AIVANTUM · ARBEITSUNTERLAGE · VERSION 0.2 · STAND 12.10.2026

Dein Prototyp funktioniert. Das beweist nichts.

Worauf das aufbaut: Diese Unterlage ist eine eigene Aufbereitung. Sie stützt sich auf veröffentlichte Forschung und Praxisberichte zur Prüfung von KI-Systemen, die zitierten Studien stehen am Ende unter „Quellen“. Auswahl, Gliederung, Beispiele und Übungen sind meine.

Kern-These

Ein Prototyp, der „funktioniert“, hat genau eines gezeigt: Auf den Fällen, die jemand ausprobiert hat, sahen die Antworten gut aus. Nicht gezeigt hat er, dass er auf den übrigen Fällen trägt, dass er den nächsten Wechsel von Modell oder Prompt übersteht und dass seine Antworten richtig sind, statt nur richtig zu klingen.

Ein Sprachmodell hat keine Anzeige für „hier bin ich unsicher“. Es formuliert Unsinn im selben souveränen Ton wie eine korrekte Auskunft. Der teure Fehler ist deshalb nicht die falsche Antwort, sondern die falsche Antwort, die niemand bemerkt hat. Die Gegeninstanz heißt Golden-Set: eine feste Sammlung echter Fälle aus eurem Alltag, jeder mit einer abgenommenen richtigen Antwort. Diese Unterlage führt dich in sieben Schritten zu deinem ersten Golden-Set und zeigt dir, wann du einer Bewertungszahl glauben darfst und wann nicht.

Eindruck oder Beleg: Beim Eindruck spielt jemand ein paar Fragen von Hand durch, es sieht gut aus, das System geht live. Bei der nächsten Änderung beginnt das Raten von vorn, weil es keinen Vorher-Zustand gibt. Beim Beleg liegt ein fester Prüfsatz bereit, er läuft automatisch, und jede Änderung erzeugt einen Unterschied, den man benennen kann: drei Fälle besser, einer schlechter, und zwar dieser.

Einstiegs-Weiche: Wo stehst du gerade?

„Wir haben ein paar Fragen durchgespielt, das sah gut aus.“ → Dir fehlt der Prüfsatz. Starte mit Schritt 1 bis 3.

„Wir wissen nicht genau, woran wir gute Antworten erkennen.“ → Dir fehlen die Kriterien. Schritt 4.

„Ich merke erst, dass etwas kaputt ist, wenn sich jemand beschwert.“ → Dir fehlt der automatische Lauf. Schritt 5.

„Wir lassen eine KI die Antworten bewerten, das reicht doch.“ → Dein Bewerter ist ungeprüft. Schritt 6.

„Wir haben Testfälle, aber die pflegt niemand.“ → Dein Set altert. Schritt 7.

Landkarte: 7 Schritte, 3 Bewegungen

Bewegung A · Verstehen (Schritte 1 und 2) → **Bewegung B · Das Set bauen (Schritte 3 bis 5)** → **Bewegung C · Dem Messen misstrauen (Schritte 6 und 7)**

- Trenne „funktioniert“ von „ist belegt“
- Lerne die vier Bausteine kennen
- Bau dein erstes Golden-Set in fünf Schritten
- Miss, was der Fachbereich versteht
- Lass den Prüflauf ohne Erinnerung laufen
- Kalibriere den Bewerter, bevor du ihm glaubst
- Halte das Set am Leben

Schritt 1 · Trenne „funktioniert“ von „ist belegt“ (A)

Eine Demo beantwortet die Frage „Kann das System das?“. Für den Betrieb brauchst du die Antwort auf eine andere Frage: Wie oft liegt es richtig, bei welchen Fällen liegt es daneben, und hat sich das seit gestern verändert? Klassische Software liefert auf dieselbe Eingabe immer dieselbe Ausgabe, ein Test ist grün oder rot. Ein Sprachmodell antwortet auf dieselbe Frage zweimal unterschiedlich, und beide Antworten können richtig sein. Deshalb prüft man hier nicht Gleichheit, sondern Eigenschaften: Stimmt der Inhalt? Ist die Aussage durch die mitgelieferte Quelle gedeckt? Fehlt etwas Wichtiges? Hat das System zu Recht abgelehnt?

Dazu kommt der Wechsel: neue Modellversionen, nachgeschärfte Prompts, geänderte Dokumente in der Wissensbasis. Jede Änderung kann ein System verbessern oder still verschlechtern. Ohne Vorher-Wert merkst du den Unterschied erst, wenn ein Kunde ihn merkt.

> Eine Prüfung beantwortet nicht „Läuft es?“, sondern „Ist es seit gestern besser oder schlechter geworden, und woran sehe ich das?“

Werkzeug · Die Demo-Probe:

(1) Auf welchen Fällen hat es funktioniert, und wer hat sie ausgesucht?

(2) Gegen welche richtige Antwort wurde verglichen, und wer hat sie festgelegt?

(3) Was passiert beim nächsten Wechsel von Modell, Prompt oder Daten, und woran würdest du eine Verschlechterung merken?

Falle & Fix: Falle, die Demo-Fälle selbst auszusuchen und ihren Erfolg als Beleg zu nehmen. Wer die Fälle aussucht, greift unbewusst zu denen, die klappen.

Fix, die Fälle aus dem echten Betrieb ziehen, auch die unbequemen, und die richtige Antwort von jemandem abnehmen lassen, der sie beurteilen kann.

Reflexion: *Welches KI-System nutzt ihr täglich, ohne je gemessen zu haben, ob es besser oder schlechter geworden ist? Was würdest du zeigen, wenn morgen jemand fragt: „Woher wisst ihr, dass das funktioniert?“*

Schritt 2 · Lerne die vier Bausteine kennen (A)

Die Prüfung von KI-Systemen, im Fachjargon kurz Evals, besteht aus vier Bausteinen. Sie bauen aufeinander auf, und die Reihenfolge ist nicht beliebig: Ohne den ersten haben die anderen drei nichts, woran sie sich messen lassen.

Werkzeug · Die vier Bausteine:

(1) Golden-Set: eine kuratierte Sammlung echter Eingaben, jede mit einer belegten richtigen Antwort oder einem klaren Prüfkriterium, erstellt oder abgenommen von Menschen mit Fachwissen.

(2) Regressionslauf: Dasselbe Set läuft automatisch bei jeder Änderung. Das eigentliche Ergebnis ist der Vergleich zur Vorversion.

(3) KI-Bewerter, im Fachjargon Judge: Ein Modell bewertet die Antworten eines anderen, und zwar dort, wo es keine eindeutige richtige Antwort gibt. Immer kalibriert, nie roh.

(4) Angriffstest: Jemand versucht gezielt, das System aus dem Tritt zu bringen, mit untergeschobenen Anweisungen, mit Fragen nach seinen internen Vorgaben oder mit Versuchen, vertrauliche Daten herauszulocken.

Falle & Fix: Falle, mit dem dritten Baustein anzufangen, weil er am bequemsten ist: Ein Modell bewertet ein anderes, und schon steht eine Zahl im Bericht.

Fix, mit dem Golden-Set beginnen. Erst wenn es steht, kann man prüfen, ob ein Bewerter überhaupt dasselbe urteilt wie ein Mensch.

Reflexion: *Welche der vier Bausteine gibt es bei euch heute schon, vielleicht unter anderem Namen? Welcher fehlt ganz?*

Schritt 3 · Bau dein erstes Golden-Set in fünf Schritten (B)

Die Fähigkeit hinter allem ist kein Wissen über Evals, sondern ein einmal vollständig gegangener Durchlauf: Fälle sammeln, richtige Antworten festlegen, Kriterien wählen, laufen lassen, Fehlschläge lesen. Welches Werkzeug du dafür nimmst, ist zweitrangig. Es gibt mehrere ausgereifte, frei verfügbare Frameworks, und für den ersten Durchlauf reicht eine Tabelle. Wer den Durchlauf einmal gemacht hat, kann ihn auf jedes Werkzeug übertragen. Wer nur darüber gelesen hat, kann es nicht.

Werkzeug · Der Erst-Aufbau in fünf Schritten:

(1) Sammle zehn bis dreißig echte Eingaben aus dem realen Anwendungsfall, keine erfundenen. Nimm die häufigen Fälle und mindestens drei, von denen ihr wisst, dass sie heikel sind.

(2) Halte je Eingabe fest, was eine gute Antwort ausmacht: entweder die richtige Antwort selbst oder ein klares Prüfkriterium, nicht beides halb.

(3) Wähle zwei bis drei Kriterien, die zu eurem Fall passen, statt die Standardliste eines Werkzeugs zu übernehmen. Schritt 4 hilft bei der Auswahl.

(4) Lass das Set einmal gegen die aktuelle Version laufen und lies jeden Fehlschlag einzeln. Der erste Lauf soll rot ausgehen, sonst misst das Set noch nichts.

(5) Häng den Lauf an jede Änderung, damit er passiert, ohne dass jemand daran denken muss.

Ein Beispiel zum Mitdenken: Ein Assistent beantwortet Kundenfragen zu Lieferungen und Rücksendungen. Fall 7 im Set lautet: „Ich habe die Ware vor 35 Tagen bekommen, kann ich sie noch zurückschicken?“ Laut Rückgaberichtlinie des Hauses gilt eine Frist von 30 Tagen, über Ausnahmen entscheidet der Kundenservice. Die abgenommene richtige Antwort: keine Zusage, ein Hinweis auf die Frist und der Weg zum Kundenservice. Ein freundliches „Klar, das ist kein Problem“ ist hier kein Stilfehler, sondern ein Fehlschlag mit Folgen. Genau solche Fälle gehören ins Set, nicht die zehn, bei denen jede Version richtig liegt.

Falle & Fix: Falle, die Testfälle vom Modell selbst erzeugen zu lassen, weil es schnell geht. Dann misst das Set, wie gut das Modell mit seinen eigenen Annahmen übereinstimmt.

Fix, künstlich erzeugte Fälle höchstens als Rohentwurf nutzen und jeden einzelnen von einem Menschen mit Fachwissen abnehmen lassen, bevor er ins Set kommt.

Reflexion: *Welcher KI-Handgriff bei euch wäre der beste erste Kandidat für zwanzig Fälle? Traust du dich, den ersten Lauf rot ausgehen zu lassen, statt das Set so lange zu glätten, bis es grün wird?*

Schritt 4 · Miss, was der Fachbereich versteht (B)

Ein Ähnlichkeitswert von 0,84 gegenüber einer Referenzantwort sagt niemandem im Fachbereich etwas. Er klingt präzise und entscheidet nichts. Miss stattdessen an Größen, über die die Menschen urteilen können, die mit dem Ergebnis arbeiten.

Werkzeug · Fünf Kriterien für den Fachbereich:

(1) Korrektheit: Stimmt die Aussage?

(2) Quellen-Deckung: Steht das Behauptete in der mitgelieferten Quelle, oder hat das Modell es ergänzt?

(3) Vollständigkeit: Fehlt etwas, das für die Entscheidung gebraucht wird?

(4) Schritt-Erfolg: Hat ein Agent den richtigen Zwischenschritt gemacht, etwa das richtige System abgefragt?

(5) Angemessene Ablehnung: Sagt das System „das weiß ich nicht“ oder „dafür bin ich nicht zuständig“, wo genau das die richtige Antwort ist?

Die richtigen Antworten müssen von Menschen kommen, die den Fall wirklich beurteilen können. Das ist der teure Teil, und genau der Teil, der das Set wertvoll macht. Ein Golden-Set ist im Kern kodifiziertes Fachwissen: die schriftliche Fassung dessen, was bisher nur die drei erfahrensten Leute im Kopf haben. Deshalb kann es niemand außerhalb eures Hauses für euch bauen, und deshalb ist es kein Nebenprodukt der IT, sondern ein Ergebnis, das bei euch bleibt.

Falle & Fix: Falle, die Kriterien aus der Voreinstellung des Werkzeugs zu übernehmen, weil sie schon da sind. Das Werkzeug hat keine Meinung darüber, was in eurem Fall „gut“ heißt.

Fix, die Kriterien aus dem Anwendungsfall ableiten und mit dem Fachbereich abstimmen, bevor zum ersten Mal gemessen wird.

Reflexion: *Wer bei euch kann eine richtige Antwort verbindlich abnehmen, und hat diese Person dafür Zeit im Kalender? Welches Wissen liegt heute nur in Köpfen und könnte als Golden-Set schriftlich werden?*

Schritt 5 · Lass den Prüflauf ohne Erinnerung laufen (B)

Zwei Prüfungen gehören zusammen. Die erste läuft vor jeder Änderung: das Golden-Set gegen die neue Version, mit einer klaren Schwelle. Sie beantwortet die Frage „Darf diese Änderung raus?“. Die zweite läuft im Betrieb: Eine feste Stichprobe echter Anfragen wird nachträglich bewertet. Sie beantwortet die Frage „Hält das System draußen, was der Test drinnen versprochen hat?“.

Beide haben eine blinde Stelle. Das Golden-Set kennt nur die Fälle, die jemand vorhergesehen hat. Der Betrieb zeigt die echten Fälle, aber ohne richtige Antwort. Zusammengeschaltet ergeben sie einen Kreislauf: Was im Betrieb auffällt, wird ein neuer Fall im Set, und das Set schützt die nächste Version davor, denselben Fehler noch einmal zu machen.

> Die Prüfung vor dem Start und der Blick aus dem Cockpit gehören zusammen. Wer nur eines hat, fliegt blind oder startet nie.

Werkzeug · Der Kreislauf in drei Teilen:

(1) Vor jeder Änderung läuft das Golden-Set mit einer harten Schwelle. Wird das Ergebnis schlechter, geht die Änderung nicht raus.

(2) Im Betrieb wird eine feste Quote echter Anfragen bewertet, Auffälligkeiten landen auf einer Prüfliste.

(3) Jeder echte Fehler wird zum neuen Fall im Golden-Set, bevor er als behoben gilt.

Falle & Fix: Falle, den Betriebsteil als Monitoring abzuhaken und sich mit Antwortzeiten und Kosten zufriedenzugeben. Die zeigen, ob das System läuft, nicht ob es richtig liegt.

Fix, mindestens ein Kriterium aus Schritt 4 auf einer Stichprobe echter Anfragen messen, auch wenn die Stichprobe klein ist.

Reflexion: *Welcher Fehler der letzten Wochen wurde behoben, ohne je zu einem dauerhaften Testfall zu werden? Wo gibt es bei euch eine Prüfung vor dem Start, aber keine Stichprobe danach?*

Schritt 6 · Kalibriere den Bewerter, bevor du ihm glaubst (C)

Wo es keine eindeutige richtige Antwort gibt, etwa bei Zusammenfassungen oder Formulierungen, lässt man gern ein zweites Modell bewerten. Das ist bequem und nützlich, und es ist die Stelle, an der am meisten Scheinsicherheit entsteht. Wer seine Prüfung automatisiert, hat ein zweites Messgerät ins Haus geholt. Dieses zweite Messgerät muss selbst geprüft werden.

Die Befundlage ist deutlich. Ein Forschungsteam der RAND Corporation hat im März 2026 vier aktuelle Bewerter-Modelle unter Stress gesetzt. Keines war über alle getesteten Aufgabenfelder hinweg verlässlich. Schon einfache Änderungen an der bewerteten Antwort, eine andere Formatierung, eine Umformulierung, mehr oder weniger Worte, haben die Urteile verschoben. Eine Arbeit auf der ACL 2026, einer der großen Fachkonferenzen für Sprachverarbeitung, zeigt am Beispiel der Bewertung von Argumenten eine zweite Schwäche: Bewerter-Modelle vermischen ihre im Training erworbenen Überzeugungen mit der Qualität dessen, was sie beurteilen sollen.

Der Ausweg ist nicht, auf Bewerter zu verzichten, dafür sind sie zu nützlich. Der Ausweg ist Kalibrierung: Der Bewerter wird gegen menschliche Urteile geprüft, bevor man ihm etwas glaubt, und er wird nur dort eingesetzt, wo die Übereinstimmung nachgewiesen ist.

> Ein Bewerter ist ein Zeuge, kein Richter.

Werkzeug · Die Judge-Kalibrierung in drei Schritten:

(1) Lass fünfzig Fälle von Menschen bewerten, danach denselben Satz vom Bewerter-Modell. Die Übereinstimmungsrate ist deine einzige Grundlage.

(2) Leg vorher fest, ab welcher Übereinstimmung du dem Bewerter traust. Liegt er darunter, setzt du ihn nicht ein, sondern überarbeitest die Bewertungsanweisung oder nimmst ein anderes Modell.

(3) Mach einen Stresstest: dieselbe Antwort einmal kürzer, einmal umformuliert, einmal anders formatiert. Ändert sich die Note, ist der Bewerter für diesen Einsatz nicht brauchbar.

Falle & Fix: Falle, eine Bewerter-Zahl als Qualitätsmaß weiterzugeben, weil sie so schön messbar aussieht.

Fix, jede Bewerter-Zahl nur zusammen mit ihrer Übereinstimmungsrate zum menschlichen Urteil berichten. Ohne diese zweite Zahl ist die erste eine Behauptung.

Reflexion: *Wo hast du zuletzt einer KI-Bewertung geglaubt, ohne zu prüfen, ob sie mit einem menschlichen Urteil übereinstimmt? Wer würde bei euch die fünfzig Vergleichsfälle bewerten?*

Schritt 7 · Halte das Set am Leben (C)

Ein eingefrorenes Golden-Set altert. Das Produkt ändert sich, die Nutzer stellen andere Fragen, die Modelle verschieben ihr Verhalten. Nach Erfahrungswerten aus der Praxis bildet ein Set, das niemand pflegt, nach sechs bis neun Monaten den Alltag nicht mehr ab. Ein lebendes Set bekommt laufend neue Fälle aus dem Betrieb, und alte Fälle werden bewusst aussortiert.

Dazu gehört der vierte Baustein. Ein System ist nicht erst dann zuverlässig, wenn es im Normalfall gute Antworten gibt, sondern wenn es standhält, sobald es jemand darauf anlegt. Jeder Angriff, der durchkommt, wird ein Testfall, genau wie jeder Fehler aus dem Betrieb.

Werkzeug · Der Datensatz-Check in vier Fragen:

(1) Könnte ein Testfall im Trainingsmaterial des Modells gelegen haben, etwa weil er aus einer öffentlichen Quelle stammt? Dann misst er Auswendiglernen statt Können und wird ersetzt. Fälle aus euren eigenen, nicht öffentlichen Daten sind die saubere Messung.

(2) Wann kam der letzte neue Fall dazu? Liegt das mehr als ein Quartal zurück, altert das Set.

(3) Versteht jemand aus dem Fachbereich jedes Kriterium ohne Erklärung? Wenn nicht, ist es eine Zahl fürs Schaufenster.

(4) Wer hat die richtigen Antworten abgenommen, und war diese Person dafür qualifiziert?

Falle & Fix: Falle, das Set einmal sorgfältig zu bauen und dann als fertig zu betrachten.

Fix, eine Person benennen, die das Set pflegt, und einen festen Termin im Quartal, an dem neue Fälle dazukommen und überholte herausfallen.

Reflexion: *Wie oft würdet ihr ein Set realistisch nachpflegen, und wer wäre dafür verantwortlich? Welche eurer KI-Anwendungen hat noch nie jemand bewusst anzugreifen versucht?*

Belegen statt glauben

Ein funktionierender Prototyp ist ein guter Anfang und ein schlechter Beweis. Den Unterschied machen vier Bausteine: ein Golden-Set aus echten Fällen mit Kriterien, die der Fachbereich versteht, ein Prüflauf, der ohne Erinnerung läuft, ein Bewerter, dem du erst glaubst, wenn er selbst geprüft ist, und ein Angriffstest, bevor es jemand anderes tut. In vielen Häusern fällt genau diese Aufgabe zwischen IT und Fachbereich durch: Die IT baut, der Fachbereich nutzt, und die Frage, wer die Qualität verantwortet, stellt niemand. Wer sie stellt, hat den wichtigsten Schritt schon gemacht.

Werkzeug · Dein Schritt für diese Woche:

(1) Nimm einen einzigen KI-Handgriff, den ihr schon produktiv nutzt.

(2) Schreib zwanzig echte Fälle mit ihrer richtigen Antwort in eine Tabelle.

(3) Lass sie einmal gegen die aktuelle Version laufen und markier jeden Fehlschlag.

(4) Lies die roten Fälle einzeln und notier, was sie gemeinsam haben.

Das erste rote Ergebnis ist kein Rückschlag, sondern der eigentliche Gewinn: Ab jetzt hast du einen Vorher-Wert, gegen den jede Änderung antreten muss. Work in progress ist dabei völlig in Ordnung.

Quellen

- RAND Corporation, Stresstest von vier Bewerter-Modellen über mehrere Aufgabenfelder, veröffentlicht im März 2026: arxiv.org/abs/2603.05399
- Findings of the Association for Computational Linguistics, ACL 2026: „Prior Beliefs Prejudice LLM-as-Judge. Evidence from Persuasion Evaluation“: aclanthology.org/2026.findings-acl.2087
- Der Fünf-Schritte-Aufbau, die Trennung von Prüfung vor dem Start und Stichprobe im Betrieb und die Alterung eines Sets nach sechs bis neun Monaten sind verdichtete Praxis aus veröffentlichten Erfahrungsberichten, keine Studienergebnisse.

Wenn du weiterkommen willst

Diese Unterlage ist ein Werkzeug, kein Ersatz für ein Gespräch. Wenn beim Durcharbeiten etwas hakt, wenn ein Schritt auf deine Lage nicht passt oder du wissen willst, wie das bei euch konkret aussähe: **schreib mir**. Antworte einfach auf die Nachricht, mit der du das hier bekommen hast — oder melde dich über aivantum.com. Ich lese und antworte selbst.

Diese Unterlage wurde mit KI-Unterstützung erstellt und redaktionell geprüft. Weitergabe zum persönlichen Gebrauch erlaubt, gewerbliche Weiterverbreitung nicht.