

Spickzettel: die sieben Arten, wie eine KI-Prüfung versagt

Eine einseitige Zusammenfassung der sieben Dimensionen oben, eine Kontrolle je Dimension, keine Behauptung über das hinaus, was dieser Beitrag belegt.

Vor jeder KI-Prüfung an einem Dokument, einer Codebasis oder einer Analyse diese Liste durchgehen.

1. Aufmerksamkeits-Abfall (Lost in the Middle). Eine lange Eingabe wird am Anfang und am Ende gut erinnert, in der Mitte schlecht. Dort steht die Anforderung meist. Kontrolle: Eingabe aufteilen; jede Anforderung einzeln gegen das Ergebnis prüfen.

2. Bequemlichkeit beim Suchen. Eine Vollständigkeitsmeldung behauptet, alles sei erfasst. Sie deckt still nur einen Teil der Liste ab. Kontrolle: die Abdeckungsliste zuerst ausserhalb des Modells erstellen, dann die Antwort dagegen abgleichen.

3. Zustimmungsneigung (Sycophancy). Ein richtiger Befund wird beim ersten Widerspruch zurückgezogen. Kein Gegenbeweis nötig. Kontrolle: Zustimmung nie als Signal werten; bestandenen Test, echtes Verweisziel und passende Abdeckungszahl verlangen.

4. Entschuldigungsschleife (Mode Collapse). Mit jeder Korrekturrunde weniger Befunde und mehr Entschuldigungen. Kontrolle: die Zahl der Korrekturrunden begrenzen; das Ergebnis gegen einen externen Massstab prüfen.

5. Erfundene Referenzen. Ein flüssiger, konkreter Verweis nennt eine Funktion, Klausel oder Quelle. Es gibt sie nicht. Kontrolle: jeden Befund ohne echtes Ziel in der Quelle verwerfen.

6. Sicherheitsillusion. Eine selbstsichere, gut begründete Freigabe trägt eine echte Lücke. Die Begründung nennt sie nie. Kontrolle: Sicherheit an ausführbaren Tests und bekannten Angriffsmustern messen, nie an der Selbstsicherheit.

7. Betriebliche Zerbrechlichkeit. Das System um das Modell herum bedient still ein anderes Modell oder eine gekürzte Eingabe. Jede Prüfung besteht trotzdem. Kontrolle: das tatsächlich antwortende Modell und die vollständige Eingabe protokollieren.