

Emerging Capabilities

SOTA2 - Group C

Anton Kesy, Étienne Muser, Katharina Schindler, Lukas Fehrenbacher, Nico Ruschmann

Offenburg University of Applied Sciences

WS 2024/2025



“Emergent Abilities of Large Language Models”

- Google Research
- Stanford University
- UNC Chapel Hill
- DeepMind

“Are Emergent Abilities of Large Language Models a Mirage?”

- CS Stanford

Was sind Emerging Capabilities?

- “abilities that are not present in smaller-scale models but are present in large-scale models”

Few-Shot-Prompting

- Arithmetik
- Transliteration
 - Wörter vom Internationalen Phonetischen Alphabet in die Standardorthographie zu übertragen
- Wortentschlüsselung
 - Ein Wort aus seinen durcheinandergewürfelten Buchstaben wiederherzustellen
- Beantwortung von Fragen in anderen Sprachen
- Wahrheitsfindung
- Verständnis von Wörtern im Kontext

- Mehrstufiges Schlussfolgern
 - Chain-of-Thought, erst bei größeren Modellen verbessert
- Befolgen von Anweisungen
 - Mischung von Aufgaben, die als Anweisungen formuliert sind
- Programmausführung
 - Zwischenausgaben vorherzusagen (Scratchpad)
 - mehrstufige Berechnungen
- Modellkalibrierung

Emergente Fähigkeiten: Warum Größe zählt?

- **Aufgabenkomplexität:** Größere Modelle bewältigen komplexe Aufgaben mit vielen Rechenschritten besser.
- **Wissensrepräsentation:** Mehr Parameter ermöglichen das Speichern und Abrufen von mehr Weltwissen.
- **Komposition:** Große Modelle kombinieren Teilfähigkeiten besser zu komplexen Lösungen.
- **Datenqualität/-menge:** Hochwertige, umfangreiche Daten fördern Fähigkeiten, auch bei kleineren Modellen.
- **Metrikwahl:** Nichtlineare Metriken können abrupte Verbesserungen vortäuschen. Glattere Kurven mit linearen Metriken.

Es kommt nicht auf die Größe an?!

- Auftretende emergente Fähigkeiten grosser Sprachmodelle könnten ein Artefakt der gewählten Bewertungsmetrik sein
- Metriken stellen stetige Verbesserung der Modellleistung mit zunehmender Skalierung verzerrt dar
- Kernargument: Fehlerrate pro Token nimmt mit Modellgrösse kontinuierlich ab
- Scaling Laws: Leistung von Deep-Learning-Modellen folgt Potenzgesetz in Bezug auf Trainingsdatengrösse, Parameteranzahl oder Rechenleistung
- Nichtlineare Metriken können kontinuierliche Verbesserung in scheinbar abrupte Veränderung umwandeln

Analyse der InstructGPT/GPT-3-Modelle

- Emergente Fähigkeiten bei arithmetischen Aufgaben verschwinden, wenn Token-Edit-Distanz statt Genauigkeit als Metrik verwendet wird
- Misst Anzahl Bearbeitungen (Einfügungen, Löschungen, Ersetzungen) zur Umwandlung der Modellausgabe in Zielausgabe → linearere Metrik

Meta-Analyse veröffentlichter Benchmarks

- Analyse des BIG-Bench-Benchmarks zeigt, dass scheinbar emergente Fähigkeiten nur bei Verwendung bestimmter Metriken auftreten
- Nichtlineare oder diskontinuierliche Metriken wie Genauigkeit oder exakte String-Übereinstimmung

Induzieren scheinbar emergenter Fähigkeiten in neuronalen Netzen bei Bilderkennungsaufgaben

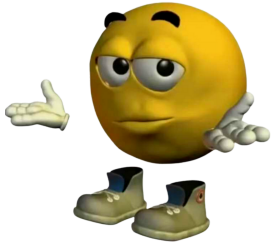
- Konstruktion neuer Metriken, die scheinbare Emergenz erzeugen

Fazit: Es kommt nicht auf die Größe an?!

- Modellgröße ist wichtiger Faktor für Leistung von LLMs, aber nicht der einzige
- Wahl der Bewertungsmetrik kann Leistungsentwicklung von Modellen stark beeinflussen
- Eindruck von emergenten Fähigkeiten kann auf Metrikwahl zurückzuführen sein

Verwendung von linearen und kontinuierlichen Metriken sollten zur adäquaten Abbildung der tatsächlichen Leistungsentwicklung von Modellen mit zunehmender Skalierung und fairem Vergleich zwischen Modellen verschiedener Größen verwendet werden

Wer hat Recht?



Medium: Are Emergent Abilities of Large Language Models a Mirage Or Not?