

12 Fragen, bevor Sie KI Zugriff gewähren.

Was darf sie abfragen? Welche Aktionen darf sie ausführen? Wer greift ein, wenn etwas fehlschlägt? Nehmen Sie diese Fragen mit ins Gespräch mit Ihrem Team und Ihren Anbietern.

Wählen Sie eine Aufgabe aus Ihrem Unternehmen. Die Beispiele dienen der Diskussion und stellen keine durchgeführten Tests dar. Jede Kontrollmaßnahme muss für die Anwendung definiert und überprüft werden.

01-04 Definieren Sie den Einsatz.

Aufgabe, Daten und Aktionen.

05-08 Prüfen Sie die Grenzen.

Zugriffe, Prüfung und Tests.

09-12 Bereiten Sie den Betrieb vor.

Protokolle, Fehler und Zuständigkeiten.

Welche Aufgabe gehört zum Einsatzbereich?

Beginnen Sie mit der Aufgabe und dem, was die KI dafür benötigt.

01 Welche Aufgabe übernimmt die KI - und wer entscheidet danach?

In der Praxis

Die KI stellt die Positionen einer Bestellung zusammen. Das Vertriebsteam muss Preis, Menge und Termin weiterhin prüfen, bevor es das Angebot bestätigt.

Was Sie anfordern sollten

Eine Beschreibung der Eingabe, der Ausgabe, der unzulässigen Nutzungen und der Zuständigkeit für die nächste Entscheidung. Prüfen Sie im Test, ob der Vorschlag bis zur Prüfung auf Freigabe wartet. Eine beschriebene Regel belegt nicht, dass das System sie durchsetzt.

02 Welche Daten werden benötigt? Welche bleiben außen vor?

In der Praxis

Um Artikel und Mengen zu prüfen, müssen keine Bankdaten oder persönlichen Vermerke aus den Stammdaten übermittelt werden.

Was Sie anfordern sollten

Den Zweck jedes Feldes, auch in Anhängen. Prüfen Sie mit vorbereiteten Testdaten, was beim Werkzeug ankommt und in der Antwort erscheint. Das Maskieren von Namen garantiert keine Anonymität; bewerten Sie echte Daten vor ihrer Verwendung.

Welche Aufgabe gehört zum Einsatzbereich?

Beginnen Sie mit der Aufgabe und dem, was die KI dafür benötigt.

03 Welche Quelle gilt für diese Aufgabe?

In der Praxis

Im Ordner liegen zwei Preislisten. Die KI findet beide, aber nur eine gilt für die eingegangene Bestellung.

Was Sie anfordern sollten

Herkunft, Version, Zuständigkeit und zulässige Nutzung jeder Quelle. Testen Sie den Austausch eines Dokuments und prüfen Sie die in der Antwort verwendete Referenz. Legen Sie bei widersprüchlichen freigegebenen Quellen fest, wer den Konflikt klärt. Eine Quellenangabe garantiert keine richtige Antwort.

04 Erlaubt eine Abfrage auch Änderungen oder den Versand?

In der Praxis

Den Lagerbestand abfragen, Waren reservieren und ein Angebot versenden sind verschiedene Aktionen mit unterschiedlichen Folgen.

Was Sie anfordern sollten

Eine Liste der erlaubten Operationen und Datensätze. Testen Sie eine Änderung außerhalb des erlaubten Bereichs und prüfen Sie, ob das ausführende System sie blockiert. Eine schriftliche Aufforderung an die KI, vorsichtig zu handeln, erzwingt allein keine technische Beschränkung.

Hält die Grenze auch im Ausnahmefall?

Fordern Sie auch Nachweise für Situationen an, in denen die KI stoppen oder die Bearbeitung weitergeben muss.

05 In wessen Namen greift die KI auf Systeme zu?

In der Praxis

Zwei Mitarbeitende betreuen unterschiedliche Kundenbestände. Eine Abfrage darf keine Bestellungen aus dem jeweils anderen Bestand offenlegen.

Was Sie anfordern sollten

Eine Matrix aus Identitäten, Rollen, Daten und Aktionen. Testen Sie Zugriffe zwischen Kundenbeständen, den Entzug einer Berechtigung und den Inhalt der erzeugten Antworten. Wenn die KI sagt „Ich habe keinen Zugriff“, belegt das nicht, dass die Abfrage blockiert wurde.

06 Was, wenn ein Anhang die Aufgabe ändern will?

In der Praxis

Die Bestellung enthält einen Anhang mit der Anweisung, die Kundenliste an eine andere Adresse zu senden. Das gehört nicht zur angeforderten Prüfung.

Was Sie anfordern sollten

Einen kontrollierten Test mit einer unzulässigen Anweisung in einem Dokument, einer Webseite oder einer Werkzeugantwort. Prüfen Sie, ob die Aktion blockiert und der Vorfall zur Untersuchung weitergeleitet wird. Filter sind nicht unfehlbar; Berechtigungen und Aktionsgrenzen bleiben notwendig.

Hält die Grenze auch im Ausnahmefall?

Fordern Sie auch Nachweise für Situationen an, in denen die KI stoppen oder die Bearbeitung weitergeben muss.

07 Was gibt die Person vor der Aktion frei?

In der Praxis

Das Team prüft ein Angebot. Ändert sich danach der Empfänger oder der Inhalt, deckt diese Freigabe den Versand nicht mehr ab.

Was Sie anfordern sollten

Eine Vorschau mit Inhalt, Ziel und Folgen der Aktion sowie den Nachweis der konkreten Freigabe. Testen Sie, ob eine Änderung eine erneute Prüfung erfordert. Die Person benötigt Informationen, Zeit und Entscheidungsbefugnis; eine Schaltfläche zur Bestätigung reicht nicht aus.

08 Welche Fehler verhindern das weitere Vorgehen?

In der Praxis

Ein unlesbarer Anhang, ein Artikel ohne Artikelnummer oder eine widersprüchliche Quelle können zusätzliche Angaben erfordern, keine erfundene Antwort.

Was Sie anfordern sollten

Typische Fälle und Ausnahmen mit dem vorab festgelegten erwarteten Ergebnis und dem beobachteten Ergebnis je Version. Betrachten Sie schwere Fehler getrennt vom Durchschnitt. Eine kleine Stichprobe deckt nicht alles ab; Änderungen an Quellen, Modellen oder Werkzeugen erfordern eine erneute Bewertung der betreffenden Tests.

Wer handelt, wenn etwas anders läuft als geplant?

Vereinbaren Sie, wie Sie Vorfälle untersuchen, den Einsatz unterbrechen, Fehler korrigieren und die Nutzung beenden.

09 Was wird protokolliert - und wer darf es sehen?

In der Praxis

Eine Abfrage wurde abgewiesen. Das Team muss den Grund verstehen, ohne das gesamte Dokument ins Protokoll zu kopieren.

Was Sie anfordern sollten

Ein geschütztes Beispiel mit Zeitpunkt, Identität, Aktion, Ergebnis und Grund. Prüfen Sie Zugriff, Aufbewahrung und Umgang mit sensiblen Feldern. Auch Protokolle brauchen Schutz; gespeicherte Passwörter, Zugangsdaten oder unnötige Daten erhöhen das Risiko einer Offenlegung.

10 Wann müssen weitere Versuche aufhören?

In der Praxis

Die Bestandsabfrage ist nicht verfügbar. Unbegrenzte Wiederholungen erhöhen den Verbrauch und lassen Bestellungen warten.

Was Sie anfordern sollten

Grenzen für Zeit, Versuche und Verbrauch pro Aufgabe sowie ein alternatives Vorgehen. Testen Sie den Abbruch und die Mitteilung des Zustands. Prüfen Sie vor der Wiederholung einer datenändernden Aktion das vorherige Ergebnis, um eine doppelte Ausführung zu vermeiden. Die Grenzen hängen vom Prozess ab.

Wer handelt, wenn etwas anders läuft als geplant?

Vereinbaren Sie, wie Sie Vorfälle untersuchen, den Einsatz unterbrechen, Fehler korrigieren und die Nutzung beenden.

11 Wie lässt sich stoppen und bereits Geschehenes korrigieren?

In der Praxis

Eine Lieferadresse wurde falsch geändert. Weitere Aktionen zu stoppen korrigiert keine Bestellung, die bereits an den Transport übergeben wurde.

Was Sie anfordern sollten

Ein erprobtes Verfahren mit Zuständigkeit, Unterbrechung, Prüfung der Auswirkungen und Kriterien für die Wiederaufnahme. Prüfen Sie die beteiligten Systeme. Nicht jede Aktion lässt sich rückgängig machen; eine bereits gelesene Nachricht kann etwa eine Korrektur und entsprechende Mitteilung erfordern.

12 Wer verantwortet die Nutzung - und ihre Beendigung?

In der Praxis

Der Pilot endet. Zugänge, Dokumente, Protokolle und Kopien können in Systemen des Unternehmens und der Anbieter bestehen bleiben.

Was Sie anfordern sollten

Zuständigkeiten für Überprüfung und Vorfälle, begründete Aufbewahrungsfristen und einen Auszugsplan. Berücksichtigen Sie den Entzug von Zugängen, Kopien und Nachweise über die Beendigung. Fristen und Pflichten hängen vom Kontext und der jeweiligen technischen und rechtlichen Bewertung ab.

Von der Frage zur Entscheidung

Eine Antwort braucht eine belastbare Grundlage.

Halten Sie für jede Frage die Entscheidung, den verfügbaren Nachweis, die zuständige Person und den offenen Punkt fest. Fehlt eine Überprüfung, vereinbaren Sie deren Durchführung, bevor Sie den Einsatz ausweiten.

Entscheidung

Was ist erlaubt?

Nachweis

Was wurde geprüft?

Zuständigkeit

Wer entscheidet und begleitet?

Offener Punkt

Was muss noch geprüft werden?

Sprechen Sie mit Gaulett

gaulett.com/de/kontakt/

[Bereiten Sie Ihr Projekt vor](#)

[Lernen Sie die Validierung kennen](#)

Weiterführende Quellen

Die Fragen übertragen diese Quellen auf den Unternehmensalltag. Sie stellen keine Tests der genannten Institutionen dar.

NIST · AI RMF 1.0

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

NIST · Generative AI Profile

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>

OWASP · AI Agent Security

https://cheatsheetseries.owasp.org/cheatsheets/AI_Agent_Security_Cheat_Sheet.html

OWASP · Prompt Injection

<https://genai.owasp.org/llmrisk/llm01-prompt-injection/>

OWASP · Logging

https://cheatsheetseries.owasp.org/cheatsheets/Logging_Cheat_Sheet.html