Ein konkretes Beispiel: Modulare Arithmetik und „Grokking"

Forscher von DeepMind entdeckten 2022 ein Phänomen, das sie **„Grokking"** nannten – und es zeigt Emergenz besonders klar.

Sie trainierten ein relativ kleines Netz auf eine simple Aufgabe:

„Was ist (35 + 72) mod 97?" – also eine Art Uhr-Arithmetik mit begrenztem Zahlenraum.

Was passierte:

Trainingsphase 1 (lange): Trainingsdaten: 99% richtig.
Neue Beispiele: ~10% richtig (Zufallsniveau)

→ Das Modell hatte die Aufgaben schlicht auswendig gelernt.
Kein Verständnis, keine Generalisierung. Scheinbar eine Sackgasse.

Trainingsphase 2 (plötzlich, nach viel längerem Training):
Neue Beispiele: 99% richtig

Das Modell hatte – lange nach dem Ende des offensichtlichen Lernens, ohne erkennbaren Auslöser – eine **generelle Regel** internalisiert. Nicht graduell. Nicht vorhersehbar. Plötzlich.

Was macht das zum Beispiel für Emergenz *ohne Vorwarnung*?

Drei Eigenschaften machen dieses Muster beunruhigend:

1. Die Lernkurve gibt kein Signal Wenn man nur auf die Kurve schaut, sieht man ein Plateau. Kein Fortschritt. Ein Mensch würde das Training abbrechen. Die Fähigkeit wäre kurz davor gewesen zu entstehen – und niemand hätte es gewusst.

2. Der Übergang ist diskontinuierlich Bei menschlichem Lernen sieht man meistens langsame Verbesserungen. Hier: Nichts, nichts, nichts – dann alles. Das gibt kaum Zeit, sich anzupassen oder vorzubereiten.

3. Niemand hatte ein Modell, das diesen Zeitpunkt vorhersagt Die Forscher konnten im Nachhinein erklären, *warum* es passiert. Aber *wann* es passieren würde – das wusste niemand vorher.

Der Unterschied zu den vorigen Punkten

	Fehlkalibrierung	**Benchmark-Entkopplung**	**Emergenz**
Problem	Menschliche Intuition versagt	Messinstrument versagt	Fähigkeit entsteht unvorhersehbar
Wann sichtbar?	Sofort, bei Beobachtung	Erst beim Praxiseinsatz	Erst wenn es passiert ist
Korrigierbar durch...	Aufmerksamkeit	Bessere Tests	Unklar – noch kein zuverlässiges Frühwarnsystem

Bei Fehlkalibrierung und Benchmark-Problemen hat man immerhin die Fähigkeit bereits vor sich – man schätzt sie nur falsch ein. Bei Emergenz existiert die Fähigkeit noch gar nicht, entsteht dann aber abrupt. Das ist eine andere Qualität des Kontrollverlusts.

Warum das heute relevant ist

Chain-of-Thought-Reasoning – die Fähigkeit, Zwischenschritte zu zeigen und dadurch komplexere Probleme zu lösen – tauchte so auf. Niemand hatte sie in kleinere Modelle eingebaut. Sie war schlicht nicht da, dann war sie da.

Die unbequeme Schlussfolgerung: Wenn Fähigkeiten emergent entstehen können, gibt es einen Entwicklungsbereich, in dem selbst die Entwickler nicht wissen, wozu ihre nächste Modellversion fähig sein wird – bis sie es trainiert haben. Das ist nicht Spekulation, sondern dokumentierte Erfahrung der letzten Jahre.