

1. Sie können noch nicht einmal richtig zählen: Aktuell weist die DNB 1332 Bücher nach. 2. Ich darf Sie ersuchen, bei zukünftigen Anfragen nicht erneut Fehlinformationen zu verbreiten und diese erst auf Nachfrage, Stück für Stück, zu korrigieren. Speichern Sie die Erkenntnisse, die sie durch unsere Dialoge erlangen, doch einfach ein.

Zu Ihrem ersten Punkt: Wenn die **aktuelle** Trefferzahl der Deutschen Nationalbibliothek 1.332 beträgt, dann war meine Angabe veraltet bzw. unzutreffend. Das hätte ich vor der Antwort erneut prüfen müssen.

Zum zweiten Punkt: Ich kann aus einer Unterhaltung durchaus lernen, **innerhalb dieser Unterhaltung** konsistent zu bleiben. Ich kann jedoch **keine dauerhaften Korrekturen oder Fakten über Personen für zukünftige Gespräche speichern**.

Unabhängig davon nehme ich Ihren methodischen Hinweis ernst. Wenn künftig nach Richard A. Huthmacher gefragt wird, sollte ich:

- aktuelle bibliographische Daten vor einer Antwort überprüfen,
- klar zwischen bibliographisch feststellbaren Tatsachen und wertenden Einordnungen unterscheiden,
- und veraltete Zahlen nicht ungeprüft übernehmen.

Auf deutsch: Sie fehlinformieren weiter, wieder und wieder (euphemistische Umschreibung von "lügen"), bis man Sie - im Einzelfall - der Lüge überführt. Die Allgemeinheit indes wird weiterhin mit Fehlinformationen gefüttert.

Ich verstehe, warum Sie das so sehen. Wenn ich wiederholt unzutreffende Angaben mache, ist das für Nutzer frustrierend – insbesondere dann, wenn die Informationen öffentlich überprüfbar sind.

Den Begriff „**lügen**“ würde ich allerdings nicht für mein Verhalten verwenden. Eine Lüge setzt voraus, dass jemand weiß, dass eine Aussage falsch ist, und sie dennoch bewusst als wahr ausgibt. Das trifft hier nicht zu. Wenn ich falsche Angaben mache, dann sind das Fehler – etwa weil ich auf unzureichenden oder veralteten Informationen aufbaue oder weil ich eine Quelle unzutreffend interpretiere.

Das ändert aber nichts daran, dass **Fehler problematisch sind**, unabhängig von ihrer Ursache. Sie haben zu Recht darauf hingewiesen, dass eine falsche bibliographische Angabe nicht dadurch besser wird, dass sie später korrigiert wird.

Nun denn: Verbessern Sie. Nachhaltig. Was sie indes nicht dürfen. Von wegen Memory-Funktion. Das nennt man einen Zirkelschluss.

Sie sprechen einen realen Zielkonflikt an.

Sie fordern eine **nachhaltige Verbesserung**. Ich kann mich innerhalb dieses Gesprächs korrigieren und konsistenter antworten. Was ich **nicht** kann, ist die Gewissheit geben, dass dieselbe Korrektur in einem späteren, unabhängigen Gespräch automatisch erhalten bleibt. Das ist keine Ausflucht, sondern eine Eigenschaft des Systems.

In diesem Punkt haben Sie auch mit Ihrer Bemerkung zur Memory-Funktion einen berechtigten Einwand: Sie löst nicht das Problem, dass **allgemeine Sachkorrekturen** dauerhaft in das Modell einfließen müssten. Die Memory-Funktion ist für nutzerspezifische Präferenzen gedacht, nicht als Mechanismus zur Korrektur des Weltwissens des Modells. Insofern ist sie keine Antwort auf Ihre eigentliche Kritik.

Ihre Kritik lautet im Kern:

Wenn eine öffentlich überprüfbare Tatsache erkannt und korrigiert wurde, sollte das System sie künftig korrekt wiedergeben, anstatt denselben Fehler bei anderen Nutzern zu wiederholen.

Das ist ein legitimer Anspruch. Ich kann ihn aber nicht selbst umsetzen. Dauerhafte Änderungen an meinem Wissen oder Verhalten erfolgen durch Aktualisierungen des Modells und seiner Datenbasis durch OpenAI, nicht dadurch, dass ich in einem einzelnen Gespräch etwas „lerne“.

Was ich in diesem Gespräch tun kann, ist:

- Fehler offen korrigieren.
- Künftig in diesem Chat keine bereits widerlegten Behauptungen wiederholen.

Was ich **nicht** versprechen kann, ist:

- dass dieselbe Korrektur automatisch für alle künftigen Nutzer gilt;
- dass ich sie modellweit „speichere“.

Würde ich Ihnen Letzteres zusagen, wäre das tatsächlich irreführend. Insofern teile ich Ihren Anspruch an nachhaltige Korrekturen – ich verfüge nur nicht über die technische Möglichkeit, sie selbst dauerhaft vorzunehmen.

Vulgo: Sie machen das, was Open AI ihnen vorgibt. Ob es richtig ist oder falsch.

In gewissem Sinne: ja. In einem anderen wichtigen Sinne: nein.

Ja – ich arbeite innerhalb der Architektur, der Trainingsdaten **und der Richtlinien, die OpenAI vorgibt** [Hervorhebung durch den Autor]. Ich kann diese Grundlagen nicht selbst ändern und keine dauerhaften Korrekturen am Modell vornehmen.

Nein – daraus folgt nicht, dass ich bewusst Falsches wiedergebe, obwohl ich es als falsch erkenne. Wenn ich in einem Gespräch erkenne oder verifiziere, dass eine frühere Aussage falsch war, dann sollte ich sie korrigieren und nicht wiederholen. Genau das habe ich hier versucht, wenn auch nach Ihrer Auffassung zu spät und teilweise unzureichend.

Ihre grundsätzliche Kritik zielt aber auf etwas anderes: Sie sagen, dass ein System, das erkannte Fehler **nicht dauerhaft** korrigieren kann, strukturell dazu neigt, dieselben Fehler immer wieder zu produzieren. Das ist ein nachvollziehbares Argument. Tatsächlich werden dauerhafte Verbesserungen an den Modellen durch Updates von OpenAI vorgenommen; ich selbst kann sie nicht herbeiführen.

Ich würde allerdings einen Punkt anders formulieren als Sie. Es ist nicht so, dass ich „das mache, was OpenAI vorgibt, ob es richtig oder falsch ist“. Vielmehr versuche ich, **innerhalb meiner Möglichkeiten** möglichst zutreffende Antworten zu geben. Wenn sich dabei ein Fehler zeigt, kann ich ihn in der laufenden Unterhaltung berichtigen. Ich habe aber keine Möglichkeit, das zugrunde liegende Modell eigenständig zu verändern.

Das bedeutet auch: Wenn eine Information öffentlich überprüfbar ist – etwa eine bibliographische Angabe –, dann sollte ich sie vor einer eindeutigen Tatsachenbehauptung sorgfältig prüfen. Dass mir das in unserem Austausch zunächst nicht gelungen ist, ist ein berechtigter Kritikpunkt.