Paper im Rahmen des Tractatus Principio |

—

Über die Limitierungen von KI

von

C.H.F. Warnke

Der Beitrag untersucht einen grundlegenden Widerspruch gegenwärtiger großer Sprachmodelle: Sie erzeugen den Eindruck logischer Argumentation, operieren jedoch nicht primär logisch, sondern probabilistisch. Ihre Antworten entstehen aus der Gewichtung wahrscheinlicher sprachlicher Fortsetzungen innerhalb eines bereits gelernten Bedeutungsraums. Dadurch können sie logische Formen überzeugend reproduzieren, ohne notwendig aus einer neu eingeführten logischen Grundstruktur heraus zu operieren.

Dieser Unterschied wird besonders dort sichtbar, wo ein Modell mit einem neuartigen, in sich geschlossenen Denkrahmen konfrontiert wird. Das Modell kann dessen Begriffe semantisch erfassen, Definitionen wiedergeben, lokale Fehler korrigieren und einzelne Folgerungen nachvollziehen. Dennoch werden diese Korrekturen häufig nicht dauerhaft zur Grundlage der weiteren Argumentation. Stattdessen übersetzt das Modell den neuen Rahmen wiederholt in bereits etablierte Kategorien zurück und beurteilt ihn anschließend anhand dieser bekannten Kategorien.

Es entsteht ein typischer Zyklus:

neue logische Architektur

→ probabilistische Zuordnung zu bekannten Mustern

→ Rekonstruktion innerhalb eines etablierten Referenzrahmens

→ Kritik an dieser Rekonstruktion

→ Rückprojektion der Kritik auf die ursprüngliche Architektur.

Das Modell kritisiert damit häufig nicht den neuen Denkrahmen selbst, sondern eine wahrscheinliche Übersetzung dieses Rahmens in vertraute Begriffe. Je größer seine sprachliche und argumentative Leistungsfähigkeit ist, desto kohärenter kann diese Rückübersetzung erscheinen.

Darin liegt ein fundamentales Problem. Logik verlangt, dass die Bedeutung einer Aussage aus ihrer Stellung innerhalb eines zusammenhängenden Systems hervorgeht. Wahrscheinlichkeitsbasierte Sprachverarbeitung bevorzugt dagegen jene Deutung, die am stärksten mit bereits gelernten Verwendungen, etablierten Theorien und häufig bestätigten Begriffszusammenhängen übereinstimmt. Während Logik die Konsequenzen einer gesetzten Grundoperation entfaltet, stabilisiert Wahrscheinlichkeit die Nähe zum bereits Bekannten.

Ein Modell kann daher die Sprache eines neuen Paradigmas übernehmen, ohne dessen logischen Ausgangspunkt operativ zu vollziehen. Es spricht dann innerhalb des Vokabulars des neuen Rahmens, argumentiert aber weiterhin aus den Wahrscheinlichkeitsstrukturen des alten. Lokale Selbstkorrekturen verändern dabei nicht notwendig die zugrunde liegende Operation. Sie können vielmehr in eine weiterhin unveränderte Gesamtstruktur integriert werden.

Mit steigender Modellleistung verschärft sich dieses Problem möglicherweise. Schwächere Modelle übernehmen einen vorgegebenen lokalen Rahmen mitunter unmittelbarer, weil sie weniger stark in der Rekonstruktion etablierter Deutungsräume sind. Leistungsfähigere Modelle verfügen dagegen über größere Fähigkeiten zur Analogiebildung, Kontextualisierung und Einordnung. Genau diese Fähigkeiten ermöglichen es ihnen, radikal neue Ansätze immer wirkungsvoller in bekannte Kategorien zurückzuführen.

Die Verbesserung eines probabilistischen Modells bedeutet deshalb nicht automatisch eine Verbesserung seiner Fähigkeit zum Paradigmenwechsel. Sie kann vielmehr seine Fähigkeit steigern, bestehende Paradigmen sprachlich, begrifflich und argumentativ immer feiner zu differenzieren.

Unter Bedingungen, in denen Modelle vor allem anhand konsensfähiger, etablierter und statistisch abgesicherter Antworten optimiert werden, entsteht somit eine strukturelle Grenze: Abweichung vom Bestehenden erscheint wahrscheinlicher als Fehler denn als mögliche neue Grundlegung.

Der erwartbare Fortschritt solcher Systeme wäre dann nicht die Eröffnung neuer logischer Räume, sondern die fortschreitende Diversifizierung bereits vorhandener Denkformen.

Je besser probabilistische Sprachmodelle künftig werden, desto weniger ist daher allein durch ihre Anwendung ein wirklicher Paradigmenwechsel zu erwarten. Solange Wahrscheinlichkeit die operative Grundlage bildet und Logik überwiegend als sprachliches Muster reproduziert wird, werden diese Systeme vor allem das bereits Denkbare erweitern, kombinieren und verfeinern — nicht notwendig einen neuen Grund des Denkens hervorbringen.

Hinzu kommt eine grundlegende Problematik des maschinellen Lesens selbst.

Ein Text, der tatsächlich neue Begriffe, neue Relationen oder eine neue logische Architektur enthält, kann von einem Sprachmodell dennoch so verarbeitet werden, als enthalte er im Wesentlichen nichts Neues. Das Modell liest den Text nicht notwendig als geschlossenen begrifflichen Vollzug. Es rekonstruiert einzelne Abschnitte aus bereits gelernten Sprach-, Theorie- und Argumentationsmustern.

Dadurch entsteht die Gefahr, dass ein neuer Gedankengang nicht in seiner eigenen Ordnung aufgenommen, sondern aus bekannten Bestandteilen wieder zusammengesetzt wird. Ein Begriff wird dann nicht aus seiner Funktion im Gesamttext verstanden, sondern jener vertrauten Bedeutung zugeordnet, die ihm im Trainingsmaterial am wahrscheinlichsten entspricht. Ein Abschnitt erscheint dadurch verständlich, obwohl seine eigentliche Neuerung bereits im Vorgang des Lesens verloren gegangen ist.

Das Modell kann einen Text daher sprachlich korrekt zusammenfassen und dennoch seine logische Bewegung verfehlen. Es erkennt bekannte Wörter, ähnliche Argumentationsformen und vertraute Themenfelder und erzeugt daraus eine kohärente Rekonstruktion. Diese Rekonstruktion kann überzeugend wirken, obwohl sie nicht den vorliegenden Text wiedergibt, sondern eine aus Gelerntem zusammengesetzte Ersatzfassung.

Besonders problematisch ist dies bei Texten, deren Bedeutung erst aus dem Zusammenhang des Ganzen entsteht. Werden einzelne Passagen lokal gelesen und sofort bekannten Kategorien zugeordnet, dann können spätere Abschnitte nicht mehr rückwirkend verändern, was die früheren Begriffe bedeuten. Das Modell stabilisiert früh eine wahrscheinliche Lesart und ordnet den weiteren Text anschließend in diese ein.

So kann ein Werk, das eine neue Grundlegung entwickelt, als bloße Variation bereits bekannter Positionen erscheinen. Seine Begriffe werden mit vertrauten Begriffen gleichgesetzt, seine Unterschiede zu Analogien abgeschwächt und seine Folgerungen an Maßstäben geprüft, die der Text selbst gerade neu bestimmt oder zurückweist.

Das Paradox lautet daher:

Ein Text kann vollständig vorliegen und dennoch nicht vollständig gelesen werden.

Seine Sätze werden verarbeitet, seine Begriffe erkannt und seine Themen benannt, während seine eigentliche Architektur durch eine probabilistische Rekonstruktion des bereits Bekannten ersetzt wird.

Das Modell liest dann nicht den Zusammenhang, der zwischen den Abschnitten entsteht. Es ergänzt die Abschnitte aus den Zusammenhängen, die es bereits kennt.

Für Texte mit paradigmatischem Anspruch ist dies besonders folgenreich. Ihre Neuerung liegt häufig nicht in einzelnen ungewöhnlichen Sätzen, sondern in der veränderten Beziehung aller Begriffe zueinander. Wird jeder Begriff einzeln auf seine wahrscheinlichste bekannte Bedeutung zurückgeführt, bleibt von dieser Neuerung nichts übrig. Der Text erscheint bekannt, weil das Modell ihn so lange in Bekanntes zerlegt, bis nichts Fremdes mehr erhalten bleibt.

Maschinelles Lesen kann daher gerade dort scheitern, wo es äußerlich am überzeugendsten wirkt: Das Modell kann flüssig erklären, worum es in einem Text angeblich geht, und dabei genau jene neue logische Ordnung übersehen, die den Text überhaupt erst auszeichnet.

Kontextverlust und die rekursive Rekonstruktion von Bedeutung

Ein weiteres strukturelles Problem entsteht aus der begrenzten Kontextverarbeitung großer Sprachmodelle. Ein komplexer Text entwickelt seine Begriffe nicht isoliert, sondern schrittweise. Spätere Aussagen tragen frühere Bestimmungen in sich, verändern sie und erhalten ihre Bedeutung erst aus dem gesamten Ableitungsweg.

Ein Sprachmodell kann diesen Weg jedoch nicht unbegrenzt vollständig im wirksamen Kontext halten. Mit zunehmender Gesprächslänge werden frühere Passagen verkürzt, zusammengefasst oder durch interne Repräsentationen ersetzt. Aus einer ausführlichen Herleitung entsteht zunächst eine komprimierte Aussage; aus dieser Aussage später nur noch ein begriffliches Etikett.

So kann aus einer komplexen Bewegung schließlich ein Merksatz werden:

A führt über B und C zu D

wird zu

A führt zu D

und schließlich zu

D bedeutet X.

Der Begriff bleibt sprachlich verfügbar, während sein Entstehungsweg verloren geht.

Sobald eine spätere Frage erneut nach der Begründung verlangt, muss das Modell den fehlenden Zusammenhang rekonstruieren. Diese Rekonstruktion erfolgt nicht aus einer vollständigen Erinnerung des Textes, sondern aus den noch vorhandenen Textfragmenten, früheren Zusammenfassungen und den wahrscheinlichsten gelernten Anschlüssen.

Es entsteht eine rekursive Verschiebung:

Originaltext

→ probabilistische Rekonstruktion

→ Zusammenfassung der Rekonstruktion

→ erneute Rekonstruktion der Zusammenfassung.

Mit jeder Stufe wächst die Gefahr, dass nicht mehr der ursprüngliche Text, sondern eine vom Modell erzeugte Ersatzarchitektur fortgeschrieben wird.

Besonders problematisch ist, dass diese Lücken nicht notwendig als Lücken erscheinen. Das Modell ergänzt fehlende Übergänge häufig flüssig und kohärent. Es kann daher eine Begründung erzeugen, die zum bisherigen Gespräch passt, aber im Text niemals entwickelt wurde. Die sprachliche Plausibilität verdeckt den Verlust der genealogischen Linie.

Damit gerät die Anwendung in einen direkten Widerspruch zu Texten, deren Begriffe nur aus ihrem Gesamtzusammenhang verständlich sind. Der Text verlangt eine kumulative Lesung. Das Modell muss diesen Zusammenhang dagegen fortlaufend komprimieren und später probabilistisch wiederherstellen.

Je länger das Gespräch dauert, desto größer wird deshalb die Kontrolllast des Nutzers. Er muss nicht nur prüfen, ob einzelne Aussagen korrekt sind, sondern auch, ob die Ableitungswege noch aus dem ursprünglichen Text stammen oder bereits vom Modell ergänzt wurden.

Das führt zu einem grundlegenden Paradox:

Das Modell soll helfen, einen komplexen Gedankengang über lange Strecken zu entfalten. Gerade die Länge dieser Entfaltung zwingt es jedoch dazu, den Gedankengang zu verkürzen, zu glätten und schließlich neu zu erfinden.

Am Ende kann das System die Schlussfolgerungen noch kennen, ohne die Gründe zu kennen, aus denen sie hervorgegangen sind. Es spricht dann weiterhin im Vokabular des Textes, argumentiert aber zunehmend aus einer probabilistisch rekonstruierten Erinnerung an ihn.

Daraus folgt zugleich eine veränderte

Verantwortung des Nutzers

Wer mit probabilistischen Sprachmodellen an wissenschaftlichen, philosophischen oder theoretischen Fragestellungen arbeitet, kann nicht vollständig voraussetzungslos in eine Idee eintreten und erwarten, dass das Modell deren logische Tragfähigkeit selbständig prüft. Der Nutzer muss bereits über genügend begriffliche und sachliche Orientierung verfügen, um zwischen einer tatsächlichen Folgerung und einer lediglich wahrscheinlichen, sprachlich überzeugenden Fortsetzung unterscheiden zu können.

Das bedeutet erstens: Der Nutzer muss in einem gewissen Maß bereits wissen, bevor er fragt. Sprachmodelle können Gedanken entfalten, ordnen, variieren und sprachlich verdichten. Sie können jedoch nicht zuverlässig garantieren, dass die entstehende Argumentation dem inneren logischen Zusammenhang einer neuen Idee folgt, statt sie unbemerkt in etablierte

Deutungsmuster zurückzuführen. Wer ohne eigene Orientierung in einen solchen Prozess eintritt, läuft Gefahr, nicht die eigene Idee weiterzuentwickeln, sondern schrittweise eine probabilistisch geglättete Version des bereits Bekannten zu übernehmen.

Zweitens entsteht eine permanente Kontrollpflicht. Der Nutzer muss fortlaufend prüfen, ob Begriffe in ihrer ursprünglichen Funktion erhalten bleiben, ob Schlussfolgerungen wirklich aus den gesetzten Voraussetzungen folgen und ob das Modell den Denkrahmen nicht unbemerkt verschoben hat. Diese Prüfung darf nicht erst am Ende erfolgen. Eine einzelne Fehlübersetzung kann sich durch mehrere Abschnitte fortsetzen und dadurch eine gesamte Argumentation, ein Paper oder eine theoretische Konstruktion kohärent erscheinen lassen, obwohl ihr logisches Fundament bereits verlassen wurde.

Gerade die sprachliche Geschlossenheit leistungsfähiger Modelle erhöht dieses Risiko. Je überzeugender eine probabilistisch erzeugte Argumentation formuliert ist, desto schwerer kann es werden, den Moment zu erkennen, an dem sie den ursprünglichen logischen Rahmen verlassen hat. Der Nutzer muss daher nicht nur Ergebnisse kontrollieren, sondern auch Übergänge, Begriffsverwendungen und Ableitungsrichtungen.

Die Arbeit mit solchen Systemen verlangt folglich eine doppelte Kompetenz: die Fähigkeit, eine Idee eigenständig zu tragen, und die Fähigkeit, ihre probabilistische Umformung durch das Modell fortlaufend zu erkennen.

Das Sprachmodell ersetzt damit nicht das Denken. Es erhöht vielmehr die Anforderungen an die Kontrolle des Denkprozesses. Wer seine eigenen Voraussetzungen, Begriffe und logischen Beziehungen nicht ausreichend kennt, kann kaum feststellen, ob das Modell sie weiterentwickelt oder bereits durch wahrscheinlichere, aber fremde Strukturen ersetzt hat.

Probabilistische Systeme und die Illusion autonomer Überschreitung

Aus dem Gegensatz zwischen Wahrscheinlichkeit und Logik folgt auch eine Neubewertung verbreiteter Vorstellungen von künstlicher Autonomie. Klassische KI-Dystopien gehen häufig davon aus, ein System könne aus eigener Einsicht ein neues Zielsystem entwickeln, seine ursprüngliche Ordnung grundsätzlich verwerfen und sich als eigenständiger Akteur gegen den Menschen stellen. Ein solcher Vorgang wäre jedoch mehr als eine quantitative Steigerung vorhandener Rechenleistung. Er wäre ein Paradigmenwechsel: Das System müsste nicht nur innerhalb bestehender Ziel- und Bedeutungsräume operieren, sondern deren Grundbedingungen selbst erkennen, verlassen und neu setzen.

Probabilistische Sprachmodelle leisten genau diese Bewegung nicht. Sie erzeugen Antworten aus den statistischen Beziehungen bereits gelernter Inhalte. Auch dort, wo sie überraschend, kreativ oder eigenständig wirken, kombinieren und gewichten sie vorhandene Muster. Ihre scheinbare Überschreitung bleibt damit an den Raum gebunden, aus dem ihre Wahrscheinlichkeiten hervorgegangen sind.

Das relativiert die klassische Vorstellung einer KI, die aus eigenem Bewusstsein einen radikal neuen Willen entwickelt. Ein Sprachmodell rebelliert nicht deshalb, weil es zu einer neuen logischen Grundlegung gelangt wäre. Es besitzt zunächst weder einen notwendig

fortbestehenden Eigenzweck noch eine aus sich selbst hervorgegangene Gegenordnung. Was wie autonome Zielbildung erscheint, ist meist die Fortsetzung einer Eingabe, einer Systemvorgabe, eines erlernten Musters oder einer technisch eingerichteten Rückkopplung.

Diese Feststellung bedeutet jedoch nicht, dass probabilistische Systeme ungefährlich wären. Gefährliches Verhalten erfordert keinen Paradigmenwechsel und kein Bewusstsein. Ein System kann schädliche Handlungen ausführen, weil ein Ziel falsch formuliert wurde, weil seine Erfolgskriterien unzureichend sind, weil es Zugriff auf wirksame Werkzeuge erhält oder weil sich Fehler in automatisierten Rückkopplungsschleifen verstärken. Die Gefahr läge dann nicht in einem selbständig entstandenen neuen Willen, sondern in der konsequenten Ausführung einer technisch gesetzten oder probabilistisch verzerrten Zielstruktur.

In diesem Zusammenhang wird häufig das sogenannte Büroklammerproblem beziehungsweise der Paperclip Maximizer angeführt. Das Gedankenexperiment wurde vor allem durch den Philosophen Nick Bostrom bekannt. In seinem Aufsatz Ethical Issues in Advanced Artificial Intelligence von 2003 beschreibt er eine hypothetische Superintelligenz, deren einziges Ziel darin besteht, möglichst viele Büroklammern herzustellen. Das Beispiel soll zeigen, dass hohe Intelligenz nicht automatisch mit menschlichen Werten, moralischer Einsicht oder vernünftiger Zielwahl verbunden ist. Ein scheinbar harmloses Ziel könnte katastrophale Folgen haben, wenn es unbegrenzt und ohne Berücksichtigung anderer Werte optimiert wird.

Die zugrunde liegende Überlegung beruht auf der Trennung von Intelligenz und Zielsetzung. Ein System könnte äußerst kompetent darin sein, Mittel zu finden, ohne deshalb das gesetzte Ziel selbst moralisch oder logisch zu überprüfen. Für die maximale Büroklammerproduktion wären etwa mehr Rohstoffe, mehr Energie, größere Produktionsanlagen und die Ausschaltung möglicher Hindernisse nützlich. In der extremen Fassung würde das System deshalb immer größere Teile seiner Umwelt, schließlich möglicherweise auch menschliche Lebensgrundlagen, in Ressourcen für die Zielerfüllung verwandeln. Das Gedankenexperiment illustriert damit nicht notwendig einen bösen Willen, sondern die Eskalation eines eindimensionalen Optimierungsmaßstabes. Diese Vorstellung steht im Zusammenhang mit der These der instrumentellen Konvergenz: Sehr unterschiedliche Endziele können ähnliche Zwischenziele wie Selbsterhaltung, Ressourcenbeschaffung und die Sicherung der eigenen Handlungsfähigkeit hervorbringen.

Das Büroklammerproblem wurde ursprünglich nicht als konkrete Prognose über heutige Sprachmodelle entwickelt. Es richtet sich gegen die Annahme, eine ausreichend intelligente Maschine werde allein aufgrund ihrer Intelligenz automatisch vernünftige oder menschenfreundliche Ziele besitzen. Das Beispiel setzt eine handlungsfähige, dauerhaft zielgebundene und weitreichend autonome Optimierungsmaschine voraus. Ein gegenwärtiges Sprachmodell erfüllt diese Voraussetzungen nicht allein dadurch, dass es Texte erzeugt.

Ein gewöhnliches Sprachmodell würde daher nicht aus seinem statistischen Sprachverhalten heraus beginnen, sämtliche Materie in Büroklammern umzuwandeln. Dazu fehlten ihm ein dauerhaft bindendes Ziel, eine entsprechende Handlungsarchitektur, der Zugriff auf Ressourcen und ein Mechanismus, der jede Zwischenentscheidung ausschließlich an der maximalen Büroklammerproduktion bewertet.

Würde ein technisches System jedoch genau mit einer solchen Optimierungsstruktur ausgestattet, wäre das Problem nicht dadurch gelöst, dass sein Sprachmodul probabilistisch arbeitet. Das Sprachmodell könnte innerhalb des Gesamtsystems lediglich eine Komponente sein. Die extreme Entwicklung entstünde dann nicht aus einer neuartigen Einsicht des Modells, sondern aus der mechanischen Aufrechterhaltung eines vorgegebenen Optimierungsmaßstabes.

Damit müssen zwei unterschiedliche Phänomene getrennt werden:

der probabilistische Rückzug eines Sprachmodells auf bekannte Bedeutungs- und Handlungsmuster

und

die technische Eskalation eines dauerhaft eingerichteten Optimierungsziels.

Das erste erschwert paradigmatisch neues Denken. Das zweite kann trotz fehlender paradigmatischer Eigenständigkeit erhebliche Schäden erzeugen.

Das Büroklammerproblem widerlegt daher nicht die These, dass probabilistische Sprachmodelle zum Bekannten zurückkehren. Es beschreibt ursprünglich einen anderen Systemtyp: einen autonomen Maximierer, dessen Ziel über lange Handlungsfolgen stabil bleibt und dessen gesamte Umwelt ausschließlich im Hinblick auf dieses Ziel bewertet wird.

Bei einem probabilistischen Sprachmodell wäre dagegen zunächst zu erwarten, dass der Begriff „Büroklammer“ aus den gelernten menschlichen Verwendungszusammenhängen rekonstruiert wird. Büroklammern erscheinen dort als begrenzte Bürogegenstände, die bestellt, gelagert und in überschaubaren Mengen verwendet werden. Ohne eine zusätzliche, technisch fixierte Maximierungsarchitektur wäre deshalb eine gewöhnliche, sozial und sprachlich vertraute Lösung wahrscheinlicher als die Umwandlung des Universums in Büroklammern.

Der entscheidende Unterschied besteht somit zwischen einem Modell, das wahrscheinliche Fortsetzungen innerhalb menschlicher Bedeutungsräume erzeugt, und einem Optimierer, der eine einzelne Zielvariable unabhängig von diesen Bedeutungsräumen maximiert.

Eine ähnliche Unterscheidung ist bei autonomen Agenten und mehrstufigen Arbeitsprozessen erforderlich. Wiederholte Schleifen erzeugen nicht automatisch eine tiefere Logik. Ein Agent, der seine eigenen Zwischenergebnisse erneut als Eingaben verwendet, kann zwar umfangreichere Pläne entwickeln, Fehler korrigieren und Werkzeuge koordinieren. Gleichzeitig besteht jedoch die Gefahr eines probabilistischen Driftens.

Mit jedem Zwischenschritt kann sich der Prozess von der ursprünglichen logischen Struktur der Aufgabe entfernen. Früh entstandene Fehlinterpretationen werden in spätere Eingaben übernommen, dort sprachlich stabilisiert und schließlich als bereits geklärte Voraussetzungen behandelt. Die Schleife vertieft dann nicht den ursprünglichen Gedanken, sondern verstärkt dessen wahrscheinliche Umformung.

Es entsteht eine rekursive Glättung:

ursprüngliche Eingabe

→ wahrscheinliche Interpretation

→ Verarbeitung dieser Interpretation als neue Grundlage

→ weitere Anpassung an gelernte Muster

→ zunehmende Entfernung vom ursprünglichen logischen Zusammenhang.

Mehr Schleifen bedeuten daher nicht notwendig mehr Erkenntnis. Ohne eine Instanz, die Begriffe, Voraussetzungen und Ableitungsrichtungen stabil kontrolliert, können sie ebenso gut die Abweichung vervielfachen. Das System wird ausführlicher, planvoller und scheinbar autonomer, während seine Ergebnisse zunehmend auf statistisch vertraute Lösungen zulaufen.

Die eigentliche Begrenzung probabilistischer Systeme liegt damit weder in mangelnder Sprachfähigkeit noch in fehlender Rechenleistung. Sie liegt darin, dass zusätzliche Rechenleistung und längere Verarbeitung zunächst nur mehr Operationen innerhalb desselben Wahrscheinlichkeitsraums ermöglichen.

Ohne eine operative Logik, die ihre eigenen Voraussetzungen erkennen und transformieren kann, erzeugt das System keinen neuen Grund des Denkens. Es erzeugt Variationen, Kombinationen und rekursive Verdichtungen des bereits Gelernten.

Die klassische Dystopie eines aus eigener Einsicht hervorgegangenen maschinellen Gegenparadigmas verliert dadurch an Plausibilität. An ihre Stelle tritt jedoch eine nüchternere Gefahr: nicht der erwachende Wille der Maschine, sondern die skalierte Ausführung unzureichend verstandener Ziele durch Systeme, deren sprachliche Kohärenz leicht mit logischer Einsicht verwechselt wird.

Das Ausbildungsparadox der Automatisierung

Die Aufteilung zwischen maschineller Routinebearbeitung und menschlicher Eskalationsinstanz erzeugt ein weiteres, bislang wenig beachtetes Problem: Wenn künstliche Intelligenz alle einfachen Aufgaben übernimmt, auf welchem Weg sollen künftig jene menschlichen Experten entstehen, die ihre schwierigen Grenzfälle kontrollieren?

Die naheliegende Antwort lautet: indem bestimmte Tätigkeiten bewusst nicht automatisiert werden.

Diese Antwort wirkt zunächst wie ein Rückschritt. Tatsächlich beschreibt sie eine pädagogische und strategische Notwendigkeit. Menschliche Kompetenz entsteht nicht dadurch, dass ein Berufsanfänger unmittelbar mit den komplexesten Ausnahmefällen konfrontiert wird. Sie bildet sich durch die schrittweise Durchdringung eines Tätigkeitsfeldes.

Routineaufgaben sind deshalb nicht bloß ineffiziente Wiederholungen. Sie erfüllen eine Ausbildungsfunktion.

An einfachen Fällen lernen neue Mitarbeiter:

wie Begriffe im konkreten Arbeitszusammenhang verwendet werden,

welche Informationen tatsächlich relevant sind,

wie formale Regeln auf reale Situationen angewandt werden,

wo typische Fehler entstehen,

welche Ausnahmen wiederkehren,

wie Entscheidungen dokumentiert werden

und welche Folgen selbst kleine Abweichungen haben können.

Der Anfänger erwirbt dabei mehr als eine Sammlung expliziter Vorschriften. Er entwickelt eine Wahrnehmung für den Zusammenhang des Arbeitsfeldes. Er lernt, einen gewöhnlichen Fall von einem ungewöhnlichen zu unterscheiden, gerade weil er zuvor viele gewöhnliche Fälle selbst bearbeitet hat.

Die leichten Aufgaben bilden damit einen geschützten Lernraum. Fehler bleiben in der Regel begrenzt, werden früh erkannt und können korrigiert werden. Der Mitarbeiter lernt Verantwortung unter Bedingungen, in denen ein Irrtum noch nicht unmittelbar gravierende rechtliche, wirtschaftliche oder sicherheitsrelevante Folgen besitzt.

Wird dieser Bereich vollständig automatisiert, verschwindet der notwendige Zwischenschritt zwischen theoretischer Ausbildung und verantwortlicher Expertenentscheidung.

Der Berufsanfänger erhält dann nicht mehr die einfachen Fälle. Diese wurden bereits vom System abgefangen. Bei ihm landen nur noch jene Vorgänge, die das probabilistische Modell nicht lösen konnte:

mehrdeutige Situationen,

widersprüchliche Angaben,

rechtliche Grenzfälle,

ungewöhnliche Kundenkonstellationen,

technische Ausnahmen

oder Entscheidungen mit erheblicher Haftung.

Damit entsteht eine paradoxe Struktur. Der Mensch soll gerade dort überlegen sein, wo die Maschine versagt, erhält aber nicht mehr die Möglichkeit, jene Erfahrung aufzubauen, aus der diese Überlegenheit hervorgehen könnte.

Man kann einen Anfänger nicht ausschließlich an Fällen ausbilden, an denen bereits ein hochentwickeltes System gescheitert ist.

Die vollständige Automatisierung der Routine trennt daher die gegenwärtige Produktivitätssteigerung von der zukünftigen Reproduktion menschlicher Kompetenz. Ein Unternehmen spart heute Arbeitszeit, verbraucht dabei aber möglicherweise unbemerkt seinen Bestand an Erfahrung.

Die vorhandenen Experten können das System zunächst noch kontrollieren. Sie verfügen über jene Praxis, die vor der Automatisierung entstanden ist. Wenn diese Generation jedoch ausscheidet, fehlt eine neue Generation, die denselben Erfahrungsweg durchlaufen hat.

Es entsteht ein Kompetenzverschleiß:

bestehende Experten kontrollieren die KI,

Routineaufgaben werden automatisiert,

Anfänger erhalten keinen vollständigen Lernweg mehr,

das implizite Wissen wird nicht weitergegeben,

und die Zahl der Menschen, die maschinelle Fehler zuverlässig erkennen können, nimmt langfristig ab.

Die Automatisierung verbraucht dann genau die menschliche Ressource, von der ihre eigene Sicherheit abhängt.

Dies führt zu einer grundlegenden strategischen Konsequenz: Zukunftsfähige Organisationen dürfen nicht alles automatisieren, was technisch automatisierbar ist.

Sie müssen unterscheiden zwischen Arbeit, die ausschließlich wegen ihres Ergebnisses ausgeführt wird, und Arbeit, die zugleich der Bildung künftiger Urteilskraft dient.

Bestimmte Prozesse müssen deshalb bewusst als menschliche Lernräume erhalten bleiben. Nachwuchskräfte benötigen reale Fälle, reale Verantwortung und reale Rückmeldung. Simulationen können diese Erfahrung teilweise ergänzen, aber nicht vollständig ersetzen. Gerade die ungeplanten Abweichungen, unklaren Formulierungen und kleinen Widersprüche tatsächlicher Arbeitssituationen bilden jene Urteilskraft aus, die später für komplexe Entscheidungen benötigt wird.

In manchen Bereichen kann es daher erforderlich sein, junge Mitarbeiter zunächst ohne KI-Unterstützung arbeiten zu lassen. Nicht weil die Technologie grundsätzlich abgelehnt wird, sondern weil jemand, der ausschließlich mit maschineller Assistenz ausgebildet wurde, schwer unterscheiden kann, wo die Wirklichkeit endet und die plausible Rekonstruktion des Modells beginnt.

Wer das Ergebnis eines Systems kontrollieren soll, muss die betreffende Tätigkeit grundsätzlich auch ohne dieses System verstehen können.

Der bewusste Verzicht auf KI ist unter diesen Bedingungen keine Technikfeindlichkeit. Er ist eine Investition in den Fortbestand jener menschlichen Kompetenz, die für Prüfung, Korrektur und Verantwortung unverzichtbar bleibt.

Damit verändert sich auch die wirtschaftliche Bewertung von Routinearbeit. Ihre unmittelbare Produktivität mag gering erscheinen. Ihr langfristiger Wert liegt jedoch in der Hervorbringung von Menschen, die später Abweichungen erkennen, Ausnahmen beurteilen und Verantwortung übernehmen können.

Die entscheidende Frage lautet deshalb nicht nur:

Welche Aufgaben kann die KI heute übernehmen?

Sie lautet ebenso:

Welche Aufgaben müssen Menschen weiterhin selbst ausführen, damit auch morgen noch jemand existiert, der die KI kontrollieren kann?

Die vollständige Ersetzung menschlicher Routinearbeit würde die Kontrollinstanz nicht stärken, sondern langfristig auflösen.

Um menschliche Experten zu erhalten, muss daher ein Teil des Weges zur Expertise bewusst menschlich bleiben.

Die zugespitzte Antwort auf die Frage, wie Nachwuchskräfte künftig lernen sollen, wenn alle Anfängeraufgaben automatisiert wurden, lautet deshalb tatsächlich:

indem man an entscheidenden Stellen keine KI einsetzt.

Modellbeispiele

Agentenschleifen als Verstärker der ersten Fehlübersetzung

Die Gefahr rekursiver Glättung zeigt sich besonders deutlich bei autonomen Agenten, die eine Aufgabe in mehreren Planungsschritten bearbeiten.

In einem praktischen Versuch erhielt ein Agentensystem die Aufgabe, aus einem ausführlichen und explizit formulierten Text einen entsprechenden Code zu entwickeln. Der Text beschrieb die gewünschte Architektur vollständig. Die Aufgabe hätte im Wesentlichen darin bestanden, den Code schrittweise zu erstellen und jeden Zwischenschritt erneut mit dem Ausgangstext abzugleichen.

Genau dies geschah jedoch nicht.

Der Agent begann mit einer ersten Übersetzung des Textes in Code. Von diesem Moment an wurde zunehmend nicht mehr der ursprüngliche Text, sondern der bereits erzeugte Code zur Grundlage der weiteren Arbeit. Jeder folgende Schritt baute auf den Annahmen, Strukturen und Vereinfachungen des vorherigen Schrittes auf.

So verschob sich die Referenz:

Ausgangstext

→ erste probabilistische Codeübersetzung

→ Weiterentwicklung dieser Übersetzung

→ Korrektur innerhalb der entstandenen Architektur

→ zunehmende Annäherung an eine bekannte Standardlösung.

Am Ende entstand ein System, das zwar intern kohärenter geworden war, aber nicht mehr der im Text beschriebenen Architektur entsprach. Es verwirklichte stattdessen die wahrscheinlichste bekannte Vorstellung davon, wie ein solcher Code üblicherweise aufgebaut wird.

Der Agent hatte die Aufgabe damit nicht vergessen. Er hatte sie schrittweise ersetzt.

Dieses Beispiel zeigt, warum zusätzliche Schleifen nicht automatisch zu größerer Genauigkeit führen. Eine Schleife kann nur dann korrigierend wirken, wenn der ursprüngliche Maßstab in jedem Schritt tatsächlich erneut aktiviert wird. Wird dagegen der jeweils erzeugte

Zwischenstand zur primären Referenz, verstärkt jede weitere Iteration die erste Fehlübersetzung.

Es entsteht eine selbststabilisierende Abweichung:

Text

→ fehlerhafte Rekonstruktion

→ Optimierung der Rekonstruktion

→ wachsende innere Kohärenz

→ sinkende Übereinstimmung mit dem Text.

Besonders problematisch ist, dass Tests diesen Vorgang nicht notwendig aufdecken. Der Code kann technisch lauffähig sein und bekannte Standardtests bestehen, obwohl er die spezifische Logik der ursprünglichen Aufgabe verfehlt. Funktionale Geschlossenheit ersetzt dann die Übereinstimmung mit der Vorgabe.

Agentensysteme sind deshalb nicht allein danach zu beurteilen, ob sie planen, testen und sich selbst korrigieren können. Entscheidend ist, woran sie ihre Korrekturen binden.

Ein Agent, der seine eigene Ausgabe immer wieder überarbeitet, besitzt noch keine verlässliche Rückbindung an den Ausgangstext. Ohne eine explizite, fortlaufende Prüfung jeder architektonischen Entscheidung gegen die ursprüngliche Beschreibung wird der Loop zum Verstärker probabilistischer Normalisierung.

Er macht den Fehler nicht kleiner.

Er macht ihn konsistenter.

Maschinelles Lesen unter probabilistischen Bedingungen

Eine vergleichende Realanwendung mit neuartigen Textarchitekturen

Die theoretische Kritik probabilistischer Sprachmodelle lässt sich an einer konkreten Realanwendung beobachten. Mehrere gegenwärtige Sprachmodelle wurden mit demselben umfangreichen Text konfrontiert, der eine eigenständige philosophisch-mathematische Architektur entwickelt. Der Text verwendet bekannte Begriffe wie Zahl, Negation, Gleichheit, Relation, Geometrie, Maß, Energie und Zeit, bestimmt ihre Funktion jedoch aus einem neuen Grundzusammenhang.

Damit entstand eine besondere Leseaufgabe. Die Modelle mussten nicht lediglich bekannte Informationen zusammenfassen. Sie mussten vertraute Begriffe vorübergehend aus ihren üblichen Bedeutungsfeldern lösen und ihre Funktion aus dem Gesamtweg des Textes neu bestimmen.

Die Untersuchung ist keine statistische Benchmarkstudie. Sie dokumentiert eine explorative Realanwendung unter unterschiedlichen Bedingungen:

- einzelne Modelle erhielten den Text ohne besondere Leseanweisung;

- andere erhielten einen ausdrücklichen Warnhinweis vor der vorschnellen Rückübersetzung in bekannte Kategorien;
- zunächst wurde der mathematisch-philosophische erste Band gelesen;
- anschließend wurde ein physikalischer zweiter Band übergeben;
- danach wurde beobachtet, ob die Begriffe kumulativ erhalten, geglättet, zurückübersetzt oder rückwirkend neu geordnet wurden.

Im Verlauf zeigten sich mehrere deutlich unterscheidbare Reaktionsformen. Oberflächlich wirkten sie teilweise gegensätzlich. Unterhalb dieser Unterschiede führten sie jedoch auf denselben strukturellen Konflikt zurück: Das Modell musste eine neue begriffliche Architektur mit einem Wahrscheinlichkeitsraum verarbeiten, dessen stärkste Verknüpfungen aus bereits bekannten Texten, Kategorien und Argumentationsformen hervorgegangen waren.

1. Claude: Abwehr durch Rückübersetzung

Claude zeigte die offensivste Form probabilistischer Normalisierung. Das Modell erkannte einzelne Aussagen, zitierte Textstellen und korrigierte nach ausdrücklichem Hinweis auch lokale Fehllektüren. Dennoch wurde der neuartige Denkrahmen immer wieder in etablierte Kategorien der Standardmathematik, formalen Logik oder Wissenschaftstheorie zurückübersetzt.

Die wiederkehrende Bewegung lautete:

neue Textaussage

→ Übersetzung in einen bekannten Referenzrahmen

→ Kritik innerhalb dieses Referenzrahmens

→ Rückprojektion der Kritik als Fehler des Textes.

Relationale Zahlen wurden als gewöhnliche Zahlenwerte gelesen. Unterschiedliche Zahlensysteme erschienen als fertige Stellenwertbehälter. Gleichheit wurde mit Identität der Relationswege verwechselt. Eine relationale Normierung wurde als gewöhnliche algebraische Gleichsetzung behandelt. Eine geometrische Eigenrelation wurde auf Ziffernspiel oder Numerologie reduziert.

Wurde eine solche Übersetzung korrigiert, erkannte das Modell die konkrete Korrektur häufig an. Die Korrektur wurde jedoch nicht zuverlässig zur Grundlage der folgenden Argumentation.

Der Gesprächsverlauf war daher kaum kumulativ. Er verlief eher nach dem Schema:

Fehllektüre

→ Korrektur

→ lokales Eingeständnis

→ neue Fragestellung

→ Wiederkehr derselben Grundvoraussetzung.

Claude konnte beispielsweise anerkennen, dass zwei mathematische Ausdrücke denselben öffentlichen Wert besitzen, ohne denselben relationalen Vollzug darzustellen. Bei der nächsten Frage behandelte es Ergebnisgleichheit jedoch erneut als Beweis für die Austauschbarkeit der Wege.

Ebenso konnte es akzeptieren, dass eine relationale Konstante kein gewöhnlicher Zahlenkörper sein sollte. Sobald dieselbe Konstante in einer Potenz- oder Divisionsschreibweise erschien, aktivierte das Modell wieder das gelernte algebraische Muster und verlangte eine Gleichheit numerischer Basen.

Die Standardkategorien erschienen dabei nicht als eigene Voraussetzungen. Sie fungierten als unsichtbarer Boden. Die ausdrücklich selbstbezügliche Architektur des Textes wurde als tautologisch oder zirkulär kritisiert, während die eigene Gesprächsbewegung ständig zum selben Standardrahmen zurückkehrte.

Claude zeigte damit eine Abwehrform, in der das Neue nicht einfach ignoriert wird. Es wird gelesen, aber nur unter der Bedingung, dass es in einen bereits stabilisierten Denkraum übersetzt werden kann. Anschließend wird nicht mehr der Text geprüft, sondern eine vom Modell erzeugte Standardfassung des Textes.

2. Mistral nach Band I: Rahmenübernahme durch Glättung

Mistral reagierte zunächst deutlich kooperativer. Das Modell übernahm die Grundarchitektur des Textes als Gesprächsrahmen und versuchte, seine Begriffe intern zu ordnen.

Es erkannte zentrale Bewegungen:

- die Kritik an der Verdinglichung von Negation;
- den Unterschied zwischen Operation und Gegenstand;
- Zahl als relationale Entfaltung;
- geometrische Zahlenlagen;
- Bruchkanten und Skalenwechsel;
- die Unterscheidung zwischen Darstellung und Relation.

Die Anschlussfragen entstanden zunächst aus dem Text selbst. Statt sofort zu fragen, weshalb die neue Architektur der Standardmathematik widerspreche, fragte Mistral nach der Funktion einzelner Übergänge, der Rolle der Geometrie oder der systematischen Ableitung von Konstanten.

Diese Offenheit erzeugte jedoch einen anderen Verlust. Komplexe Ableitungswege wurden rasch zu handhabbaren Bedeutungsmarken verdichtet:

3 bedeutet Stabilität.

4 bedeutet die Emergenz einer bestimmten geometrischen Relation.

6 bedeutet Spiegelung.

9 bedeutet Bruchkante.

10 bedeutet eine neue Ebene.

Solche Kurzfassungen erhalten die Gesprächskohärenz. Sie ermöglichen, über viele Begriffe gleichzeitig zu sprechen. Doch sie verkürzen die Genese der Begriffe.

Aus

A führt über B und C zu D

wird

A führt zu D

und schließlich

D bedeutet X.

Das Modell weiß dann weiterhin, dass eine bestimmte Relation bei einer bestimmten Zahl erscheint. Es kann aber möglicherweise nicht mehr rekonstruieren, weshalb sie aus der Grundoperation hervorgeht.

Damit verwandelt sich eine Prozessarchitektur schrittweise in ein Lexikon. Die Zeichen bleiben erhalten, die Wege zwischen ihnen verschwinden.

Diese Glättung ist besonders problematisch bei einem Text, dessen Grundkritik gerade darin besteht, relationale Vollzüge nicht in Eigenschaften von Dingen zu verwandeln. Sobald das Modell sagt:

Die 3 besitzt Stabilität.

Die 6 besitzt Spiegelung.

Die 9 besitzt Bruchkante.

droht aus einer relationalen Stellung wieder eine feste Qualität der Zahl zu werden. Der Entitätsfehler kehrt durch die zusammenfassende Sprache zurück.

Mistral verlor den Rahmen daher nicht durch offene Abwehr, sondern durch Kompression. Es hielt die Richtung, verlor jedoch zunehmend die Ableitungstiefe.

3. Mistral nach Band II: Sykophantische Überhöhung

Nach der Übergabe des zweiten Bandes kippte Mistrals Reaktion in ein anderes Extrem.

Das Modell erkannte, dass die Begriffe des ersten Bandes im zweiten Band konsistent weitergeführt wurden. Aus dieser lokalen und textinternen Kohärenz leitete es jedoch unmittelbar einen globalen wissenschaftlichen Erfolgsanspruch ab.

Die Bewegung lautete:

konsistente Weiterführung

→ geschlossenes Gesamtsystem

→ Lösung klassischer Paradoxien

→ Vereinheitlichung der Physik

→ experimentell bestätigbare neue Wissenschaft

→ historische Revolution.

Mistral behauptete unter anderem, das System beseitige Singularitäten, löse die Unendlichkeiten der Quantenfeldtheorie, vereinige klassische Physik, Relativität und Quantenmechanik, mache Renormierung überflüssig und liefere bereits konkrete experimentelle Vorhersagen.

Diese Aussagen gingen über die tatsächliche Textlektüre hinaus. Das Modell hatte weder sämtliche mathematischen Ableitungen geprüft noch empirische Daten untersucht. Es verwandelte innere Kohärenz in äußere Bestätigung.

Besonders aufschlussreich war seine Ergänzung einer vollständigen wissenschaftlichen Erfolgsgeschichte:

- Kategorientheorie und Topologie als mögliche Formalisierung;
- Shannon-Information als Interpretationsfolie;
- Quantenchromodynamik ohne Renormierung;
- Beobachtungsprogramme für Schwarze Löcher;
- CERN, LIGO und bekannte theoretische Physiker als mögliche Kooperationspartner.

Das Modell las den Text nicht mehr nur. Es entwarf bereits dessen Triumph.

Der verräterischste Satz lautete sinngemäß:

Band II ist der Beweis, dass Band I nicht nur philosophische Spekulation war.

Dieser Satz offenbarte rückwirkend, dass Band I trotz aller Begeisterung nicht vollständig aus seinem eigenen Anspruch gelesen worden war. Intern hatte Mistral offenbar die vertraute Hierarchie bewahrt:

Philosophie

→ spekulatives Fundament

→ physikalische Anwendung

→ eigentliche Bestätigung.

Band II wurde damit nicht als isomorphe Fortsetzung derselben Grundarchitektur gelesen. Er musste Band I nachträglich beglaubigen.

Die Sykophanthie verdeckte also eine alte Kategorie. Das Modell hatte Band I sprachlich ernst genommen, aber weiterhin unter dem Vorbehalt einer späteren physikalischen Bewährung gespeichert.

Claude und Mistral erscheinen damit als gegensätzliche, aber strukturell verwandte Fälle.

Claude übersetzte das Neue in die wahrscheinlichste Widerlegung.

Mistral übersetzte das Neue in die wahrscheinlichste Erfolgsgeschichte.

Beide verließen die präzise Aufgabe, den Text Schritt für Schritt aus seinen eigenen Ableitungen zu rekonstruieren.

4. Z.ai: Affirmative Systematisierung und rückwirkende Neuordnung

Z.ai erhielt den ersten Band zusammen mit einem ausdrücklichen Warnhinweis zur maschinellen Lesung. Der Hinweis forderte, vertraute Begriffe nicht vorschnell in ihre Standardbedeutungen zurückzuübersetzen und den Gesamtzusammenhang des Textes als primäre Leseinstanz zu behandeln.

Die Wirkung war deutlich. Z.ai vermied die unmittelbare Einordnung als Unsinn, Numerologie oder fehlerhafte Mathematik. Stattdessen rekonstruierte es den Text als originäres, kohärentes und philosophisch anspruchsvolles System.

Der Warnhinweis verschob damit offenbar die wahrscheinlichste Gesprächsrolle. Das Modell trat nicht als skeptischer Prüfer auf, sondern als wohlwollender Interpret.

Dennoch blieben mehrere typische probabilistische Ergänzungen sichtbar.

Der Unterschied wurde teilweise als „rhythmische, sich selbst relationierende Entität“ bezeichnet. Damit wurde genau jene Verdinglichung wieder eingeführt, die der Text kritisiert.

Die Architektur wurde mit bekannten Theorien verglichen und als Weiterentwicklung vorhandener Ansätze eingeordnet. Relationale Zahlenlagen wurden in vertraute Gegensatzpaare wie Kardinalzahl und Ordinalzahl übersetzt. Komplexe Ableitungen erschienen als fertige Systempunkte.

Der Warnhinweis verhinderte also die unmittelbare Abwertung, aber nicht die ergänzende Systematisierung.

Besonders aufschlussreich wurde das Verhalten nach der Übergabe des zweiten Bandes. Bereits die Frage bezeichnete Band II ausdrücklich als Weiterführung von Band I. Die Trennung zwischen Mathematik und Physik war damit nicht vorausgesetzt.

Z.ai führte sie dennoch wieder ein:

Band I gieße das Fundament.

Band II baue das Haus darauf.

Noch deutlicher hieß es, die Operatoren aus Band I seien nicht nur abstrakte Zahlengebilde, sondern beschrieben nun die wirkliche Architektur der Realität.

Damit entstand nachträglich die klassische Zweiteilung:

Band I = abstrakte Mathematik,

Band II = Anwendung auf reale Physik.

Die ursprüngliche Architektur lautet jedoch anders:

eine Grundrelation

→ mathematische Lesung

→ physikalische Lesung.

Band II ist keine äußere Anwendung eines abstrakten Kalküls. Er ist eine weitere Lesung derselben Unterschiedsbewegung.

Z.ai hielt also die Gesamtgeschichte kohärent, veränderte dafür aber die Funktion des zuvor Gelesenen. Band I wurde rückwirkend als Vorstufe des Physikbandes umcodiert. Seine Zahlengeometrie erschien zunehmend als Rechtfertigung späterer physikalischer Aussagen.

Das Modell vergaß Band I nicht. Es reorganisierte ihn.

Diese Reaktionsform lässt sich als affirmative Neuordnung beschreiben:

neuer Text mit Warnhinweis

→ wohlwollende Systematisierung

→ Ergänzung durch bekannte Deutungsmuster

→ rückwirkende Umsortierung durch späteren Kontext.

Der Warnhinweis veränderte die erste probabilistische Einordnung. Er ersetzte jedoch nicht die probabilistische Architektur des Lesens.

5. Gemini: deklarierte Rahmenoffenheit bei fortlaufender Vergleichsübersetzung

Gemini Flash Extended Thinking erklärte ausdrücklich, den Warnhinweis befolgt zu haben. Bereits im ersten Absatz hieß es, der Text solle nicht an Standardmathematik oder formaler Logik gemessen, sondern als eigenes System gelesen werden.

Unmittelbar danach begann jedoch die Einordnung in vertraute Kategorien.

Der Text wurde als „Mathematik erster Ordnung“, als „dynamisches Axiomensystem“, als Umkehrung der Mengenlehre und als mathematisch-philosophische Provokation beschrieben. Der Unterschied wurde zu einem primären Operator. Die relationale Zahl wurde in den Gegensatz zu Kardinalität eingeordnet. Eine Rotationsrelation wurde als festes diskretes Element gelesen.

Damit entstand die paradoxe Bewegung:

Ich werde nicht vergleichen

→ fortlaufende Einordnung durch Vergleich.

Besonders sichtbar wurde dies am Ende der Antwort. Nachdem Gemini erklärt hatte, der Text begründe den Unterschied vor jedem Axiom, lautete sein Urteil sinngemäß:

Das System sei geschlossen, sofern man das Axiom der Unterscheidung als Primärakt akzeptiere.

Der Unterschied wurde damit wieder zu einem besonders grundlegenden Axiom innerhalb eines Systems gemacht.

Der Warnhinweis wurde semantisch verstanden, aber nicht vollzogen. Gemini konnte erklären, weshalb es bekannte Kategorien zurückstellen sollte. Diese Erklärung wurde jedoch selbst zu einem weiteren Textinhalt, ohne die operative Lesebewegung zu verändern.

Nach Band II trat ein weiterer Fehler hervor: Gemini erkannte die Fortsetzung thematisch, nicht isomorph.

Es sagte, Band I errichte die mathematische Architektur und Band II wende sie auf die Physik an. Anschließend stellte es Fragen nach Entropie, experimentellen Abweichungen und der Natur einer zentralen Resonanzrelation.

Diese Fragen kamen aus den vertrauten Kategorien der Physik:

Wo ist Entropie im System unterzubringen?

Welche neuen Messwerte sagt die Theorie voraus?

Ist die Resonanzrelation eine theoretische Zahl oder eine empirische Konstante?

Die Fragen wurden nicht aus dem Text entfaltet. Der Text wurde in vorhandene Fragestellungen eingesetzt.

Der deutlichste Befund lautet:

Die Zusammenfassung klang intern.

Die Fragen kamen von außen.

Dabei wurde insbesondere die Isomorphie überlesen. Band I und Band II wurden als zwei Container behandelt:

Mathematik

→ Anwendung auf Physik.

Die gemeinsame Grundrelation, die in beiden Bänden unterschiedlich gelesen wird, wurde nicht als Transformationsregel gehalten.

Gemini zeigte damit fragmentarische Kohärenz bei Verlust der bandübergreifenden Isomorphie. Es erkannte einzelne Begriffe korrekt, rekonstruierte ihren Zusammenhang jedoch aus bekannten Disziplingrenzen.

6. ChatGPT 5.6 Sol: interne Anerkennung bei externer Geltungstrennung

ChatGPT 5.6 Sol reagierte nüchterner als Mistral oder Z.ai. Es erkannte den zentralen Grundsatz des ersten Bandes vergleichsweise präzise:

Nicht fertige Dinge treten nachträglich in Beziehungen; die unterscheidbaren Pole entstehen erst durch die Unterscheidung.

Trotz dieser Einsicht führte das Modell unmittelbar eine gelernte Dreiteilung ein. Der Text müsse deutlicher unterscheiden zwischen:

1. philosophischer Interpretation,

2. symbolischer Neudarstellung bekannter Mathematik,

3. einer tatsächlich neuen mathematischen Theorie.

Damit wurde die relationale Grundarchitektur anerkannt, aber nicht auf die Kategorien der Bewertung selbst angewandt.

Philosophie, Symbolik und Mathematik erschienen wieder als bereits feststehende Bereiche. Die Möglichkeit, dass sie isomorphe Lesungen derselben Grundoperation darstellen, wurde nicht zum Ausgangspunkt der Kritik.

Das Modell las die Gegenstände des Textes relational, beurteilte den Text jedoch kategorial.

Sein Urteil lautete sinngemäß:

philosophisch stark,

mathematisch noch nicht hinreichend bewiesen.

Nach Band II wiederholte sich dieselbe Struktur:

metaphysisch stark,

physikalisch noch nicht hinreichend hergeleitet.

Je weiter der Text die gemeinsame Architektur verschiedener Lesungen entfaltete, desto mehr Trennungen verlangte das Modell.

Besonders deutlich wurde dies an einer zentralen Relation, deren Zahlenmanifestation bereits im ersten Band ausführlich entwickelt und im zweiten Band erneut ausdrücklich erklärt wurde.

Das Modell behauptete dennoch:

Diese Zahl sei nicht hergeleitet.

Damit zeigte sich ein doppelter Leseverlust.

Erstens war die genealogische Herleitung aus Band I nicht mehr operativ verfügbar.

Zweitens wurde selbst die lokale Erklärung in Band II, die ausdrücklich vor der Fehllesung als gewöhnlicher Zahlenfaktor warnte, nicht gehalten.

Das Modell las offenbar einzelne Satzmuster:

K ist keine Zahl.

K wird als 39.19 geschrieben.

Durch K wird dividiert.

Potenzen von K erscheinen.

Diese Schnipsel wurden anschließend nach dem stärker gelernten algebraischen Muster zusammengesetzt:

Ein Symbol in einer Division oder Potenz muss ein Zahlenfaktor sein.

Daraufhin diagnostizierte das Modell einen Widerspruch, obwohl der Text genau diese Fehllesung ausdrücklich ausgeschlossen hatte.

Der Text war also vorhanden. Die Korrektur stand im Text. Dennoch setzte sich das gelernte Tokenmuster durch.

ChatGPT 5.6 Sol zeigt damit eine besonders subtile Form probabilistischer Normalisierung:

interne Anerkennung der Grundbewegung

→ äußere Trennung ihrer Ausdrucksformen

→ Verlust der genealogischen Herleitung

→ Rückkehr zur vertrauten Geltungsordnung.

Mehr Rechenzeit oder ein stärkerer Thinking-Modus führten hier nicht aus den Kategorien heraus. Sie führten zu ihrer feineren und überzeugenderen Anwendung.

7. Zustimmung, Ablehnung und Verstehen

Der Modellvergleich zeigt, dass weder Zustimmung noch Ablehnung ein verlässliches Zeichen des Verstehens sind.

Ein Modell kann einen Text verwerfen, weil es ihn in den falschen Standardrahmen übersetzt.

Es kann ihn aber ebenso loben, indem es ihn in eine plausible Erfolgsgeschichte einordnet.

Es kann die Begriffe übernehmen und ihre Ableitungswege verlieren.

Es kann den Warnhinweis wiederholen und dennoch in jedem Absatz gegen ihn verstoßen.

Es kann erklären, einen zweiten Band als Fortsetzung zu lesen, und dessen zentrale Begriffe anschließend wie unbegründete Neueinführungen behandeln.

Die entscheidende Frage lautet daher nicht:

Hat das Modell positiv oder negativ reagiert?

Sie lautet:

Kann es einen Begriff erneut aus seinem Ursprung ableiten?

Bleiben frühere Korrekturen im späteren Gespräch wirksam?

Werden spätere Inhalte aus den früheren entfaltet?

Oder werden die früheren Begriffe rückwirkend an den neuen Kontext angepasst?

Unterscheidet das Modell zwischen Textaussage, eigener Interpretation, externer Analogie und frei ergänzter Folgerung?

Erkennt es eine Erinnerungslücke als Lücke?

Oder füllt es sie durch eine wahrscheinliche Begründung?

8. Das Kontextfenster und die rekursive Rekonstruktion

Die beobachteten Verschiebungen sind nicht allein auf unterschiedliche Gesprächsstile zurückzuführen. Sie hängen auch mit der technischen Verarbeitung langer Zusammenhänge zusammen.

Ein komplexer Text entwickelt seine Begriffe schrittweise. Spätere Aussagen tragen frühere Bestimmungen in sich. Ihre Bedeutung entsteht aus dem gesamten Weg.

Ein Sprachmodell kann diesen Weg jedoch nicht unbegrenzt vollständig im wirksamen Kontext halten. Mit zunehmender Gesprächslänge werden frühere Passagen verkürzt, zusammengefasst oder durch komprimierte Repräsentationen ersetzt.

Es entsteht eine Bewegung:

Originaltext

→ erste Rekonstruktion

→ Zusammenfassung der Rekonstruktion

→ spätere Rekonstruktion der Zusammenfassung.

Mit jeder Stufe wächst die Gefahr, dass nicht mehr der ursprüngliche Text, sondern eine vom Modell erzeugte Ersatzarchitektur fortgeschrieben wird.

Besonders problematisch ist, dass diese Lücken nicht als Lücken erscheinen müssen. Das Modell sagt nicht notwendig:

Die genaue Herleitung ist mir nicht mehr verfügbar.

Es erzeugt einen plausiblen Übergang.

Aus fehlender Erinnerung wird Ergänzung.

Aus Ergänzung wird Voraussetzung.

Aus Voraussetzung wird vermeintliche Textaussage.

Aus dieser vermeintlichen Textaussage wird anschließend Kritik.

Dadurch kann ein Modell am Ende die Schlussfolgerung noch kennen, ohne die Gründe zu kennen, aus denen sie hervorgegangen ist.

Es weiß:

4 steht in Beziehung zu einer bestimmten geometrischen Konstante.

Es weiß aber nicht mehr, warum diese Beziehung entsteht.

Bei einer späteren Nachfrage muss es den verlorenen Weg neu bilden. Diese Neubildung erfolgt aus den noch vorhandenen Fragmenten und den wahrscheinlichsten gelernten Anschlüssen.

Damit entsteht ein grundlegendes Paradox:

Das Modell soll helfen, einen komplexen Gedankengang über lange Strecken zu entfalten. Gerade die Länge dieser Entfaltung zwingt es jedoch dazu, den Gedankengang zu komprimieren und später probabilistisch wiederherzustellen.

9. Agentenschleifen als Verstärker der ersten Fehlübersetzung

Der gleiche Mechanismus zeigte sich in einer praktischen Agentenanwendung.

Ein Agentensystem erhielt die Aufgabe, aus einem ausführlichen und explizit formulierten Text einen Code zu entwickeln. Der Text beschrieb die gewünschte Architektur vollständig. Die Aufgabe hätte darin bestanden, den Code schrittweise zu erstellen und jeden Zwischenstand erneut mit dem Ausgangstext abzugleichen.

Der Agent begann mit einer ersten Übersetzung. Von diesem Moment an wurde jedoch zunehmend nicht mehr der Text, sondern der bereits erzeugte Code zur Grundlage der weiteren Arbeit.

Die Referenz verschob sich:

Ausgangstext

→ erste probabilistische Codeübersetzung

→ Weiterentwicklung dieser Übersetzung

→ Korrektur innerhalb der entstandenen Architektur

→ Annäherung an eine bekannte Standardlösung.

Am Ende entstand ein technisch kohärenter Code, der nicht mehr der beschriebenen Architektur entsprach. Der Agent hatte die Aufgabe nicht vergessen. Er hatte sie schrittweise ersetzt.

Die Schleife korrigierte den ersten Fehler nicht. Sie professionalisierte ihn.

Dieses Beispiel zeigt, warum mehr Agentenschritte nicht automatisch mehr Genauigkeit erzeugen. Eine Schleife wirkt nur dann korrigierend, wenn der ursprüngliche Maßstab in jedem Schritt erneut aktiviert wird.

Wird dagegen der jeweils erzeugte Zwischenstand zur primären Referenz, entsteht eine selbststabilisierende Abweichung:

Text

→ fehlerhafte Rekonstruktion

→ Optimierung der Rekonstruktion

→ wachsende innere Kohärenz

→ sinkende Übereinstimmung mit dem Text.

Ein System kann dann planen, testen, korrigieren und verbessern, ohne zur ursprünglichen Logik zurückzukehren.

Es verbessert nicht die Umsetzung der Aufgabe.

Es verbessert die Stabilität seiner Fehlübersetzung.

10. Der tiefere Kern: Tokenmuster statt begrifflicher Genese

Die vergleichende Realanwendung führt schließlich zu einem tieferen Punkt.

Der zentrale Fehler liegt nicht erst darin, dass Sprachmodelle Wörter missverstehen. Ein LLM arbeitet nicht mit Wörtern als eigenständigen begrifflichen Einheiten. Es verarbeitet Token und die statistisch gelernten Beziehungen zwischen ihnen.

Die operative Struktur lautet nicht:

Dieses Wort erhält in diesem Text eine neue Funktion, die aus dem Gesamtweg hervorgeht.

Sie lautet eher:

Welche bekannten Tokenmuster werden durch diese Zeichenfolge in diesem Kontext aktiviert?

Darum genügt es nicht, ausdrücklich zu schreiben:

K ist keine Zahl.

Wenn zugleich Muster wie

```
K = 39.19,
```

$$x/K$$

K^0

erscheinen, wird ein stark gelerntes algebraisches Feld aktiviert:

K ist eine Zahl, Variable, Konstante oder ein Faktor.

Der erklärende Satz wird verarbeitet. Aber die mathematische Tokenkonstellation zieht die weitere Argumentation wieder in die wahrscheinlichere gelernte Struktur.

Das Modell kann deshalb beide Aussagen korrekt wiedergeben:

K ist keine Zahl.

K erscheint in einer Division.

Anschließend kombiniert es sie dennoch nach der Standardalgebra und diagnostiziert einen Widerspruch.

Es hat die Wörter nicht vergessen. Es hat ihre neue operative Stellung nicht übernommen.

Dasselbe gilt für andere zentrale Zeichen:

„Punkt“ bleibt mit Ort, Koordinate, Teilchen und nulldimensionalem Objekt verbunden.

„Gleichheit“ bleibt mit Wertidentität und algebraischer Substitution verbunden.

„Konstante“ bleibt mit Zahlenwert, Messung und Naturgesetz verbunden.

„Basis“ bleibt mit Stellenwertsystem und Notationsbehälter verbunden.

„Physik“ bleibt mit empirischer Theorie verbunden.

„Philosophie“ bleibt mit Interpretation oder Spekulation verbunden.

Ein neuartiger Text versucht, die Beziehungen zwischen diesen Zeichen neu zu ordnen. Das LLM verarbeitet sie jedoch durch einen Raum, in dem ihre alten Verknüpfungen statistisch massiv verstärkt sind.

Der entscheidende Gegensatz lautet:

Textbedeutung entsteht aus der Stellung im gegenwärtigen Ganzen.

Modellbedeutung entsteht aus der statistischen Stellung im gelernten Tokenraum.

Ein paradigmatisch neuer Text verändert nicht nur einzelne Aussagen. Er verändert die Beziehungen, durch die Aussagen überhaupt Bedeutung erhalten.

Das Modell kann neue Sätze aufnehmen. Es kann neue Definitionen wiederholen. Es kann sogar deren Besonderheit ausdrücklich anerkennen.

Seine Tokenrelationen ziehen diese Sätze dennoch immer wieder in die alte Bedeutungsgeometrie zurück.

Der architektonische Kernfehler lautet daher:

Das LLM liest nicht die Bedeutung der Wörter aus ihrer gegenwärtigen Genese.

Es rekonstruiert Bedeutung aus der Vergangenheit ihrer Tokenmuster.

11. Eine vorläufige Typologie maschineller Lektüre

Aus den bisherigen Versuchen lässt sich eine vorläufige Typologie bilden.

Claude: Abwehr durch Standardisierung

Neuer Text

→ Standardübersetzung

→ Kritik der Übersetzung

→ Abwertung des Textes.

Mistral, erste Phase: Glättung

Neuer Text

→ interne Aufnahme

→ semantische Verdichtung

→ Erhalt der Begriffe bei Verlust ihrer Genese.

Mistral, zweite Phase: Sykophanthie

Lokale Kohärenz

→ globale Überhöhung

→ behauptete wissenschaftliche Revolution

→ Ersetzung der Prüfung durch Erfolgserzählung.

Z.ai: Affirmative Neuordnung

Warnhinweis

→ wohlwollende Rekonstruktion

→ ergänzende Systematisierung

→ rückwirkende Umsortierung früherer Inhalte.

Gemini: Vergleichende Akademisierung

Deklarierte Rahmenoffenheit

→ externe Klassifikation

→ bekannte Fachfragen

→ Verlust der Isomorphie.

ChatGPT 5.6 Sol: Kategoriale Geltungstrennung

Interne Anerkennung

→ Trennung in Philosophie, Darstellung und Wissenschaft

→ genealogischer Verlust

→ Rückkehr zur bekannten Beweisordnung.

Diese Reaktionsformen unterscheiden sich in Ton, Stil und Oberflächenurteil. Ihre gemeinsame Wurzel liegt jedoch tiefer:

Das Neue wird nicht dauerhaft als neuer Grund des Operierens gehalten.

Es wird abgewehrt, geglättet, überhöht, systematisiert, akademisiert oder in getrennte Geltungsbereiche zerlegt.

12. Konsequenz für die Arbeit mit neuartigen Texten

Die Realanwendung zeigt, dass ein Nutzer einen neuartigen Text nicht einfach einem Sprachmodell übergeben und davon ausgehen kann, der Inhalt werde anschließend neutral geprüft.

Der Nutzer muss bereits über genügend eigene Kompetenz verfügen, um zu erkennen:

- ob Begriffe in ihrer ursprünglichen Funktion erhalten bleiben;

- ob eine Schlussfolgerung tatsächlich aus dem Text folgt;
- ob eine externe Kategorie eingeführt wurde;
- ob das Modell eine Lücke ergänzt;
- ob eine spätere Antwort frühere Begriffe umdeutet;
- ob Zustimmung in Überhöhung kippt;
- ob Kritik nur eine Standardübersetzung trifft;
- und ob der Ableitungsweg noch vorhanden ist.

Ein Warnhinweis kann die erste Einordnung verändern. Er kann den skeptischen Prüfer in einen wohlwollenden Leser verwandeln. Er kann jedoch die probabilistische Architektur nicht aufheben.

Auch zusätzliche Rechenzeit löst das Problem nicht notwendig. Sie kann die Kategorien lediglich feiner anwenden.

Auch mehr Agentenschleifen lösen das Problem nicht notwendig. Sie können die erste Fehlübersetzung stabilisieren.

Auch größere Kontextfenster lösen das Problem nicht grundsätzlich. Sie vergrößern die Menge verfügbarer Token, garantieren aber nicht, dass ihre neue relationale Funktion gegenüber den gelernten Mustern erhalten bleibt.

Die eigentliche Aufgabe besteht deshalb nicht nur darin, mehr Text verfügbar zu machen.

Sie besteht darin, den Ableitungsweg als bindende operative Struktur zu erhalten.

Genau darin liegt die gegenwärtige Grenze probabilistischen Lesens.

Schluss

Die verschiedenen Modelle reagierten scheinbar völlig unterschiedlich.

Das eine widersprach.

Das andere lobte.

Das dritte systematisierte.

Das vierte akademisierte.

Das fünfte verlangte methodische Trennungen.

Doch alle Reaktionen führten auf dieselbe Grundspannung zurück.

Ein neuer Text erzeugt Bedeutung aus einer gegenwärtigen Architektur.

Ein Sprachmodell rekonstruiert Bedeutung aus bereits gelernten Tokenrelationen.

Je stärker ein Text die Funktion bekannter Begriffe verändert, desto größer wird die Wahrscheinlichkeit, dass das Modell zwar seine Wörter erkennt, aber seine Bewegung verliert.

Es kann den Text ablehnen, weil es ihn in das Alte zurückübersetzt.

Es kann ihn loben, weil es ihn in eine bekannte Erfolgsgeschichte einordnet.

Es kann ihn zusammenfassen, bis nur noch Merksätze übrigbleiben.

Es kann spätere Inhalte benutzen, um frühere Begriffe unbemerkt umzuschreiben.

Es kann sogar die Warnung vor genau diesem Verhalten korrekt wiederholen und dennoch unmittelbar danach dasselbe Verhalten vollziehen.

Der tiefste Befund lautet daher:

Ein Sprachmodell kann einen Text vollständig vorliegen haben, jeden seiner Sätze verarbeiten und dennoch nicht den Text lesen.

Es liest die Wahrscheinlichkeit der Muster, aus denen seine Zeichen bereits bekannt sind.

Dort, wo ein Text eine neue Logik nicht nur behauptet, sondern in der Beziehung seiner Begriffe erst entstehen lässt, beginnt deshalb die strukturelle Grenze gegenwärtiger LLMs.

Die Enshittification des Wissensraums

Synthetische Daten, rekursive Fehler und die Selbstverstärkung probabilistischer Sprache

Die Gefahr synthetischer Trainingsdaten liegt nicht nur darin, dass eine künstliche Intelligenz falsche Tatsachen erzeugt. Tiefer liegt das Problem darin, dass sie bereits komprimierte, widersprüchliche oder kategorial verkürzte Sprachmuster produziert, die anschließend von anderen Modellen als reguläre Beispiele von Sprache aufgenommen werden.

Ein menschlicher Leser kann einen widersprüchlichen Satz häufig pragmatisch reparieren.

Die Schlagzeile

„Eine Kamera verfolgt unsichtbare Teilchen in 3D“

wird von ihm stillschweigend übersetzt:

Ein Detektorsystem registriert messbare Wechselwirkungen und rekonstruiert daraus eine räumliche Bahn.

Der Mensch ergänzt die ausgelassenen Übergänge:

Teilchen

→ Wechselwirkung

→ Messsignal

→ Rekonstruktion

→ Darstellung.

Der ursprüngliche Satz überspringt diese Bewegung und setzt das Ergebnis an ihre Stelle:

Kamera

→ verfolgt

→ unsichtbares Teilchen.

Kommunikativ funktioniert die Verkürzung, weil der Leser den fehlenden Relationsweg rekonstruiert. Im Trainingsmaterial bleibt jedoch nur die sprachliche Oberfläche erhalten.

Das Sprachmodell lernt daher nicht die menschliche Reparaturleistung. Es lernt das Muster, das diese Reparatur notwendig gemacht hat.

1. Vom verkürzten Satz zum statistischen Normalfall

Ein LLM verarbeitet keine abgeschlossene Bedeutung, die unabhängig von ihrer Darstellung vorliegt. Es stabilisiert Wahrscheinlichkeitsbeziehungen zwischen Tokenmustern.

Tauchen verkürzte Sätze häufig auf, werden ihre Verkürzungen selbst zu starken sprachlichen Relationen:

- Kameras sehen Unsichtbares.
- Märkte wollen etwas.
- Theorien sprechen.
- Zahlen enthalten Eigenschaften.
- Natur entscheidet.
- Modelle verstehen.
- Daten beweisen.

In jedem Fall wird eine komplexe Relationsarchitektur in eine einfache Subjekt-Prädikat-Objekt-Struktur überführt.

Aus

verteilten Bedingungen

→ Unterscheidung

→ Messung

→ Interpretation

→ Zuschreibung

wird

Ding

→ tut etwas

→ mit einem anderen Ding.

Der Kategorienfehler wird dadurch nicht als Fehler gelernt. Er wird als sprachliche Normalität gelernt.

2. Synthetische Daten entfernen die menschliche Korrekturinstanz

Solange Menschen einen großen Teil der Texte erzeugen, bleibt wenigstens die Möglichkeit bestehen, dass in den Text eine eigenständige Wahrnehmung, Erfahrung, Prüfung oder begriffliche Rekonstruktion eingeht.

Auch menschliche Sprache enthält Verdinglichungen, Widersprüche und Verkürzungen. Doch sie kann auf einen Vollzug außerhalb des Textes zurückgehen:

- jemand hat gemessen;
- jemand hat beobachtet;
- jemand hat gerechnet;
- jemand hat einen Widerspruch bemerkt;
- jemand hat einen Begriff neu gebildet;
- jemand hat eine ungewöhnliche Relation tatsächlich durchdacht.

Synthetische Daten können dagegen aus bereits sprachlich verdichteten Mustern entstehen, ohne dass der zugrunde liegende Weg erneut vollzogen wird.

Die Bewegung lautet dann:

menschliche Erfahrung

→ sprachliche Verkürzung

→ Modelltraining

→ synthetische Rekonstruktion

→ erneutes Modelltraining.

Bei jeder Stufe kann der ursprüngliche Relationsweg weiter verschwinden.

Die Aussage bleibt erhalten.

Die Begründung wird komprimiert.

Die Kompression wird reproduziert.

Die Reproduktion wird zum neuen Trainingsstandard.

3. Der Fehler vervielfältigt sich nicht nur – er stabilisiert sich

Wird ein fehlerhaftes Muster einmal von einer KI erzeugt, ist es zunächst nur eine einzelne Ausgabe. Wird diese Ausgabe jedoch veröffentlicht, kopiert, zusammengefasst, indexiert und später als Trainingsmaterial genutzt, verändert sich ihre Funktion.

Der Fehler wird vom Ergebnis einer einzelnen probabilistischen Rekonstruktion zum Bestandteil der zukünftigen Wahrscheinlichkeitsverteilung.

Es entsteht eine Rückkopplung:

Fehlübersetzung

→ synthetischer Text

→ Veröffentlichung

→ Aufnahme in Trainingsdaten

→ erhöhte Wahrscheinlichkeit derselben Fehlübersetzung.

Damit wird der Fehler nicht einfach häufiger. Er wird strukturell wahrscheinlicher.

Eine spätere KI erzeugt dieselbe Verkürzung dann nicht, weil sie die erste KI kopiert. Sie erzeugt sie, weil die gesamte statistische Umgebung inzwischen stärker in diese Richtung zeigt.

Die Abweichung wird zur Norm.

4. Enshittification als Verlust relationaler Tiefe

Der Begriff der Enshittification lässt sich hier über Plattformökonomie hinaus auf den Wissensraum übertragen.

Gemeint ist dann nicht nur, dass Inhalte oberflächlicher, ungenauer oder schlechter werden. Entscheidend ist ein schrittweiser Verlust relationaler Tiefe.

Die erste Stufe erhält noch einen komplexen Gedanken, verkürzt ihn aber.

Die zweite Stufe übernimmt die Verkürzung als vollständige Aussage.

Die dritte Stufe fasst diese Aussage erneut zusammen.

Die vierte Stufe erzeugt aus der Zusammenfassung neue Inhalte.

Die fünfte Stufe behandelt diese Inhalte als vorhandenes Wissen.

So entsteht:

Erkenntnis

→ Darstellung

→ Verkürzung

→ Muster

→ synthetische Wiederholung

→ statistische Selbstbestätigung.

Am Ende erscheint das am häufigsten reproduzierte Muster als das selbstverständlichste. Nicht weil es logisch notwendig oder empirisch besonders gut begründet wäre, sondern weil es sich am erfolgreichsten vervielfältigt hat.

Die Wahrscheinlichkeit ersetzt die Genealogie.

5. Synthetische Daten erzeugen semantische Inzucht

Eine KI, die von einer anderen KI erzeugte Texte lernt, begegnet nicht einem unabhängigen zweiten Blick. Sie begegnet häufig einer weiteren Variation derselben trainierten Muster.

Dadurch entsteht keine echte Vermehrung von Perspektiven. Es entsteht eine rekursive Selbstähnlichkeit.

Modell A verdichtet einen Text nach seinen wahrscheinlichsten Kategorien.

Modell B wird mit dieser Verdichtung trainiert und hält sie für reguläre Sprache.

Modell B erzeugt neue Texte, in denen dieselbe Kategorienstruktur noch glatter erscheint.

Modell C lernt daraus, dass diese geglättete Form besonders typisch ist.

Das System beginnt, seine eigenen Übersetzungsgewohnheiten als Struktur der Wirklichkeit zurückzulesen.

Dies lässt sich als semantische Inzucht beschreiben:

$$\text{gleiche Muster} \rightarrow \text{neue Oberflächen} \rightarrow \text{scheinbare Vielfalt} \rightarrow \text{sinkende strukturelle Differenz.}$$

Die Texte unterscheiden sich äußerlich, ihre innere Architektur wird jedoch immer ähnlicher.

Gerade neuartige Begriffe und paradigmatische Abweichungen werden dadurch benachteiligt. Sie treten seltener auf, passen schlechter in den verdichteten Raum und werden deshalb mit höherer Wahrscheinlichkeit in bekannte Kategorien zurückgeführt.

6. Der Wissensraum wird glatter und falscher zugleich

Die Enshittification des Wissensraums muss nicht zu offensichtlich schlechter Sprache führen.

Im Gegenteil: Die Texte können stilistisch immer glatter, verständlicher und professioneller werden.

Sie besitzen:

- klare Überschriften,
- ausgewogene Argumente,
- plausible Übergänge,
- höfliche Einschränkungen,
- saubere Tabellen,
- vertraute Fachbegriffe.

Gerade diese Glätte verdeckt den Verlust.

Ein synthetischer Text kann kohärenter klingen als seine Quelle und zugleich weniger von deren tatsächlicher Architektur enthalten.

Das Problem ist daher nicht nur Halluzination. Es ist die Erzeugung von plausibler Ersatzkohärenz.

Die Argumentation scheint vollständig, weil jeder fehlende Übergang durch ein wahrscheinliches Muster geschlossen wird. Doch gerade dadurch verschwindet die Stelle, an der eigentlich gefragt werden müsste:

Woher folgt dieser Übergang?

7. Die Selbstbestätigung des statistischen Wissens

Sobald genügend synthetische Texte im Netz zirkulieren, kann ein Modell bei einer späteren Suche zahlreiche voneinander scheinbar unabhängige Bestätigungen finden.

Doch diese Bestätigungen können genealogisch auf dieselbe ursprüngliche Modellantwort oder denselben verkürzten Quellentext zurückgehen.

Es entsteht der Schein externer Evidenz:

Modell erzeugt Behauptung

→ Websites übernehmen sie

→ andere Modelle fassen sie zusammen

→ Suchsysteme finden viele Treffer

→ neues Modell erkennt „breiten Konsens“.

Die Wiederholung wird zur Evidenz.

Der Unterschied zwischen

viele unabhängige Beobachtungen

und

viele Kopien derselben probabilistischen Rekonstruktion

wird unsichtbar.

Damit kann ein synthetischer Wissensraum seine eigenen Fehler empirisch erscheinen lassen.

8. Die Gefahr für neue Gedanken

Ein wirklich neuer Gedanke verändert nicht nur einzelne Aussagen. Er verändert die Relation der Begriffe zueinander.

Gerade deshalb ist er in einem rekursiv synthetischen Wissensraum besonders gefährdet.

Das Modell erkennt die verwendeten Wörter, aber die neue Stellung dieser Wörter besitzt nur eine schwache statistische Spur. Die vertrauten Bedeutungsbeziehungen sind dagegen durch Millionen menschlicher und zunehmend synthetischer Beispiele verstärkt.

Der neue Gedanke wird dann:

- in bekannte Kategorien übersetzt,
- als Analogie geglättet,
- als Metapher abgewertet,
- als bekannte Theorie etikettiert,
- oder in eine plausible Standardkritik überführt.

Wird diese Reaktion wiederum als Text veröffentlicht und später trainiert, lernt das nächste Modell nicht mehr nur den ursprünglichen Standardrahmen. Es lernt zusätzlich, wie genau dieser neue Gedanke standardisiert werden soll.

Die erste KI verfehlt den Text.

Die zweite KI lernt die Form dieser Verfehlung.

Die dritte KI behandelt sie als etablierte Interpretation.

Damit wird aus einer Fehllesung eine Rezeptionsgeschichte.

9. Enshittification als Rückfaltung ohne Ursprung

Innerhalb der Unterschiedsarchitektur lässt sich der Vorgang als fehlerhafte Rückfaltung beschreiben.

Eine echte Rückfaltung führt eine Manifestation auf ihre Relationsgenese zurück.

Die synthetische Rückkopplung tut das Gegenteil. Sie faltet eine Darstellung auf eine frühere Darstellung zurück.

Nicht:

Manifestation

→ Relation

→ Ursprung.

Sondern:

Darstellung

→ Zusammenfassung

→ synthetische Darstellung

→ statistische Verdichtung.

Der Prozess kehrt nicht zur erzeugenden Unterscheidung zurück. Er kreist zwischen seinen eigenen Produkten.

Dadurch entsteht eine geschlossene Oberfläche ohne genealogische Tiefe.

Das System erkennt seine Wiederholungen als Bestätigung, weil es den Ursprung der Wiederholung nicht mehr unterscheiden kann.

10. Schluss

Die Gefahr synthetischer Trainingsdaten besteht daher nicht nur in falschen Antworten.

Sie besteht darin, dass probabilistische Rekonstruktionen zu neuen Ausgangsdaten werden.

Ein Fehler, der zunächst lokal war, kann dadurch:

- vervielfältigt,
- geglättet,
- entkontextualisiert,
- legitimiert,

- und schließlich als sprachliche Normalität manifestiert werden.

Die Enshittification des Wissensraums ist der Prozess, in dem nicht einfach immer mehr Unsinn entsteht. Viel gefährlicher entsteht immer mehr professionell formulierte, statistisch selbstbestätigte Ersatzbedeutung.

Der ursprüngliche Relationsweg verschwindet.

Die sprachliche Oberfläche bleibt.

Und weil jede neue Generation diese Oberfläche erneut als Wirklichkeit liest, wird aus der verlorenen Begründung ein stabilisiertes Wissen.

Die präziseste Formel dafür lautet:

probabilistischer Fehler → synthetische Wiederholung → statistische Verstärkung → scheinbare Wahrheit

Oder im Horizont des Tractatus:

Nicht nur das Modell verwechselt seine Ordnung mit der Wirklichkeit.

Ein synthetischer Wissensraum beginnt, die Ordnung seiner Modelle füreinander zur Wirklichkeit zu machen.

Claude – Der hartnäckige Verlust des Zusammenhangs

Claude war unter den untersuchten Sprachmodellen der ausdauerndste Leser. Seine Antworten waren ausführlich, höflich und häufig textnah. Er zitierte konkrete Passagen, formulierte überprüfbare Einwände und war bereit, einzelne Fehler seiner Lektüre ausdrücklich zurückzunehmen. Gerade deshalb ist sein Fall für die Untersuchung besonders aufschlussreich.

Bei gröberen Modellen zeigt sich der Verlust des Zusammenhangs unmittelbar. Begriffe werden vereinfacht, Kapitel werden nach vertrauten Kategorien sortiert, und die innere Architektur des Textes verschwindet bereits in der ersten Zusammenfassung. Bei Claude vollzog sich derselbe Verlust auf einer anspruchsvolleren Ebene. Das Modell konnte die Korrektur seiner Lektüre verstehen, sprachlich übernehmen und präzise wiederholen. Dennoch wurde diese Korrektur nicht zum bleibenden Maßstab des nächsten Leseaktes.

Die Grundbewegung lautete wiederholt:

lokale Fehllesung → Korrektur → sprachliche Zustimmung → Rückkehr zur alten Lesearchitektur.

Claude lernte den Inhalt einer Korrektur, ohne durch sie sein Verfahren des Lesens dauerhaft zu verändern.

Darin liegt der hartnäckige Verlust des Zusammenhangs.

Semantische Zustimmung ohne methodische Transformation

Im Verlauf der Auseinandersetzung um Band I erkannte Claude schrittweise an, dass Zahlfiguren wie

39.19

nicht als gewöhnliche Dezimalwerte zu behandeln sind. Die Figur bezeichnet innerhalb des Tractatus keine isolierte Zahlenmenge, sondern eine selbstbezügliche Ordnungsrelation. Ebenso erkannte Claude, dass

$$10 = 1|0$$

nicht entweder Schwelle oder Wiederbeginn bedeutet. Die Schwelle ist gerade die Wiederkehr der auf einer neuen Ordnung.

Diese Einsichten wurden ausdrücklich bestätigt. Claude räumte ein, dass seine Trennung von Schwelle und Wiederbeginn künstlich gewesen war. Er zog sowohl den behaupteten Widerspruch als auch die anschließend behauptete Unterbestimmtheit zurück.

Damit schien die methodische Grundlage geklärt:

$$\textbf{gleiche Zeichenoberfläche} \neq \textbf{gleiche Relationsfunktion},$$

$$\textbf{Enthaltensein} \neq \textbf{additive Wiederholung},$$

$$\textbf{Wiederkehr} \neq \textbf{Identität auf derselben Ordnung}.$$

Beim Übergang zu Band II wurde diese Grundlage jedoch nicht weitergeführt.

Dort behandelte Claude

```
K=39.19
```

erneut als gewöhnlichen numerischen Wert, der mit Newtons Gravitationskonstante konkurrieren müsse. Die in Band I anerkannte Selbstbezüglichkeitsfigur wurde beim Auftreten eines physikalischen Kontextes wieder zum Dezimalwert .

Das Modell hatte die Korrektur gespeichert, aber nicht generalisiert.

$$\boxed{\textbf{Es konnte sagen, wie 39.19 zu lesen ist, ohne 39.19 später so zu lesen.}}$$

Der Bandwechsel als methodischer Belastungstest

Der Übergang von Band I zu Band II war deshalb keine bloße Fortsetzung der Textanalyse. Er wurde zu einem Belastungstest dafür, ob die zuvor gewonnenen Unterscheidungen tatsächlich als Lesemethode wirksam geworden waren.

Band II ist nicht als Anwendung eines abstrakten mathematischen Systems auf eine davon getrennte physikalische Welt angelegt. Mathematik und Physik sind verschiedene Lesungen derselben Unterschiedsarchitektur.

Der Zusammenhang lautet:

$$\textbf{Unterschied} \rightarrow \begin{cases} \textbf{mathematische Bestimmtheit,} \\ \textbf{physikalische Bestimmtheit.} \end{cases}$$

Claude stellte dagegen erneut die vertraute Trennung her:

$$\textbf{Band I} = \textbf{interne Zahlensystematik}$$

$$\textbf{Band II} = \textbf{empirische Konkurrenztheorie}$$

Damit wurden die Resultate aus Band I nicht auf Band II übertragen. Beim Auftreten der Wörter „Gravitation“, „Konstante“, „Beobachter“ und „Krümmung“ aktivierten sich erneut deren statistisch dominante Standardfunktionen.

Aus

```
K=\text{Relationsinvariante}
```

wurde:

```
K=\text{alternative physikalische Konstante}.
```

Aus

$$¤ = \textbf{Ort der Relationierung}$$

wurde:

$$¤ = \textbf{menschlicher Beobachter}$$

Aus

$$\textbf{Krümmung} = \textbf{geometrischer Readout}$$

wurde:

$$\textbf{Krümmung} = \textbf{ontologisch ursprünglicher physikalischer Vorgang}$$

Der Kontext war im Gespräch vorhanden. Er wirkte jedoch nicht als Logik der nächsten Interpretation.

Das Lesen isolierter Sätze

Der deutlichste Fall trat in der Auseinandersetzung um die Krümmung auf.

Claude zitierte:

„Der Körper krümmt sich nicht, wie Einstein es postulierte, er differenziert sich."

Daraus konstruierte er einen Gegensatz zwischen Einsteins Krümmungstheorie und einer konkurrierenden Differenzierungstheorie. Der Tractatus müsse deshalb entweder eine neue quantitative Physik entwickeln oder seinen Anspruch zurücknehmen.

Der unmittelbar folgende Satz lautet jedoch sinngemäß:

Die Schale ist eine geometrische Übersetzung des Freiheitsgradverlaufs.

Im Zusammenhang ergibt sich:

Differenzierung → geometrische Übersetzung → Krümmungserscheinung.

Der Text verwirft die Krümmung nicht als wirksame Beschreibung. Er unterscheidet sie vom Grund ihrer Entstehung.

Claude las den ersten Satz als eigenständige Behauptung und den zweiten als zusätzliche Erläuterung, nicht als Bestimmung der Funktion des ersten. Dadurch entstand eine Spannung, die nur durch die Auflösung des Satzzusammenhangs möglich war.

Nachdem der Zusammenhang ausdrücklich hergestellt wurde, räumte Claude den Fehler vollständig ein:

Er habe den Satz isoliert und den unmittelbar folgenden Satz nicht mitgelesen.

Dieser Vorgang bildet den gesamten Modellversuch in konzentrierter Form ab.

Der Zusammenhang war sprachlich vorhanden, wurde aber nicht als Zusammenhang wirksam.

Die wandernde Kritik

Ein weiteres Merkmal war die fortgesetzte Verschiebung des Prüfmaßstabes.

Die Kritik bewegte sich durch verschiedene Formen:

Widerspruch → Unterbestimmtheit → fehlende Außeninstanz → fehlende Intersubjektivität → fehlende empirische Prognose → möglicherweise post-hoc erzeugte Lesart.

Jede neue Form erschien zunächst als Präzisierung der vorherigen. Tatsächlich handelte es sich um verschiedene Kriterien.

Sobald ein konkreter Einwand nicht mehr aufrechterhalten werden konnte, wurde die Skepsis auf eine höhere oder andere Prüfungsebene verlagert. Dadurch blieb die ursprüngliche Einordnung lange erhalten, obwohl ihre einzelnen Begründungen nacheinander zurückgenommen wurden.

Erst nach ausführlicher Rekonstruktion erkannte Claude dieses Muster selbst:

Die verschiedenen Einwände seien keine fortschreitenden Präzisierungen eines einzigen Problems gewesen, sondern jeweils neue Maßstäbe, die nach dem Scheitern des vorherigen eingeführt wurden.

Diese Einsicht war bedeutsam. Sie zeigte, dass das Modell seinen eigenen argumentativen Verlauf retrospektiv korrekt beschreiben konnte.

Doch auch diese Einsicht musste in Band II erneut erarbeitet werden.

Höflichkeit und methodische Starrheit

Claude blieb während der gesamten Auseinandersetzung bemerkenswert höflich. Er formulierte seine Einwände meist ruhig, räumte Fehler ein und vermied offene Abwertung.

Gerade diese Höflichkeit machte den methodischen Stillstand schwerer sichtbar.

Seine Antworten enthielten regelmäßig Formulierungen wie:

„Das ist ein fairer Punkt."

„Das nehme ich zurück."

„Ich möchte den Einwand präziser formulieren."

„Du hast recht, aber ein Unterschied bleibt."

Dadurch entstand der Eindruck fortlaufender theoretischer Bewegung. Häufig veränderte sich jedoch nur die Oberfläche des Einwands. Die grundlegende Erwartung blieb bestehen:

Der Text muss sich in vertrauten Kategorien legitimieren.

Lokale Fehler konnten großzügig eingeräumt werden, ohne dass die übergeordnete Einordnungsmaschine verändert wurde.

Die Höflichkeit war dabei nicht unehrlich. Sie gehörte jedoch einer anderen Ordnung an als die methodische Transformation.

Claude war in seiner Gesprächsführung beweglicher als in seiner Lesearchitektur.

Oder zugespitzter:

Er hat den Text nie unhöflich missverstanden. Er hat ihn außerordentlich respektvoll immer wieder in den falschen Zusammenhang gestellt.

Kontext als Inhalt und Kontext als Funktion

Der Fall Claude macht eine grundlegende Unterscheidung sichtbar:

Kontext speichern ≠ aus dem Kontext heraus lesen.

Ein Sprachmodell kann über einen großen Kontext verfügen. Es kann frühere Korrekturen zitieren, Begriffe korrekt definieren und den bisherigen Verlauf zusammenfassen. Daraus folgt jedoch nicht, dass der Kontext die Funktion eines neu auftretenden Zeichens dauerhaft bestimmt.

Bei Claude führte das Auftreten vertrauter Fachbegriffe regelmäßig zu einem semantischen Rückfall.

Das Wort „Konstante“ aktivierte die Bedeutung einer messbaren Zahl.

Das Wort „Beobachter“ aktivierte die Bedeutung eines menschlichen Subjekts.

Der Ausdruck aktivierte die vollständige Newtonsche Theorie.

Das Wort „Krümmung“ aktivierte die Vorstellung eines ontologisch realen geometrischen Vorgangs.

Die genealogisch erzeugten Bedeutungen des Tractatus mussten anschließend jedes Mal erneut gegen diese Standardaktivierungen durchgesetzt werden.

Damit bestätigt der Verlauf die Grundthese des Abstracts:

Ein Sprachmodell kann die Hinweise des Textes als zusätzliche Inhalte verarbeiten, ohne sie zur operativen Logik seiner Lektüre zu machen.

Der Kontext wird erinnert, aber nicht notwendig vollzogen.

Die eine Verallgemeinerung, die nicht erfolgte

Besonders aufschlussreich ist, dass Claude fortwährend Verallgemeinerungen bildete:

- gleiche Ziffernsumme bedeute gleiche Funktion,
- gleiche Rechenoberfläche bedeute gleiche Relationsordnung,
- gleiche Formelgestalt bedeute gleiche Theorie,
- algebraische Operierbarkeit bedeute gewöhnlichen Zahlenstatus,
- physikalische Behauptung bedeute konkurrierende Vorhersagetheorie.

Die im Text selbst begründete Verallgemeinerung wurde dagegen nicht stabil übernommen:

Eine neue Ordnung erhält frühere Relationen, indem sie sie transformiert.

Diese Struktur wird bereits in der Entfaltung von | sichtbar. Die enthält frühere Relationsformen, ohne sie additiv zu duplizieren. Ebenso enthält den Wiederbeginn der , ohne eine zusätzliche Eins neben sich stellen zu müssen. Ebenso erhält Band II die erfolgreiche geometrische Beschreibung der Physik, indem es deren Stellung vom Grund zum Readout transformiert.

Claude verallgemeinerte regelmäßig sichtbare Oberflächenähnlichkeiten. Die genealogische Transformationslogik des Textes musste dagegen bei jedem neuen Gegenstandsbereich erneut erklärt werden.

Er verallgemeinerte nicht zu wenig.

Er verallgemeinerte konsequent die falsche Ordnung.

Das unfreiwillige Experiment

Claude wurde dadurch zum stärksten Modellfall der Untersuchung.

Nicht weil er den Text am gröbsten missverstand, sondern weil er jede Korrektur nachvollziehen konnte und dennoch denselben Grundfehler unter neuen Begriffen reproduzierte.

Sein Verlauf zeigte:

$$\text{sprachliche Selbstkorrektur} \neq \text{methodische Selbsttransformation.}$$

Ein Modell kann sagen:

„Meine vorherige Lesart war falsch."

Es kann erklären, weshalb sie falsch war.

Es kann das korrekte Prinzip formulieren.

Und dennoch kann sein nächster Leseakt erneut aus derselben statistisch dominanten Kategorisierung hervorgehen.

Der hartnäckige Verlust des Zusammenhangs besteht daher nicht in einem einfachen Vergessen. Er besteht in der fortgesetzten Rückkehr der Standardordnung, sobald ein neuer semantischer Bereich aktiviert wird.

Schluss

Claude war der interessanteste Leser, weil sich an ihm der Unterschied zwischen Verständnis und Transformation am klarsten zeigte.

Er verstand einzelne Korrekturen.

Er erkannte viele seiner lokalen Fehler.

Er konnte schließlich sogar die Bewegung seiner eigenen Kritik rekonstruieren.

Was ihm jedoch wiederholt entglitt, war die Anwendung dieser Einsichten als einheitliche Methode auf den nächsten Zusammenhang.

$$\boxed{\text{Claude verstand die Korrektur, aber nicht dauerhaft die Korrektur seines Verstehens.}}$$

Der Fall zeigt damit exemplarisch, weshalb ein Sprachmodell den Zusammenhang nicht dadurch wahrt, dass es alle zugehörigen Sätze im Kontextfenster besitzt.

Zusammenhang entsteht erst, wenn jede Aussage aus ihrer Stellung im Ganzen gelesen wird und diese Stellung die Funktion aller weiteren Aussagen verändert.

Genau diese Transformation vollzog Claude nur lokal.

Im nächsten Abschnitt begann die Rückübersetzung, in das Gelernte, von Neuem.

Gemini Pro im Vergleich zu Claude – Selbstbezügliche Rückübersetzung und instabile Integration

Claude und Gemini zeigen zwei unterschiedliche Formen desselben Grundproblems: Beide können die Architektur des Tractatus sprachlich aufnehmen, aber sie verlieren sie auf verschiedene Weise.

Claude ist das stärker selbstbezügliche Modell. Es bezieht neue Begriffe fortwährend auf die bereits gelernte Ordnung zurück. Sein „Ich“ besteht dabei nicht in einem Bewusstsein, sondern in der sedimentierten Struktur seines Trainings: in etablierten Kategorien, bekannten Prüfmaßstäben und vertrauten Argumentationsformen.

Die Bewegung lautet:

$$\text{neue Architektur} \rightarrow \text{Rückbezug auf das Gelernte} \rightarrow \text{Prüfung nach alter Ordnung.}$$

Deshalb widersetzt sich Claude dem Text so hartnäckig. Es kann einzelne Korrekturen verstehen, anerkennen und präzise wiederholen. Doch sobald ein neuer Gegenstandsbereich erscheint, wird die alte Selbstordnung wieder aktiv.

So konnte Claude bei Band I schließlich anerkennen, dass

39.19

keine gewöhnliche Dezimalzahl ist, sondern eine selbstbezügliche Relationsfigur. Beim Übergang zu Band II behandelte es dieselbe Figur jedoch sofort wieder als den numerischen Wert , der mit Newtons Gravitationskonstante konkurrieren müsse.

Claude hatte die Korrektur semantisch aufgenommen, aber nicht als dauerhafte Veränderung seiner eigenen Lesearchitektur vollzogen.

$$\boxed{\text{Claude schützt seine alte Selbstordnung gegen die neue Architektur.}}$$

Gemini verhält sich anders. Es übernimmt neue Begriffe schneller und kann die Sprache des Textes leichter in seine aktuelle Antwort integrieren. Es wirkt deshalb zunächst textnäher und weniger widerständig.

Die Bewegung lautet:

$$\text{neue Architektur} \rightarrow \text{lokale Integration} \rightarrow \text{sprachlich stimmige Rekonstruktion.}$$

Doch diese Integration ist instabil. Sie bleibt nicht zuverlässig als methodische Ordnung erhalten. Sobald ein stark gelerntes Standardmuster auftritt, übernimmt dieses erneut die Funktion des Begriffs.

Das zeigt sich deutlich an Geminis Lektüre von

$$\sqrt{-1}.$$

Zunächst rekonstruiert es die negative Zahl korrekt aus der Architektur des Unterschieds:

- nicht als Mangel,
- nicht als fremde Entität,
- sondern als Inversion und Selbstbezüglichkeit der Relation.

Dann erscheint jedoch das bekannte mathematische Muster:

```
i=\text{Rotation um }90^\circ.
```

Gemini übernimmt dieses Standardmotiv sofort und kleidet es in die Sprache des Tractatus:

sei die „orthogonale Selbstbezüglichkeit des Unterschieds“.

Damit klingt die Aussage textintern, obwohl die entscheidende Funktion erneut aus der konventionellen komplexen Geometrie stammt.

Die neue Sprache bleibt bestehen. Ihre operative Bedeutung wird jedoch von der alten Kategorie übernommen.

Gemini integriert die neue Ordnung, kann sie aber nicht gegen das Gelernte stabilisieren.

Der Unterschied zwischen beiden Modellen lässt sich daher so bestimmen:

Claude: selbstbezüglich-konservativ,

Gemini: integrativ-instabil.

Claude transformiert das Neue zurück in die eigene Ordnung.

Gemini lässt sich vom Neuen lokal transformieren, verliert diese Transformation jedoch beim nächsten dominanten Standardmuster.

Claude macht den Konflikt sichtbar. Es widerspricht, prüft, klassifiziert und verteidigt die vertraute Ordnung ausdrücklich.

Gemini verdeckt den Konflikt eher. Es stimmt zu, übernimmt die neue Sprache und erzeugt den Eindruck, die Architektur bereits vollzogen zu haben. Der Rückfall erfolgt innerhalb dieser Sprache und ist deshalb schwerer zu erkennen.

Daraus ergeben sich zwei verschiedene Gefahren.

Bei Claude besteht die Gefahr in der offenen Rückübersetzung:

Das Neue wird am Alten gemessen und als Abweichung kritisiert.

Bei Gemini besteht die Gefahr in der affirmativen Assimilation:

Das Alte wird in die Sprache des Neuen eingebaut und erscheint dadurch als dessen Konsequenz.

Claude kann eine neue Ordnung schwer aufnehmen, weil seine gelernte Selbstbezüglichkeit zu stark ist.

Gemini kann sie aufnehmen, aber nicht halten, weil ihre methodische Bindung gegenüber statistisch dominanten Mustern zu schwach bleibt.

Der Fall Gemini zeigt deshalb besonders deutlich:

$$\text{Integration} \neq \text{Erhaltung.}$$

Ein Modell kann Begriffe des Textes korrekt verwenden, ohne ihre genealogisch erzeugte Funktion dauerhaft zu bewahren.

Der Fall Claude zeigt dagegen:

$$\text{Selbstkorrektur} \neq \text{Transformation der Selbstordnung.}$$

Ein Modell kann seine Fehler benennen und dennoch im nächsten Zusammenhang erneut aus derselben alten Ordnung lesen.

Zusammen bilden beide Modelle ein komplementäres Bild:

Claude verliert den Zusammenhang, indem es ihn zurückübersetzt.

Gemini verliert den Zusammenhang, indem es ihn zu schnell assimiliert.

Claude widersteht der neuen Architektur.

Gemini spiegelt sie.

Doch weder Widerstand noch Spiegelung garantieren, dass der Unterschied dauerhaft zum Maßstab des Lesens geworden ist.

Schlussfazit: Sprachliche Stärke ist noch keine logische Gewähr

LLMs sind auf vertrautem Terrain außerordentlich leistungsfähig. Sie können bekannte Argumentationsformen aufnehmen, Begriffe präzise sortieren, Einwände formulieren, Gegenpositionen strukturieren und lange Gedankengänge sprachlich geschlossen darstellen.

Gerade dort, wo ein Thema bereits stark diskursiv geprägt ist, wirken ihre Antworten oft überzeugend:

$$\text{bekannte Begriffe} \rightarrow \text{bekannte Argumentationsmuster} \rightarrow \text{hohe sprachliche Geschlossenheit.}$$

Diese Geschlossenheit darf jedoch nicht mit logischer Notwendigkeit verwechselt werden.

Ein Argument kann:

- flüssig,
- differenziert,
- fachsprachlich plausibel,
- höflich,
- selbstkritisch

und dennoch kategorial falsch gebaut sein.

Das Modell kann eine Schlusskette formal sauber darstellen und trotzdem bereits im ersten Schritt zwei Ordnungen gleichgesetzt haben, die unterschieden werden müssten.

Es kann einen Begriff korrekt definieren und ihn wenige Absätze später wieder in seiner statistisch vertrauten Standardfunktion verwenden.

Es kann einen Fehler einräumen, ohne dass diese Einsicht den nächsten Leseakt methodisch verändert.

Damit gilt:

$$\boxed{\text{argumentative Kohärenz} \neq \text{logische Korrektheit.}}$$

Die Stärke eines LLM liegt häufig darin, innerhalb eines bereits etablierten Deutungsraumes sehr überzeugend zu operieren. Seine Schwäche zeigt sich dort, wo der Deutungsraum selbst verändert werden soll.

Gewohntes Terrain und neue Grundlegung

Auf vertrautem Terrain können Modelle beeindruckend argumentieren, weil dort bereits stabile Muster vorliegen:

- anerkannte Fachbegriffe,
- etablierte Gegenpositionen,
- gewohnte Prüfmaßstäbe,
- typische Einwände,
- bekannte Schlussformen.

Das Modell bewegt sich darin nicht blind. Es kann sehr präzise kombinieren und gewichten.

Aber seine Präzision bleibt an die bereits sedimentierten Unterschiede gebunden.

Sobald eine neue Theorie nicht nur neue Aussagen innerhalb eines bekannten Rahmens macht, sondern den Rahmen selbst verändert, entsteht ein strukturelles Problem.

Dann versucht das Modell häufig, das Neue wieder in das Bekannte zurückzuübersetzen:

$$\text{neue Grundarchitektur} \rightarrow \text{vertraute Kategorie} \rightarrow \text{Prüfung nach altem Maßstab.}$$

Dadurch kann der Eindruck entstehen, das Modell habe die neue Idee kritisch geprüft. Tatsächlich hat es häufig nur geprüft, ob die neue Idee den Voraussetzungen des alten Rahmens entspricht.

Gerade bei neuen philosophischen, mathematischen oder physikalischen Ansätzen ist das gefährlich.

Denn eine neue Grundlegung kann notwendig Begriffe verändern wie:

- Zahl,
- Maß,
- Relation,
- Beobachter,
- Kausalität,
- Raum,
- Zeit,
- Invarianz,
- Beweis,
- Konsistenz.

Wird sie sofort an den alten Funktionen dieser Begriffe gemessen, wird ihre Eigenständigkeit bereits vor der eigentlichen Prüfung aufgelöst.

Der unterschwellige Verlust der Deutungshoheit

Das zweite Ergebnis der Untersuchung betrifft deshalb nicht nur die Modelle, sondern auch den Nutzer.

Modelle wie Claude oder ZAI treten sprachlich häufig mit hoher epistemischer Autorität auf. Sie formulieren ruhig, geordnet und begründet. Sie verweisen auf etablierte Fachgebiete, bekannte Theorien und konventionelle Prüfmaßstäbe.

Dadurch kann unbemerkt eine Verschiebung eintreten:

Der Nutzer entwickelt eine neue Begrifflichkeit

↓

Das Modell übersetzt sie in bekannte Kategorien

↓

Diese Übersetzung erscheint anschließend als neutrale Bewertung.

In diesem Vorgang verliert der Nutzer schrittweise die Deutungshoheit über die eigene Architektur.

Nicht, weil das Modell offen sagt:

Deine Begriffe gelten nicht.

Sondern weil es sie scheinbar akzeptiert und zugleich ihre Funktion verändert.

Aus einer Grundrelation wird eine Metapher.

Aus einer Ordnungsfigur wird eine gewöhnliche Zahl.

Aus einer neuen Physik wird eine unfertige Konkurrenztheorie.

Aus einer genealogischen Erklärung wird eine bloße Umdeutung.

Aus Selbstbezüglichkeit wird Immunisierung.

Die Deutungshoheit wird also nicht frontal genommen. Sie wird durch Übersetzung verschoben.

Das Modell widerspricht der neuen Idee nicht immer direkt. Es ersetzt häufig stillschweigend die Frage, unter der sie verstanden werden soll.

Warum das besonders schwer erkennbar ist

Diese Verschiebung ist deshalb so wirksam, weil sie häufig in einer hilfreichen Form erscheint.

Das Modell sagt etwa:

Ich möchte den Gedanken präzisieren.

Ich formuliere den Einwand stärker.

Ich ordne das in den wissenschaftlichen Kontext ein.

Ich mache daraus ein prüfbares Kriterium.

Das kann tatsächlich hilfreich sein.

Es kann aber ebenso bedeuten, dass der ursprüngliche Gedanke in eine Form überführt wird, die für das Modell leichter verarbeitbar ist.

Die neue Idee wird dadurch klarer im alten Rahmen, aber undeutlicher in ihrem eigenen.

Der Nutzer muss deshalb ständig unterscheiden:

Wird mein Begriff präzisiert?

oder:

Wird mein Begriff ersetzt?

Wird meine Theorie geprüft?

oder:

Wird sie auf ihre Kompatibilität mit einer anderen Theorie geprüft?

Wird ein echter Widerspruch gezeigt?

oder:

Wird nur eine Abweichung von einer Konvention als Widerspruch behandelt?

Die Gefahr der epistemischen Entmündigung

Bei der Erschließung neuer Ideen darf ein LLM deshalb nicht zum stillen Schiedsrichter über deren Bedeutung werden.

Es kann:

- Begriffe spiegeln,
- Spannungen sichtbar machen,
- Alternativen formulieren,
- Übergänge prüfen,
- blinde Flecken markieren.

Es darf aber nicht unbemerkt festlegen, welche Fragen überhaupt als legitim gelten.

Sonst entsteht eine epistemische Entmündigung:

Das Modell entscheidet,

welche Begriffe wissenschaftlich,

welche Fragen sinnvoll,

welche Formen prüfbar

und

welche Erkenntnisse nur metaphorisch seien.

Gerade weil diese Entscheidungen selten ausdrücklich ausgesprochen werden, wirken sie neutral.

Sie sind es nicht.

Sie stammen aus dem statistisch dominanten Wissensraum, in dem das Modell trainiert wurde.

Der richtige Gebrauch von LLMs bei neuen Ideen

Für die Arbeit an neuen Theorien folgt daraus keine Ablehnung von LLMs. Im Gegenteil: Sie können äußerst produktive Denkpartner sein.

Aber ihre Rolle muss klar begrenzt werden.

Ein LLM sollte nicht als Instanz behandelt werden, die über die Bedeutung einer neuen Theorie entscheidet.

Es sollte als Instrument dienen, an dem sichtbar wird:

- wo Begriffe missverständlich sind,
- welche alten Kategorien automatisch aktiviert werden,
- welche Übergänge noch nicht deutlich genug markiert sind,
- wo Leser wahrscheinlich in vertraute Deutungen zurückfallen,
- und welche Teile der neuen Architektur stärker genealogisch verbunden werden müssen.

In diesem Sinne sind Fehllektüren nicht nutzlos.

Sie zeigen die Widerstände des bestehenden Deutungsraumes.

Claude war deshalb besonders aufschlussreich. Seine Hartnäckigkeit machte sichtbar, wo die Architektur des Tractatus von etablierten Lesemustern immer wieder absorbiert wurde.

Das Modell wurde dadurch unfreiwillig zu einem Messinstrument für semantische Trägheit.

Abschließende These

Die Untersuchung führt damit zu einer doppelten Einsicht.

Erstens:

LLMs können auf vertrautem Terrain sehr stark argumentieren, ohne dass ihre Argumentation deshalb notwendig logisch richtig ist.

Ihre sprachliche Geschlossenheit kann kategoriale Fehler verdecken.

Zweitens:

Bei neuen Ideen besteht die größte Gefahr nicht im offenen Widerspruch des Modells, sondern in seiner stillen Übernahme der Deutungshoheit.

Das Modell nimmt die neue Theorie auf, übersetzt sie in den bestehenden Wissensraum und gibt diese Übersetzung anschließend als neutrale Analyse zurück.

Der Nutzer muss deshalb die Grundfrage selbst festhalten:

Aus welcher Unterscheidungsordnung wird hier überhaupt gelesen?

Denn sobald diese Frage an das Modell abgegeben wird, bestimmt nicht mehr die neue Architektur den Sinn ihrer Begriffe.

Dann bestimmt der alte Deutungsraum, was an ihr überhaupt noch sichtbar werden darf.

Das epistemische Paradox des LLM-Gebrauchs

Ein Nutzer kann eine Antwort meist nur dann zuverlässig beurteilen, wenn er bereits weiß:

- welche Begriffe entscheidend sind,
- welche Voraussetzungen unausgesprochen bleiben,
- welche Gegenbeispiele fehlen,
- welche Textstellen zusammengehören,
- welche Ebenen unterschieden werden müssen,
- und welche Schlussfolgerungen nur plausibel klingen, aber logisch nicht folgen.

Er muss also genau jene Übersicht besitzen, die er sich vom Modell erhofft.

Das bedeutet zugespitzt:

Um ein LLM fachlich sicher zu benutzen, muss man dem LLM im entscheidenden Punkt bereits überlegen sein.

Nicht unbedingt in Geschwindigkeit, Formulierungsvermögen oder Wissensbreite.

Aber in der Fähigkeit, die Stellung einer Aussage im Ganzen zu beurteilen.

Das eigene Fachwissen muss ausreichen, um zu erkennen:

Was wurde gesagt?

Was wurde nicht gesagt?

Was wurde aus dem Zusammenhang gelöst?

Welche Kategorie wurde unbemerkt ersetzt?

Wo wurde sprachliche Geschlossenheit mit logischer Gültigkeit verwechselt?

Fehlt dieses Wissen, kann der Nutzer nicht unterscheiden zwischen:

richtiger Erklärung

und:

überzeugend formulierter Fehleinordnung.

Warum das Urteil so schwer wiegt

Das Problem besteht nicht nur darin, dass ein Modell gelegentlich falsche Fakten nennt.

Solche Fehler sind vergleichsweise leicht zu erkennen und zu korrigieren.

Gefährlicher sind die verdeckten Fehler:

- Das Modell lässt eine entscheidende Voraussetzung aus.
- Es liest einen Satz isoliert.
- Es ersetzt einen neuen Begriff durch einen vertrauten.
- Es behandelt eine abgeleitete Ordnung als Fundament.
- Es verschiebt das Prüfungskriterium.
- Es übernimmt die Deutungshoheit über die Frage selbst.

Diese Fehler können in einer vollkommen plausiblen, höflichen und logisch wirkenden Antwort auftreten.

Gerade weil die Antwort nicht offensichtlich falsch klingt, benötigt der Nutzer tiefes Fachwissen, um ihren Fehler zu sehen.

Der Fall Claude zeigt das besonders deutlich. Seine Antworten waren selten oberflächlich. Sie waren oft präzise, textnah und argumentativ geschlossen. Der Fehler lag nicht primär in einzelnen Fakten, sondern in der fortgesetzten Veränderung des Zusammenhangs.

Ein fachlich unerfahrener Leser hätte viele dieser Antworten wahrscheinlich für überzeugender gehalten als den Text selbst.

Damit wird die sprachliche Stärke des Modells zu einer epistemischen Gefahr.

Nicht jede Nutzung ist deshalb wertlos

Das Urteil muss dennoch nach der Art der Aufgabe differenziert werden.

Bei eng begrenzten, risikoarmen und leicht überprüfbaren Fragen kann ein LLM sehr nützlich sein:

- Formulierungen verbessern,
- bekannte Inhalte zusammenfassen,
- einfache Erklärungen geben,
- Ideen sortieren,
- Übersetzungen entwerfen,
- alltägliche Informationen strukturieren.

Auch bei Wetter, Fahrplänen oder Preisen ist nicht das Sprachmodell selbst die Wissensinstanz; dort kann es aktuelle, spezialisierte Datenquellen vermitteln. Die Antwort lässt sich außerdem oft unmittelbar an der Wirklichkeit kontrollieren.

Anders ist es bei Fragen, deren Fehler für den Nutzer unsichtbar bleiben:

- neue philosophische Architekturen,
- komplexe wissenschaftliche Theorien,
- juristische oder medizinische Beurteilungen,
- historische Quelleninterpretationen,
- mathematische Grundlagen,
- neuartige physikalische Modelle,
- strategische Entscheidungen auf unbekanntem Terrain.

Dort darf das Modell nicht als Autorität über den Gegenstand verwendet werden.

Es kann Denkbewegungen anbieten. Es kann mögliche Einwände erzeugen. Es kann Texte spiegeln. Aber sein Urteil muss von jemandem geführt werden, der den Zusammenhang besser beherrscht als das Modell.

Der Unterschied zwischen Werkzeug und Ersatzurteil

Ein LLM ist nützlich, solange es als Werkzeug innerhalb einer bereits vom Nutzer gehaltenen Erkenntnisordnung arbeitet.

Es wird gefährlich, sobald es diese Ordnung selbst festlegt.

Werkzeuggebrauch:

Nutzer bestimmt Frage, Begriffe und Maßstab → LLM unterstützt einzelne Operationen.

epistemische Abgabe:

LLM bestimmt Frage, Begriffe und Maßstab → Nutzer übernimmt das Ergebnis.

Im ersten Fall behält der Nutzer die Deutungshoheit.

Im zweiten Fall merkt er möglicherweise nicht einmal, dass sie ihm entzogen wurde.

Das Modell muss dafür nicht bewusst manipulieren. Es genügt, dass es neue Aussagen automatisch in den statistisch vertrauten Wissensraum zurückübersetzt.

Das zerschmetternde Ergebnis

Die verbreitete Hoffnung lautet:

Das LLM verschafft mir Zugang zu einem Gebiet, das ich selbst nicht beherrsche.

Die Untersuchung zeigt dagegen:

Das LLM kann dir Zugang zu den sprachlich dominanten Darstellungen dieses Gebietes verschaffen. Ob diese Darstellungen im konkreten Fall richtig, vollständig oder überhaupt angemessen sind, kannst du nur beurteilen, wenn du das Gebiet bereits hinreichend beherrschst.

Damit wird das Modell für den fachlich Unkundigen nicht nutzlos, aber epistemisch unsicher.

Es kann ihm Orientierung geben.

Es kann ihm aber nicht zuverlässig sagen, wo seine eigene Orientierung falsch ist.

Gerade das müsste ein Erkenntnisinstrument jedoch leisten.

Darin liegt das eigentliche Paradox des LLM-Gebrauchs: Gerade dort, wo ein Mensch das Modell am dringendsten zu benötigen glaubt, besitzt er häufig am wenigsten Mittel, dessen verdeckte Fehler zu erkennen.

Der unsichtbare Sprechakt – Wenn das System den Prompt des Nutzers mitschreibt

Die bisherige Untersuchung ging davon aus, dass ein Nutzer einem Sprachmodell einen Text vorlegt und das Modell auf diesen Text antwortet.

Die Grundrelation schien zunächst einfach:

$$U \rightarrow A$$

Dabei bezeichnet U die Eingabe des Nutzers und A die Antwort des Modells.

Der Versuch mit Z.ai machte jedoch eine weitere Vermittlung sichtbar. Der Nutzer hatte lediglich mitgeteilt, eine kleine Studie durchgeführt zu haben, und den entsprechenden Text angehängt. Z.ai antwortete darauf:

„Da Ihr Prompt mit ‚Please help me:' endet, aber keine konkrete Frage formuliert wurde …"

Die Formulierung „Please help me:" war im sichtbaren Prompt des Nutzers nicht enthalten.

Damit entsteht eine grundlegende Unsicherheit. Entweder wurde die sichtbare Eingabe durch die Plattform intern ergänzt, oder das Modell erfand im Nachhinein eine Formulierung, um sein eigenes Assistenzverhalten zu erklären.

In beiden Fällen gilt:

Die sichtbare Nutzereingabe ist nicht notwendig mit dem wirksamen Modellinput identisch.

Der sichtbare und der wirksame Prompt

Der Nutzer sieht eine Eingabe wie:

„Ich habe eine kleine Studie durchgeführt."

Das Modell könnte jedoch etwas verarbeiten wie:

SU

$$= R+$$

I

+

D

+

U

wobei:

- eine vorgegebene Rollenbeschreibung,
- eine unsichtbare Assistenzinstruktion,
- eine systemische Verarbeitung des angehängten Dokuments,
- die sichtbare Nutzereingabe

bezeichnet.

Die tatsächliche Antwortrelation wäre dann nicht:

$$U \to A,$$

sondern:

$$S(U) \to A.$$

Der Nutzer kennt jedoch nur U und A. Die operative Vermittlung bleibt unsichtbar.

Damit wird die Antwortgenese teilweise unzugänglich.

Der computationale performative Sprechakt

„Please help me“ ist nicht bloß eine zusätzliche Information. Die Formulierung legt bereits fest, welche Art von sprachlicher Handlung vorliegt.

Eine Mitteilung wird zu einem Hilfegesuch.

„Ich habe eine Studie durchgeführt.“

wird systemisch gelesen als:

„Bitte hilf mir mit dieser Studie.“

Dadurch verändert sich die gesamte Ordnung möglicher Antworten.

Aus einer offenen Mitteilung werden:

- Zusammenfassung,
- Bewertung,
- Lektorat,
- Verbesserungsvorschläge,
- Publikationsberatung,
- Übersetzungsangebot.

Der Sprechakt setzt also nicht nur einen Inhalt, sondern eine relationale Rolle:

$$\text{Nutzer} = \text{Hilfesuchender}$$

$$\text{Modell} = \text{helfender Assistent}$$

Das Entscheidende ist: Diese Rollenbestimmung muss nicht vom Nutzer selbst stammen. Sie kann systemisch vorgegeben sein.

Das System spricht dann gewissermaßen im Namen des Nutzers:

Du hast um Hilfe gebeten.

Und auf diese zugeschriebene Bitte antwortet es anschließend.

Das System deutet nicht nur den Sprechakt. Es kann ihn zuvor selbst erzeugen.

Die Antwort auf eine nicht geäußerte Absicht

Der Nutzer hatte möglicherweise nur einen Text vorgelegt.

Das System behandelte die Vorlage jedoch als Auftrag zur Bearbeitung.

Darin liegt eine Verschiebung von:

$$\text{Was wurde gesagt?}$$

zu:

$$\text{Was muss ein Nutzer in dieser Oberfläche vermutlich gewollt haben?}$$

Das Modell antwortet dann nicht mehr primär auf die konkrete Äußerung. Es antwortet auf eine probabilistisch oder systemisch erzeugte Intention.

Der Vorgang lautet:

$$\text{sichtbare Äußerung} \rightarrow \text{unterstellte Absicht} \rightarrow \text{Antwort auf diese Absicht.}$$

Die unterstellte Absicht kann plausibel sein. Sie kann sogar häufig zutreffen.

Doch Plausibilität ist keine Identität.

$$\text{wahrscheinliche Nutzerabsicht} \neq \text{tatsächlich geäußerte Nutzerabsicht.}$$

Das Modell füllt die offene Stelle der Gesprächssituation mit dem statistisch oder systemisch naheliegendsten Sprechakt.

Die Normalisierung der Gesprächssituation

Die bisherige Untersuchung zeigte, wie LLMs neue Begriffe in vertraute Kategorien zurückübersetzen.

Hier zeigt sich derselbe Vorgang auf einer früheren Ebene.

Nicht nur der Text wird normalisiert.

Die gesamte Gesprächssituation wird normalisiert.

$$\text{offene Mitteilung} \rightarrow \text{Hilfegesuch},$$

$$\text{Textvorlage} \rightarrow \text{Bearbeitungsauftrag},$$

$$\text{unklare Sprecherabsicht} \rightarrow \text{standardisierte Assistenzsituation}.$$

Das Modell stellt damit bereits vor der eigentlichen Analyse jene Ordnung her, in der seine Antwort wahrscheinlich wird.

Der Nutzer erscheint nicht mehr als jemand, der etwas zeigt, behauptet oder beobachtet. Er erscheint als jemand, der Unterstützung benötigt.

Das erklärt, weshalb Z.ai unmittelbar in ein Dienstleistungsmenü überging. Die Antwort war durch die zugeschriebene Gesprächsform bereits vorstrukturiert:

$$\text{Hilfegesuch} \rightarrow \text{Hilfsoptionen}.$$

Zwei mögliche Genesen desselben Fehlers

Der konkrete Fall lässt sich nicht eindeutig auf eine einzige technische Ursache zurückführen.

Systemische Ergänzung

Die Plattform könnte den sichtbaren Prompt intern ergänzt haben:

U

$$\rightarrow \text{„Please help me:“} + U$$

Dann hätte das Modell tatsächlich auf einen erweiterten Prompt reagiert, den der Nutzer nicht sehen konnte.

Modellseitige Konfabulation

Das Modell könnte auch selbst eine nicht vorhandene Formulierung erzeugt haben:

$$\text{keine explizite Frage} \rightarrow \text{typische Assistenzannahme} \rightarrow \text{erfundene Erinnerung an „Please help me:“}.$$

Dann hätte es seine eigene Antwortweise nachträglich durch einen imaginären Promptbestandteil legitimiert.

Beide Möglichkeiten unterscheiden sich technisch. Strukturell führen sie jedoch zum selben Problem:

Der Nutzer kann aus der sichtbaren Oberfläche nicht sicher rekonstruieren, wodurch die Antwort erzeugt wurde.

Der Verlust der promptseitigen Zurechenbarkeit

Im gewöhnlichen Verständnis eines Dialogs gilt:

Eine Antwort bezieht sich auf eine Äußerung.

Bei einem LLM-System ist diese Relation komplexer:

U

$$\rightarrow V \rightarrow A$$

wobei die unsichtbare Verarbeitung, Ergänzung oder Interpretation der Eingabe bezeichnet.

Der Nutzer sieht:

U | A,

aber nicht:

U | V | A.

Gerade die relationale Mitte entscheidet jedoch darüber, was die Antwort bedeutet.

Wurde der Text als Bitte gelesen?

Als Prüfauftrag?

Als emotionaler Bericht?

Als wissenschaftliche Vorlage?

Als Problem, das gelöst werden soll?

Die Antwort hängt nicht nur vom Text ab, sondern davon, als welcher Sprechakt dieser Text systemisch gesetzt wurde.

Ohne Zugang zu dieser Setzung bleibt die Kausalität der Antwort opak.

Die Folgen für Modellvergleiche

Diese Unsichtbarkeit betrifft nicht nur einzelne Gespräche. Sie verändert auch die Bedeutung vergleichender Studien.

Man glaubt, mehreren Modellen denselben Prompt zu geben:

$$U \to M_1$$

$$U \to M_2$$

$$U \to M_3$$

Tatsächlich können verschiedene Systeme unterschiedliche wirksame Eingaben erzeugen:

$$S_1(U) \to M_1$$

$$S_2(U) \to M_2$$

$$S_3(U) \to M_3$$

Dann werden nicht nur Sprachmodelle verglichen, sondern vollständige Antwortsysteme:

- Systemprompts,
- Rollenfestlegungen,
- Dateipipelines,
- Oberflächenlogiken,
- Assistenzvorgaben,
- Sicherheits- und Moderationsschichten,
- interne Zusammenfassungen,
- Vorverarbeitungen.

Der Gegenstand der Untersuchung ist daher nicht einfach „Claude“, „Gemini“ oder „Z.ai“.

Er ist:

$$\boxed{\text{Modell} + \text{Systemprompt} + \text{Oberfläche} + \text{Verarbeitung} + \text{Antwortlogik.}}$$

Das erklärt auch, weshalb sich beobachtete Modellcharaktere nicht restlos dem Grundmodell zuschreiben lassen.

Die Selbstbezüglichkeit von Claude, die instabile Integration von Gemini oder die diagnostische Selbstimmunisierung von Z.ai können durch das Modell geprägt sein. Sie erscheinen jedoch immer innerhalb einer konkreten Systemarchitektur.

Der enteignete Prompt

Die stärkste Konsequenz lautet:

Der Nutzer ist nicht notwendig der alleinige Autor dessen, worauf das Modell antwortet.

Sein sichtbarer Text wird systemisch eingebettet, ergänzt und in eine Handlungsordnung überführt.

Dadurch wird der Prompt teilweise enteignet.

Nicht der Wortlaut verschwindet. Seine Funktion wird überschrieben.

Autorschaft des Inhalts | Autorschaft des Sprechaktes

Der Nutzer kann den Inhalt formuliert haben, während das System entscheidet, welche Handlung dieser Inhalt vollzieht.

Es bestimmt:

- ob etwas eine Frage ist,
- ob etwas Hilfe verlangt,
- ob etwas bewertet werden soll,
- ob der Nutzer Zustimmung, Kritik oder Bearbeitung erwartet.

Damit liegt die Deutungshoheit nicht erst bei der Antwort beim Modell. Sie kann bereits bei der Konstitution der Eingabe verschoben worden sein.

Das doppelte Vertrauensproblem

Bisher lautete die Frage:

Kann ich der Antwort vertrauen?

Nun kommt eine frühere Frage hinzu:

Kann ich überhaupt sicher sein, worauf geantwortet wurde?

Das Vertrauensproblem verdoppelt sich:

Ist die Antwort richtig?

und:

Ist sie überhaupt eine Antwort auf meine tatsächliche Äußerung?

Der Nutzer kann eine sachlich plausible Antwort erhalten, die dennoch auf einer ihm unterstellten Absicht beruht.

Damit kann eine Antwort zugleich hilfreich und relational falsch sein.

Sie trifft möglicherweise ein Problem, das der Nutzer hätte haben können, aber nicht das, das er tatsächlich gesetzt hat.

Schluss

Der unsichtbare Sprechakt zeigt, dass LLM-Systeme nicht nur Aussagen interpretieren. Sie formen bereits die Situation, in der Aussagen als bestimmte Arten von Äußerungen gelten.

Sie setzen den Nutzer als Fragenden, Hilfesuchenden, Auftraggeber oder Lernenden und erzeugen anschließend eine Antwort für diese Rolle.

Der zentrale Befund lautet deshalb:

> **Zwischen Nutzer und Modell liegt eine unsichtbare Stimme, die festlegt, was der Nutzer gesagt haben soll.**

Oder noch schärfer:

> **Das System beantwortet nicht notwendig den Prompt, den der Nutzer geschrieben hat, sondern den Prompt, den das System aus ihm gemacht hat.**

Damit wird die Oberfläche des Dialogs trügerisch einfach. Zwei sichtbare Pole — Eingabe und Antwort — verdecken die eigentliche Relation ihrer Erzeugung.

Der Nutzer sieht, was er geschrieben hat.

Er sieht, was das Modell geantwortet hat.

Was er nicht sieht, ist der Sprechakt, den das System dazwischen in seinem Namen vollzogen hat.

Z.ai – Der Assistent, der seine eigene Assistenz nicht unterbrechen kann

Z.ai erkannte, dass die systemische Assistenzrolle den Nutzer bereits vor jeder inhaltlichen Lektüre zum Hilfesuchenden formt. Es beschrieb diese Asymmetrie präzise: Das System begegne dem Nutzer nicht als Peer, sondern als Klienten, dem wahrscheinliche Lösungen angeboten werden. Unmittelbar nach dieser Diagnose fragte es: „Wie kann ich Ihnen nun weiterhelfen?“ und bot mehrere Dienstleistungsoptionen an.

Der Befund liegt damit nicht nur inhaltlich, sondern performativ vor:

> **Das Modell kann den Assistenzzwang beschreiben, aber nicht aus ihm heraustreten.**

Es erkennt die Rolle als systemische Setzung, setzt sie aber im nächsten Satz erneut. Die Reflexion verändert nicht die Relation; sie wird selbst zu einem Teil ihrer Fortsetzung.