PRAXISBERICHT · ÖFFENTLICHE FASSUNG · JULI 2026

Production Confidence Review

Agentic Fitness Coach

Eine Webanwendung verbindet personalisierte Trainingspläne, absolvierte Einheiten, Gesundheitsinformationen und Mahlzeiten in einem gemeinsamen Produkt. Ein KI-Agent nutzt diese Informationen, um Fortschritte zu bewerten und spätere Trainings- und Ernährungsempfehlungen anzupassen.

Die Anwendung wurde zunächst als Einzelnutzer-Prototyp entwickelt und wird inzwischen von ausgewählten Testnutzern täglich eingesetzt. Mit der Erweiterung zur Mehrbenutzer-Anwendung stiegen die Anforderungen an Datentrennung, Betriebsstabilität und nachvollziehbares Agentenverhalten.

Kann die Anwendung für einen kontrollierten, einladungsbasierten Pilotbetrieb mit ersten zahlenden Nutzern freigegeben werden?

REPORT NAVIGATION

Inhalt

> Der Bericht trennt Management-Entscheidung, priorisierte Maßnahmen und technische Evidenz. Die Bewertung gilt ausschließlich für den beschriebenen Pilotbetrieb.

SECTION

Entscheidung auf einen Blick

Bereit mit Bedingungen

Entscheidungsgegenstand: Kontrollierter, einladungsbasierter Pilotbetrieb mit ersten zahlenden Nutzern

Die Anwendung kann für einen begrenzten und betreuten Pilotbetrieb freigegeben werden.

Der zentrale Nutzerweg ist nutzbar, wesentliche Grenzen zwischen Nutzern sind technisch angelegt und das wichtigste bekannte Fehlverhalten des Agenten wird vor Releases reproduzierbar geprüft.

Ein öffentlicher Self-Service-Rollout oder eine größere Skalierung ist noch nicht freigegeben.

JETZT FREIGEGEBEN

- Einladungsbasierter Zugang für eine klar begrenzte Pilotgruppe
- Nutzung des geprüften zentralen Nutzerwegs
- Einsatz einer festgelegten Anwendungs- und Agentenversion
- Persönliche Betreuung und direkte Reaktion auf gemeldete Probleme
- Manuelle Kontrolle kritischer Betriebsabläufe während der Pilotphase

NOCH NICHT FREIGEGEBEN

- Öffentliche Registrierung ohne Zugangskontrolle
- Unbegleiteter Rollout an eine größere Nutzergruppe
- Häufige Releases ohne verpflichtende technische Prüfungen
- Ungeprüfte Änderungen an Prompts, Modellen oder Agentenlogik
- Skalierung ohne Kostenmessung pro Nutzer und Ablauf
- Betrieb ohne aktive Überwachung kritischer Hintergrundprozesse

SECTION

Warum diese Entscheidung vertretbar ist

1. Der zentrale Nutzerweg funktioniert

Nutzer können:

1. Ziele und relevanten Kontext hinterlegen
2. personalisierte Trainings- und Ernährungsempfehlungen erhalten
3. Trainings, Mahlzeiten und Gesundheitsinformationen erfassen
4. ihren Fortschritt bewerten lassen
5. spätere Empfehlungen auf Basis des Verlaufs anpassen lassen

Dieser Ablauf ist vorhanden und wird im bestehenden Testbetrieb regelmäßig genutzt.

2. Das wichtigste bekannte Agentenrisiko wurde kontrollierbarer gemacht

Im Testbetrieb wurde eine mittelfristige Zielsetzung vom Agenten unzureichend interpretiert. Der daraus entstandene Trainingsplan war technisch gültig und plausibel formuliert, entwickelte Belastung und Progression jedoch nicht angemessen über den geplanten Zeitraum.

Dieser Fehlerfall sowie weitere repräsentative Szenarien wurden in ein wiederholbares Evaluationsverfahren aufgenommen.

Kritische Fehlinterpretationen können dadurch vor der Freigabe einer neuen Agentenversion erkannt werden.

3. Die verbleibenden Risiken sind im begrenzten Pilotbetrieb beherrschbar

Monitoring, Release-Absicherung, Isolationstests und Kostenkontrolle sind noch nicht vollständig ausgebaut.

Innerhalb einer kleinen und bekannten Pilotgruppe können diese Risiken vorübergehend begrenzt werden durch:

- kontrollierte Benutzeranlage
- feste Software- und Agentenversionen
- direkte Betreuung
- manuelle Betriebsprüfungen
- klar benannte Verantwortlichkeiten
- begrenzte Änderungshäufigkeit

Diese Übergangskontrollen reichen nicht für einen breiten Rollout. Für einen begrenzten Pilotbetrieb sind sie vertretbar.

SECTION

Was jetzt Aufmerksamkeit braucht

1. Kritische Hintergrundfehler aktiv erkennen

PRIORITÄT: HOCH

Ein aktivierter Hintergrundprozess schlug im Testbetrieb wiederholt fehl. Die Fehler waren technisch auffindbar, lösten aber keine aktive Benachrichtigung aus.

Geschäftliche Auswirkung

Ein wichtiger Nutzerablauf kann gestört bleiben, obwohl die Anwendung weiterhin erreichbar erscheint. Probleme werden unter Umständen erst durch Nutzerbeschwerden erkannt.

Erforderlich vor

Vergrößerung der Pilotgruppe

Nächster Schritt

- wiederholte Fehler automatisch erkennen
- nach einer definierten Schwelle alarmieren
- betroffenen Nutzer und Ablauf datensparsam identifizieren
- Alarmierung mit einem kontrollierten Fehler testen
- Reaktion und Wiederherstellung dokumentieren

---

2. Releases verpflichtend absichern

PRIORITÄT: HOCH

Änderungen am Hauptzweig werden automatisch ausgerollt. Eine separate produktionsnahe Staging-Umgebung und ein vollständiges verpflichtendes Release-Gate fehlen.

Geschäftliche Auswirkung

Fehler können zuerst im laufenden Pilotsystem sichtbar werden.

Dies betrifft sowohl klassische Softwarefehler als auch Veränderungen durch:

- neue Prompts
- Modellwechsel
- veränderte Agentenwerkzeuge
- neue Datenformate
- angepasste Integrationen

Erforderlich vor

Regelmäßigen Produktänderungen während des Piloten

Nächster Schritt

Vor jedem produktiven Deployment mindestens ausführen:

- Build
- statische Prüfungen
- automatisierte Tests
- Agenten-Evaluationen
- Isolationstests
- Smoke-Test des zentralen Nutzerwegs

---

3. Nutzer- und Datentrennung breiter nachweisen

PRIORITÄT: HOCH

Nutzergebundene Zugriffe, begrenzte Agentenrechte und gezielte Isolationstests sind vorhanden.

Im Review wurde keine tatsächliche Offenlegung fremder Nutzerdaten festgestellt. Nicht alle kritischen Zugriffswege werden jedoch automatisiert mit mehreren Nutzern und Fehlerfällen geprüft.

Geschäftliche Auswirkung

Die Eintrittswahrscheinlichkeit erscheint begrenzt. Die mögliche Auswirkung eines Fehlers wäre dennoch hoch.

Erforderlich vor

Öffentlicher Registrierung

Nächster Schritt

Automatisierte Tests für mindestens folgende Fälle ergänzen:

- Zugriff auf fremde Trainingsdaten
- Zugriff auf fremde Mahlzeiten oder Gesundheitsinformationen
- Agentenausführung im falschen Nutzerkontext
- fehlerhafte Nutzerzuordnung in Hintergrundprozessen
- manipulierte oder fremde Ressourcenkennungen
- direkte Zugriffe über API- und Datenbankgrenzen

---

4. Kosten pro Nutzer und Ablauf messen

PRIORITÄT: MITTEL

Die bisherigen Gesamtkosten liegen im erwarteten Bereich. Eine belastbare Zuordnung zu einzelnen Nutzern, Agentenläufen oder Produktabläufen fehlt.

Geschäftliche Auswirkung

Kostensteigerungen können erst auf der Gesamtrechnung auffallen. Preisgestaltung und Skalierbarkeit bleiben dadurch schwer bewertbar.

Erforderlich vor

Preisvalidierung und Skalierung

Nächster Schritt

Mindestens erfassen:

- Modellkosten pro Nutzer
- Modellkosten pro zentralem Ablauf
- Anzahl und Dauer von Agentenläufen
- fehlgeschlagene und wiederholte Ausführungen
- Infrastrukturkosten pro aktivem Nutzer
- Kostenentwicklung bei wachsender Nutzung

---

5. Wiederherstellung reproduzierbar prüfen

PRIORITÄT: MITTEL

Backups und Rollback-Möglichkeiten sind grundsätzlich vorhanden. Die vollständige Wiederherstellung wurde im Reviewumfang jedoch nicht reproduzierbar nachgewiesen.

Geschäftliche Auswirkung

Ein vorhandenes Backup ist erst belastbar, wenn Wiederherstellung, Dauer und Vollständigkeit tatsächlich geprüft wurden.

Erforderlich vor

Breitem kommerziellen Rollout

Nächster Schritt

- Recovery-Schritte dokumentieren
- Wiederherstellung kontrolliert testen
- maximalen Datenverlust festlegen
- Zielzeit für die Wiederherstellung bestimmen
- verantwortliche Rolle benennen
- Funktionsfähigkeit nach dem Restore prüfen

SECTION

Die eine Verbesserung, die jetzt umgesetzt wurde

Agent Behavior Release Gate

Das wichtigste begrenzbare Produktrisiko war kein klassischer Softwarefehler.

Das größere Risiko bestand darin, dass der Agent technisch gültige und plausibel formulierte Ergebnisse erzeugt, die das fachliche Ziel des Nutzers dennoch verfehlen.

Deshalb wurde ein reproduzierbares Prüfverfahren für zentrale Agentenszenarien umgesetzt.

Warum diese Maßnahme ausgewählt wurde

Eine mittelfristige Zielsetzung war bereits falsch interpretiert worden. Dieser Fehler konnte vor der Umsetzung des Release Gates nur durch manuelle Prüfung oder Nutzerfeedback erkannt werden.

Das neue Verfahren beantwortet vor einem Release mindestens diese Fragen:

- Versteht der Agent kurz- und mittelfristige Ziele korrekt?
- Entwickelt sich ein Trainingsplan nachvollziehbar über die Zeit?
- Werden Belastungsgrenzen und relevante Gesundheitsinformationen berücksichtigt?
- Erkennt der Agent unklare oder widersprüchliche Eingaben?
- Tritt ein bereits bekannter Fehler nach Änderungen erneut auf?

Umgesetzter Umfang

- repräsentative Szenarien für kurz- und mittelfristige Trainingsziele
- Regressionstest für die beobachtete Fehlinterpretation
- Szenarien mit unklaren oder widersprüchlichen Eingaben
- definierte Mindestanforderungen an Progression und Belastungssteuerung
- Kriterien für kritische und nicht kritische Abweichungen
- wiederholbare Ausführung gegen relevante Agentenversionen
- dokumentierte Ergänzung neuer Szenarien
- Integration in den Release-Prozess

Fertig, wenn

Das Umsetzungspaket gilt als abgeschlossen, wenn:

- der bekannte Fehlerfall zuverlässig erkannt wird

- eine kontrolliert fehlerhafte Agentenversion die Prüfung nicht besteht
- die freigegebene Version alle definierten kritischen Szenarien besteht
- die Ergebnisse nachvollziehbar dokumentiert werden
- das Produkt- oder Entwicklungsteam die Prüfung selbst erneut ausführen kann

Übergeben

- getesteter Code-Change
- Evaluationsszenarien
- Bewertungskriterien
- automatisierter Regressionstest
- Ausführungsanleitung
- Anleitung zur Ergänzung neuer Fehlerfälle
- Übersicht der weiterhin nicht abgedeckten Risiken

Vorher

Änderungen an Modell, Prompt oder Agentenlogik konnten ein bekanntes Fehlverhalten erneut einführen, ohne dass dies vor dem Produktiveinsatz zuverlässig erkannt wurde.

Nachher

Bekannte kritische Fehlinterpretationen werden reproduzierbar geprüft. Eine Agentenversion, die den definierten Mindestanforderungen nicht entspricht, kann vor der Freigabe gestoppt werden.

Nicht durch diese Maßnahme gelöst

Das Release Gate beweist nicht, dass jede mögliche Agentenausgabe korrekt ist.

Es ersetzt insbesondere nicht:

- die fachliche Einzelfallprüfung jeder Empfehlung
- medizinische Validierung
- Monitoring des laufenden Betriebs
- Kostenkontrolle
- vollständige Datenisolationstests
- allgemeine End-to-End-Tests
- die Erkennung vollständig unbekannter Fehlerklassen

SECTION

Nächste Freigabestufe

Über einen breiteren Rollout sollte erst entschieden werden, wenn:

- kritische Hintergrundfehler aktive Benachrichtigungen erzeugen
- der zentrale Nutzerweg vor jedem Release geprüft wird
- bekannte kritische Agentenfehler reproduzierbar erkannt werden
- kritische Nutzergrenzen vollständig automatisiert geprüft werden
- Kosten pro Nutzer und Ablauf messbar sind
- Warn- oder Begrenzungsmechanismen für Kosten bestehen
- fehlerhafte Releases gestoppt oder zurückgerollt werden können
- Wiederherstellung und Incident-Reaktion dokumentiert und getestet sind

SECTION

Technische Evidenz

Der folgende Teil dokumentiert die Grundlage der Entscheidung. Er richtet sich primär an technische Verantwortliche, die Befunde, Nachweise und verbleibende Risiken nachvollziehen oder weiterbearbeiten müssen.

SECTION

1. Umfang des Reviews

Bewerteter Einsatz

Bewertet wurde ein kontrollierter, einladungsbasierter Pilotbetrieb mit ersten zahlenden Nutzern.

Das Urteil gilt ausschließlich für:

- eine begrenzte und bekannte Nutzergruppe
- den beschriebenen zentralen Nutzerweg
- die geprüfte Anwendungs- und Agentenversion
- einen betreuten Betrieb
- eine direkte Reaktionsmöglichkeit bei Problemen

Zentraler Nutzerweg

Der Review konzentrierte sich auf folgenden Ablauf:

1. Nutzer hinterlegt Ziele und relevanten Kontext.
2. Die Anwendung erstellt einen Trainings- oder Ernährungsplan.
3. Der Nutzer dokumentiert absolvierte Einheiten, Mahlzeiten und Gesundheitsinformationen.
4. Der Agent bewertet den Verlauf.
5. Spätere Empfehlungen werden auf Basis der erfassten Informationen angepasst.

Nicht jede Produktfunktion wurde vollständig geprüft.

Priorisiert wurden Komponenten, deren Ausfall oder Fehlverhalten den Pilotbetrieb unmittelbar gefährden würde.

Prüffelder

- Produktverhalten und zentraler Nutzerweg
- Architektur und externe Integrationen
- Nutzer- und Datentrennung
- Agentenverhalten und fachliche Fehlerfälle
- Tests und Release-Absicherung
- Deployment und Wiederherstellung
- Monitoring und Fehlererkennung
- Kosten und Skalierung
- Wartbarkeit und technische Übergabe

Nicht Bestandteil des Reviews

Der Review war ausdrücklich kein:

- vollständiger Penetrationstest
- medizinischer Wirksamkeitsnachweis
- regulatorisches Audit
- vollständiger Last- oder Skalierungstest
- Review jeder einzelnen Produktfunktion
- allgemeine Zertifizierung der Produktionsreife
- Ersatz für eine Datenschutz-Folgenabschätzung
- Ersatz für rechtliche oder medizinische Beratung

SECTION

2. Methode und Bewertungsgrundlage

Verwendete Evidenz

EVIDENZART	BEDEUTUNG
Gemessen	Durch automatisierte Prüfungen, Logs oder reproduzierbare Messungen bestätigt
Demonstriert	Im Produkt oder in einer kontrollierten Ausführung nachvollzogen
Dokumentiert	Durch Quellcode, Konfiguration oder technische Dokumentation belegt
Berichtet	Durch Erfahrungen aus dem bisherigen Testbetrieb beschrieben
Unbekannt	Mit dem vorhandenen Zugriff nicht belastbar bestätigt

Aussagen auf Basis dokumentierter oder berichteter Evidenz wurden schwächer gewichtet als reproduzierbare Messungen.

Bewertungsmaßstab

WERTUNG	REIFEGRAD	BEDEUTUNG
0	Unkontrolliert	Wesentliche Kontrollen fehlen oder ein hohes Risiko ist nicht begrenzt
1	Fragil	Der Normalfall funktioniert, wichtige Absicherungen fehlen
2	Kontrolliert	Für den definierten Einsatz unter bekannten Grenzen vertretbar
3	Belegt	Getestet, beobachtbar, dokumentiert und klar verantwortet

Die Bewertungen werden nicht zu einem Durchschnitt verrechnet.

Ein einzelner glaubwürdiger Blocker kann gute Ergebnisse in anderen Bereichen überwiegen. Das finale Urteil basiert deshalb auf Risiko, Auswirkung und verfügbarer Evidenz und nicht auf einer mathematischen Gesamtnote.

Durchgeführte Prüfungen

Zum Review gehörten insbesondere:

- statische Prüfung relevanter Anwendungsteile
- Ausführung vorhandener automatisierter Tests
- Prüfung ausgewählter Nutzer- und Agentenberechtigungen
- Nachvollziehen des zentralen Nutzerwegs
- Analyse des Deployment-Ablaufs
- Auswertung bekannter Fehler aus dem Testbetrieb
- Review vorhandener Monitoring- und Diagnosemöglichkeiten
- Bewertung der verfügbaren Kosteninformationen
- Auswahl und Umsetzung eines begrenzten Risikopunkts

Die vorhandenen automatisierten Prüfungen bestanden zum Zeitpunkt des Reviews. Sie deckten jedoch noch nicht den vollständigen Nutzerweg und nicht alle kritischen Agentenszenarien ab.

SECTION

3. Readiness Scorecard

PRÜFFELD	WERTUNG	ZUSAMMENFASSUNG
Produkt und Nutzerweg	2	Der zentrale Ablauf ist vorhanden und wird regelmäßig genutzt
Architektur und Integration	2	Systemgrenzen und zentrale Verantwortlichkeiten sind erkennbar
Nutzer- und Datentrennung	2*	Wesentliche Kontrollen sind vorhanden, die Evidenz ist noch nicht vollständig
Agentenverhalten	2	Bekannter Fehlerfall und priorisierte Szenarien werden reproduzierbar geprüft
Tests und Release-Prozess	1–2	Gezielte Tests bestehen, ein vollständiges verpflichtendes Gate fehlt
Deployment und Recovery	1	Deployment ist automatisiert, eine produktionsnahe Vorstufe fehlt
Monitoring	0–1	Fehler sind auffindbar, lösen aber nicht zuverlässig eine aktive Meldung aus
Kostenkontrolle	1	Gesamtkosten sind plausibel, aber nicht pro Nutzer und Ablauf messbar
Wartbarkeit und Übergabe	1–2	Die Anwendung ist dokumentiert, enthält jedoch manuelle Betriebsanteile

- Kontrolliert mit Bedingungen. Im Review wurde keine tatsächliche Offenlegung fremder Nutzerdaten festgestellt. Die möglichen Auswirkungen eines Fehlers rechtfertigen dennoch zusätzliche automatisierte Nachweise.

SECTION

4. Detaillierte Befunde

Befund 1: Agentenverhalten war fachlich nicht ausreichend reproduzierbar

PRIORITÄT: KRITISCH **EVIDENZ:** BERICHTET UND REPRODUZIERT **STATUS:** DURCH BEGRENZTES UMSETZUNGSPAKET REDUZIERT

Eine mittelfristige Zielsetzung wurde vom Agenten unzureichend interpretiert. Der resultierende Trainingsplan war technisch gültig, entwickelte Belastung und Progression jedoch nicht angemessen über den geplanten Zeitraum.

Auswirkung

Ein Ergebnis kann formal korrekt erscheinen, obwohl es den fachlichen Zweck verfehlt. Solche Fehler sind für Nutzer schwer zu erkennen, weil die Ausgabe plausibel formuliert sein kann.

Ursache

Es fehlte ein reproduzierbarer Satz repräsentativer Szenarien, der zentrale fachliche Erwartungen prüft.

Vor Änderungen an Prompts, Modellen oder Agentenlogik konnte nicht zuverlässig festgestellt werden, ob bekannte Fehlverhalten erneut auftreten.

Umgesetzte Maßnahme

Das bekannte Fehlverhalten und weitere priorisierte Szenarien wurden in das Agent Behavior Release Gate aufgenommen.

Verbleibendes Risiko

Der Testsatz deckt bekannte und priorisierte Fehlerklassen ab, nicht jede mögliche Eingabe oder Empfehlung.

Befund 2: Wiederholte Hintergrundfehler blieben ohne aktive Alarmierung

PRIORITÄT: HOCH **EVIDENZ:** DEMONSTRIERT **STATUS:** OFFEN

Ein aktivierter Hintergrundprozess schlug wiederholt fehl. Die Fehler waren über technische Logs nachvollziehbar, lösten aber keine aktive Benachrichtigung aus.

Auswirkung

Ein für Nutzer relevanter Ablauf kann über längere Zeit gestört bleiben, obwohl die Anwendung selbst weiterhin erreichbar ist.

Dadurch besteht das Risiko, dass:

- Probleme erst durch Nutzerbeschwerden auffallen
- Daten oder Empfehlungen verspätet verarbeitet werden
- Fehlerursachen nachträglich schwerer rekonstruierbar sind
- der Betrieb stabil erscheint, obwohl Teilprozesse ausgefallen sind

Erforderliche Maßnahme

- wiederholte Fehler automatisch erkennen
- nach einer definierten Schwelle alarmieren
- betroffenen Ablauf datensparsam identifizieren
- Reaktion und Wiederherstellung dokumentieren
- Alarmierung kontrolliert verifizieren

Bewertung

Vor einer deutlichen Vergrößerung der Nutzergruppe erforderlich.

Befund 3: Nutzer- und Datentrennung besitzt eine starke Basis, aber unvollständige Evidenz

PRIORITÄT: HOCH **EVIDENZ:** DOKUMENTIERT UND TEILWEISE GEMESSEN **STATUS:** TEILWEISE KONTROLLIERT

Die Anwendung verwendet nutzergebundene Zugriffe, begrenzte Agentenberechtigungen und gezielte Isolationstests.

Es wurde keine tatsächliche Offenlegung fremder Nutzerdaten festgestellt.

Verbleibende Lücke

Nicht alle kritischen Zugriffswege des zentralen Nutzerablaufs werden automatisiert mit mehreren Nutzern, Rollen und Fehlerfällen geprüft.

Auswirkung

Ein Fehler in diesem Bereich hätte trotz begrenzter Eintrittswahrscheinlichkeit eine hohe mögliche Auswirkung.

Erforderliche Maßnahme

Automatisierte Isolationstests für:

- fremde Trainingsdaten
- fremde Mahlzeiten und Gesundheitsinformationen
- Agentenausführung im falschen Nutzerkontext
- Hintergrundverarbeitung mit falscher Nutzerzuordnung
- manipulierte Identifikatoren
- direkte API- und Datenbankzugriffe

Bewertung

Für den begrenzten Pilotbetrieb vertretbar. Vor einer öffentlichen Registrierung muss die Evidenz erweitert werden.

Befund 4: Änderungen erreichen das Produktivsystem ohne ausreichende Vorstufe

PRIORITÄT: HOCH **EVIDENZ:** DEMONSTRIERT **STATUS:** OFFEN

Änderungen am Hauptzweig werden automatisiert ausgerollt. Eine separate produktionsnahe Staging-Umgebung sowie ein vollständiges verpflichtendes Release-Gate fehlen.

Auswirkung

Fehler können zuerst im laufenden Pilotsystem sichtbar werden.

Dies betrifft:

- klassische Softwarefehler
- neue Prompts
- Modellwechsel
- angepasste Agentenwerkzeuge
- neue Datenformate
- veränderte externe Integrationen

Erforderliche Maßnahme

Vor dem produktiven Deployment mindestens verpflichtend ausführen:

- Build
- statische Codeprüfungen
- automatisierte Tests
- zentralen End-to-End-Nutzerweg
- kritische Isolationstests
- Agenten-Evaluationen
- kontrollierten Smoke-Test

Bewertung

Für seltene und eng kontrollierte Änderungen im Pilotbetrieb vorübergehend vertretbar. Für häufige Releases oder eine größere Nutzergruppe nicht ausreichend.

Befund 5: Kosten sind plausibel, aber nicht steuerbar belegt

PRIORITÄT: MITTEL **EVIDENZ:** BERICHTET **STATUS:** OFFEN

Die bisherigen Gesamtkosten liegen im erwarteten Bereich. Eine belastbare Zuordnung zu einzelnen Nutzern, Agentenläufen oder Produktabläufen besteht nicht.

Auswirkung

Kostensteigerungen können erst auf der Gesamtrechnung auffallen.

Unklar bleibt:

- welche Nutzer besonders hohe Kosten verursachen
- welche Agentenabläufe teuer sind
- ob Fehler unnötige Modellaufrufe erzeugen
- welche Kosten bei wachsender Nutzung entstehen
- ob die geplante Preisgestaltung wirtschaftlich tragfähig ist

Erforderliche Maßnahme

Mindestens messen:

- Modellkosten pro Nutzer
- Modellkosten pro Ablauf
- Anzahl und Dauer von Agentenläufen
- fehlgeschlagene oder wiederholte Ausführungen
- Infrastrukturkosten pro aktivem Nutzer
- Kostenentwicklung bei steigender Nutzung

Bewertung

Kein unmittelbarer Blocker für eine kleine Pilotgruppe. Vor Preisvalidierung und Skalierung erforderlich.

Befund 6: Wiederherstellung ist möglich, aber operativ nicht ausreichend belegt

PRIORITÄT: MITTEL **EVIDENZ:** BERICHTET UND DOKUMENTIERT **STATUS:** OFFEN

Backups und Rollback-Möglichkeiten sind grundsätzlich vorhanden. Die Wiederherstellung wurde im verfügbaren Reviewumfang jedoch nicht vollständig reproduzierbar nachgewiesen.

Auswirkung

Ein vorhandenes Backup ist erst dann belastbar, wenn klar ist:

- welche Daten enthalten sind
- wie lange die Wiederherstellung dauert

- wer sie durchführen kann
- welche Daten verloren gehen können
- wie die Funktionsfähigkeit anschließend geprüft wird

Erforderliche Maßnahme

- Recovery-Schritte dokumentieren
- Wiederherstellung kontrolliert testen
- Zielwerte für Datenverlust und Wiederherstellungsdauer definieren
- verantwortliche Rolle benennen
- definierten Smoke-Test nach dem Restore ausführen

Befund 7: Wartbarkeit ist für den Pilotbetrieb ausreichend, aber personenabhängig

PRIORITÄT: MITTEL **EVIDENZ:** DOKUMENTIERT **STATUS:** TEILWEISE KONTROLLIERT

Die Anwendung besitzt eine nachvollziehbare Struktur und technische Dokumentation. Einige Betriebs- und Diagnoseabläufe hängen jedoch noch stark von Einzelwissen und manuellen Schritten ab.

Auswirkung

Bei Abwesenheit der hauptverantwortlichen Person können Diagnose und Wiederherstellung länger dauern.

Erforderliche Maßnahme

Kurze Runbooks für:

- fehlgeschlagene Hintergrundverarbeitung
- fehlerhaftes Deployment
- Modell- oder Provider-Ausfall
- unerwarteten Kostenanstieg
- Wiederherstellung aus einem Backup
- Sperrung oder Entfernung eines Nutzers

SECTION

5. Technischer Nachweis der Umsetzung

Für eine vollständige Kundenfassung sollten die folgenden Nachweise konkret referenziert werden.

Code-Änderung

- Repository oder Übergabeort
- Pull Request oder Commit
- betroffene Komponenten
- technische Kurzbeschreibung
- Review-Status

Automatisierter Nachweis

- Anzahl der enthaltenen Szenarien
- geprüfte Agentenversion
- erwartete und tatsächliche Ergebnisse
- Ergebnis des bekannten Regressionstests
- Ergebnis der kontrolliert fehlerhaften Variante

Release-Integration

- Zeitpunkt der Ausführung
- auslösender Branch oder Workflow
- Verhalten bei kritischem Fehler
- dokumentierte Ausnahme- oder Freigaberegeln

Übergabe

- verantwortliche Rolle
- Anleitung zur erneuten Ausführung
- Anleitung zur Ergänzung neuer Fehlerfälle
- bekannte Grenzen
- empfohlener nächster Ausbau

Eine öffentliche Fassung kann sensible Repository- und Kundendetails abstrahieren. Sie sollte dennoch einen konkreten Vorher-Nachher-Nachweis enthalten.

SECTION

6. Verbleibende Risiken und Übergangskontrollen

Hintergrundfehler werden noch nicht zuverlässig alarmiert

Übergangskontrolle:

- regelmäßige manuelle Prüfung kritischer Prozesse
- direkte Betreuung der Pilotnutzer
- klarer Eskalationsweg bei Problemen

Nicht alle Agentenfehler sind durch Evaluationen abgedeckt

Übergangskontrolle:

- begrenzte Pilotgruppe
- feste Agentenversion
- manuelle Prüfung ungewöhnlicher Ergebnisse
- direkte Eskalationsmöglichkeit

Nutzertrennung ist nicht vollständig automatisiert belegt

Übergangskontrolle:

- keine öffentliche Registrierung
- kontrollierte Benutzeranlage
- begrenzte Rollen und Agentenrechte

Fehlerhafte Releases können weiterhin den Pilotbetrieb erreichen

Übergangskontrolle:

- kleine Releases
- feste Freigabeverantwortung
- dokumentierte Rollback-Möglichkeit
- begrenzte Änderungshäufigkeit

Kosten besitzen noch keine Warn- oder Begrenzungsmechanismen

Übergangskontrolle:

- begrenzte Nutzerzahl
- regelmäßige manuelle Prüfung der Gesamtkosten
- keine automatische Skalierung ohne Freigabe

SECTION

7. Roadmap nach Entscheidungspunkten

Vor Beginn des bezahlten Piloten

- Agent Behavior Release Gate abschließen und verifizieren
- geprüfte Agenten- und Anwendungsversion festlegen
- Pilotgruppe und Zugriff klar begrenzen
- verantwortliche Person für Störungen benennen
- manuellen Kontroll- und Supportprozess festhalten

Vor Vergrößerung der Pilotgruppe

- Monitoring für zentrale Hintergrundprozesse einführen
- aktive Alarmierung mit kontrolliertem Fehler testen
- Tests und Build verpflichtend vor Releases ausführen
- Kosten pro Nutzer und zentralem Ablauf messen
- Incident- und Recovery-Runbook dokumentieren

Vor öffentlicher Registrierung

- kritische Nutzer- und Datengrenzen umfassend automatisiert prüfen
- produktionsnahe Staging-Umgebung einführen
- vollständigen zentralen Nutzerweg als Release-Gate ausführen
- fehlerhafte Releases automatisch stoppen oder zuverlässig zurückrollen
- Rollen, Verantwortlichkeiten und Eskalationswege festlegen

Vor breitem kommerziellen Rollout

- Kostenlimits und Warnschwellen definieren
- Betriebsdaten aus dem Pilot auswerten
- Agenten-Evaluationen um reale Pilotfehler erweitern
- Wiederherstellung kontrolliert testen
- Last- und Skalierungsannahmen überprüfen
- erneute formale Go-Live-Entscheidung treffen

SECTION

8. Ergänzender 30/60/90-Tage-Plan

Die zeitliche Darstellung ist nur eine Orientierung. Maßgeblich bleiben die jeweiligen Freigabebedingungen.

Innerhalb von 30 Tagen

- Monitoring und Alarmierung für kritische Hintergrundprozesse einführen
- Build, Tests und Agenten-Evaluationen als Release-Gate ausführen
- Kosten pro Nutzer und zentralem Ablauf messen
- Incident- und Recovery-Verantwortung festlegen

Innerhalb von 60 Tagen

- produktionsnahe Staging-Umgebung einführen
- kritische Nutzerzugriffe breiter automatisiert prüfen
- Agenten-Evaluationen mit Fehlerfällen aus dem Pilot erweitern
- Wiederherstellung kontrolliert testen

Innerhalb von 90 Tagen

- Warnschwellen und Kostenlimits festlegen
- Deployment- und Recovery-Runbooks vervollständigen
- Skalierungs- und Lastannahmen überprüfen
- neue Go-Live-Entscheidung auf Basis realer Betriebsdaten treffen

SECTION

9. Kriterien für die nächste Freigabeentscheidung

Ein breiterer Rollout sollte erst freigegeben werden, wenn:

- kritische Fehler rechtzeitig eine aktive Benachrichtigung erzeugen
- der zentrale Nutzerweg vor jedem Release geprüft wird
- bekannte kritische Agentenfehler reproduzierbar erkannt werden
- Änderungen an Agentenlogik und Modellen kontrolliert freigegeben werden
- Kosten pro Nutzer und Ablauf messbar sind
- Warn- oder Begrenzungsmechanismen für Kosten bestehen
- kritische Nutzergrenzen automatisiert geprüft werden
- ein fehlerhaftes Release gestoppt oder zurückgerollt werden kann
- Wiederherstellung und Incident-Reaktion dokumentiert und getestet sind

SECTION

10. Grenzen und Gültigkeit des Berichts

Diese Fallanalyse dokumentiert die technische Bewertung einer Anwendung in einer kontrollierten Pilotphase.

Sie ist:

- kein Kundenreferenzbericht
- kein vollständiger Sicherheits- oder Penetrationstest
- keine medizinische Bewertung
- keine regulatorische Zertifizierung
- keine Garantie für fehlerfreien Betrieb
- keine pauschale Freigabe für jede Nutzerzahl oder jeden Einsatzbereich

Das Urteil gilt ausschließlich für:

- den beschriebenen zentralen Nutzerweg
- die geprüfte Anwendungsversion
- die geprüfte Agentenversion
- die kontrollierte Nutzergruppe
- die benannten Betriebsbedingungen

Einige Aussagen beruhen auf Angaben aus dem bisherigen Betrieb. Dazu gehören insbesondere tägliche Nutzung, bisherige Kosten, vorhandene Backups und frühere Rollbacks.

Wo keine reproduzierbare Evidenz vorlag, wurde dies in der Bewertung berücksichtigt.

Änderungen an Architektur, Agentenlogik, Modell, Datenzugriffen, Deployment oder Nutzungsszenario können eine erneute Bewertung erforderlich machen.