Is Scaling Enough to Achieve AGI?

A Critical Analysis of the Scaling Hypothesis

Spring 2026
CIS5590 - Topics in Computer Science
Presenter: Che-Wei, Hsu and Ming-Chun, Lee

Agenda

What we'll cover today

01

What is Scaling?

Model size, data, and compute

02

Evidence FOR Scaling → AGI

4 key papers supporting the hypothesis

03

Evidence AGAINST Scaling → AGI

4 papers highlighting limitations

04

Key Debate & My Position

Quantitative gains vs. qualitative intelligence

What is Scaling?

The three pillars of modern AI growth

Model Size

More parameters → learn more complex patterns

Training Data

More data → wider range of captured information

Compute

More GPU/TPU power → training large models at scale

Scaling Laws: performance improves predictably as these three factors increase — a power-law relationship (Kaplan et al., 2020)

Evidence FOR: Scaling Might Be Enough

Four landmark papers supporting the hypothesis

Sparks of AGI (2023)

GPT-4 handles math, coding, medicine & law near human-level — without task-specific training. Suggests early signs of general intelligence.

Scaling Laws (2020)

Model performance follows a reliable power-law as size, data & compute increase — scaling is systematic and predictable.

Emergent Abilities (2022)

New capabilities (e.g., multi-step reasoning) appear suddenly once models pass a certain scale — qualitative, not just quantitative gains.

Chinchilla (2022)

Optimal balance of model size and data outperforms larger, undertrained models. Scaling should be principled, not just bigger.

Evidence FOR: Scaling Might Be Enough

An additional supporting perspective

GPT-4 Technical Report (OpenAI, 2023)

OpenAI's own technical report on GPT-4 documents that the model achieves human-level performance on a wide range of professional and academic benchmarks — including the Uniform Bar Exam (top 10%), AP exams, and the SAT. This was achieved without benchmark-specific tuning, relying purely on large-scale pretraining. The report demonstrates that scaling produces not just incremental gains but qualitative leaps in capability across structured domains.

Summary of FOR evidence

Across five papers: scaling is systematic (Scaling Laws), unlocks emergent abilities (Emergent Abilities), can be optimized (Chinchilla), and produces near-human professional performance (Sparks of AGI, GPT-4 Report). Together they suggest scaling is a credible path toward increasingly general AI.

Evidence AGAINST: Scaling Is Not Enough

Four studies revealing critical limitations

TruthfulQA (2021)

Larger models can be LESS truthful — best model: 58% vs humans: 94%. Scaling amplifies pattern imitation, not factual reliability.

Illusion of Thinking (Apple, 2025)

Model performance collapses at high problem complexity. Models fail to apply stable reasoning procedures even when required.

Chinchilla (2022)

Simply growing parameter count without proportional data is insufficient. Exposes a flawed understanding of what 'scaling' means.

Scaling Is Not Enough (2026)

Structural 'informational linkage gaps' cannot be solved by more data or compute. Scaling improves interpolation, not true abstraction.

Evidence AGAINST: Scaling Is Not Enough

An additional critical perspective

On the Measure of Intelligence (François Chollet, 2019)

François Chollet argues that current AI benchmarks — including those that large language models excel at — measure skill acquisition from exposure to vast training data, not true general intelligence. He proposes that genuine intelligence should be measured by efficiency of learning: the ability to solve novel problems with minimal prior exposure. By this definition, scaling primarily rewards memorization and interpolation rather than flexible reasoning. Even the most capable scaled models struggle dramatically on the ARC benchmark, which was designed to require genuine abstraction.

Summary of AGAINST evidence

Across five papers: scaling makes models less truthful (TruthfulQA), reasoning collapses at complexity (Illusion of Thinking), naive parameter growth fails (Chinchilla), structural limits exist (Scaling Is Not Enough), and benchmarks measure skill — not intelligence (Chollet). This is a consistent pattern of fundamental limitations.

The Key Debate

Does scaling produce qualitative intelligence, or just quantitative gains?

✓ Scaling SUPPORTERS say…

- **Continuous improvement** in performance as scale grows
- **Emergent capabilities** arise unpredictably at large scale
- **General intelligence** could gradually emerge

VS

✗ Scaling CRITICS argue…

- **Better benchmarks ≠ true intelligence**
- **Reasoning & truthfulness** still fundamentally weak
- **Structural limits** can't be overcome by scale alone

Our Position

"Scaling alone is not enough to achieve AGI."

Scaling is powerful —

GPT-4 shows impressive cross-domain performance (Sparks of AGI).

But patterns ≠ understanding —

Models imitate training data, including falsehoods (TruthfulQA).

Reasoning is fragile —

Performance collapses at higher problem complexity (Illusion of Thinking).

New ideas needed —

To achieve AGI, we need architectures beyond simply 'scaling up'.

Conclusion

Scaling: essential but insufficient

Scaling has driven remarkable AI progress — but it primarily improves pattern recognition, not genuine reasoning, truthfulness, or cross-domain abstraction.

Scaling works:	Reliable, predictable performance gains across tasks
Fundamental gaps remain:	Truthfulness, robustness, and structural reasoning limits persist
Next steps:	New architectures, reasoning modules, and neuro-symbolic approaches are needed

Thank you.