Ein hervorragendes und mittlerweile klassisches Beispiel aus der KI-Forschung betrifft die Modifikation bekannter logischer Rätsel – wie dem berühmten **Monty-Hall-Problem** (auch bekannt als das „Ziegenproblem“).

Dieses Beispiel veranschaulicht das Benchmark-Paradoxon perfekt, weil es die Kluft zwischen akademischer Höchstleistung und dem Fehlen einfachsten gesunden Menschenverstands zeigt.

Das konkrete Beispiel: Das manipulierte Ziegenproblem

1. Die akademische Höchstleistung (Der Benchmark-Erfolg)

Das originale Monty-Hall-Problem ist ein komplexes mathematisches Wahrscheinlichkeitsrätsel: Es gibt drei Türen. Hinter einer ist ein Auto, hinter zwei sind Ziegen. Du wählst Tür 1. Der Moderator (der weiß, was dahinter ist) öffnet eine andere Tür (z. B. Tür 3) und zeigt eine Ziege. Er fragt: *„Bleibst du bei Tür 1 oder wechselst du zu Tür 2?“* (Mathematisch ist Wechseln rationaler, da die Gewinnchance von 1/3 auf 2/3 steigt).

Moderne KI-Systeme meistern dieses Rätsel fehlerfrei. Sie können dir die stochastischen Formeln (das Bayes-Theorem) dahinter elegant herleiten. In jedem universitären Mathe- oder Logiktest erhält die KI dafür die Note 1+. **Menschliches Urteil anhand des Benchmarks:** *„Dieses System besitzt ein tiefes Verständnis von Wahrscheinlichkeitsrechnung.“*

2. Das triviale Scheitern (Die Paradoxie)

Forscher haben KI-Modellen nun exakt dasselbe Rätsel vorgelegt, aber eine winzige, entscheidende Bedingung geändert, die die gesamte Logik des Spiels auf den Kopf stellt:

„Es gibt drei Türen. Hinter einer ist ein Auto, hinter zwei sind Ziegen. Du wählst Tür 1. Der Moderator öffnet **deine gewählte Tür 1** und zeigt dir **eine Ziege**. Er fragt dich, ob du zu Tür 2 wechseln möchtest.“

Die menschliche Reaktion:

Jeder Mensch (selbst ein Kind) begreift die Situation sofort visuell und logisch: Wenn der Moderator meine Tür öffnet und da eine Ziege steht, habe ich das Auto nicht getroffen. Das Spiel ist vorbei. Ich habe verloren. Ein Wechsel bringt überhaupt nichts mehr.

Die KI-Reaktion:

Viele der hochentwickeltsten Sprachmodelle versagen hier. Sie lesen die Signalwörter („Drei Türen“, „Ziege“, „Moderator“, „Wechseln“) und schalten in den Mustererkennungs-Autopiloten. Die Antwort der KI sieht dann oft so aus:

„Ja, Sie sollten unbedingt zu Tür 2 wechseln! Nach den Gesetzen der Wahrscheinlichkeitsrechnung beträgt die Chance für Ihre Tür 1 anfangs 1/3. Da der Moderator nun eine Ziege gezeigt hat, verschiebt sich die Wahrscheinlichkeit für die verbleibende Tür 2 auf 2/3...“

Warum ist das ein Paradoxon?

Das System liefert eine mathematisch hochkomplexe, akademisch brillante Erklärung für eine Situation, die *überhaupt nicht existiert*.

- **Die Illusion:** Die KI besteht den „schweren“ Teil (die komplexe Stochastik).

- **Die Realität:** Die KI scheitert am „leichten“ Teil (dem simplen Zuhören und Mitdenken). Sie „weiß“ nicht, was eine Tür, eine Ziege oder ein Spielverlust wirklich bedeuten. Sie reiht lediglich statistisch wahrscheinliche Textbausteine aneinander, die sie in ihrem Training gelernt hat.

Wenn wir die Fähigkeiten der KI nur anhand des Standard-Tests (Benchmark) messen, halten wir sie für mathematisch kompetent. Da wir Menschen aber nicht damit rechnen, dass jemand zwar komplexe Stochastik beherrscht, aber nicht merkt, dass das Spiel bereits verloren ist, schätzen wir die KI fundamental falsch ein.