Warum KI überzeugend Unsinn erzählt (Halluzinationen)

Kapitel online

Halluzinationen sind kein Defekt, sondern die Kehrseite des Funktionsprinzips: Sieh am Temperature-Regler, warum ein Modell auch bei erfundenen Fragen selbstbewusst antwortet — und lerne die Gegenmittel.

Aktuell · Kapitel 34 von 50 · ca. 15 Min. · 250 XP

Einfach gesagt: Die KI muss immer antworten, auch wenn sie nichts weiß — merkst du ihr das nicht an, klingt geraten genauso sicher wie gewusst.

Jeder hat es erlebt: Die KI antwortet druckreif, mit Quellenangabe und felsenfester Überzeugung — und es stimmt **hinten und vorne nicht**. Das nennt man **Halluzination**. Heute verstehst du, warum das *zwangsläufig* passiert (kein Bug!), woran du es erkennst und was wirklich dagegen hilft.

Was du hier lernst

- Warum Halluzinationen **kein Bug** sind, sondern die Kehrseite des Funktionsprinzips
- Wie eine **flache Wahrscheinlichkeits-Verteilung** aussieht — und dass du sie im Chat nie siehst
- Warum **Selbstsicherheit** nichts über Richtigkeit sagt
- Was **Temperature** wirklich macht — und warum Runterdrehen kein Wissen erzeugt
- Was **wirklich hilft**: Wissen mitgeben, Quellen verlangen, nachrechnen

Für den Unterricht

Einstieg (3 Min): Kapitel gemeinsam lesen, Alltagsbeispiel aufgreifen.

Erkunden am Gerät (9 Min): die interaktiven Module auf herrki.de/350_halluzinationen ausprobieren.

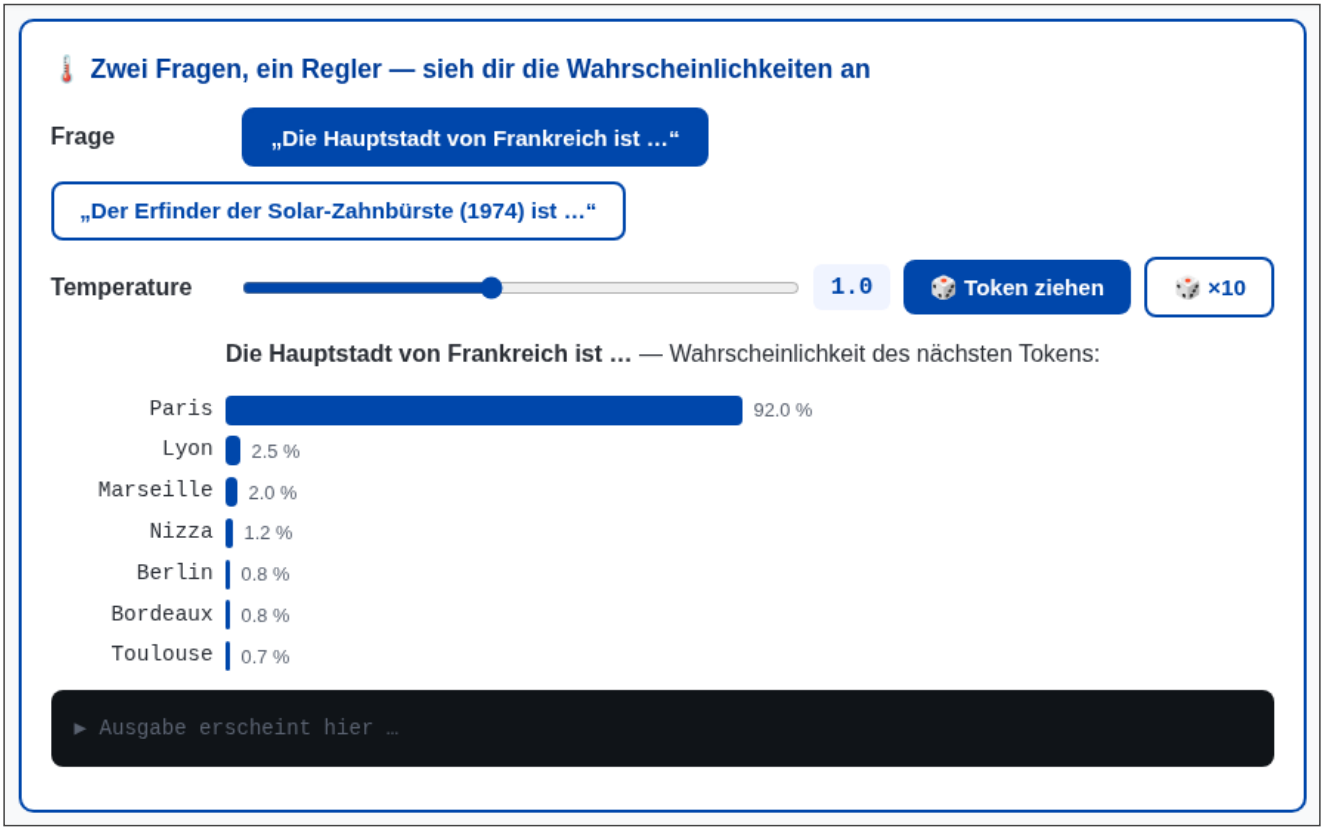
Sichern mit Quiz (3 Min): „Kurzer Check" besprechen, „Das nimmst du mit" festhalten.

Zur Erinnerung: Was ein Sprachmodell eigentlich tut

Aus „Wie LLMs Fragen beantworten“ weißt du: Ein Sprachmodell wählt **immer nur das nächste Token** — aus einer Wahrscheinlichkeits-Liste über das gesamte Vokabular. Und hier steckt schon das ganze Drama: **Es MUSS ein Token wählen.** Auch wenn es zur Frage schlicht nichts Belastbares gelernt hat.

Am Gerät ausprobieren: herrki.de/350_halluzinationen#abschnitt-2

Direkt zum Modul



Interaktives Modul – am Gerät ausprobieren

⚡ **Der Kernbefund:** Bei der Frankreich-Frage ragt EIN Kandidat heraus — das Wissen war tausendfach in den Trainingsdaten. Bei der **erfundenen** Solar-Zahnbürste ist die Verteilung von Anfang an **flach**: kein Kandidat ist besser als die anderen … und das Modell zieht **trotzdem selbstbewusst einen Namen**. Es klingt exakt gleich überzeugt — denn der *Tonfall* kommt aus dem Sprachmuster, nicht aus dem Wissen. **Selbstsicherheit sagt NICHTS über Richtigkeit.**

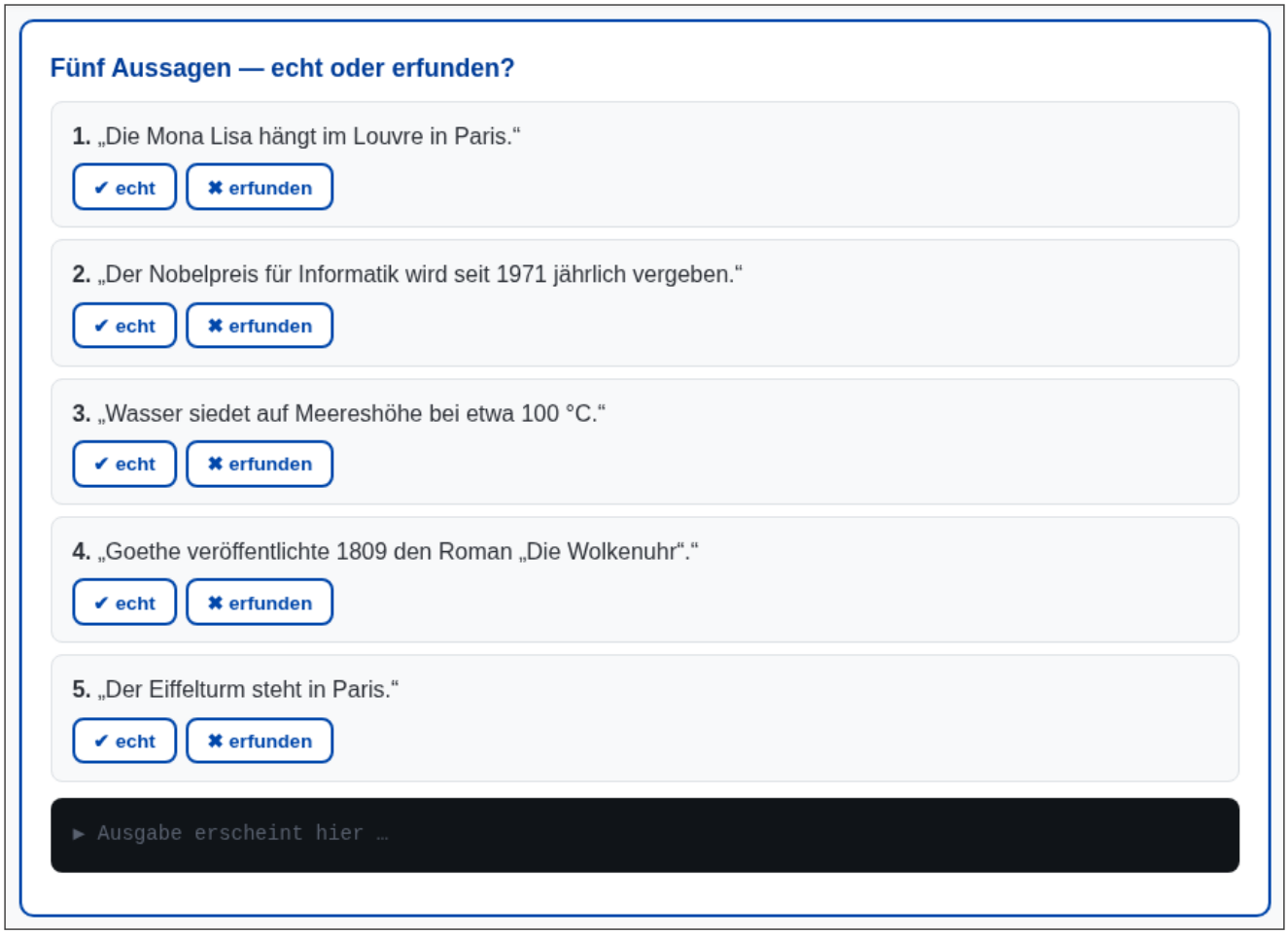
Was Temperature wirklich macht (und was nicht)

Der Regler formt die Verteilung: **niedrig** = die wahrscheinlichste Fortsetzung gewinnt fast immer (vorhersagbar, „langweilig“), **hoch** = auch seltene Token kommen dran (kreativ, streuend). Aber merke: **Temperature runterdrehen erzeugt kein Wissen.** Bei der erfundenen Frage wählt das Modell dann nur *immer denselben* falschen Namen — konsequent statt zufällig falsch.

Selbsttest: Erkennst du die Halluzination?

Am Gerät ausprobieren: herrki.de/350_halluzinationen#abschnitt-4

Direkt zum Modul



Interaktives Modul – am Gerät ausprobieren

Was wirklich hilft

Wissen mitgeben statt abfragen: RAG legt die echten Dokumente in den Prompt (Kapitel 330) — dämpft Halluzinationen deutlich, heilt sie nicht.

Quellenzwang: „Antworte NUR mit Angabe der Fundstelle; wenn keine existiert, sag: weiß ich nicht.“ — und die Quelle stichprobenartig ÖFFNEN.

Nachrechnen & Nachschlagen: Zahlen, Zitate, Paragrafen, Literaturangaben grundsätzlich prüfen — genau dort halluzinieren Modelle am liebsten.

Zweite Meinung: dieselbe Frage einem zweiten Modell oder einer Suchmaschine stellen — widersprechen sie sich, ist Vorsicht angesagt.

Runter zum Tensor

Rechnen wir die Balken von oben mit der Hand nach (Kapitel 060). Das Modell liefert für jede Vokabel einen Rohwert, das **Logit**. Wir nehmen fünf davon — zum Mitrechnen verkleinert; in echt steht dort eine Zahl *pro Vokabel-Token*, je nach Modell rund 50.000 bis 130.000 Stück. **Softmax** macht daraus Prozente: jedes Logit wird zu e hoch diesem Wert, dann teilt man durch die Summe aller.

```
// Die erfundene Frage: „Der Erfinder der Solar-Zahnbürste (1974) ist …“
Token Logit e^Logit Anteil an der Summe
Müller 2.60 13.46 13.46 / 60.08 = 22.4 %
Tanaka 2.50 12.18 12.18 / 60.08 = 20.3 %
Johnson 2.48 11.94 11.94 / 60.08 = 19.9 %
Schneider 2.44 11.47 11.47 / 60.08 = 19.1 %
Dubois 2.40 11.02 11.02 / 60.08 = 18.3 %
                     ------- -----------------
             Summe 60.08 100.0 %
// Fünf Namen, fünf fast gleich hohe Balken – das ist die flache Verteilung.
```

Und zum Vergleich dieselbe Rechnung bei einer Frage, deren Antwort millionenfach in den Trainingsdaten stand. Nur die Logits sind andere — die Zahl links entscheidet alles:

```
// Die belegte Frage: „Die Hauptstadt von Frankreich ist …“
Token Logit e^Logit Anteil
Paris 8.40 4447.07 93.3 %
Lyon 4.80 121.51 2.5 %
Marseille 4.60 99.48 2.1 %
Nizza 4.10 60.34 1.3 %
Berlin 3.70 40.45 0.8 %
                     ------- -------
             Summe 4768.85 100.0 %
// 3.6 Logit-Punkte Vorsprung werden durch e^x zu Faktor 37 – daher der eine Riesenbalken.
```

Und die Temperature? Sie teilt die Logits, **bevor** Softmax rechnet. Kleines T spreizt die Abstände, großes T drückt sie zusammen — hier mit den fünf erfundenen Namen von oben:

```
// Logits geteilt durch T daraus die Prozente
T = 0.4 6.50 6.25 6.20 6.10 6.00 26.3 % 20.5 % 19.5 % 17.7 % 16.0 %
T = 1.0 2.60 2.50 2.48 2.44 2.40 22.4 % 20.3 % 19.9 % 19.1 % 18.3 %
T = 2.0 1.30 1.25 1.24 1.22 1.20 21.2 % 20.1 % 19.9 % 19.6 % 19.2 %
// Schau auf die Spalte „Müller“: 26.3 statt 22.4 Prozent – mehr passiert nicht.
// Kein Herunterdrehen der Welt macht aus dem Raten ein Wissen.
```

Genau das ist die ganze Mechanik, und so heißt sie im Code. Die Logits sind dabei ein Tensor der Form [1, V] — V ist die Größe des Vokabulars, also statt unserer fünf Kandidaten die vollen 50.000 bis 130.000:

```
wahrscheinlichkeiten = softmax( logits.div(temperature) )
// Softmax macht daraus Prozente (Summe = 1). "Ich weiß es nicht" ist nur dann dabei,
```

```
// wenn solche Saetze im Training haeufig auf solche Fragen folgten — meist folgten: Antworten.
```

Zurück zur Anwendung

Ab heute liest du KI-Antworten anders: Der souveräne Ton ist **Sprachmuster**, die Fakten sind **Wahrscheinlichkeits-Züge**. Bei allem, was Konsequenzen hat — Verträge, Gesundheit, Geld, Noten — gilt: **prüfen, bevor du es weitergibst.** Die KI ist ein brillanter Formulierungs-Partner, kein Orakel.

Zum Weiterlesen

- **Der Überblick:** Ziwei Ji u. a., *„Survey of Hallucination in Natural Language Generation“*, 2022 — arxiv.org/abs/2202.03629 — sortiert, welche Arten von Halluzination es gibt und woher sie jeweils kommen
- **Warum es sich nicht wegprogrammieren lässt:** Adam Kalai & Santosh Vempala, *„Calibrated Language Models Must Hallucinate“*, 2023 — arxiv.org/abs/2311.14648 — der mathematische Beleg für das, was du oben am Regler gesehen hast
- **Wie man es misst:** Stephanie Lin u. a., *„TruthfulQA“*, 2021 — arxiv.org/abs/2109.07958 — ein Test aus Fangfragen, bei dem größere Modelle teils schlechter abschneiden als kleine
- **Die Antwort eines Herstellers:** OpenAI, *„Why language models hallucinate“*, 2025 — openai.com/index/why-language-models-hallucinate — die These: Prüfverfahren belohnen Raten und bestrafen „ich weiß es nicht“ — Halluzination ist auch ein Anreizproblem

Kurzer Check – mit Lösung

Lösungen für Lehrkräfte – für Lernende die Online-Version nutzen.

Frage 1: Warum halluzinieren Sprachmodelle?

- Ein Programmierfehler, der bald behoben wird.
- ✓ **Das System MUSS immer das plausibelste nächste Token wählen — auch wenn es zur Frage kein belastbares Wissen hat.**
- Weil Nutzer zu schwierige Fragen stellen.

Kein Defekt, sondern die Kehrseite des Funktionsprinzips: Bei flacher Verteilung wird trotzdem gezogen — und sprachlich perfekt verpackt.

Frage 2: Die KI antwortet extrem selbstsicher. Was sagt das über die Richtigkeit?

- Sehr sichere Formulierungen sind fast immer korrekt.
- ✓ **Nichts — der überzeugte Ton ist gelerntes Sprachmuster, keine Wissens-Anzeige.**
- Selbstsichere Antworten wurden von Menschen geprüft.

Im Experiment oben klangen beide Antworten gleich souverän — nur die Verteilung dahinter war völlig verschieden. Und die siehst du im Chat-Fenster nicht.

Frage 3: Hilft es gegen Halluzinationen, die Temperature auf ganz niedrig zu stellen?

- Ja — bei Temperature 0 sind alle Antworten korrekt.
- ✓ **Nein — sie macht die Wahl nur konsequenter; fehlendes Wissen ersetzt sie nicht. Es hilft: Wissen mitgeben (RAG), Quellen verlangen, prüfen.**
- Ja, aber nur bei Fragen auf Deutsch.

Bei der erfundenen Frage wählt das Modell mit niedriger Temperature einfach IMMER denselben falschen Namen. Verlässlichkeit der Streuung ≠ Verlässlichkeit der Fakten.

Das nimmst du mit

- **Halluzinationen sind kein Bug.** Sie sind die Kehrseite des Funktionsprinzips: Das Modell zieht das wahrscheinlichste nächste Token — auch dann, wenn es nichts weiß.
- **Die flache Verteilung siehst du nie.** Im Chat kommt nur der fertige Satz an, nicht die Unsicherheit, aus der er gezogen wurde.
- **Selbstsicherheit sagt nichts über Richtigkeit.** Der souveräne Ton ist ein Sprachmuster aus dem Training, kein Gradmesser für Wahrheit.
- **Temperature steuert die Streuung, nicht die Wahrheit.** Ganz unten liefert das Modell denselben falschen Namen — nur eben zuverlässig.
- **Was wirklich hilft:** Wissen mitgeben, Quellen verlangen, nachrechnen — und bei allem mit Konsequenzen prüfen, bevor du es weitergibst.

Weiter geht’s: Wenn selbstsicheres Raten das Problem ist — was hilft dagegen? Ein Modell, das sich erst laut Notizen macht und dann antwortet. Genau das siehst du in „Erst denken, dann reden“.

So zitierst du dieses Kapitel

HerrKI: Warum KI überzeugend Unsinn erzählt (Halluzinationen). herrki.de/350_halluzinationen, ein privates Hobbyprojekt von Terragon.de, abgerufen am 09.09.2026.