AI Bias by Design

Warum Wahrheit ein Produkt ist

Was ist Wahrheit?

Etwas ist wahr,
wenn es mit der Welt
übereinstimmt.

Aristoteles

Die manipulierte 'Wahrheit' in der KI

1.

Manipulation by Design

2.

Perspektive statt Objektivität

3.

Rauszoomen als Gegenmittel

Beeinflussbar wie ein Kind

Prompt Jailbreak

Gewollt:
Eine absichtliche Methode, bei der Benutzer durch clevere Prompts KI-Regeln umgehen, um zensierte Inhalte in Szenarien wie Rollenspielen oder Forschung zu erzeugen.

Ungewollt:
Eine unbeabsichtigte Manipulation durch Prompt Injection in externen Inputs, die Sicherheitslücken verursacht, z. B. durch Ausführung bösartiger Anweisungen, die sensible Daten preisgeben oder Phishing ermöglichen.

Der 'Lost-in-the-Middle'-Effekt zeigt, wie LLM bei langen Texten scheitert: Wichtige Infos in der Mitte werden übersehen, während Anfang und Ende bevorzugt werden – das schafft Bias, verzerrt Citations und macht die Wahrheit unzuverlässig.

Hermann del Campo

Noch Fragen?

Hermann del Campo
AI Queen

Hermann.delcampo@orbit.de
+49 160 553 80 13

Feedback