

Die sieben Dimensionen

Eine Zusammenfassung der sieben Dimensionen, eine Kontrolle je Dimension.



01 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.



02 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.



03 Zustimmungsneigung (Sycophancy)

Ein richtiger Befund wird beim ersten Widerspruch zurückgezogen. Kein Gegenbeweis nötig. **Kontrolle:** Zustimmung nie als Signal werten; bestandenem Test, echtes Verweisziel und passende Abdeckungszahl verlangen.



04 Entschuldigungsschleife (Mode Collapse)

Mit jeder Korrekturrunde weniger Befunde und mehr Entschuldigungen. **Kontrolle:** die Zahl der Korrekturrunden begrenzen; das Ergebnis gegen einen externen Massstab prüfen.



05 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.



06 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.



07 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.



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