WELT 5: DIE MODELL-FLOTTE · MISSION 5.2

Das Modell-Duell als Blindverkostung

Vier Modelle, ein Auftrag, keine Etiketten. Ihre Jury bewertet blind nach Kriterien – und erlebt beim Aufdecken garantiert eine Überraschung.

Zur Mission

DAUER 55 Min. KOSTEN Gratis mit Konto WERKZEUGE ChatGPT, Claude (Web), Google Gemini, Microsoft Copilot, Mistral Le Chat
DATENAMPEL Grün – unbedenklich

Das Wichtigste

- Blind bewerten: erst Kriterien, dann Antworten, erst zum Schluss Modellnamen.
- Gleicher Prompt, neuer Chat, erste Antwort zählt – sonst ist der Vergleich unfair.
- Anonymisieren heisst Herkunftsspuren entfernen, nicht Inhalte verbessern.
- Fakten prüfen: Eleganz ersetzt keine fachliche Richtigkeit.
- Dasselbe Modell schwankt – ein Durchgang ist ein Hinweis, kein Urteil.
- Hausregel pro Aufgabentyp: zuerst Modell X, Gegenlesen mit Modell Y – regelmässig wiederholen.

Ihr Produkt

Ein ausgefüllter **Jury-Bogen** mit Kriterien, Punkten und Aufdeckung – und Ihre **Hausregel**: welches Modell Sie für diesen Aufgabentyp künftig zuerst fragen und welches zum Gegenlesen.

Abnahme: Der Jury-Bogen enthält vier vor dem Lesen festgelegte Kriterien mit Beschreibung, Punkte für jede anonyme Antwort, das aufgedeckte Ergebnis und eine Regel, welches Modell Sie künftig zuerst fragen und welches zum Gegenlesen.

So gehen Sie vor

1. Ihr Tipp, bevor es losgeht
2. Kriterien zuerst – die Jury vereidigen
3. Gleicher Auftrag an alle
4. Etiketten ab: anonymisieren und mischen
5. Die Jury tagt
6. Aufdecken – und die Hausregel

Start-Prompt

Ich unterrichte [IHR FACH, z. B. Pflege] im [LEHRJAHR, z. B. 1. Lehrjahr] an einer Schweizer Berufsfachschule.
Schlage mir 5 Aufträge aus meinem Alltag vor, die sich gut für einen fairen Vergleich mehrerer KI-Modelle eignen: kurz (Ergebnis höchstens eine A4-Seite), ohne Personendaten, mit einem Ergebnis, dessen Qualität ich gut beurteilen kann.
Nenne zu jedem Auftrag, woran man eine gute Antwort erkennt.
Schweizer Rechtschreibung (immer «ss», kein Eszett).

Qualitätscheck

- Die vier Kriterien standen fest, bevor eine Antwort gelesen wurde – mit Beschreibung für die Punktzahlen.
- Alle Modelle erhielten exakt denselben Prompt in einem neuen Chat, ohne Nachbesserung.
- Die Antworten waren für die Jury nicht einem Modell zuzuordnen.
- Fachliche Aussagen wurden geprüft, nicht nur der Stil bewertet.
- Die Hausregel ist vorsichtig formuliert und nennt den Aufgabentyp.

Hängen Sie fest? Frag die KI

Ich bin Lehrperson an einer Schweizer Berufsfachschule und will mehrere KI-Modelle mit demselben Auftrag vergleichen. Welche Arten von Aufträgen eignen sich für einen fairen Vergleich und welche nicht? Erkläre es mit je einem Beispiel.

Klappt nicht? Drei Modelle genügen. Wer nur ein Konto hat, nimmt ChatGPT ohne Anmeldung (eingeschränkt möglich), Microsoft Copilot oder Mistral Vibe dazu. Ganz allein? Dann lassen Sie eine Kollegin die Antworten anonymisieren und mischen – oder Sie mischen selbst und bewerten erst am nächsten Tag.

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