSQL, eine Datenbank je Service	Standard T3; klare Datenhoheit je Baustein
Belege und Archiv	Azure Blob Storage mit Unveränderbarkeitsrichtlinie	Aufbewahrung R7
Dokumentextraktion	Azure AI Document Intelligence mit eigenen Modellen je Belegtyp	Strukturierte Extraktion mit Konfidenzwerten (ADR-003)
Sprachmodell	Azure OpenAI als regionale Bereitstellung, hinter eigenem Port	Nur für Normalisierung und Assistenz (ADR-004)
Oberfläche	React-Anwendung mit Backend-for-Frontend	Ein Cockpit für drei Nutzergruppen
Plattform	AKS, Argo CD, Kustomize, Terraform	Vorgaben T1 und T2

Bereich	Wahl	Kurzbegründung
Observability	OpenTelemetry, Prometheus, Loki, Tempo, Grafana	Plattformstandard; fachliche und technische Kennzahlen in einem Werkzeug
Identität	Microsoft Entra ID, Workload Identity	Keine Passwörter und Schlüssel in Konfigurationen

5 Bausteinsicht

LEA besteht aus acht Services, gekoppelt über Kafka-Ereignisse. Die Fallsteuerung führt den Ablauf, der Prüfkern beurteilt, alle anderen liefern zu oder führen aus.

5.1 Ebene 1: Gesamtsystem

Die Fallsteuerung führt den Ablauf, Kafka koppelt die Services



Bausteinsicht Ebene 1 · 8 Services, 1 Ereignisbus

Services oberhalb von Kafka erzeugen und prüfen den Fall, jene unterhalb führen den Entscheid aus. Nicht gezeichnet sind die Assistenz-API für Begründungsentwürfe (cockpit an docai) und die Bestand-API, die `rules` und `settlement` über `integration` aufrufen.

Die Zerlegung folgt fachlichen Fähigkeiten und der Änderungsgeschwindigkeit: Regeln ändern sich wöchentlich, Extraktionsmodelle monatlich, Abrechnungslogik jährlich. Jeder Service gehört genau einem Team.

Baustein	Verantwortung	Team
Eingang (`intake`)	Nimmt XML und Dokumente an, validiert, prüft auf Viren und Duplikate, legt das Original ab. Überführt XML direkt in das kanonische Rechnungsmodell.	Eingang und Dokumente
Dokumentenverständnis (`docai`)	Klassifiziert Dokumentbelege, extrahiert Felder, normalisiert sie in das kanonische Rechnungsmodell und liefert eine Konfidenz je Feld. Kapselt alle AI-Dienste.	Eingang und Dokumente
Fallsteuerung (`case`)	Führt je Rechnung einen Fall als BPMN-Prozess. Stösst Prüfung und Abrechnung an, erzeugt Aufgaben, überwacht Fristen.	Prüfung und Regeln
Prüfkern (`rules`)	Prüft eine Rechnung gegen einen Regelstand und liefert Befunde mit Entscheidvorschlag. Verwaltet Regelstände und Referenzdaten.	Prüfung und Regeln
Abrechnung (`settlement`)	Berechnet die Kostenbeteiligung, führt das Kostenbeteiligungskonto, erzeugt Leistungsabrechnung, Zahlungsauftrag und Rechnungsantwort.	Abrechnung und Cockpit
Cockpit (`cockpit`)	Weboberfläche für Sachbearbeitung, vertrauensärztlichen Dienst und Tarifteam. Ein Backend-for-Frontend bündelt die Zugriffe.	Abrechnung und Cockpit
Integration (`integration`)	Adapter zu Bestand, Finanzsystem, Output-Management und Data Warehouse. Übersetzt deren Modelle in jene von LEA.	Abrechnung und Cockpit
Audit und Archiv (`audit`)	Schreibt die Entscheidhistorie unveränderbar fort, archiviert Belege, beantwortet Nachweisabfragen.	Abrechnung und Cockpit

Wichtige Schnittstellen:

Schnittstelle	Art	Von → an	Inhalt
`BelegEingegangen`	Ereignis	intake → docai, audit	Beleg-ID, Kanal, Verweis auf das Original
`RechnungBereit`	Ereignis	intake oder docai → case	Rechnung im kanonischen Modell, bei Dokumenten mit Konfidenz je Feld
Prüf-API	REST, synchron	case → rules	Prüfauftrag hinein, Prüfergebnis heraus; idempotent über die Fall-ID
`FallEntschieden`	Ereignis	case → settlement, audit	Entscheid, Befunde, Regelstand-ID, Entscheidweg (dunkel oder manuell)
`AbrechnungErstellt`, `ZahlungsauftragErstellt`	Ereignis	settlement → integration, audit	Beträge, Empfänger, Verweis auf den Fall
Bestand-API	REST, gepuffert	rules, settlement → integration	Person und Deckungen zum Stichtag
Aufgaben- und Fall-API	REST	cockpit → case	Arbeitskörbe, Aufgaben übernehmen und abschliessen
Regelstand-API	REST	cockpit → rules	Regelstände bearbeiten, testen, freigeben
Assistenz-API	REST	cockpit → docai	Entwurf eines Begründungstexts zu einem Befund

5.2 Ebene 2: Prüfkern

Der Prüfkern ist hexagonal gebaut. Im Inneren liegt ein funktionaler Kern ohne Ein- und Ausgabe, ohne Uhr und ohne Zufall. Dieselben Eingaben ergeben immer dasselbe Prüfergebnis.

Komponente	Aufgabe
Prüf-API	Nimmt den Prüfauftrag an, lädt den Kontext, ruft den Kern, speichert das Ergebnis.
Kontextlader	Holt Deckung (über Integration), Stammdaten des Leistungserbringers und die Leistungshistorie der Person für Mengen- und Kumulationsregeln.
Funktionaler Kern	`pruefe(rechnung, kontext, regelstand)` durchläuft vier Stufen: formal, Deckung, Tarif, Auffälligkeit. Ergebnis: Befunde mit Regel-ID, Position, Schweregrad, Begründung und Betragswirkung sowie ein Entscheidvorschlag.
Regelstand-Verwaltung	Hält Regelstände mit Gültigkeit und Status (Entwurf, getestet, freigegeben). Importiert Tarifkataloge und Register.
Regel-Testbank	Lässt einen Entwurf gegen Referenzfälle und gegen die Rechnungen der letzten 90 Tage laufen und zeigt jede Abweichung zum gültigen Regelstand.
Ergebnisspeicher	Legt jedes Prüfergebnis mit Hash der Eingaben und Regelstand-ID ab.

Regeltyp	Beispiel	Darstellung
Katalogregel	Tarifposition am Behandlungsdatum gültig, Taxpunkte korrekt	Daten aus dem Tarifkatalog
Kombinationsregel	Position nicht kumulierbar mit einer anderen; Mengenlimit je Sitzung oder Zeitraum	Regeltabellen
Deckungsregel	Produkt deckt die Leistungsart; Leistungserbringer am Behandlungsdatum zugelassen	Regeltabellen plus Deckung aus Bestand
Auffälligkeitsregel	Gleiche Leistung auf zwei Rechnungen; Menge weit über dem Üblichen für die Fachrichtung	Parametrisierte Regeln mit Schwellenwerten

Regeltyp	Beispiel	Darstellung
Routingregel	Ab welchem Betrag, welcher Konfidenz oder Befundkombination ein Mensch prüft	DMN-Tabelle in der Fallsteuerung, nicht im Prüfkern

5.3 Ebene 2: Dokumentenverständnis

Das Dokumentenverständnis ist eine Pipeline aus sechs Schritten. Die Schritte 2 bis 4 rufen AI-Dienste auf, alle über denselben Port.

1. Vorverarbeitung: Seiten trennen und drehen, Bildqualität prüfen. Unlesbare Belege gehen mit einem Hinweis an die Versicherten zurück.
2. Klassifikation: Belegtyp bestimmen, zum Beispiel Rückforderungsbeleg einer Arztpraxis, Therapierechnung, Apothekenquittung oder kein Leistungsbeleg.
3. Extraktion: Azure AI Document Intelligence liest Kopf- und Positionsdaten mit einem eigenen Modell je Belegtyp.
4. Normalisierung: Leistungserbringer und Tarifpositionen werden gegen die Referenzdaten abgeglichen. Bei unklaren Positionstexten schlägt das Sprachmodell eine Katalogposition vor; übernommen wird nur, was im Katalog existiert.
5. Konfidenzbewertung: Jedes Feld erhält eine Konfidenz, kalibriert an den Korrekturen der Sachbearbeitung. Für kritische Felder (Person, Leistungserbringer, Datum, Tarifposition, Betrag) gelten strengere Schwellen.
6. Übergabe: `RechnungBereit` mit kanonischem Modell, Konfidenzen und Modellversionen.

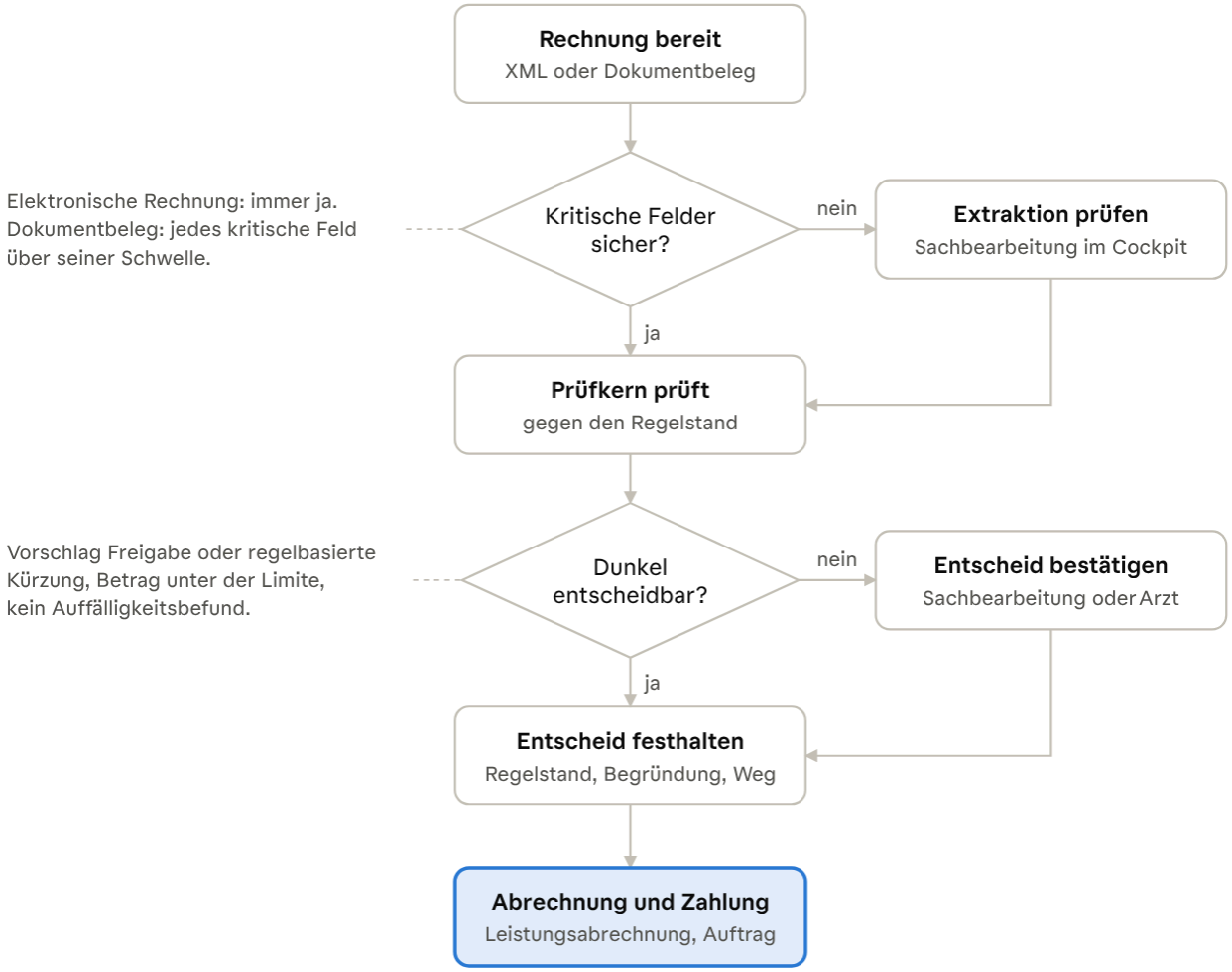
Der AI-Port protokolliert jeden Aufruf mit Modellversion und Dauer, begrenzt Parallelität und öffnet bei Störungen einen Circuit Breaker (siehe 6.3).

6 Laufzeitsicht

Vier Szenarien zeigen das Zusammenspiel: der dunkle Normalfall, der Dokumentbeleg mit menschlicher Prüfung, der Ausfall eines AI-Dienstes und die Freigabe einer Regeländerung.

6.1 Elektronische Rechnung, dunkel verarbeitet

Menschen prüfen nur, wo Daten unsicher oder Entscheide belastend sind



Ablauf Rechnungsprüfung · 6 Schritte, 2 Entscheidpunkte

Der gerade Weg nach unten ist die Dunkelverarbeitung. Die beiden Abzweigungen nach rechts sind die einzigen Stellen, an denen ein Mensch eingreift. Die Schritte im Einzelnen:

1. `intake` holt die Rechnung beim Intermediär ab, validiert sie gegen das XML-Schema und prüft auf Duplikate.
2. `intake` speichert Original und kanonisches Modell und schreibt `RechnungBereit` in seine Outbox, alles in einer Transaktion. Erst danach quittiert es dem Intermediär.
3. `case` startet den Prozess «Rechnungsprüfung» und ruft die Prüf-API auf.
4. `rules` lädt Deckung, Stammdaten und Leistungshistorie, wählt den Regelstand zum Behandlungsdatum, führt den Kern aus und speichert das Prüfergebnis.
5. Die Routingtabelle in `case` wertet das Ergebnis aus. Dunkel entschieden wird, wenn der Vorschlag «Freigabe» oder «regelbasierte Kürzung» lautet, der Betrag unter CHF 20'000 liegt und kein Auffälligkeitsbefund vorliegt.

6. `case` veröffentlicht `FallEntschieden`. `settlement` berechnet die Kostenbeteiligung, bucht auf dem Kostenbeteiligungskonto und erzeugt Leistungsabrechnung, Zahlungsauftrag und Rechnungsantwort.
7. `integration` übergibt Zahlungsauftrag und Dokumentauftrag an die Nachbarsysteme. `audit` schreibt die Entscheidhistorie fort.

Im Normalbetrieb dauert der Durchlauf unter 10 Sekunden. Das Ziel von 5 Minuten lässt Raum für Lastspitzen und Wiederholversuche. 2 % der dunkel entschiedenen Fälle gehen nachgelagert in eine Stichprobenkontrolle, ohne die Zahlung aufzuhalten.

6.2 Dokumentbeleg mit menschlicher Prüfung

1. Eine versicherte Person lädt in der App das Foto eines Rückforderungsbelegs hoch. `intake` prüft auf Viren, legt das Original ab und veröffentlicht `BelegEingegangen`.
2. `docai` durchläuft die Pipeline aus 5.3. Bei zwei Rechnungszeilen liegt die Konfidenz für die Tarifposition unter der Schwelle.
3. `docai` veröffentlicht `RechnungBereit` mit den Konfidenzen. `case` startet den Fall und erzeugt die Aufgabe «Extraktion prüfen».
4. Die Sachbearbeiterin sieht im Cockpit Beleg und extrahierte Felder nebeneinander, unsichere Felder sind markiert. Sie korrigiert eine Position und bestätigt.
5. `case` ruft die Prüf-API mit den bestätigten Daten auf. Der Kern meldet den Befund «Mengenlimit überschritten» und schlägt eine Kürzung vor.
6. Der Entscheid belastet die versicherte Person und beruht auf Dokumentdaten. Deshalb erzeugt `case` die Aufgabe «Entscheid bestätigen». Das Cockpit schlägt einen Begründungstext vor, die Sachbearbeiterin passt ihn an und bestätigt.
7. Weiter wie in 6.1 ab Schritt 6. Die Korrektur aus Schritt 4 geht in die Auswertung der Extraktionsqualität ein (8.4).

Dokumentbelege laufen nur dunkel, wenn alle kritischen Felder über der Schwelle liegen und der Prüfkern «Freigabe» vorschlägt.

6.3 Ausfall eines AI-Dienstes

1. Aufrufe an Document Intelligence laufen in Zeitüberschreitungen. Nach fünf Fehlern innert 30 Sekunden öffnet der Circuit Breaker im AI-Port.
2. `docai` pausiert seinen Verbraucher. Die Belege bleiben im Topic, keiner geht verloren. Die XML-Verarbeitung läuft unverändert weiter.
3. Alle 60 Sekunden prüft ein Testaufruf, ob der Dienst wieder antwortet.
4. Überschreitet der Rückstau 2′000 Belege oder 30 Minuten, geht ein Alarm an die Bereitschaft.

5. Ab vier Stunden Rückstau kann die Betriebsleitung Leistungen im Cockpit die manuelle Erfassung einschalten. `case` erzeugt dann Erfassungsaufgaben mit Beleg und leerer Maske.
6. Antwortet der Dienst wieder, schliesst der Circuit Breaker schrittweise. `docai` arbeitet den Rückstau mit begrenzter Parallelität ab.

Nachrichten, die sich dreimal nicht verarbeiten lassen, landen mit Fehlergrund in einem Dead-Letter-Topic. Die Betriebsansicht im Cockpit zeigt sie an und erlaubt die erneute Einspeisung.

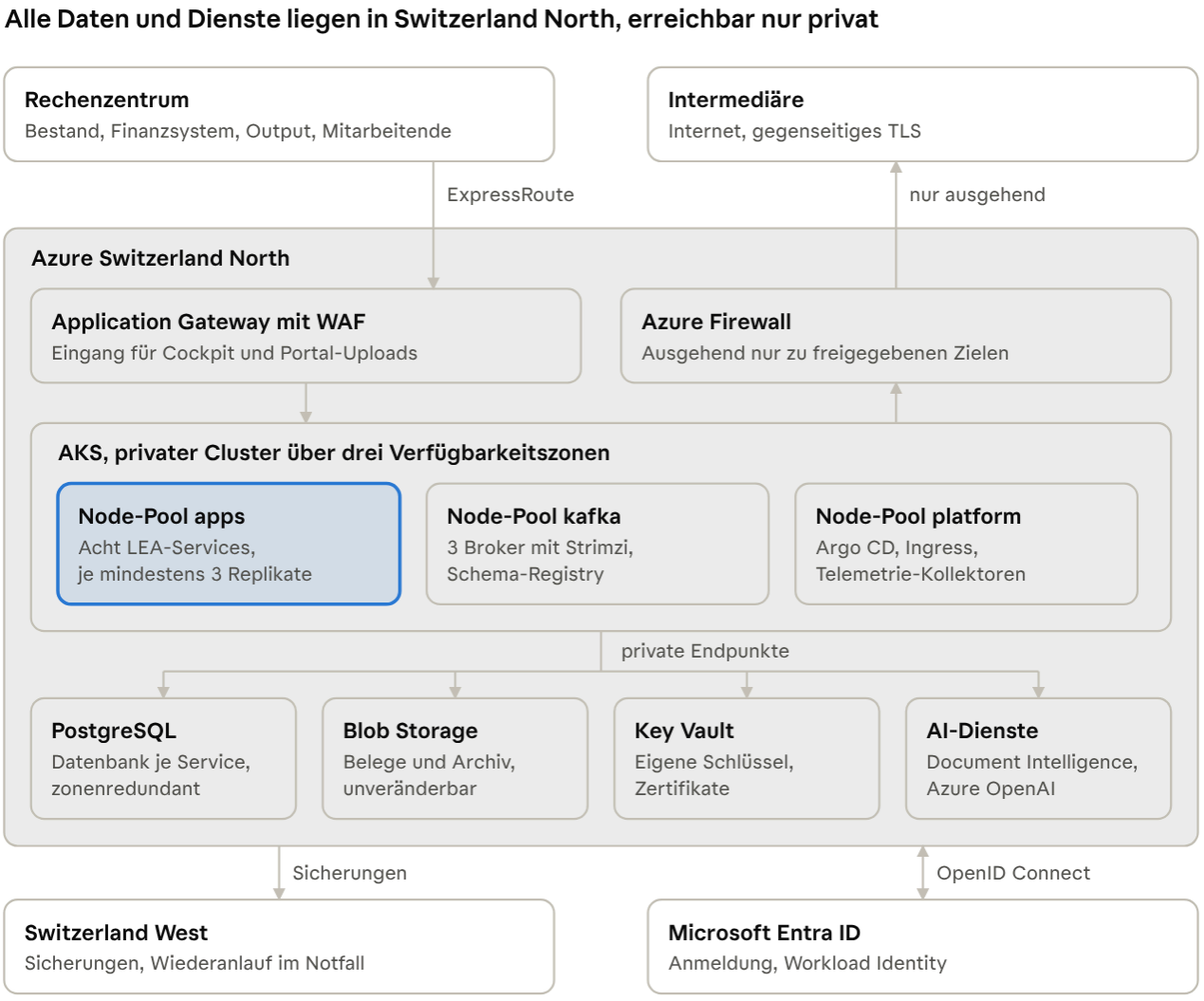
6.4 Regeländerung freigeben

1. Ein Tarifspezialist importiert einen neuen Katalog oder ändert eine Regel. Es entsteht ein Regelstand im Entwurf mit Gültigkeitsdatum.
2. Die Regel-Testbank lässt den Entwurf gegen die Referenzfälle und die Rechnungen der letzten 90 Tage laufen. Der Bericht zeigt jede Abweichung, zum Beispiel «412 Rechnungen würden neu gekürzt».
3. Eine zweite Person prüft den Bericht und gibt frei (Vier-Augen-Prinzip, O4).
4. Der freigegebene Regelstand ist unveränderbar und wird ab seinem Gültigkeitsdatum angewendet. Ein Software-Deployment ist nicht nötig.
5. Korrekturen erzeugen immer einen neuen Regelstand. Bereits gefällte Entscheide verweisen weiter auf den Stand, mit dem sie gefällt wurden.

7 Verteilungssicht

LEA läuft in einem privaten AKS-Cluster in Switzerland North, verteilt über drei Verfügbarkeitszonen. Alle Plattformdienste sind nur über private Endpunkte erreichbar.

7.1 Produktionsumgebung



Verteilungssicht Produktion · Region Switzerland North

Der einzige Weg hinein führt aus dem Unternehmensnetz über das Application Gateway, der einzige Weg hinaus über die Firewall zu den Intermediären.

Knoten oder Dienst	Was dort läuft	Auslegung
AKS, Node-Pool «apps»	Die acht LEA-Services	Mindestens drei Replikate je Service, verteilt über die Zonen. `docai` und `rules` skalieren nach Kafka-Rückstau auf bis zu zwölf.
AKS, Node-Pool «kafka»	Kafka mit Strimzi, Schema-Registry	Drei Broker, einer je Zone; Replikationsfaktor 3, mindestens zwei synchrone Replikate

Knoten oder Dienst	Was dort läuft	Auslegung
AKS, Node-Pool «platform»	Argo CD, Ingress, Telemetrie-Kollektoren	Betrieben vom Plattformteam
Azure Database for PostgreSQL	Eine Datenbank je Service mit eigenem Login, verteilt auf drei Server	Zonenredundante Hochverfügbarkeit, Wiederherstellung auf einen Zeitpunkt
Azure Blob Storage	Belegspeicher für Originale; Archiv mit Unveränderbarkeitsrichtlinie über 10 Jahre	Zonenredundant; Sicherungskopie in Switzerland West
Azure Key Vault	Eigene Schlüssel für Datenbanken und Speicher, Zertifikate	Zugriff nur über Workload Identity
Azure AI Document Intelligence, Azure OpenAI	Extraktion und Sprachmodell	Private Endpunkte, öffentlicher Zugriff abgeschaltet
Application Gateway mit Web Application Firewall	Eingang für Cockpit und Portal-Uploads aus dem Unternehmensnetz	Kein direkter Zugang aus dem Internet
Azure Firewall	Ausgehender Verkehr	Nur freigegebene Ziele, etwa die Intermediäre
ExpressRoute	Verbindung ins Rechenzentrum zu Bestand, Finanzsystem und Output-Management	Redundant ausgelegt

7.2 Umgebungen

Umgebung	Zweck	Daten
dev	Entwicklung, ein Namespace je Team	Synthetisch
test	Automatisierte Integrations- und Vertragstests	Synthetisch
abnahme	Fachtests, Lasttests, Vergleichsläufe der Regel-Testbank	Anonymisierte Rechnungsdaten, synthetische Belegbilder
prod	Produktion inklusive Schattenbetrieb	Echte Daten

Produktion hat einen eigenen Cluster in einer eigenen Subscription. Die drei übrigen Umgebungen teilen sich einen Cluster.

7.3 Verfügbarkeit und Notfall

Ereignis	Verhalten	Ziel
Ausfall einer Zone	Die verbleibenden zwei Zonen tragen die Last ohne Eingriff.	Keine Unterbrechung
Ausfall der Region	Wiederanlauf in Switzerland West aus Infrastruktur-Code und Sicherungen. Kafka wird neu aufgebaut; die Outbox-Tabellen halten Ereignisse 14 Tage und befüllen die Topics neu.	Wiederanlauf innert 8 Stunden, Datenverlust höchstens 1 Stunde
Ausfall der AI-Dienste	Siehe 6.3	XML-Verarbeitung unberührt

Im Notfallbetrieb in Switzerland West ist die Verfügbarkeit der AI-Dienste nicht zugesichert. Dokumentbelege werden dann manuell erfasst oder zurückgestellt (Risiko RK6).

7.4 Auslieferung

1. Ein Merge in den Hauptzweig startet die Pipeline: bauen, testen, Code und Abhängigkeiten prüfen, Container-Image scannen, Stückliste (SBOM) erzeugen, Image signieren.
2. Die Pipeline trägt die neue Image-Version im GitOps-Repository ein, im Kustomize-Overlay der Zielumgebung.
3. Argo CD gleicht den Cluster mit dem Repository ab. Eine Zulassungsrichtlinie lässt nur signierte Images zu.
4. Der Rollout erfolgt stufenweise, zuerst 10 % der Instanzen. Steigt die Fehlerquote, wird automatisch zurückgerollt.
5. Infrastrukturänderungen (Terraform) haben einen eigenen Lieferweg mit Plan-Review durch das Plattformteam.

8 Querschnittliche Konzepte

Acht Konzepte gelten in allen Bausteinen. Die ersten fünf tragen die Qualitätsziele 1 und 2 und sind deshalb ausführlicher beschrieben.

8.1 Fachliches Modell

Begriff	Bedeutung	Führender Baustein
Beleg	Das eingegangene Original, XML oder Dokument, unverändert	intake
Rechnung	Inhalt des Belegs im kanonischen Modell: Kopf (Person, Leistungserbringer, Behandlungszeitraum, Vergütungsart, Gesetz) und Positionen (Tarif, Tarifposition, Menge, Datum, Betrag)	intake, docai
Fall	Bearbeitungsvorgang zu einer Rechnung mit Status, Aufgaben und Entscheid	case
Regelstand	Unveränderbare Version aller Regeln und Kataloge mit Gültigkeitszeitraum	rules
Prüfergebnis	Befunde und Entscheidvorschlag zu einer Rechnung unter einem Regelstand	rules
Entscheid	Freigabe, Kürzung oder Rückweisung mit Entscheidweg und Begründung	case
Abrechnung	Vergüteter Betrag, Kostenbeteiligung, Auszahlung und Empfänger	settlement

Das kanonische Rechnungsmodell liegt als versioniertes Avro-Schema in der Schema-Registry. Änderungen müssen abwärtskompatibel sein.

8.2 Regelversionierung und Reproduzierbarkeit

Jeder Entscheid lässt sich neu berechnen, weil alle Eingaben des Prüfkerns unveränderbar gespeichert sind.

- Ein Regelstand hat eine ID, einen fachlichen Gültigkeitszeitraum und einen Freigabezeitpunkt. Er wird nie geändert, nur abgelöst.
- Für jede Rechnungsposition gilt der Regelstand, der am Behandlungsdatum gültig und zum Prüfzeitpunkt freigegeben war.
- Der Entscheid speichert den Hash der Rechnung, einen Schnappschuss des Kontexts (Deckung, Auszug der Leistungshistorie), die Regelstand-ID und die Version des Prüfkerns. Bei Dokumentbelegen kommen Modellversionen und Konfidenzen dazu.
- `audit` rechnet jede Nacht 1′000 zufällig gewählte Entscheide nach. Jede Abweichung löst einen Alarm aus.

- Ändert sich der Code des Kerns, muss das Ergebnis für bestehende Regelstände gleich bleiben. Das sichern rund 12'000 Referenzfälle mit erwartetem Prüfergebnis.

8.3 Sicherheit und Datenschutz

Thema	Umsetzung
Datenhaltung	Alle Daten liegen in Switzerland North, Sicherungen in Switzerland West. Kein Dienst verarbeitet Daten ausserhalb der Schweiz.
Verschlüsselung	TLS auf allen Verbindungen, gegenseitige Authentisierung zwischen den Services über das Service Mesh. Ruhende Daten sind mit eigenen Schlüsseln aus dem Key Vault verschlüsselt.
Identität	Mitarbeitende melden sich über Entra ID mit Mehrfaktor-Authentisierung an. Dienste nutzen Workload Identity; es gibt keine statischen Geheimnisse.
Berechtigung	Rollen (Sachbearbeitung, Teamleitung, Vertrauensarzt, Tarifpflege, Betrieb, Revision) kombiniert mit Attributen (Mandant, Team). Geprüft im Backend-for-Frontend und nochmals im Service.
Abschirmung des vertrauensärztlichen Dienstes	Medizinische Zusatzunterlagen liegen in einem eigenen Speicherbereich mit eigenem Schlüssel. Die Sachbearbeitung sieht nur die Beurteilung der Leistungspflicht.
Mandantentrennung	Jeder Datensatz trägt das Kennzeichen OKP oder Zusatzversicherung. Zugriffe und Auswertungen sind danach getrennt.
Datenminimierung gegenüber AI-Diensten	Document Intelligence erhält das Belegbild. Das Sprachmodell erhält nur Positionstexte und Katalogkandidaten, keine Namen, Adressen oder Versichertennummern.
Zugriffsprotokoll	Jeder lesende Zugriff auf Falldaten wird mit Person, Fallbezug und Zeitpunkt protokolliert. Technische Logs enthalten nur IDs, keine Personendaten.

Thema	Umsetzung
Aufbewahrung und Löschung	10 Jahre im Archiv (R7). Operative Datenbanken werden 24 Monate nach Fallabschluss ausgedünnt; Löschläufe sind protokolliert.
Datenschutz-Folgenabschätzung	Liegt vor und wird bei jeder neuen AI-Verwendung nachgeführt.

8.4 AI-Governance

Jede AI-Verwendung ist inventarisiert, hat einen messbaren Qualitätsmassstab und einen Rückfallweg. Die Gliederung folgt den Erwartungen der FINMA-Aufsichtsmitteilung 08/2024 (R6).

Verwendung	Technik	Wirkung auf den Entscheid	Risikoklasse	Rückfall
Belegklassifikation	Document Intelligence, eigenes Klassifikationsmodell	Bestimmt das Extraktionsmodell	Mittel	Manuelle Zuordnung
Feldextraktion	Document Intelligence, eigene Extraktionsmodelle	Liefert die Rechnungsdaten; unter der Schwelle prüft ein Mensch	Hoch	Manuelle Erfassung
Normalisierung von Positionen	Sprachmodell	Schlägt eine Katalogposition vor; der Katalog validiert	Hoch	Aufgabe «Extraktion prüfen»
Begründungsentwurf	Sprachmodell	Keine; ein Mensch redigiert und verantwortet den Text	Tief	Textbausteine
Auffälligkeitsregeln	Statistische Schwellenwerte, kein lernendes Modell	Führt nur zu manueller Prüfung	Mittel	Regel abschalten

- Messung: Ein Referenzdatensatz mit 5′000 von Menschen geprüften Belegen, geschichtet nach Belegtyp und Sprache (Deutsch, Französisch, Italienisch). Ziel für dunkel verarbeitete Belege: mindestens 99 % korrekte kritische Felder.
- Freigabe: Ein neues Modell, ein neuer Prompt oder eine neue Schwelle geht erst nach einem Vergleichslauf gegen den Referenzdatensatz in Produktion. Entwicklung und Freigabe liegen bei verschiedenen Personen.

- Überwachung: Korrekturquote je Feld und Belegtyp, Verteilung der Konfidenzen, Anteil dunkel. Weicht die Korrekturquote über sieben Tage um mehr als zwei Prozentpunkte ab, geht ein Alarm an das Team.
- Versionierung: Jedes Ergebnis trägt Modell-, Prompt- und Schwellenversion.
- Dokumentation: Je Verwendung ein Steckbrief im Repository mit Zweck, Daten, Modell, Kennzahlen, Grenzen, Tests und Begründung der Risikoklasse.
- Unabhängige Überprüfung: Das Risikomanagement prüft Verwendungen der Klasse «Hoch» jährlich.
- Schutz des Sprachmodells: Belegtexte gelten als nicht vertrauenswürdige Eingabe. Das Modell hat keine Werkzeuge, antwortet in einem festen JSON-Schema, und nur existierende Katalogpositionen werden übernommen.

8.5 Menschen im Entscheidweg

Situation	Wer entscheidet	Kennzeichnung
Elektronische Rechnung ohne Befund	System	Automatisiert
Elektronische Rechnung mit eindeutigem Regelbefund, etwa ungültige Tarifposition	System; die Rechnungsantwort nennt Regel und Begründung	Automatisiert, Überprüfung auf Verlangen
Dokumentbeleg, alle kritischen Felder sicher, Vorschlag «Freigabe»	System	Automatisiert
Dokumentbeleg mit Kürzung oder Rückweisung	Mensch	Manuell
Auffälligkeitsbefund oder Betrag ab CHF 20′000	Mensch; ab CHF 5′000 Auszahlung zusätzlich eine zweite Person	Manuell
Medizinische Frage	Vertrauensärztlicher Dienst	Manuell

Automatisierte Entscheide sind auf der Leistungsabrechnung gekennzeichnet (R4). Verlangt eine Person die Überprüfung, eröffnet LEA einen manuellen Fall mit Verweis auf den ursprünglichen Entscheid.

8.6 Ereignisverarbeitung und Konsistenz

- Transactional Outbox: Zustandsänderung und Ereignis werden in derselben Datenbanktransaktion geschrieben. Ein Relay veröffentlicht aus der Outbox nach Kafka.
- Zustellung mindestens einmal. Verbraucher sind idempotent und erkennen Wiederholungen an der Ereignis-ID.
- Reihenfolge: Partitionierung nach Fall-ID.
- Schemas: Avro in der Schema-Registry, nur abwärtskompatible Änderungen. Die Pipeline prüft die Verträge.
- Fehler: drei Wiederholungen mit wachsendem Abstand, danach Dead-Letter-Topic.
- Mehrstufige Abläufe koordiniert die Fallsteuerung. Kompensationen wie das Storno einer Abrechnung sind als BPMN-Pfad modelliert.
- Synchrone Aufrufe (Prüf-API, Bestand-API) haben Zeitlimit, Wiederholung, Circuit Breaker und begrenzte Parallelität.

8.7 Observability

- Eine Trace-ID begleitet die Rechnung vom Eingang bis zur Abrechnung. Die Fall-ID ist Attribut in Traces und Logs.
- Fachliche Kennzahlen: Dunkelverarbeitungsquote je Kanal, Durchlaufzeit, Rückstau je Arbeitskorb, Korrekturquote der Extraktion, Befunde je Regel.
- Technische Kennzahlen: Kafka-Rückstau, Fehlerquoten, Antwortzeiten der Nachbarsysteme, Zustand der Circuit Breaker.
- Alarmiert wird auf den Verbrauch des Fehlerbudgets der Service-Level-Ziele, nicht auf Einzelmetriken.

8.8 Test und Qualitätssicherung

Ebene	Gegenstand	Zeitpunkt
Unit- und Property-Tests	Funktionaler Kern, Abrechnungslogik	Jeder Commit
Golden-Master-Tests	Rund 12′000 Referenzfälle mit erwartetem Prüfergebnis	Jeder Commit, jede Regeländerung
Vertragstests	Ereignis-Schemas und REST-Verträge	Pipeline
Prozesstests	BPMN-Pfade inklusive Fristen und Eskalation	Pipeline

Ebene	Gegenstand	Zeitpunkt
AI-Evaluation	Referenzdatensatz mit 5'000 Belegen	Jede Änderung an Modell, Prompt oder Schwelle
Last- und Störungstests	20 Rechnungen pro Sekunde über 30 Minuten; Zonenausfall; AI-Ausfall	Monatlich in der Abnahmeumgebung
Schattenbetrieb	Entscheid von LEA gegen Entscheid des Altsystems, je Rechnung	Laufend bis zur Umschaltung

9 Architekturentscheidungen

Neun Entscheide prägen LEA. Alle haben den Status «akzeptiert». Die vollständigen ADRs mit Diskussion liegen im Repository.

ADR-001: Die Fallsteuerung orchestriert mit BPMN

- Kontext: Die Rechnungsprüfung kennt menschliche Aufgaben, Fristen, Eskalationen und Kompensationen. Reine Choreografie über Ereignisse würde diese Ablauflogik über alle Services verstreuen.
- Entscheid: Ein BPMN-Prozess je Fall in `case`, mit eingebetteter Flowable-Engine. Die übrigen Services bleiben über Ereignisse entkoppelt.
- Verworfen: Choreografie nur mit Kafka; eine selbst gebaute Zustandsmaschine.
- Konsequenzen: Der Ablauf ist sichtbar und für den Fachbereich lesbar, Aufgabenverwaltung ist enthalten. Dafür wird `case` zur zentralen Komponente, deren Durchsatz laufend zu testen ist (RK5). Laufende Prozessinstanzen brauchen Migrationsregeln bei Modelländerungen.

ADR-002: Der Prüfkern ist ein funktionaler Kern, Regeln sind Daten

- Kontext: Regeln ändern sich häufig, und jeder Entscheid muss reproduzierbar sein. Die Genehmigung des Tarifsystems ist bis Ende 2028 befristet; weitere Anpassungen sind absehbar (BAG).
- Entscheid: Ein eigener Kern in Java aus reinen Funktionen. Regeln liegen als typisierte, versionierte Tabellen vor. Der Kern kennt keine Ein- und Ausgabe.
- Verworfen: Eine Regel-Engine mit Regeln als Skripten, weil das Tarifteam sie nicht selbst pflegen kann und die Reproduktion über Versionen aufwendig wird. DMN für alle Regeln, weil Mengen- und Zeitraumregeln über die Leistungshistorie darin unhandlich sind.

- Konsequenzen: Deterministisch, schnell und mit Referenzfällen testbar. Neue Regeltypen brauchen weiterhin Code. Architekturtests verhindern, dass Ein- und Ausgabe in den Kern wandert.

ADR-003: Dokumentextraktion mit Document Intelligence, Sprachmodell nur zur Normalisierung

- Kontext: 18 % der Belege sind Dokumente, rund 1.4 Mio. pro Jahr. Layouts variieren, und die Fallsteuerung braucht eine Konfidenz je Feld.
- Entscheid: Azure AI Document Intelligence mit eigenen Modellen je Belegtyp. Das Sprachmodell ordnet nur unklare Positionstexte einer Katalogposition zu. Beides liegt hinter dem AI-Port.
- Verworfen: Ein Sprachmodell mit Bildeingabe für die gesamte Extraktion, weil es keine kalibrierten Feldkonfidenzen liefert und jeder Modellwechsel die ganze Extraktion betrifft. Selbst betriebene Texterkennung mit eigenem Modell, weil dem Team dafür die Erfahrung fehlt.
- Konsequenzen: Konfidenz je Feld und planbare Kosten je Beleg. Jeder Belegtyp braucht gelabelte Trainingsbelege. Die Bindung an Azure wird durch den Port und den Referenzdatensatz gemildert.

ADR-004: Sprachmodell als regionale Bereitstellung in der Schweiz, mit Ausstiegsweg

- Kontext: T8 verlangt Verarbeitung in der Schweiz. Azure OpenAI ist seit September 2023 in Switzerland North verfügbar (Microsoft). Nur regionale Bereitstellungen verarbeiten in der Region. Ob Nachfolgemodelle dort regional bereitstehen, sagt Microsoft nicht zu (Microsoft Q&A, Mai 2026).
- Entscheid: Regionale Bereitstellung in Switzerland North. Die Verwendungen sind so geschnitten, dass ein kleines Modell genügt und ein Ausfall verkraftbar ist. Ein Modellwechsel braucht die AI-Evaluation aus 8.4. Als Ausstiegsweg wird halbjährlich ein offenes Modell auf GPU-Knoten in AKS gegen den Referenzdatensatz geprüft.
- Verworfen: Bereitstellung in einer EU-Datenzone, weil die Verarbeitung die Schweiz verlässt. Ausschliesslich selbst betriebenes Modell, weil GPU-Betrieb und Modellpflege vom ersten Tag an anfallen würden.
- Konsequenzen: T8 ist erfüllt. Die verfügbare Modellgeneration hinkt nach, und Abkündigungen können einen Wechsel erzwingen (RK1).

ADR-005: Kafka mit Strimzi auf AKS statt eines verwalteten Dienstes

- Kontext: Ereignisse koppeln alle Services. R5 verlangt eine dokumentierte Exit-Strategie.

- Entscheid: Apache Kafka, betrieben mit dem Strimzi-Operator in einem eigenen Node-Pool.
- Verworfen: Azure Event Hubs mit Kafka-Protokoll, weil es an Azure bindet und sich nicht deckungsgleich mit Apache Kafka verhält.
- Konsequenzen: Portabel, und dieselbe Technik läuft in allen Umgebungen bis zum Entwicklerrechner. Betrieb und Upgrades liegen beim Plattformteam (TS4).

ADR-006: Transactional Outbox statt verteilter Transaktionen

- Kontext: Der Zustand in PostgreSQL und das Ereignis in Kafka müssen zusammenpassen.
- Entscheid: Eine Outbox-Tabelle je Service, ein Relay veröffentlicht, Verbraucher sind idempotent.
- Verworfen: Veröffentlichen direkt nach dem Commit, weil bei einem Absturz Ereignisse verloren gehen. Durchgängiges Event Sourcing, weil der Umstellungsaufwand für die Teams den Nutzen übersteigt.
- Konsequenzen: Kein Ereignis geht verloren, und Topics lassen sich aus der Outbox neu befüllen (7.3). Idempotenz ist in jedem Verbraucher Pflicht.

ADR-007: Acht fachlich geschnittene Services

- Kontext: Drei Teams mit 21 Personen. `docai` und `rules` haben ein anderes Lastprofil und einen anderen Änderungsrhythmus als der Rest.
- Entscheid: Acht Services, zwei bis vier je Team, jeder intern hexagonal aufgebaut. Ein neuer Service entsteht nur bei eigenem Skalierungs- oder Teamgrund.
- Verworfen: Ein modularer Monolith, weil die AI-Abhängigkeit isoliert und die Prüfung getrennt skalierbar sein soll. Feingranulare Microservices je Prüfstufe, weil sie Betriebsaufwand und verteilte Fehlerbilder ohne Gegenwert bringen.
- Konsequenzen: Jedes Team liefert unabhängig. Dafür gelten die Regeln verteilter Systeme aus 8.6 und 8.7.

ADR-008: AI fällt keine belastenden Entscheide

- Kontext: Qualitätsziel 1 und R4. Extraktionsfehler lassen sich nicht ausschliessen.
- Entscheid: Kürzungen und Rückweisungen entstehen nur aus einer deterministischen Regel auf strukturiert gelieferten Daten oder nach Bestätigung durch einen Menschen. Bei Dokumentbelegen ist jeder belastende Entscheid zu bestätigen.
- Verworfen: Dunkle Kürzung bei hoher Konfidenz. Sie brächte geschätzt drei Prozentpunkte mehr Automatisierung, aber Fehlkürzungen träfen Versicherte direkt.

- Konsequenzen: Ein AI-Fehler führt zu Mehrarbeit oder im schlechtesten Fall zu einer Überzahlung, nie zu einer ungerechtfertigten Ablehnung. Das Überzahlungsrisiko begrenzen Schwellen und die Stichprobe. Der Automatisierungsgrad bei Dokumenten bleibt bei rund 60 %.

ADR-009: Deckungsdaten synchron aus Bestand lesen, nicht replizieren

- Kontext: Nach T5 bietet Bestand eine REST-API, aber keine Ereignisse. Die Deckung muss stichtagsgenau sein.
- Entscheid: Synchroner Aufruf über `integration` mit 15 Minuten Puffer. Antwortet Bestand nicht, wartet der Fall und versucht es erneut; es gibt keinen Entscheid auf veralteten Daten.
- Verworfen: Eine nächtliche Replik in LEA, weil bis zu 24 Stunden alte Deckungen bei Mutationen zu Fehlentscheiden führen.
- Konsequenzen: Die Deckung ist immer aktuell. Verfügbarkeit und Antwortzeit von Bestand begrenzen den Durchsatz von LEA (RK4, TS2).

10 Qualitätsanforderungen

Zwölf Szenarien machen die fünf Qualitätsziele aus 1.2 prüfbar. Jedes hat eine Messgrösse; gemessen wird in Produktion oder in der Abnahmeumgebung.

10.1 Qualitätsbaum

Qualitätsziel	Teilmerkmal	Szenarien
1 Nachvollziehbarkeit und Korrektheit	Reproduzierbarkeit, Begründbarkeit	Q1, Q2
1 Nachvollziehbarkeit und Korrektheit	Fachliche Richtigkeit, Extraktionsgüte	Q3, Q4
2 Vertraulichkeit	Datenhaltung, Zugriffsschutz	Q5, Q6
3 Automatisierungsgrad	Dunkelverarbeitung, Durchlaufzeit, Durchsatz	Q7, Q8
4 Änderbarkeit	Regeln, Modelle	Q9, Q10
5 Robustheit	Fehlertoleranz, Wiederherstellbarkeit	Q11, Q12

10.2 Qualitätsszenarien

ID	Auslöser	Erwartete Reaktion	Messgrösse
Q1	Die Revision verlangt die Herleitung eines Entscheids aus dem Vorjahr.	`audit` liefert Eingaben, Regelstand, Befunde und Entscheidweg und rechnet den Entscheid nach.	Antwort in unter 1 Minute; Nachrechnung identisch
Q2	Ein Leistungserbringer erhält eine Rückweisung.	Die Rechnungsantwort nennt je Befund Regel, Position und Begründung.	100 % der Rückweisungen mit Regel-ID und Begründungstext
Q3	Die Stichprobenkontrolle bewertet 2 % der dunkel entschiedenen Fälle nach.	Die Entscheide halten der Nachprüfung stand.	Fehlentscheide unter 0.5 %
Q4	Dokumentbelege werden dunkel verarbeitet.	Die kritischen Felder stimmen mit dem Beleg überein.	Mindestens 99 % korrekt, gemessen am Referenzdatensatz und an der Stichprobe
Q5	Ein Team will einen Dienst ausserhalb der Schweiz einbinden.	Die Plattformrichtlinie verweigert das Anlegen der Ressource.	Keine Ressource ausserhalb der zugelassenen Regionen; täglicher Bericht
Q6	Eine Sachbearbeiterin öffnet einen Fall ausserhalb ihres Mandanten.	Der Zugriff wird verweigert und protokolliert.	100 % verweigert; monatliche Auswertung an den CISO
Q7	Normalbetrieb über einen Kalendermonat.	LEA entscheidet ohne manuellen Eingriff.	Mindestens 85 % über alle Kanäle, 90 % elektronisch, 60 % Dokumente
Q8	Januar-Spitze mit 60′000 Belegen am Tag und Lieferungen von bis zu 5′000 Rechnungen auf einmal.	`docai` und `rules` skalieren nach Rückstau.	95 % der dunklen Fälle in unter 5 Minuten; Lasttest mit 20 Rechnungen pro Sekunde über 30 Minuten ohne wachsenden Rückstau

ID	Auslöser	Erwartete Reaktion	Messgrösse
Q9	Die Tariforganisation publiziert eine neue Katalogversion.	Das Tarifteam importiert, testet und gibt frei.	Produktiv in höchstens 10 Arbeitstagen, ohne Software-Deployment
Q10	Das eingesetzte Sprachmodell wird abgekündigt.	Das Team stellt hinter dem AI-Port um und durchläuft die AI-Evaluation.	Umstellung in höchstens 15 Arbeitstagen; Änderungen nur in `docai`
Q11	Document Intelligence fällt zwei Stunden aus.	Dokumentbelege stauen sich, XML läuft weiter, der Rückstau wird danach abgebaut.	Kein verlorener Beleg; XML-Durchlaufzeit unverändert; Rückstau innert 4 Stunden abgebaut
Q12	Eine Verfügbarkeitszone fällt aus.	Die übrigen Zonen übernehmen ohne Eingriff.	Unterbrechung unter 60 Sekunden, kein Datenverlust. Bei Regionsausfall: Wiederanlauf innert 8 Stunden, Datenverlust höchstens 1 Stunde, jährlich geübt

11 Risiken und technische Schulden

Das grösste Risiko für das Automatisierungsziel ist die Extraktionsgüte bei Dokumentbelegen (RK2), das grösste für die Korrektheit ein fehlerhafter Regelstand (RK3).

11.1 Risiken

ID	Risiko	Eintritt	Auswirkung	Massnahmen
RK1	Das genutzte Sprachmodell wird abgekündigt, bevor ein Nachfolger in Switzerland North regional bereitsteht.	Mittel	Mittel	AI-Port, AI-Evaluation, halbjährlich geprüftes Ausstiegsmodell (ADR-004). Alle Verwendungen des Sprachmodells dürfen ausfallen.

ID	Risiko	Eintritt	Auswirkung	Massnahmen
RK2	Die Extraktionsgüte bleibt unter dem Ziel, vor allem bei Handyfotos, handschriftlichen Ergänzungen und französischen oder italienischen Belegen. Die Dunkelquote für Dokumente bleibt unter 60 %.	Hoch	Mittel	Schwellen je Belegtyp, Nachtraining aus Korrekturen, Prüfung der Bildqualität schon in der App, stufenweise Zielwerte.
RK3	Ein fehlerhafter Regelstand kürzt oder bezahlt tausende Rechnungen falsch.	Tief	Hoch	Regel-Testbank mit Abweichungsbericht, Vier-Augen-Freigabe, Ablösung durch neuen Regelstand, vorbereiteter Prozess für Massenkorrekturen.
RK4	Antwortzeit oder Verfügbarkeit von Bestand begrenzt den Durchsatz in der Januar-Spitze.	Mittel	Mittel	Puffer, gemeinsame Lasttests mit dem Bestand-Team, Drosselung in `integration` (ADR-009).
RK5	Die Flowable-Engine wird bei 60′000 Prozessstarts am Tag zum Engpass.	Tief	Hoch	Monatliche Lasttests, kurzlebige Prozessinstanzen, Historie asynchron auslagern.
RK6	In der Notfallregion Switzerland West sind Kapazität und AI-Dienste nicht zugesichert.	Tief	Mittel	Manuelle Erfassung als Rückfall, jährliche Klärung der Kapazität mit Microsoft, jährliche Wiederanlaufübung.
RK7	Sachbearbeitende bestätigen Vorschläge des Systems ungeprüft.	Mittel	Mittel	Unsichere Felder markieren, Kontrollfälle einstreuen, Stichprobe auch auf manuell bestätigte Fälle, Schulung.
RK8	Ein Beleg enthält Text, der das Sprachmodell manipulieren soll.	Tief	Tief	Keine Werkzeuge, festes Antwortschema, Validierung gegen den Katalog (8.4).

ID	Risiko	Eintritt	Auswirkung	Massnahmen
RK9	Das Tarifsystem ändert sich stärker als erwartet; seine Genehmigung ist bis Ende 2028 befristet.	Hoch	Mittel	Regeln als Daten (ADR-002), Szenario Q9.

11.2 Technische Schulden

ID	Schuld	Folge	Abbau
TS1	Altsystem und LEA laufen parallel. Das Kostenbeteiligungskonto wird im Schattenbetrieb doppelt geführt.	Abgleichaufwand, Abweichungen	Umschaltung je Rechnungstyp bis Ende 2027, danach Stilllegung des Altsystems
TS2	Bestand liefert keine Ereignisse (ADR-009).	Synchrone Kopplung, Durchsatz hängt an Bestand	Mit der Ablösung von Bestand ab 2029 auf Ereignisse umstellen
TS3	Für Altfälle wird XML 4.5 weiter angenommen.	Zwei Übersetzer in `intake`	Entfernen, sobald keine Storni im alten Format mehr eintreffen
TS4	Das Kafka-Wissen liegt bei zwei Personen im Plattformteam (ADR-005).	Betriebsrisiko bei Ausfällen und Upgrades	Dritte Person ausbilden, Betriebshandbücher, halbjährliche Neubewertung verwalteter Angebote
TS5	Die Regel-Testbank vergleicht Entscheide, aber nicht die Beträge der Kostenbeteiligung.	Lücke bei Änderungen an der Abrechnung	Erweiterung auf `settlement` im 1. Quartal 2027
TS6	Auffälligkeitsregeln sind einfache Schwellenwerte.	Geringe Trefferquote, viele unnötige manuelle Prüfungen	Bewusst so für Release 1; lernendes Modell wird nach zwölf Monaten Betriebsdaten evaluiert

12 Glossar

Begriff	Bedeutung
AI-Port	Schnittstelle in `docai`, hinter der alle AI-Dienste liegen. Macht Modelle austauschbar und protokolliert jeden Aufruf.
Ambulante Pauschalen	Tarifstruktur, die ambulante Behandlungen pauschal vergütet. Gilt seit dem 1. Januar 2026 zusammen mit TARDOC.
Circuit Breaker	Schutzschalter, der Aufrufe an einen gestörten Dienst vorübergehend unterbindet.
Datenannahmestelle	Zertifizierte Stelle des Versicherers, die bei DRG-Rechnungen als einzige die medizinischen Datensätze entgegennimmt (Art. 59a KVV).
Dead-Letter-Topic	Kafka-Topic für Nachrichten, die sich wiederholt nicht verarbeiten liessen.
Dunkelverarbeitung	Verarbeitung vom Eingang bis zum Entscheid ohne manuellen Eingriff.
Fall	Bearbeitungsvorgang zu genau einer Rechnung.
Forum Datenaustausch	Verein, der die XML-Standards für den administrativen Datenaustausch im Schweizer Gesundheitswesen publiziert.
generalInvoiceRequest, generalInvoiceResponse	XML-Meldungen des Forum Datenaustausch für Rechnung und Rechnungsantwort, aktuell Version 5.0.
GLN	Global Location Number. Identifiziert Leistungserbringer und Versicherer im Datenaustausch.
Golden-Master-Test	Test, der ein gespeichertes erwartetes Ergebnis mit dem aktuellen vergleicht.
Hexagonale Architektur	Aufbau, bei dem die Fachlogik im Kern liegt und Technik über Ports und Adapter angebunden ist.

Begriff	Bedeutung
Intermediär	Dienstleister, der elektronische Rechnungen zwischen Leistungserbringern und Versicherern übermittelt.
Kostenbeteiligung	Anteil, den Versicherte selbst tragen: Franchise und Selbstbehalt.
KVG, KVV	Bundesgesetz und Verordnung über die Krankenversicherung.
Leistungsabrechnung	Dokument an die versicherte Person mit vergütetem Betrag und Kostenbeteiligung.
OAAT	Organisation ambulante Arzttarife AG. Tariforganisation, die TARDOC und die ambulanten Pauschalen pflegt.
OKP	Obligatorische Krankenpflegeversicherung nach KVG.
Outbox	Tabelle, in die ein Service Ereignisse in derselben Transaktion wie seine Zustandsänderung schreibt.
Referenzdatensatz	5′000 von Menschen geprüfte Belege, an denen jede AI-Änderung gemessen wird.
Regelstand	Unveränderbare Version aller Regeln und Kataloge mit Gültigkeitszeitraum.
Rückforderungsbeleg	Rechnung, die Versicherte im Tiers garant zur Rückerstattung einreichen.
Schattenbetrieb	LEA verarbeitet produktive Rechnungen parallel zum Altsystem; wirksam ist nur dessen Entscheid.
TARDOC	Einzelleistungstarif für ambulante ärztliche Leistungen, seit dem 1. Januar 2026 in Kraft.
Taxpunktwert	Frankenbetrag je Taxpunkt, je Kanton vereinbart oder festgesetzt.
Tiers garant	Die versicherte Person bezahlt die Rechnung und fordert den Betrag beim Versicherer zurück.

Begriff	Bedeutung
Tiers payant	Der Versicherer bezahlt den Leistungserbringer direkt.
Vertrauensärztlicher Dienst	Ärztinnen und Ärzte, die den Versicherer in medizinischen Fragen der Leistungspflicht beraten.
VVG	Versicherungsvertragsgesetz. Grundlage der Zusatzversicherungen.
Workload Identity	Verfahren, mit dem sich ein Dienst im Cluster ohne gespeichertes Geheimnis gegenüber Azure ausweist.
WZW	Wirksamkeit, Zweckmässigkeit, Wirtschaftlichkeit: Voraussetzungen für die Leistungspflicht nach KVG.
ZSR	Zahlstellenregister der SASIS. Die ZSR-Nummer identifiziert Leistungserbringer gegenüber den Versicherern.

Quellen

Die realen Rahmenbedingungen des Fallbeispiels stützen sich auf diese Seiten, abgerufen am 16. November 2026:

- BAG: Ambulanter Arzttarif – TARDOC und ambulante Pauschalen seit 1. Januar 2026, Befristung bis Ende 2028, OAAT
- Forum Datenaustausch: XML-Standard Rechnung – generalInvoiceRequest und generalInvoiceResponse 5.0
- FMH: Rechtlicher Leitfaden, Rechnungsstellung KVG – Tiers garant und Tiers payant, Datenannahmestelle
- Onlinekommentar zu Art. 21 DSG – automatisierte Einzelentscheide, Kennzeichnung durch Bundesorgane
- Pro Infirmis: Rechtsratgeber Datenschutz – Krankenversicherer als Bundesorgane (OKP) und als Private (Zusatzversicherung)
- FINMA: Rundschreiben 2018/3 Outsourcing – Geltung für Versicherungsunternehmen
- FINMA: Aufsichtsmitteilung 08/2024 – Governance und Risikomanagement beim Einsatz von KI

- Microsoft: Azure OpenAI Service in Switzerland North – Verfügbarkeit seit 18. September 2023, Datenhaltung, kein Training mit Kundendaten
- Microsoft Q&A: Nachfolge für GPT-4o-Standard-Bereitstellungen in Switzerland North – keine Zusage für regionale Nachfolgemodelle (Mai 2026)
- arc42: Template-Übersicht – die zwölf Kapitel