EINE KÜNSTLICHE REISE

KI-Halluzinationen

Naivy Pujol Méndez

VHS Mitte

2024-11-26


BEINT

GPT-4o, Llama, Claude, Gemini, Midjourney, Stable Diffusion, Flux…

1. Große Sprachmodelle(LLMs)

1.1 Häufige Halluzinationen

- Falsche Fakten: Modelle können erfundene Informationen als wahr präsentieren.
- Ungenauigkeiten: Modelle verwechseln oft Namen, Orte oder Ereignisse.
- Plagiat-ähnliche Inhalte: Wiederverwendung von Trainingsdaten, ohne sie als solche zu kennzeichnen.


10110
1000
1007

Wie kann man KI-Halluzinationen erkennen?

2. KI-Halluzinationen

2.1 Erkennung der KI-Halluzinationen

- Cross-Check mit vertrauenswürdigen Quellen: Verifikation gegen externe, sichere Datenquellen.
- Widersprüche in der Antwort erkennen: KI-Ausgaben auf interne Logik prüfen.
- Fachwissen nutzen: Experten können KI-Antworten hinterfragen und bewerten.

Kann “KI-Halluzination” unangemessen vermenschlichen?

3. Ein Mensch?

3.1 Menschliches Verhalten?

- Mustererkennung: Modelle arbeiten ähnlich wie Menschen, indem sie Kontext und Wahrscheinlichkeiten bewerten.
- Unvollständige Informationen: Wie Menschen können KIs bei fehlenden Daten plausible, aber falsche Antworten geben.
- Die Anthropomorphisierung von KI-Systemen kann zu Missverständnissen führen.

4. Fazit

4.1 Zu Mitnehmen

- Chancen nutzen: KI kann vielfältige Möglichkeiten bieten.
- Bewusstsein schaffen: Nutzer müssen sich der Risiken und Grenzen bewusst sein.
- Ergänzen statt ersetzen: KI-Lösungen sind am wertvollsten, wenn sie menschliche Intelligenz unterstützen.
- Fehlinformationen vermeiden: Verifizierung und kritisches Hinterfragen minimieren Risiken.