
Das Erdbeer-Beispiel

Stell dir vor, du testest ein modernes Sprachmodell mit zwei Aufgaben:

Aufgabe 1: *„Schreib ein Gedicht über Vergänglichkeit im Stil von Rilke."*

Das Modell liefert etwas Beeindruckendes – kohärent, stilistisch stimmig, emotional berührend. Viele Menschen denken dann: *„Das versteht Sprache und Gefühle wirklich."* → **Überschätzung**

Aufgabe 2: *„Wie viele ‚r' stecken im Wort ‚Erdbeere'?"*

Das Modell antwortet falsch oder zählt unzuverlässig. Viele denken dann: *„Es kann nicht mal einfach zählen – wie soll es dann wirklich denken können?"* → **Unterschätzung**

Warum ist das *systematisch* und nicht nur *zufällig*?

Der Punkt ist: Die Fehleinschätzung folgt einem vorhersehbaren Muster.

Aufgabe sieht aus wie...	Tatsächliche Schwierigkeit	Reaktion
Komplex, kreativ, „menschlich"	Für KI oft leicht (Mustererkennung auf riesigen Textmengen)	**Überschätzung**
Simpel, mechanisch, „trivial"	Für KI oft schwer (erfordert symbolisches Zählen)	**Unterschätzung**

Das ist das eigentlich Beunruhigende: Menschen nutzen die **Oberfläche einer Aufgabe** als Heuristik für die Fähigkeit des Systems – aber diese Heuristik ist für KI systematisch falsch kalibriert.

Bei einem Menschen wäre sie richtig: Wer Rilke imitieren kann, kann auch zählen. Bei einem Sprachmodell gilt dieser Zusammenhang nicht.

Die Konsequenz

Wenn die Fehleinschätzung nur zufällig wäre, würden sich Fehler im Durchschnitt herausmitteln. Bei systematischer Fehlkalibrierung passiert das nicht – man überschätzt *vorhersehbar* in einer Klasse von Situationen und unterschätzt *vorhersehbar* in einer anderen. Das macht es schwer, aus Erfahrung zu lernen, weil die Korrekturen in die falsche Richtung ziehen können.