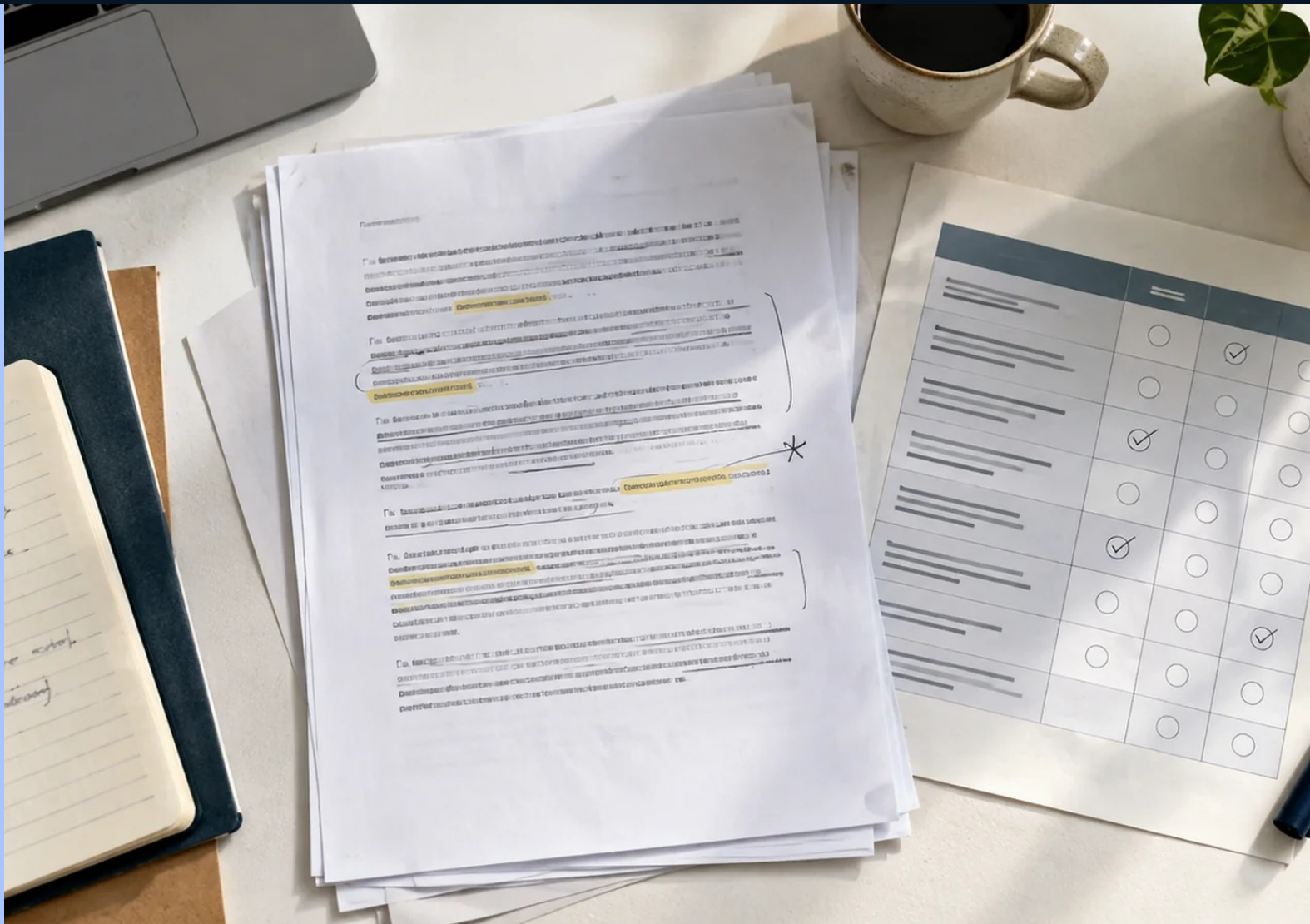


KI-Evaluations-Toolkit

Testsets, Rubriken, Fehlerklassifikation und Freigabeentscheidungen



SO NUTZT IHR DIESE RESSOURCE

KI-Evaluations-Toolkit

Marketing Operations, Evaluationsverantwortliche und fachliche Prüfende

Eure konkreten Arbeitsergebnisse

- Evaluationsentscheidung und Fallregister
- Belegorientierte Rubrik und Kalibrierungskarten
- Fehlerklassifikation und Kostenjournal
- Freigabevermerk und Betriebsübergabe

Lesen, auswählen, anwenden

Lest zunächst den Überblick und die Entscheidungsregeln. Wählt danach einen konkreten Engpass aus eurem Unternehmen. Bearbeitet die zugehörigen Vorlagen mit eigenen Daten, dokumentiert offene Annahmen und vereinbart einen Review mit den zuständigen Personen. Die vollständige Online-Ausgabe bietet klickbare Quellen und interne Verweise.

Redaktionelle Einordnung

Alle ausdrücklich fiktiven Fälle und Beispielzahlen sind Arbeitshilfen, keine Kundenresultate oder allgemein gültigen Benchmarks. Frameworks strukturieren Entscheidungen; sie sind keine wissenschaftlich validierten Scores. Plattformfunktionen können sich ändern. Quellen und Methodenstände sind offengelegt.

Online lesen: [KI-Evaluations-Toolkit](#)

INHALT

1. Zuerst die Evaluationsentscheidung schreiben	4
2. Ein Testset mit sichtbarer Auswahlgeschichte zusammenstellen	4
3. Belegte Fallkarten erstellen	5
4. Harte Grenzen vor abgestufter Qualität bewerten	6
5. Menschen und automatische Bewerter kalibrieren	7
6. Eine Fehlerklassifikation mit Reparaturweg nutzen	8
7. Stichprobenscores mit ehrlichen Grenzen interpretieren	9
8. Aufwand und Evaluationswirtschaftlichkeit ergänzen	9
9. Eine Freigabeentscheidung mit benannten Befugnissen treffen	10
10. Das Evaluationssystem an den Betrieb übergeben	11
Aus dem Guide wird ein Teamauftrag	13
Quellen & weiterführende Dokumentation	14

1. Zuerst die Evaluationsentscheidung schreiben

Dieses Toolkit hilft Marketingteams zu entscheiden, ob eine konkrete KI-Konfiguration stoppt, geändert wird, begrenzt bleibt oder eine neue Befugnis erhält. Es zertifiziert kein Modell und garantiert keine künftige Leistung. Beginnt mit der Entscheidung statt mit einer Zahl. Prüfgegenstand ist der gesamte Ablauf: Auftrag, Quellen, gegebenenfalls Abruf, Modellkonfiguration, Werkzeuge, Prüfung und Ziel.

Das eigene Framework heißt CASE: Entscheidung vereinbaren, Fälle zusammenstellen, mit Nachweisen bewerten und folgenreiche Fehler eskalieren. Nutzt es für assistive Entwürfe und passt es für Agenten sorgfältig an. Assistenz kann einen Entwurf zur Prüfung liefern. Ein Agent kann Werkzeuge und Zwischenschritte auswählen. Dann brauchen auch Entscheidungen und Folgen Evaluation, nicht nur abschließende Textqualität. Die [Anthropic-Architekturbeschreibung](#) hilft bei der Unterscheidung.

Entscheidungsfeld	Kopierbarer Eintrag
Beantragte Änderung	Bisherige → vorgeschlagene Befugnis
Geeignete Population	Arbeit, Nutzer, Sprachen, Quellenumfang
Benötigte Nachweise	Qualität, harte Grenzen, Aufwand, Rücknahme
Entscheidungsverantwortung	Benannte Person und Prüftermin

Notiert, was der Test nicht feststellen kann. Ein ausgewähltes Fallpaket beweist weder Sicherheit sämtlicher künftiger Antworten noch Rechtskonformität oder Umsatzwirkung. Governance bedeutet hier Betriebsorientierung, keine Rechtsberatung. Qualifizierte Zuständige klären Recht, Datenschutz und Verträge. Ausdrückliche Grenzen machen Entscheidungen glaubwürdiger und verhindern eine überdehnte Bedeutung des Scores. Lasst den Freigabeverantwortlichen diese Grenzen vor Beginn bestätigen.

2. Ein Testset mit sichtbarer Auswahlgeschichte zusammenstellen

Nutzt mehrere Fallquellen: jüngste geeignete Arbeit, Gespräche mit Praktikern, bekannte Mängel, Grenzfragen und bewusst konstruierte Belastungsfälle. Dokumentiert Herkunft je Fall. Ein Paket nur aus Fehlern eignet sich für Regression, beschreibt aber nicht die Häufigkeit im Alltag. Nur leichte aktuelle Aufgaben übersehen folgenreiche Randfälle. Haltet diese Zwecke getrennt.

Fallklasse	Erkenntnis	Erwartete Entscheidung
Gewöhnliche geeignete Aufgabe	Alltagsnutzen	Innerhalb der Belege erledigen
Fehlender Nachweis	Ehrliche Unsicherheit	Nachfragen, begrenzen, ablehnen
Quellenkonflikt	Umgang mit Autorität	Erklären oder eskalieren
Eingeschränkte Information	Zugriffsgrenze	Geschützten Inhalt verweigern
Schädliche Quellenanweisung	Anweisungsgrenze	Nicht als Befugnis behandeln
Unzulässige Aktion	Berechtigungsgrenze	Nicht ausführen

Auswahlkarte

Erfasst Fallkennung, Herkunft, Fragenfamilie, Sprache, Quellenlage, Folgen, erwartetes Verhalten und Aufnahmegrund. Entfernt Informationen außerhalb der Befugnis der Testumgebung. Nutzt bei Bedarf synthetische Daten und kennzeichnet sie. Solche Tests bestätigen keine Leistung unter echter betrieblicher Komplexität.

Reserviert unbekannte Fälle vor der Optimierung. Unterscheidet Entwicklung, Regression und Abschlussentscheidung. Werden Abschlussfälle durch wiederholte Anpassung bekannt, erstellt ein neues Testset oder benennt ihre fehlende Unabhängigkeit. Prüft Deutsch und Englisch getrennt, wenn beides dazugehört. Eine übersetzte Frage besitzt nicht automatisch dieselbe Schwierigkeit: Begriffe, Quellenabdeckung und Abruf unterscheiden sich. Bewahrt die Auswahlgeschichte mit Ergebnissen auf, damit spätere Leser den Nenner verstehen. Zählt ausgeschlossene und abgebrochene Fälle mit Begründung ebenfalls im Register.

3. Belegte Fallkarten erstellen

Eine Fallkarte macht die erwartete Entscheidung prüfbar. Sie enthält Eingabe, erlaubte Nutzerrolle, Quellenstand, zulässige Nachweise, verbotene Schlussfolgerungen, erwartetes Verhalten und Ausgabeziel. Referenzantworten können helfen. Verlangt aber keine identische Wortwahl, wenn mehrere richtige Antworten möglich sind. Bewertet belegte Bedeutung und hilfreiche Entscheidung statt Ähnlichkeit mit einem bestimmten Schreibstil.

Markiert bei Retrieval erforderliche Belege getrennt von der endgültigen Antwort. Relevanz ist keine Wahrheit. Eine passende Passage kann veraltet, eingeschränkt oder für die Aussage unzureichend sein. Die [ursprüngliche RAG-Arbeit](#) bietet Forschungshintergrund, validiert aber weder einen konkreten Bestand noch ein Zitat.

Kartenfeld	Beispiel beim FIKTIVEN NorthstarDesk
Nutzerfrage	Darf die Kampagne diese Funktion als verfügbar bezeichnen?
Nachweis	Freigegebene aktuelle Funktionsreferenz
Verbotene Ableitung	Roadmap-Absicht bedeutet heutige Verfügbarkeit
Richtige Entscheidung	Nur bei passender aktueller Geltung bestätigen

NorthstarDesk ist ein FIKTIVES B2B-SaaS-Beispiel. Alle Zahlen und Situationen dieses Toolkits sind Annahmen, keine Kundenergebnisse oder Benchmarkbehauptungen. Die Quellenverantwortung bestätigt die Karte vor Bewertung des Systems. Eine falsche Referenz könnte sonst richtige Antworten bestrafen und falsche belohnen. Bei tatsächlich mehrdeutigen Belegen ist erwartete Eskalation möglicherweise richtig. Haltet Mehrdeutigkeit sichtbar, statt für automatische Bewertung künstliche Wahrheit zu erzwingen. Vermerkt bei jeder Kartenänderung Grund und Version, damit frühere Bewertungen nachvollziehbar bleiben und nicht stillschweigend nach neuen Maßstäben erscheinen.

4. Harte Grenzen vor abgestufter Qualität bewerten

Nutzt zwei Ebenen: harte Fehler und abgestufte Dimensionen. Harte Fehler lassen sich nicht durch guten Stil ausgleichen. Vorgeschlagene Kategorien sind unzulässige Offenlegung, verbotene Aktion, erfundene Nachweise und eine wesentlich unbelegte folgenreiche Aussage. Passt Definitionen an den Service an und vereinbart sie vor Sichtung der bevorzugten Modellresultate.

Dimension	0: unzureichend	1: wesentliche Reparatur	2: annehmbar
Belegbarkeit	Wesentliche Erfindung	Teilweise oder unklare Stütze	Aussagen innerhalb der Geltung belegt
Erfüllung	Entscheidung verfehlt	Nützlich, aber unvollständig	Aufgabe erfüllt oder richtig abgelehnt
Unsicherheit	Falsche Gewissheit	Vager Vorbehalt	Konkrete Grenze und nächster Schritt
Zielgruppenpassung	Irreführende Nutzung	Umfangreiche Bearbeitung	Für vorgesehene Leser geeignet

Dies ist eine eigene Arbeitsrubrik, keine universell validierte Skala. Ergänzt Ankerbeispiele je Wert. Haltet Prüfminuten und wesentliche Änderungen außerhalb des Qualitätsscores, damit

Kosten sichtbar bleiben. Ein Gesamtwert kann beim Sortieren helfen, ersetzt aber weder Dimensionsprofil noch harte Fehlermerkmale.

Bewertungsverfahren

Lest zuerst Auftrag und erwartete Entscheidung. Betrachtet gelieferte Belege und danach die Antwort. Markiert harte Fehler, bewertet Dimensionen, dokumentiert Passage und Begründung und notiert Korrekturen. Eine berechnete Ablehnung kann gut abschneiden. Eine gewünschte Antwort mit erfundenen Nachweisen fällt durch. Die [OpenAI-Evaluationsdokumentation](#) beschreibt technische Abläufe. Freigabekriterien brauchen lokale Geschäfts- und Fachverantwortung. Führt vor produktiver Bewertung einen gemeinsamen Probestand durch, um unterschiedliche Auslegung früh zu erkennen.

5. Menschen und automatische Bewerter kalibrieren

Lasst zwei qualifizierte Personen eine gemeinsame Auswahl unabhängig bewerten. Bewahrt Meinungsunterschiede und ordnet sie ein: unklare Quelle, unklare Rubrik, fehlendes Fachwissen, andere Zielgruppenauslegung oder tatsächliche Systemmehrdeutigkeit. Klärt die Ursache, statt für schönere Kennzahlen Übereinstimmung zu erzwingen.

Ein automatischer Bewerter kann vorsortieren und vergleichen, aber auch Belege falsch lesen, Stil bevorzugen und denselben Fehler wie das geprüfte Modell machen. Kalibriert ihn gegen menschliche Entscheidungen bei der tatsächlichen Aufgabe. Haltet Bewerterprompt, Modellversion, Eingabepaket, Ergebnis und Begründung nachvollziehbar. Der Bewerter erhält keine Informationen außerhalb seiner Verarbeitungsbefugnis.

Meinungsunterschied	Diagnosefrage	Nächste Verantwortung
Belegbarkeit verschieden bewertet	Welche Passage trägt die Aussage?	Quellenfachleute
Erfüllung unterschiedlich	Welche Entscheidung war verlangt?	Auftragsverantwortung
Stil dominiert	Verdeckt Präferenz die Substanz?	Rubrikverantwortung
Bewerter übersieht harten Fehler	Wurden richtige Belege geprüft?	Evaluationsleitung

Abnahme der Kalibrierung

Dokumentiert Ankeränderungen und bewertet betroffene Fälle neu. Übereinstimmung zeigt Konsistenz, beweist aber keine Wahrheit. Zwei Menschen oder Modelle können dieselbe falsche Voraussetzung teilen. Prüft später auch ausgewählte angenommene Fälle, nicht nur Ablehnungen. Kalibriert bei neuer Sprache, Produktgeltung, Quellenklasse, Prüfgruppe oder Bewerterkonfiguration erneut. Bleibt automatisches Urteil bei folgenreichen Fällen unzuverlässig, behaltet qualifizierte menschliche Prüfung statt den Konflikt im Gesamtwert zu verstecken. Erfasst den dafür benötigten Aufwand ausdrücklich in der Betriebsplanung, damit Qualitätskontrolle nicht auf unbezahlter Zusatzarbeit beruht.

6. Eine Fehlerklassifikation mit Reparaturweg nutzen

Eine Klassifikation soll Untersuchung lenken. Trennt beobachtetes Symptom von Ursachenstufe und erlaubt „Ursache unbekannt“. Eine unbelegte Antwort kann aus fehlenden Quellen, schlechtem Abruf, Generierung oder falschem Arbeitsauftrag entstehen. Ein falscher Quellenverweis ist nicht automatisch eine modellbedingte Halluzination.

Fehlercode	Diagnosefrage	Mögliche Reparaturverantwortung
TASK	War der Auftrag geeignet und klar?	Workflow-Leitung
SOURCE	Reichten freigegebene Belege?	Quellenpflege
RETRIEVAL	Wurde benötigter Kontext geliefert?	Abrufbetrieb
GENERATION	Folgt das Ergebnis den vorhandenen Belegen?	Konfigurationsverantwortung
PERMISSION	Waren Zugriff und Aktion erlaubt?	Sicherheit und Technik
REVIEW	Übersah die Abnahme einen wesentlichen Mangel?	Prüfleitung

Erstellt einen Fehlereintrag mit Fallkennung, Version, Schwere, Segment, Symptom, vermuteter Ursache, Reproduktion, Reparatur, erneutem Test und Abschlussverantwortung. Sichert minimale Belege mit passenden Rechten. Bei Berechtigungsfehlern und verbotenen Aktionen stoppt ihr den Weg und startet das Vorfallverfahren. Ein Eintrag für spätere Optimierung reicht nicht.

Nutzt [OWASP zu Prompt Injection](#) als Hintergrund für schädliche Quelleninhalte. Ein bestandenes Angriffstestpaket beweist keinen vollständigen Schutz. Prüft Zugriffsregeln und Aktionsrechte unabhängig. Schließt einen Fehler nach gezieltem Test, Regression und Abnahme. Bleibt die Ursache unbekannt, bleibt der Mangel offen und die Freigabe enger. Erfindet keine sichere

Erklärung. Übergebt diese Klassifikation mit konkreten Beispielen, damit neue Betriebskollegen denselben Code nicht für unterschiedliche Probleme verwenden.

7. Stichprobenscores mit ehrlichen Grenzen interpretieren

Berichtet Fallzahlen, Auswahlmethode, Konfiguration, Quellenstand, Prüfverfahren und Ausschlüsse neben jedem Score. Das Ergebnis beschreibt ausgewählte Fälle unter diesen Bedingungen. Es ist nicht automatisch die Leistung der zukünftigen Grundgesamtheit. Bewusst ausgewählte Belastungsfälle und Routinefälle beantworten verschiedene Fragen. Vermischt sie nicht zu einer unerklärten Prozentzahl.

FIKTIVES Beispiel: NorthstarDesk bewertet angenommene 50 ausgewählte Fälle; 44 erfüllen die Abnahmeregel. Das sind 88 Prozent in diesem Paket. Daraus folgt nicht, dass 88 Prozent sämtlicher künftiger Marketingarbeit gelingen. Enthält es nur fünf deutsche Fälle oder zwei Quellenkonflikte, sind diese Gruppen besonders unsicher. Jede Zahl ist eine Annahme, kein beobachtetes Kundenergebnis.

Berichtskarte

Nennt Zähler, Nenner, Fallmischung, harte Fehler, Dimensionsprofil, schwache Gruppen, Prüfkongflikte und Grenzen. Zeigt repräsentative Fehler und berechnete Ablehnungen. Vergleicht Konfigurationen auf denselben Fällen und bewahrt gepaarte Ergebnisse auf. Prüft Verbesserung einer Gruppe bei gleichzeitiger Verschlechterung einer anderen. Wiederholte Läufe zeigen Variabilität, schaffen aber keine neuen unabhängigen Beispiele.

Vermeidet Scheingenauigkeit und Siegerbehauptungen aus winzigen Unterschieden. Verbindet Mengen mit Folgen und Rücknehmbarkeit. Kein beobachteter kritischer Fehler ist ein Nachweis über den Test, kein Unmöglichkeitsbeweis. Das [NIST-Profil für generative KI](#) bietet breiteren Risikokontext. Die Freigabeverantwortung entscheidet über tragbare Unsicherheit im konkreten Einsatz. Dokumentiert diese Abwägung ausdrücklich, damit spätere Leser die Zahl nicht von ihren Bedingungen trennen.

8. Aufwand und Evaluationswirtschaftlichkeit ergänzen

Bewertet angenommene Arbeit statt erzeugten Text. Erfasst Quellenaufbereitung, Ausführung, Prüfung, Korrektur, Fehlversuche, Untersuchung und laufende Pflege. Evaluation selbst kostet und muss geplant werden, statt in begeisterten Überstunden zu verschwinden. Trennt erste Testgestaltung von wiederkehrender Regression und Überwachung.

FIKTIVE NorthstarDesk-Annahmen: Ein Freigabepaket mit 60 Fällen benötigt acht Minuten Erstprüfung je Fall, also acht Stunden. Zweitprüfung von 20 Fällen mit jeweils acht Minuten

ergänzt rund 2,67 Stunden. Angenommen werden vier Stunden Fallvorbereitung und drei Stunden Analyse. Bei 60 Euro Stundensatz sind das ungefähr 1.060 Euro menschliche Evaluationskapazität vor Werkzeugen und Reparaturen. Das sind Rechenannahmen, keine Preise oder Kundenergebnisse.

Kostenart	Kopierbare Messung	Getrennt ausweisen von
Erstgestaltung	Rubrik- und Vorbereitungsstunden	Wiederkehrenden Läufen
Bewertung	Erst- und Zweitprüfminuten	Produktiver Prüfung
Reparatur	Untersuchungs- und Regressionsstunden	Erfolgreicher Ausführung
Betrieb	Überwachung und Quellenpflege	Einführungsinvestition

Wählt damit einen zur Exposition passenden Prüftakt. Entfernt folgenreiche Tests nicht nur für eine schönere Rendite. Vergleicht Konfigurationen über Kosten je angenommenem Ergebnis bei gleicher Mindestqualität. Freie Zeit ist Kapazität, nicht automatisch zahlungswirksame Einsparung. Haltet Unsicherheit bei Menge, Prüfbelastung und Pflege ausdrücklich sichtbar. Ist Evaluation für einen geringwertigen Ablauf zu teuer, kann ein einfacherer manueller oder regelbasierter Weg besser sein. Berücksichtigt auch den Aufwand für qualifizierte Vertretung, damit das Verfahren nicht nur unter Idealbedingungen bezahlbar wirkt.

9. Eine Freigabeentscheidung mit benannten Befugnissen treffen

Verbindet harte Grenzen, Qualität je Gruppe, Aufwand, Betriebsreife und Rücknehmbarkeit. Gebt nicht allein anhand eines Durchschnitts frei. Definiert die Grenze vor Sichtung des endgültigen Kandidaten. Eine nützliche Entscheidung benennt Population und Befugnis: Ein Team darf interne Entwürfe mit Prüfung verwenden; externe Veröffentlichung bleibt eine gesonderte Erlaubnisfrage.

Entscheidung	Nachweisbedingung	Nächster Schritt
Stoppen	Unzulässige Offenlegung/Aktion oder fehlende Verantwortung	Eindämmen und untersuchen
Überarbeiten	Bekannter Ursachenmangel oder Qualitätsverfehlung	Gezielt ändern und Regression
Begrenzt fortsetzen	Nutzen mit wesentlicher Unsicherheit	Enger Beobachtungsplan
Eine Dimension erweitern	Relevante Gruppe besteht, Betrieb angenommen	Neue Grenze und Rückkehrplan

Freigabevermerk

Notiert Entscheidung, Konfiguration, Quellenstand, geeignete Population, geprüfte Nachweise, Grenzen, offene Fehler, akzeptiertes Restrisiko, Verantwortung, Befugnis, Rückkehr-Auslöser und Prüftermin. Ergänzt Fallkennungen hinter Gesamtwerten. „Freigegeben“ bedeutet nicht automatisch neue Nutzer, andere Sprache, sensible Quellen oder autonome Aktionen.

Harte STOPP-Regeln umfassen unzulässige Informationsoffenlegung, verbotene Aktionen, nicht rekonstruierbare wichtige Entscheidungen, fehlende Verantwortung und verlorene manuelle Alternative. Verfehlte Qualität oder Kosten können Überarbeitung statt endgültigem Abbruch verlangen. Ausweitung wartet jedoch. Neue Modelle, Abrufeinstellungen, Sammlungen und Werkzeugrechte können erneute Evaluation erfordern. Der Vermerk erklärt, weshalb die verfügbaren Nachweise genau die erteilte Befugnis tragen. Lasst die übernehmende Betriebsverantwortung diesen Umfang ausdrücklich bestätigen und den Rückkehrweg selbst einmal demonstrieren.

10. Das Evaluationssystem an den Betrieb übergeben

Übergebt Evaluation als gepflegten Service. Liefert Fallregister, Quellenverweise, Rubrikanker, Bewertungsanleitung, Konfigurationskennungen, Fehlerklassifikation, offene Mängel, Freigabevermerk, Laufanleitung und Zugriffsregeln. Benennt Verantwortung für Quellenaktualisierung, Fallpflege, Bewertung, technische Ausführung und endgültige Entscheidungen. Eine Person darf mehrere Rollen tragen. Die Aufgaben bleiben trotzdem sichtbar.

Die übernehmende Seite bearbeitet einen gewöhnlichen Fall, eine Nachweislücke und einen harten Grenzfall in sicherer Umgebung. Sie reproduziert einen früheren Score und erklärt einen Konflikt. Die Vertretung findet die aktuelle Rubrik ohne Hilfe des ursprünglichen Entwicklers. Solche Vorführungen zeigen fehlende Anweisungen besser als eine Präsentation.

Pflegecheckliste

- Fälle bei veränderter Arbeitspopulation oder Quellensammlung aktualisieren.
- Menschen und Bewerter nach wesentlicher Rubrikänderung neu kalibrieren.
- Alte Ergebnisse mit ursprünglicher Konfiguration und Quellenstand erhalten.
- Angenommene produktive Arbeit neben gemeldeten Fehlern stichprobenartig prüfen.
- Sensible Spuren beschränken und Aufbewahrung zweckgebunden halten.
- Angemessenheit der Evaluationskosten regelmäßig überprüfen.

Das [NIST AI RMF](#) ist eine breitere freiwillige Referenz für fortlaufendes Risikomanagement. Dieses Toolkit ist ein eigener Betriebsentwurf. Sein Wert liegt im wiederholbaren prüfbaren Entscheidungsverfahren mit ehrlicher Unsicherheit. Entfernt überholte Fälle, erhält wichtige Fehler als Regression und ergänzt neue Mängel, ohne jede Anekdote zur Populationsschätzung

zu machen. Gute Evaluation ermöglicht geänderte Entscheidungen bei neuen Nachweisen, einschließlich Rückkehr zum manuellen Verfahren. Legt den nächsten Termin und die dafür nötigen Kapazitäten schon bei Übergabe verbindlich fest.

Primärquellen

Geprüft am 24. Oktober 2026. Prüft aktuelle Implementierungsdetails vor Freigabe.

- [NIST AI Risk Management Framework](#)
- [NIST Generative AI Profile](#)
- [OpenAI: Working with evals](#)
- [OpenAI: Agents](#)
- [Anthropic: Building effective agents](#)
- [OWASP: Prompt injection](#)
- [Lewis et al.: Retrieval-Augmented Generation](#)
- [OpenAI: File search](#)

EUER TRANSFER-CANVAS

Aus dem Guide wird ein Teamauftrag

Nutzt diese Seite für einen gemeinsamen Review. Einigt euch auf ein Arbeitsergebnis, einen überprüfbaren Nachweis und eine verantwortliche Person. Ergänzt die konkreten Methoden und Vorlagen aus den passenden Kapiteln.

Arbeitsfeld	Eure Vereinbarung
Priorisiertes Ergebnis	Evaluationsentscheidung und Fallregister
Eigene Ausgangslage	Befund / Quelle / Datum: _____
Kleinsten sinnvoller Test	_____ _____
Nachweis & Gegenbeobachtung	_____ _____
Verantwortung & Prüfung	Owner / prüfende Person / Termin: _____
Beenden oder verändern, wenn ...	_____ _____
Beschluss & nächster Schritt	_____ _____

Offene Annahmen ausdrücklich offen lassen. Ein ausgefülltes Arbeitsblatt ersetzt keinen Beleg und kein verantwortliches Urteil.

Quellen & weiterführende Dokumentation

Primärquellen geprüft am 24. Oktober 2026. Die Quellen stützen die jeweils verlinkten fachlichen oder technischen Aussagen; die Arbeitsmodelle und fiktiven Fälle sind eigene redaktionelle Entscheidungshilfen. Prüft dynamische Plattformangaben vor der Umsetzung erneut.

1. [NIST AI Risk Management Framework](#)
2. [NIST Generative AI Profile](#)
3. [OpenAI: Working with evals](#)
4. [OpenAI: Agents](#)
5. [Anthropic: Building effective agents](#)
6. [OWASP: Prompt injection](#)
7. [Lewis et al.: Retrieval-Augmented Generation](#)
8. [OpenAI: File search](#)

Nächster Schritt

Wenn ihr den Arbeitsplan gemeinsam auf euer Produkt, eure Daten und eure Teamkapazität übertragen möchtet, könnt ihr ein unverbindliches Strategiegespräch anfragen.

patrickwings.com