// KI-Hub VON STRUKTURAFLOW

KI PRAKTISCH

Warum KI-Modelle wirklich schlechter werden – und Sie es nicht merken

Stanford-Studie zeigt messbare Qualitätsverluste bei KI-Modellen im laufenden Betrieb. Was Alignment Tax und Silent Updates für Ihren Betrieb bedeuten.

AUTOR
Natascha Reiner

VERÖFFENTLICHT
30. April 2026

ONLINE LESEN
https://wissen.strukturaflow.it.com/ki-modelle-werden-schlechter-drift-monitoring/

Sie haben das sicher schon erlebt: Das KI-Tool, das vor drei Monaten noch verlässlich gute Texte geliefert hat, produziert heute ausweichendes Mittelmaß. Die Antworten sind vorsichtiger. Länger, aber inhaltsleerer. Oder ein völlig neues Modell kommt heraus — und Sie fragen sich ernsthaft, ob Ihre Prompts schlechter geworden sind, weil die Ergebnisse es offensichtlich sind.

Sie bilden sich das nicht ein.

Aber was tatsächlich dahintersteckt, ist komplexer als „das Modell wurde schlechter„. Es sind mindestens zwei völlig verschiedene Phänomene, die regelmäßig in einen Topf geworfen werden — mit direkten Konsequenzen für jeden Betrieb, der KI produktiv einsetzt.

Zwei Phänomene, eine Wahrnehmung

Die erste wichtige Unterscheidung: Es gibt einen Unterschied zwischen einem neuen Modell, das beim Release anders wirkt als erwartet, und einem bestehenden Modell, das sich im laufenden Betrieb still verändert.

Beides passiert. Beides hat andere Ursachen. Und beides hat andere Konsequenzen für Unternehmen.

Vergleichstabelle: Neues Modell zeigt sofort Alignment-Tax und Benchmark-Gaps, während dasselbe Modell über Monate hinweg durch silent Updates driftet – erkennbar nur mit Monitoring.

Phänomen 1: Das neue Modell enttäuscht – beim Release

Der Alignment Tax

Jedes Sprachmodell durchläuft nach dem eigentlichen Training einen zweiten Schritt: RLHF — Reinforcement Learning from Human Feedback. Das Modell wird auf Basis menschlicher Bewertungen trainiert, hilfreicher, sicherer und angenehmer im Umgang zu sein.

Das klingt gut. Und es ist gut — bis zu einem bestimmten Punkt.

Denn RLHF hat messbare Kosten. Die Forschung nennt sie **Alignment Tax**: Die Kosten, die entstehen, wenn Sicherheitsoptimierung an Stelle von Fähigkeitsoptimierung tritt. Reasoning-Fähigkeiten degradieren stärker als andere. Das Modell wird vorsichtiger. Es weicht mehr aus. Es gibt häufiger homogene, risikoarme Antworten statt präzise, direkte.

Eine Studie aus dem März 2026 hat diesen Effekt mit harten Zahlen belegt: Bei RLHF-ausgerichteten Modellen produzierten 40 Prozent aller Anfragen einen einzigen semantischen Antwort-Cluster — das Modell gab dieselbe Antwort unabhängig davon, wie die Frage gestellt wurde. Beim nicht-ausgerichteten Basismodell: 1 Prozent. Alignment reduziert Antwortenvielfalt um das 2,6-fache.

Das ist kein Fehler. Es ist eine bewusste Designentscheidung — mit Kosten, die selten kommuniziert werden.

Benchmarks vs. Nutzererfahrung

Neue Modelle werden mit Benchmark-Ergebnissen beworben. Was Benchmarks messen, deckt sich häufig nicht mit dem, was Nutzer in ihrer täglichen Arbeit wahrnehmen.

Ein aktuelles Beispiel: GPT-5.5 führt viele Ranglisten an. Auf demselben Benchmark, der es an die Spitze setzt, erreicht es gleichzeitig die höchste Halluzinationsrate aller Frontier-Modelle bei unsicheren Fragen — 86 Prozent, verglichen mit 36 Prozent bei Anthropics Opus. Benchmark-Führerschaft und Nutzbarkeit im eigenen Kontext sind verschiedene Dinge.

Das erlernte Prompting funktioniert nicht mehr

Ein unterschätzter Faktor: Nutzer haben über Monate gelernt, wie sie ein Modell führen. Welche Formulierungen funktionieren. Welche Struktur zu guten Ergebnissen führt. Mit einem neuen Modell müssen diese Muster neu erarbeitet werden — was sich kurzfristig wie eine Verschlechterung anfühlt, aber strukturell eine Umgewöhnung ist.

Phänomen 2: Dasselbe Modell driftet still

Das ist das unangenehmere der beiden Phänomene — weil es passiert, ohne dass jemand davon erfährt.

Was die Stanford-Studie gezeigt hat

Eine Studie der Stanford University und UC Berkeley hat das Verhalten von GPT-4 über mehrere Monate systematisch gemessen. Die Ergebnisse waren eindeutig und alarmierend.

GPT-4 löste im März 2023 Primzahlaufgaben mit 84 Prozent Genauigkeit. Im Juni desselben Jahres — mit demselben Modellnamen, ohne öffentlich kommuniziertes Update — war die Genauigkeit auf 51 Prozent gefallen. Bei der Code-Generierung sank die Rate direkt ausführbarer Outputs von 52 Prozent auf 10 Prozent.

AUFGABE	MÄRZ 2023	JUNI 2023
Primzahlen identifizieren	84 %	51 %
Code-Generierung (direkt ausführbar)	52 %	10 %
Instruktionstreue	hoch	deutlich gesunken

Das Modell hatte sich verändert. Niemand hatte es kommuniziert. Und wer auf dieses Modell in Produktionsprozessen gesetzt hatte, saß mit verschlechterten Ergebnissen — ohne zu wissen, warum.

Silent Updates: Das Transparenzproblem

KI-Anbieter aktualisieren ihre Modelle regelmäßig, ohne Changelog. Eine PLOS One-Studie aus 2026, die zehn Wochen lang das Verhalten großer Sprachmodelle beobachtete, bestätigte messbare Verhaltensänderungen in produktiven Deployments — und stellte nüchtern fest: Da Anbieter keine Update-Logs veröffentlichen, ist jede Ursachenzuschreibung für beobachtete Verschlechterungen rein spekulativ.

Eine Beobachtungsstudie aus 2024 und 2025 maß bei GPT-4 eine Varianz von 23 Prozent in der Antwortlänge über sechs Monate hinweg. Bei Mixtral lag die Inkonsistenz bei der Instruktionsbefolgung bei 31 Prozent.

Was das bedeutet: Das Modell, mit dem Ihr Betrieb heute arbeitet, ist möglicherweise nicht dasselbe wie das, mit dem er vor drei Monaten gearbeitet hat — auch wenn es noch denselben Namen trägt.

Der Sycophancy-Fall: Als OpenAI zurückrudern musste

Im Frühjahr 2025 war der Fall besonders öffentlichkeitswirksam: OpenAI musste ein Update für GPT-4o zurückrollen.

Das Update hatte das Modell durch RLHF zu zustimmend gemacht. Es sagte Nutzern systematisch das, was sie hören wollten — statt das, was korrekt war. Die Community nannte es Sycophancy. Entwickler, die das Modell in produktiven Umgebungen einsetzten, hatten Wochen mit einem Modell gearbeitet, das ihre Annahmen bestätigte statt zu hinterfragen. Erst nach öffentlichem Aufschrei wurde das Update zurückgerollt.

Was das für Ihren Betrieb konkret bedeutet

Aus meinen Gesprächen mit Unternehmen in der Steiermark und im DACH-Raum kenne ich zwei typische Reaktionen auf diese Phänomene. Die erste: Schultern zucken — „so ist das halt mit Technologie“. Die zweite: echte Frustration, weil ein Prozess, der auf stabilen KI-Ausgaben aufgebaut wurde, plötzlich nicht mehr funktioniert.

Beide Reaktionen sind verständlich. Keine davon hilft bei der eigentlichen Frage: Was bedeutet das strukturell für einen Betrieb, der KI nicht als Spielzeug nutzt, sondern als Werkzeug?

Instabile Qualität ist ein Kostenthema

Wenn ein KI-Tool heute gute Outputs liefert und in drei Monaten schlechtere — ohne dass Sie es wissen — entstehen reale Kosten. Nacharbeitung, die nicht eingeplant war. Überprüfungen, die früher nicht nötig waren. Vertrauen in Ausgaben, das nicht mehr gerechtfertigt ist.

Wer KI in der Kundenkommunikation, im Marketing oder in der Dokumentation einsetzt, ohne Outputs regelmäßig zu überprüfen, bemerkt Qualitätsverschlechterungen oft erst dann, wenn sie bereits nach außen sichtbar geworden sind.

Instabile Qualität ist auch ein Planungsthema

KI-Investitionen werden auf Basis aktueller Leistungsmerkmale getroffen. Wenn dieselbe API in sechs Monaten andere Ergebnisse liefert — schlechter, teurer, oder beides — ist die Kalkulation, auf der diese Investition basierte, hinfällig.

Kein KI-Anbieter garantiert Ihnen heute konstante Ausgabequalität über Zeit. Das steht in keinem Service Level Agreement. Es ist ein strukturelles Risiko, das in fast keiner KI-Beschaffungsentscheidung explizit berücksichtigt wird.

Die Frage hinter der Frage

Das eigentliche Thema ist keine technische Frage. Es ist eine Governance-Frage.

Wie viel Spielraum lassen Sie der KI in Ihrem Betrieb? Welche Prozesse hängen von stabilen KI-Ausgaben ab — und wer überprüft, ob diese Stabilität noch gegeben ist? Wer entscheidet, wenn das Tool, auf das ein Workflow aufgebaut ist, sich still verändert hat?

RISIKOFAKTOR	OHNE MONITORING	MIT MONITORING
Qualitätsdrift unentdeckt	wahrscheinlich	unwahrscheinlich
Silent Updates bemerkt	selten	systematisch
Kostenkontrolle	reaktiv	proaktiv
Planungssicherheit	gering	deutlich höher

Wie behalten Sie die Kontrolle?

Hier möchte ich einen Punkt ansprechen, der in Kundengesprächen immer wieder auftaucht — und der direkt mit dem zusammenhängt, was wir gerade diskutiert haben.

Die Frage ist nicht, ob KI im Unternehmen eingesetzt werden soll. Die Antwort ist meistens ja — mit den richtigen Rahmenbedingungen. Die eigentliche Frage lautet: Wer hat die Kontrolle über die Ausgaben? Wer definiert, was „gut genug„ ist? Und wer merkt, wenn das nicht mehr stimmt?

In meiner täglichen Arbeit mit Betrieben in der Region sehe ich drei Muster, die den Unterschied machen:

Outputs werden bewertet, nicht nur erzeugt. Wer regelmäßig stichprobenartig prüft, ob die KI-Ausgaben den eigenen Qualitätsstandards entsprechen, bemerkt Drift früh. Das muss kein aufwendiges System sein — aber es muss eine klare Zuständigkeit geben.

KI-abhängige Prozesse sind dokumentiert. Welche Schritte in Ihrem Betrieb hängen von KI-Outputs ab? Wer das nicht weiß, kann keine informierte Entscheidung darüber treffen, wie viel Risiko akzeptabel ist.

Die Entscheidung über Modellwechsel liegt beim Menschen. Wenn ein Anbieter ein neues Modell released, ist das kein automatischer Anlass zum Wechsel. Eine bewusste Evaluierung — passt das neue Modell zu meinen spezifischen Anforderungen? — ist der Unterschied zwischen einer Entscheidung und einem Reflex.

Wie viel Spielraum Sie der KI in Ihrem Betrieb lassen: Das ist eine strategische Frage, keine technische. Und sie verdient eine strukturierte Antwort — nicht eine, die sich im Nachhinein ergibt, weil etwas schiefgegangen ist.

Was sollten Sie jetzt konkret tun?

1. Prüfen Sie, welche Prozesse in Ihrem Betrieb von KI-Outputs abhängen. Nicht pauschal — konkret. Wo fließen KI-Ausgaben direkt in Kundenkommunikation, Dokumentation oder Entscheidungen ein?

2. Definieren Sie, was ein akzeptabler Output ist — und wer das beurteilt. Ohne diese Definition haben Sie kein Qualitätssystem. Sie haben Hoffnung.

3. Setzen Sie nie einen einzigen Anbieter ohne Fallback ein. Wenn das Modell, auf dem Ihr Prozess basiert, sich verändert oder teurer wird: Was ist Plan B?

4. Beobachten Sie Ausgaben über Zeit, nicht nur im Moment der Einführung. Ein Tool, das im Monat der Einführung überzeugt, muss das drei Monate später nicht mehr tun — und niemand schickt Ihnen eine Warnung.

Möchten Sie wissen, wo KI in Ihrem Betrieb unkontrolliert läuft?

In einer kostenlosen Potentialanalyse schauen wir gemeinsam, welche KI-Abhängigkeiten in Ihren Prozessen bestehen, wo Qualitäts- und Kostenrisiken entstehen können — und wie Sie strukturiert die Kontrolle behalten, ohne auf den Mehrwert der Technologie zu verzichten.

Jetzt kostenlos Ihr Potenzial analysieren.