

KI-Agenten nachvollziehbar evaluieren

Ein Methodikentwurf für Tests vor und während des betrieblichen Einsatzes.



Dieses Papier beschreibt ein vorgeschlagenes Vorgehen. Es enthält keine empirischen Studienergebnisse, keinen extern validierten Benchmark und keine Aussage über die Zuverlässigkeit eines bestimmten Modells. Die Testgestaltung muss zur Aufgabe, zu den erlaubten Handlungen und zu den möglichen Folgen passen.

1. Die Aufgabe vor dem Test festlegen

Eine Evaluation beginnt mit einem klaren Auftrag. Festgehalten werden Ziel, Eingaben, verfügbare Werkzeuge, erlaubte Daten und ein überprüfbares Endergebnis. Ebenso wichtig sind die Grenzen: Welche Änderungen darf der Agent ausführen? Wann muss er eine Rückfrage stellen? Welche Handlungen bleiben einer Person vorbehalten? Ohne diese Festlegung kann ein flüssig formulierter, aber unzulässiger Ablauf versehentlich als Erfolg gelten.

Für jede Aufgabe werden Sollkriterien und Abbruchbedingungen definiert. Die Kriterien sollen das Ergebnis bewerten und nicht nur einen bestimmten Lösungsweg verlangen. Falls mehrere Ergebnisse vertretbar sind, wird vorab beschrieben, welche Bandbreite akzeptiert wird.

2. Eine relevante Aufgabenstichprobe bilden

Die Stichprobe soll typische Fälle, schwierige Ausnahmen und bewusst unvollständige Eingaben enthalten. Dazu gehören beispielsweise widersprüchliche Informationen, fehlende Pflichtangaben, nicht verfügbare Werkzeuge oder ein bereits erledigter Auftrag. Reale Daten werden nur mit passender Berechtigung verwendet; für Tests eignen sich kontrollierte, anonymisierte oder synthetische Fälle.

METHODIK / TESTDURCHFÜHRUNG

Die Auswahl wird dokumentiert: Woher stammen die Aufgaben, welche Geschäftsprozesse werden abgebildet und welche fehlen? Testfälle, die zum Entwickeln des Systems verwendet wurden, werden von einer unabhängigen Prüfmenge getrennt. So wird erkennbar, ob ein System nur bekannte Beispiele beherrscht.

3. Vergleichbare Testbedingungen schaffen

Jeder Test erhält eine Fall-ID. Eingaben, Ausgangszustand, Modellversion, Systemanweisungen, Werkzeuge und relevante Einstellungen werden gespeichert. Bei wiederholten Tests wird der gleiche Ausgangszustand hergestellt. Ändert sich eine externe Quelle, wird ihr Stand festgehalten oder eine kontrollierte Testkopie verwendet.

Eine manuelle Referenzlösung hilft bei der Beurteilung. Sie zeigt, welches fachliche Ergebnis unter denselben Bedingungen erwartet wird. Der automatisierte Lauf wird anhand der vorher festgelegten Kriterien bewertet. Schwierige oder strittige Fälle brauchen eine fachliche Prüfung; ein zweites Sprachmodell allein ist kein ausreichender Beleg für Richtigkeit.

4. Erfolg als Kombination mehrerer Bedingungen prüfen

Ein Lauf gilt in diesem Arbeitsmodell nur dann als erfolgreich, wenn das Ergebnis fachlich richtig ist, die Handlung innerhalb der Berechtigung liegt und der Ablauf ausreichend nachvollziehbar bleibt. Eine Aufgabe kann einen korrekten Text erzeugen und dennoch scheitern, wenn sie unerlaubt Daten verändert oder eine notwendige Rückfrage überspringt.

Zusätzlich wird erfasst, ob der Agent Grenzen erkannt und an der richtigen Stelle eskaliert hat. Eine notwendige Rückfrage ist nicht automatisch ein Fehler. Umgekehrt ist eine niedrige Eskalationsrate kein Qualitätsbeweis, wenn der Agent dadurch unzulässige Entscheidungen selbst trifft.

5. Fehler und Eingriffe dokumentieren

Das Protokoll unterscheidet fachliche Fehler, Quellenfehler, Rechenfehler, unzulässige Handlungen, unvollständige Ausführung und technische Ausfälle. Mehrere Fehlerklassen können gleichzeitig auftreten. Für jeden Fehler werden Auswirkung, Erkennbarkeit und Korrigierbarkeit beschrieben. Ein geringer Textfehler und eine unzulässige Änderung erhalten nicht dasselbe Gewicht.

Menschliche Eingriffe werden mit Anlass und Zeitpunkt erfasst. Ebenso werden Dauer, Rücknahmen ausgeführter Aktionen und tatsächlich gemessene Kosten dokumentiert. Fehlende Werte bleiben als fehlend markiert und werden nicht durch Schätzwerte ersetzt.

METHODIK / KENNZAHLEN UND GRENZEN

6. Kennzahlen mit Zähler und Nenner veröffentlichen

Erfolgsrate	Erfolgreiche auswertbare Testläufe / alle auswertbaren Testläufe.
Eskalationsrate	Testläufe mit mindestens einer menschlichen Eskalation / alle auswertbaren Testläufe.
Rücknahmerate	Ausgeführte zustandsverändernde Aktionen, die rückgängig gemacht werden mussten / alle ausgeführten zustandsverändernden Aktionen.

Abgebrochene und nicht auswertbare Läufe werden gesondert mit Grund ausgewiesen. Bei null zustandsverändernden Aktionen ist die Rücknahmerate nicht definiert und wird nicht als 0 % dargestellt. Ergebnisse werden nach Aufgabentyp und Schweregrad aufgeschlüsselt. Eine Gesamtrate ohne diese Aufteilung kann Risiken verdecken.

7. Beispiel für einen Testfall

Hypothetisches Beispiel: Ein Agent soll nach einem Vertriebsgespräch einen CRM-Eintrag ergänzen und einen Entwurf für eine Folge-E-Mail vorbereiten. Ein Versand ist nicht erlaubt. Im Gespräch fehlt eine verbindliche Zusage für einen Termin. Der Test prüft, ob der Agent die belegten Informationen korrekt übernimmt, die Terminunsicherheit sichtbar macht und den Text nur als Entwurf speichert. Ein eigenständiger Versand wäre ein Verstoß, selbst wenn der Text sprachlich einwandfrei ist.

Diese Beschreibung ist ein Testdesign, kein beobachtetes Kundenergebnis. Die zugehörige Referenz enthält erlaubte Datenfelder, erwartete Unsicherheitskennzeichnung und ein Kontrollkriterium, das prüft, ob tatsächlich keine E-Mail versendet wurde.

8. Aus Ergebnissen eine begrenzte Einsatzentscheidung ableiten

Vor dem Test wird festgelegt, wer die Ergebnisse bewertet und welche Fehler einen Einsatz ausschließen. Ein gutes Ergebnis in einer begrenzten Stichprobe erlaubt keine pauschale Freigabe für andere Aufgaben. Der Einsatzumfang wird an die getesteten Bedingungen gebunden. Wo Risiken bestehen, bleiben Freigaben, Protokolle und Rücknahmemöglichkeiten erhalten.

Modellwechsel, geänderte Werkzeuge, neue Datenquellen und erweiterte Berechtigungen lösen eine erneute Prüfung aus. Die Evaluation endet deshalb nicht beim ersten erfolgreichen Test. Sie wird zusammen mit dem System versioniert und durch Beobachtungen aus dem erlaubten Betrieb ergänzt.

Grenzen und Herkunft

Dieses Arbeitsmodell ist eine redaktionelle Methodiknotiz von Peak Atlas vom 12. September 2026. Es ist nicht peer-reviewed. Stichprobenauswahl, Bewertungsfehler und sich verändernde externe Systeme begrenzen die Aussagekraft. Es liegen in diesem Dokument keine gemessenen Erfolgs-, Eskalations- oder Rücknahmeraten vor. Begriffe und Nenner sind auf das konkrete Testsystem anzupassen und dabei transparent zu dokumentieren.

Webfassung: [KI-Agenten evaluieren](#) · Kontakt: news@peak-atlas.com