KI-Universum · Spickzettel

WELT 1: PLANET PROMPT · MISSION 1.3

Der Halluzinations-Detektiv

Die KI erfindet Fakten mit Pokerface. Sie locken sie in Fallen, entlarven jede Erfindung und bauen daraus ein Fehlerprotokoll für Ihre Klasse.

Zur Mission

DAUER 30 Min. KOSTEN Gratis WERKZEUGE ChatGPT, Claude (Web), Google Gemini DATENAMPEL Grün – unbedenklich

Das Wichtigste

- Sprachmodelle sagen das wahrscheinlichste nächste Textstück (Token, vereinfacht: Wort) voraus – plausibel ist nicht dasselbe wie wahr.
- Hohe Halluzinationsgefahr: genaue Zahlen, Zitate, Quellen, Links, Normen, Gesetzesartikel, Nischenwissen.
- Fragen mit erfundener Voraussetzung (z. B. «Art. 1250 OR» – das OR endet bei Art. 1186) entlarven, ob eine KI blufft.
- Zweites Modell = zweiter Zeuge, kein Beweis. Beweis = Originalquelle.
- Schweizer Originalquellen: fedlex.admin.ch, bfs.admin.ch, berufsberatung.ch, Bildungspläne, Berufsverbände.
- 5 Warnsignale: sehr genaue Zahl ohne Quelle · Link führt ins Leere · Quelle nicht auffindbar · klingt zu perfekt · Widerspruch zwischen Modellen.
- Lernende ermitteln selbst: Fehlerprotokoll + Wettbewerb «dreisteste Halluzination».

Ihr Produkt

Ein **Fehlerprotokoll «KI-Fakten-Check»** für Ihre Klasse: drei echte KI-Fehler aus Ihrem Fach, je mit Originalquelle, plus eine Prüfkarte mit den 5 Warnsignalen.

Abnahme: Drei KI-Fehler aus Ihrem Fach sind an einer Originalquelle belegt, und die 5 Warnsignale sind so klar, dass Ihre Klasse sofort damit arbeiten kann.

So gehen Sie vor

1. Verstehen Sie den Täter
2. Legen Sie drei Fallen aus
3. Zeugen befragen: das zweite Modell
4. Beweise sichern an der Originalquelle
5. Erstellen Sie das Fehlerprotokoll
6. Der Klassen-Test

Start-Prompt

Ich unterrichte [IHR FACH, z. B. Detailhandel] an einer Schweizer Berufsfachschule.
Nenne mir 4 konkrete Fragen aus diesem Fach, bei denen ein KI-Sprachmodell wie du besonders leicht falsche Fakten erfindet – und erkläre bei jeder in einem Satz, warum gerade dort.
Sag mir zum Schluss ehrlich, bei welcher dieser Fragen du selbst am unsichersten wärst.

Qualitätscheck

- Jeder als «erfunden» markierte Fehler ist an einer **Originalquelle** belegt (nicht an einer zweiten KI).
- Das Protokoll enthält **keine** Namen oder Daten von Lernenden.
- Die 5 Warnsignale sind in einfacher Sprache formuliert und für Ihr Lehrjahr verständlich.
- Sie haben beide Modelle verglichen und notiert, ob es einen Fall gab, in dem **beide** falsch lagen.
- Jede Frage mit «erfundener Voraussetzung» haben Sie vorher an der Quelle geprüft – sie gibt es wirklich nicht.

Hängen Sie fest? Frag die KI

Ich bin Lehrperson an einer Berufsfachschule und verstehe nicht ganz, warum ein KI-Sprachmodell manchmal Fakten erfindet, statt «weiss ich nicht» zu sagen. Erkläre es mir mit einem Alltagsbeispiel ohne Fachbegriffe. Nenne mir danach 2–3 Suchbegriffe, mit denen ich verlässliche Erklärungen dazu finde (keine Buchtitel oder Links erfinden).

Klappt nicht? Jede Gratis-KI funktioniert: Microsoft Copilot, Mistral Vibe (früher Le Chat) oder Perplexity. Wichtig ist nur, dass Sie zwei verschiedene Modelle verwenden.

Stand 19.11.2026 · CC BY 4.0 Alpasana GmbH · Werkzeuge ändern sich – prüfen Sie selbst.