Emergence

Phase Transitions · Emergent Abilities · The Mirage Debate · Grokking

What is emergence?

Emergence: complex behaviors arise from simple components interacting, producing properties that **no individual component possesses**

Neurons

86 billion cells firing electrical signals

↓

Consciousness, memory, emotions

Water molecules

H_2O with simple hydrogen bonds

↓

Wetness, waves, surface tension

Parameters

Billions of weights doing matrix multiplies

↓

Reasoning, code generation, humor

"The whole is greater than the sum of its parts"
— the central mystery of complex systems

Emergence in nature: ant colonies

Individual ant

- Follow pheromone trails
- Drop pheromone when carrying food
- Avoid obstacles
- Random walk when no signal
- No map, no plan, no leader

~250,000 neurons per ant

× **millions**

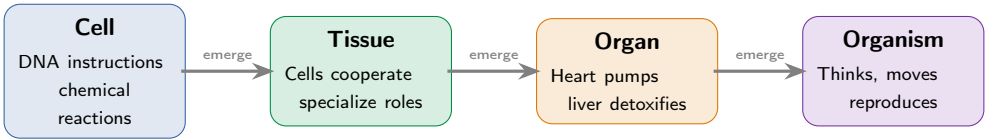
Colony behavior

- Optimal foraging paths
- Complex nest architecture
- Division of labor
- Collective defense strategies
- Temperature regulation

Colony acts as a "superorganism"

No ant knows the colony's structure. No ant plans the foraging route.
Intelligence emerges from **local interactions** at scale — sound familiar?

Emergence in nature: cells to organs



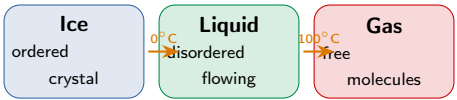
No single heart cell "pumps." No single neuron "thinks."
The function **only exists** at the level of the organized whole.

More examples:

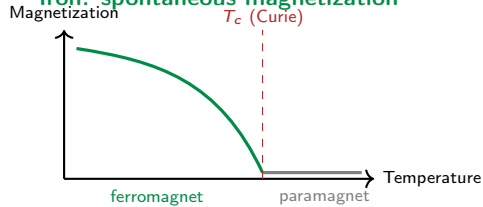
- **Bird flocking:** 3 simple rules → mesmerizing patterns
- **Stock markets:** individual trades → bubbles and crashes
- **Traffic jams:** local braking → backward-moving waves
- **Cities:** individual choices → neighborhoods

Phase transitions: emergence in physics

Water: temperature as control parameter



Iron: spontaneous magnetization



The pattern:

Control parameter crosses
a **critical threshold** →
qualitatively new behavior appears

The LLM analogy:

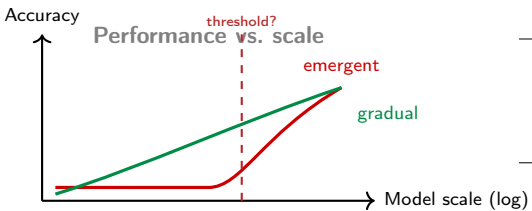
Scale (params / FLOPs) crosses
a critical threshold →
new abilities appear?

Phase transitions are **sharp** in the thermodynamic limit.
Do LLMs undergo similar sharp transitions as they scale, or is it more gradual?

Emergent abilities in LLMs

Wei et al. (2022): "Emergent Abilities of Large Language Models"

An ability is **emergent** if it is **absent in smaller models** but **present in larger models**



Ability	Scale threshold
Few-shot learning	∼175B (GPT-3)
Chain-of-thought	∼100B
Multi-digit arithmetic	∼13B
Word unscrambling	∼137B
Instruction following	∼100B

137 abilities documented

Key claim: we **cannot predict** what abilities will emerge at the next scale.
This made scaling feel like exploring uncharted territory.

Chain-of-thought: the poster child of emergence

Wei et al. (2022): CoT prompting on GSM8K (grade-school math)

Small model (~10B)

Q: "Roger has 5 tennis balls. . . "

CoT: "Let me think. . . the fish swam over the purple. . . **42**"

Incoherent chain → wrong answer

CoT actually **hurts** performance!

scale up

Large model (540B PaLM)

Q: "Roger has 5 tennis balls. . . "

CoT: "He buys 2 cans of 3 each, so $2 \times 3 = 6$. Total $= 5 + 6 =$ **11**"

Coherent reasoning → correct

GSM8K: 18% → **57%** with CoT

"Let's think step by step" — the same prompt goes from **harmful** to **transformative** depending on model scale. This is what made emergence so compelling.

Zero-shot CoT (Kojima et al., 2022): no examples needed, just "Let's think step by step"

Are emergent abilities a mirage?

Schaeffer et al. (2023) — NeurIPS Outstanding Paper Award

Emergent abilities are not a property of the model —
they are an artifact of the **evaluation metric**

Exact-match accuracy

"12345" vs "12346"

Score: **0** (all or nothing)

Discontinuous → sharp "emergence"

Token edit distance

"12345" vs "12346"

Score: **0.8** (4 out of 5 right)

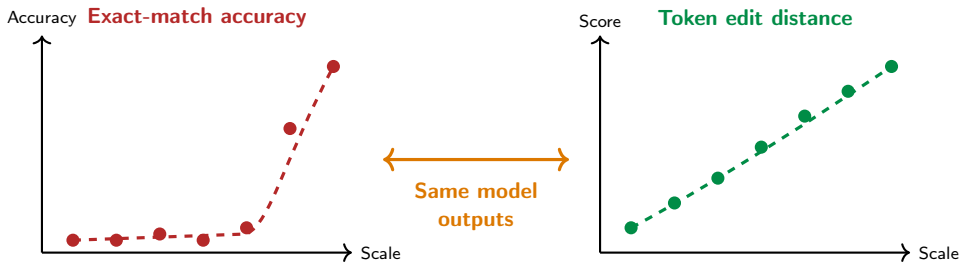
Continuous → smooth improvement

Same model outputs, different metric → emergence disappears.

92% of claimed emergent abilities on BIG-Bench used just 2 discontinuous metrics.

"Nothing in this paper should be interpreted as claiming that LLMs cannot display emergent abilities.
We argue that the *evidence* is flawed." — Schaeffer et al.

The metric illusion



A 5-digit addition model getting 4/5 digits right scores **0%** on exact match but **80%** on edit distance. The "emergence" is in the ruler, not the model.

The AND-gate: why some emergence might be real

For **compositional tasks** (solve step A **AND** B **AND** C), even smooth per-step improvement produces a **sharp transition** in overall success

3-step reasoning task

If each step accuracy = p:

Overall = p^3

This AND-gate effect is a **genuine computational phenomenon**, not a measurement artifact

p (per step)	p^3 (overall)	
0.5	0.125	random
0.7	0.343	mediocre
0.9	0.729	decent
0.95	0.857	

With **10 steps** at p=0.9:

$0.9^{10} = 0.35$ (still failing!)

At p=0.99: $0.99^{10} = 0.90$ (works)

Where the debate stands today

Schaeffer critique

Many claimed emergent abilities were metric artifacts.

92% used discontinuous metrics

Valid for specific claims

Middle ground

Capabilities develop continuously but with **nonlinear** scaling.

Pre-training loss is a better predictor than size

Current mainstream view

Compositional tasks

Multi-step reasoning shows genuine sharp transitions via the AND-gate effect.

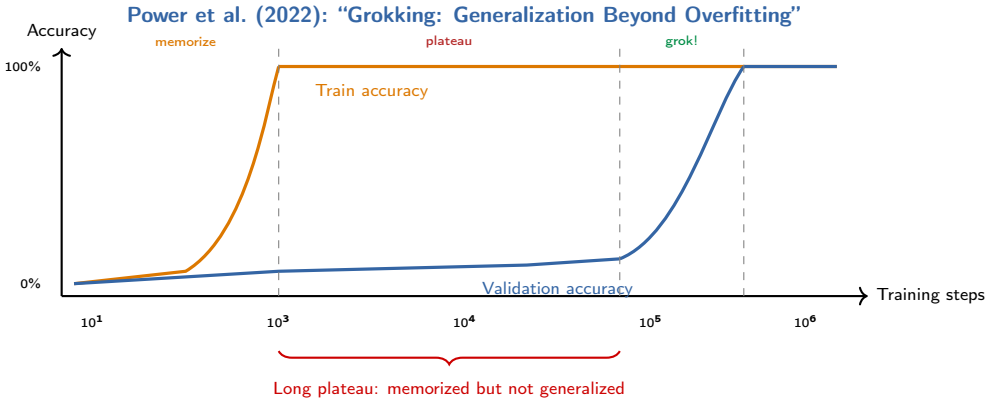
Real emergence, not metric

Strongest case for real emergence

Key shift: from "emergence is unpredictable magic" to "capabilities scale nonlinearly, and we're learning to predict them better"

Pre-training loss matters more than parameter count. When loss drops below a critical threshold, downstream performance improves sharply. Loss = the real control parameter.

Grokking: delayed generalization



Division mod 97: the model memorizes in $\sim 10^3$ steps but generalizes only at $\sim 10^6$ steps.
1000× more training than needed for memorization. Weight decay is critical.

Grokking and emergence: the common pattern

Phase transition

Control: temperature
Order: magnetization

Below T_c: ordered
Above T_c: disordered

LLM emergence

Control: scale / loss
Order: task accuracy

Below threshold: random
Above threshold: competent

Grokking

Control: training steps
Order: generalization

Before grok: memorized
After grok: generalized

All three share: **long plateau** $\rightarrow$ **sharp transition** $\rightarrow$ **new regime**

The mechanism hypothesis:

The network discovers a **simpler** internal representation that generalizes beyond memorization

Grokking is universal:

Humayun et al. (ICML 2024): occurs in CNNs, ResNets, and practical settings — not just toy tasks

Weak vs. strong emergence

Weak emergence

Unexpected but **derivable** from low-level rules (given enough simulation)

- Ant colony behavior
- Traffic jams
- LLM capabilities
- Weather patterns

Most scientific emergence

Strong emergence

Not deducible even in principle from the low-level domain

- Consciousness?
- Subjective experience?

Chalmers (2006): "uncomfortably like magic"

Highly debated / possibly none

LLM emergence is almost certainly **weak** — capabilities arise from known components (attention, gradient descent, data) in ways that are surprising but not fundamentally mysterious.

The surprise comes from **scale**: we can describe the parts but didn't predict the composite behavior

Why emergence matters for AI

1. Scaling strategy

If abilities emerge at scale, it justifies spending billions on larger models. But if it's gradual, we can predict what we'll get — less gambling.

2. Safety

Unpredictable emergence means dangerous capabilities could appear without warning. Gradual scaling is easier to monitor and control.

3. Evaluation

Metric choice matters enormously. Continuous metrics give a more honest picture of model progress. Don't confuse your ruler with reality.

4. Understanding

We built these systems but don't fully understand them. Emergence = humility about what billions of parameters can do.

Whether emergence is "real" or a measurement artifact has **practical consequences**:
it changes how we train, evaluate, deploy, and regulate AI systems.

Further reading

Emergence in LLMs

- Wei et al. (2022), "Emergent Abilities of Large Language Models" — the original 137-ability survey
- Schaeffer et al. (2023), "Are Emergent Abilities of LLMs a Mirage?" — NeurIPS Outstanding Paper

Grokking & Phase Transitions

- Power et al. (2022), "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets"
- Humayun et al. (ICML 2024), "Deep Networks Always Grok and Here is Why"

Emergence in General

- Bedau (1997), "Weak Emergence" — philosophical framework
- Chalmers (2006), "Strong and Weak Emergence" — consciousness and irreducibility
- Mitchell (2009), *Complexity: A Guided Tour* — accessible intro to complex systems

Questions?