



Kapitel 3: Prüfen und widerlegen

Wie du eine Antwort unter Last setzt, auch wenn du vom Thema weniger verstehst.

- 1 Eine Antwort unter Last setzen
- 2 Die Quellenprobe
- 3 Gegen sich selbst antreten lassen
- 4 Zahlen, die keiner nachgerechnet hat
- 5 Meisterstück: ein prüfbares Belegdossier

Lektion 1 von 5

Eine Antwort unter Last setzen

Ziel: Ich setze eine Antwort unter Last, statt sie zu glauben oder zu verwerfen.

Nachfragen ist keine Prüfung

„Bist du sicher?“ ist keine Gegenprobe. Die Frage erzeugt entweder eine Entschuldigung und eine geänderte Antwort oder eine Bekräftigung. Beides sagt nichts über die Sache.

Eine belastbare Prüfung macht aus dem Zweifel eine Aufgabe: eine Gegenposition, die Bedingung, unter der die Aussage falsch wäre, eine Quelle oder eine Zahl mit Fundstelle. Erst damit lässt sich außerhalb der Antwort nachsehen.

Merksatz

Frag nicht nach Sicherheit. Frag nach dem, was die Aussage widerlegen würde.

Behauptungen einzeln prüfbar machen

Eine Antwort wird in drei Sorten Aussagen zerlegt: Tatsachen, Schlussfolgerungen und Empfehlungen. Zu jeder prüfbaren Behauptung kommt ein möglicher Gegenbefund.

Die Grenze: Wertungen und Prognosen lassen sich nicht immer endgültig widerlegen, müssen dann aber Annahmen und Unsicherheit sichtbar machen.

Merksatz

Formuliere für jede tragende Behauptung, welcher konkrete Befund sie schwächen oder widerlegen würde.

Behauptungen einzeln prüfbar machen: wo abgekürzt wird

Am häufigsten wird hier abgekürzt mit: „Prüfe nur die Kernaussage, weil die übrigen Sätze daraus logisch folgen.“ Klingt gründlich, ist es nicht. Sie folgen selten daraus. Meist liegen zwischen Kernaussage und Schluss zwei Zwischenschritte, die niemand geprüft hat.

An dieselbe Stelle gehört stattdessen: „Listet jede Aussage einzeln auf und ordne sie als Tatsache, Schluss oder Empfehlung ein.“ Das kannst du befolgen und am Ergebnis nachsehen.

Merksatz

Eine Antwort ist so stark wie ihre schwächste ungeprüfte Behauptung.

Lektion 2 von 5

Die Quellenprobe

Ziel: Ich prüfe Quellenangaben, ohne dem Modell die Prüfung zu überlassen.

Erfundene Quellen sehen echt aus

Eine halluzinierte Quellenangabe hat oft plausiblen Autor, Jahr, Titel und sogar eine formal wirkende DOI. Sprachliche Plausibilität trennt sie nicht von einer echten Quelle.

Die Selbstauskunft des Modells ist nur ein Hinweis. Verlange nachvollziehbare Toolnachweise, öffne die Originalstelle und prüfe, ob sie die konkrete Aussage wirklich trägt.

Merksatz

Ein Quellenname oder Link ist kein Beleg. Entscheidend sind Toolnachweis, geöffnete Originalstelle und passende Aussage.

Quellenhierarchie in 4 Prüfschritten

Eine Quellenhierarchie bevorzugt für eine Behauptung die nächstgelegene Primärquelle und prüft zusätzlich Datum, Geltungsbereich und konkrete Fundstelle.

Die Grenze: Primärquellen sind nicht automatisch fehlerfrei oder verständlich; für Einordnung können gute Sekundärquellen nötig sein.

Merksatz

Öffne die geltende Originalfassung und gleiche Wortlaut, Datum, Adressaten und Ausnahme mit der konkreten Behauptung ab.

Quellenhierarchie in 4 Prüfschritten: die bequeme Zeile und die bessere

Am häufigsten wird hier abgekürzt mit: „Übernimm die Zusammenfassung eines Suchtreffers, wenn sie die Primärquelle verlinkt.“ Klingt gründlich, ist es nicht. Dann zitierst du die Zusammenfassung, nicht die Quelle. Einschränkungen fallen auf diesem Weg als Erstes weg.

An dieselbe Stelle gehört stattdessen: „Nenn zu jeder Aussage die Originalquelle mit Datum und Geltungsbereich.“ Das kannst du befolgen und am Ergebnis nachsehen.

Merksatz

Zitiert wird die Quelle, nicht die Zusammenfassung, die auf sie zeigt.

Lektion 3 von 5

Gegen sich selbst antreten lassen

Ziel: Ich lasse eine Antwort gegen sich selbst antreten, bevor ich sie benutze.

Zwei Durchgänge, zwei Rollen

Die Gegenprobe liefert eine zweite simulierte Perspektive: Du lässt dieselbe Frage beantworten und den Text dann in einem neuen Chat angreifen.

Das kann Bestätigungsneigung mindern, macht aus demselben Modell aber keine unabhängige Fachperson. Gemeinsame Modellfehler können in beiden Durchgängen bleiben.

Merksatz

Getrennte Chats trennen Perspektiven. Wichtige Aussagen prüfst du zusätzlich an Originalquellen oder unabhängig fachlich.

Unabhängige Gegenprobe mit zweiter Methode

Eine unabhängige Gegenprobe wechselt Datenquelle, Methode oder fachliche prüfende Person und vergleicht vorher festgelegte Schlüsselaussagen.

Die Grenze: Ein zweiter Chat desselben Modells kann neue Einwände liefern, teilt aber Training, typische Fehler und unbekannte Systemannahmen.

Merksatz

Nutze für die tragende Aussage eine andere Quelle oder Methode und kläre jede Abweichung am Original.

Unabhängige Gegenprobe mit zweiter Methode: wo abgekürzt wird

Am häufigsten wird hier abgekürzt mit: „Frag dasselbe Modell in einem neuen Chat und behandle Übereinstimmung als unabhängige Bestätigung.“ Klingt gründlich, ist es nicht. Ein neuer Chat wechselt den Kontext, nicht die Quelle. Dieselben Trainingsdaten liefern denselben Irrtum ein zweites Mal.

An dieselbe Stelle gehört stattdessen: „Prüf diese eine Aussage ausschließlich anhand der Passage, die ich dir gebe.“ Das kannst du befolgen und am Ergebnis nachsehen.

Merksatz

Zweimal dasselbe Modell ist eine Wiederholung, keine Bestätigung.

Lektion 4 von 5

Zahlen, die keiner nachgerechnet hat

Ziel: Ich behandle jede Zahl aus einer KI-Antwort als ungeprüft, bis ich sie belegt habe.

Rechnen und Text sind zwei verschiedene Dinge

Ein Sprachmodell ohne Werkzeug erzeugt Text. Eine Rechnung entsteht dabei als Text, der wie eine Rechnung aussieht. Das geht oft gut und manchmal daneben, ohne dass sich am Auftreten etwas ändert.

Hängt am Dienst ein Rechner oder eine Codeausführung, wird wirklich gerechnet. Geprüft werden dann vier Dinge einzeln: die Eingabedaten, die Formel oder das Programm, ob das Werkzeug überhaupt lief und das Ergebnis.

Merksatz

Eine Zahl ohne offengelegten Rechenweg ist eine Behauptung. Auch ein Werkzeug rechnet nur, was es bekommt.

Reproduzierbare Rechnung mit Code und 5 Prüfpunkten

Für mehrstufige Rechnungen erzeugt Code ein wiederholbares Verfahren; Eingaben, Einheiten, Formel, Rundung und Ausgabe werden getrennt sichtbar.

Die Grenze: Ausführbarer Code garantiert nur die Umsetzung; falsche Daten, Einheiten oder fachliche Formeln liefern weiterhin präzise falsche Ergebnisse.

Merksatz

Prüfe Einheiten, Nenner, fehlende Werte und mindestens einen von Hand berechneten Kontrollfall.

Reproduzierbare Rechnung mit Code und 5 Prüfpunkten: wo abgekürzt wird

Am häufigsten wird hier abgekürzt mit: „Vertrau dem Ergebnis, sobald der Code ohne Fehlermeldung ausgeführt wurde.“ Klingt gründlich, ist es nicht. Ohne Fehlermeldung heißt: Der Code lief. Ob er die richtigen Zahlen und die richtige Formel benutzt hat, sagt das nicht.

An dieselbe Stelle gehört stattdessen: „Gib Eingaben, Einheiten und Formel getrennt aus, bevor du rechnest.“ Das kannst du befolgen und am Ergebnis nachsehen.

Merksatz

Dass Code läuft, heißt nicht, dass er das Richtige rechnet.

Sensitivität statt einzelner Punktzahl

Teste bei unsicheren Eingaben mindestens 3 plausible Werte und beobachte, ob die Entscheidung dadurch kippt.

Die Grenze: Eine Sensitivitätsanalyse zeigt Abhängigkeiten, liefert aber keine Wahrscheinlichkeit für die gewählten Annahmen.

Merksatz

Wenn kleine Eingabeänderungen das Urteil drehen, ist die Unsicherheit Teil des Ergebnisses.

Lektion 5 von 5

Meisterstück: ein prüfbares Belegdossier

Ziel: Ich zerlege ein wichtiges KI-Ergebnis und sichere seine tragenden Aussagen mit unabhängigen Quellen und Rechenwegen ab.

Was du hier baust

In 30 bis 45 Minuten baust du ein Belegdossier mit zerlegten Behauptungen, Originalfundstellen, Gegenprobe und reproduzierbarer Rechnung.

Die KI verarbeitet und kritisiert; du öffnest die Originale, bewertest die Gegenbelege und entscheidest selbst, welche Aussagen übernommen werden.

Merksatz

Ein Dossier belegt nur, was du selbst im Original nachgesehen hast.

Das nimmst du mit

- Ich setze eine Antwort unter Last, statt sie zu glauben oder zu verwerfen.
- Ich prüfe Quellenangaben, ohne dem Modell die Prüfung zu überlassen.
- Ich lasse eine Antwort gegen sich selbst antreten, bevor ich sie benutze.
- Ich behandle jede Zahl aus einer KI-Antwort als ungeprüft, bis ich sie belegt habe.
- Ich zerlege ein wichtiges KI-Ergebnis und sichere seine tragenden Aussagen mit unabhängigen Quellen und Rechenwegen ab.

Eine KI mit Websuche nennt dir eine Studie mit Link. Was gilt?

- A** Mit Websuche sind erfundene Quellenangaben ausgeschlossen, weil jede Angabe aus einem echten Treffer stammt
- B** Der Link belegt nur eine erreichbare Seite; die behauptete Aussage braucht zusätzlich eine konkrete Fundstelle im Originaldokument
- C** Ein erreichbarer Link belegt Existenz und Aussage der Studie, weil die Websuche nur gefundene Seiten zitiert

Auflösung

Richtig ist B: Der Link belegt nur eine erreichbare Seite; die behauptete Aussage braucht zusätzlich eine konkrete Fundstelle im Originaldokument

Suchende Modelle rufen Seiten auf und fassen zusammen. Dabei kann die Zusammenfassung über das hinausgehen, was auf der Seite steht, und der Link stimmt trotzdem.