Bounded Regret

Future ML Systems Will Be Qualitatively Different

JAN 11, 2022 • 7 MIN READ

In 1972, the Nobel prize-winning physicist Philip Anderson wrote the essay "More Is Different". In it, he argues that quantitative changes can lead to qualitatively different and unexpected phenomena. While he focused on physics, one can find many examples of More is Different in other domains as well, including biology, economics, and computer science. Some examples of More is Different include:

- **Uranium.** With a bit of uranium, nothing special happens; with a large amount of uranium packed densely enough, you get a nuclear reaction.
- **DNA.** Given only small molecules such as calcium, you can't meaningfully encode useful information; given larger molecules such as DNA, you can encode a genome.
- **Water.** Individual water molecules aren't wet. Wetness only occurs due to the interaction forces between many water molecules interspersed throughout a fabric (or other material).
- **Traffic.** A few cars on the road are fine, but with too many you get a traffic jam. It could be that 10,000 cars could traverse a highway easily in 15 minutes, but 20,000 on the road at once could take over an hour.
- **Specialization.** Historically, in small populations, virtually everyone needed to farm or hunt to survive; in contrast, in larger and denser communities, enough food is produced for large fractions of the population to specialize in non-agricultural work.

While some of the examples, like uranium, correspond to a sharp transition, others like specialization are more continuous. I'll use **emergence** to refer to qualitative changes that arise from quantitative increases in scale, and **phase transitions** for cases where the change is sharp.

In this post, I'll argue that emergence often occurs in the field of AI, and that this should significantly affect our intuitions about the long-term development and deployment of AI systems. We should expect weird and surprising phenomena to emerge as we scale up systems. This presents opportunities, but also poses important risks.

Emergent Shifts in the History of AI

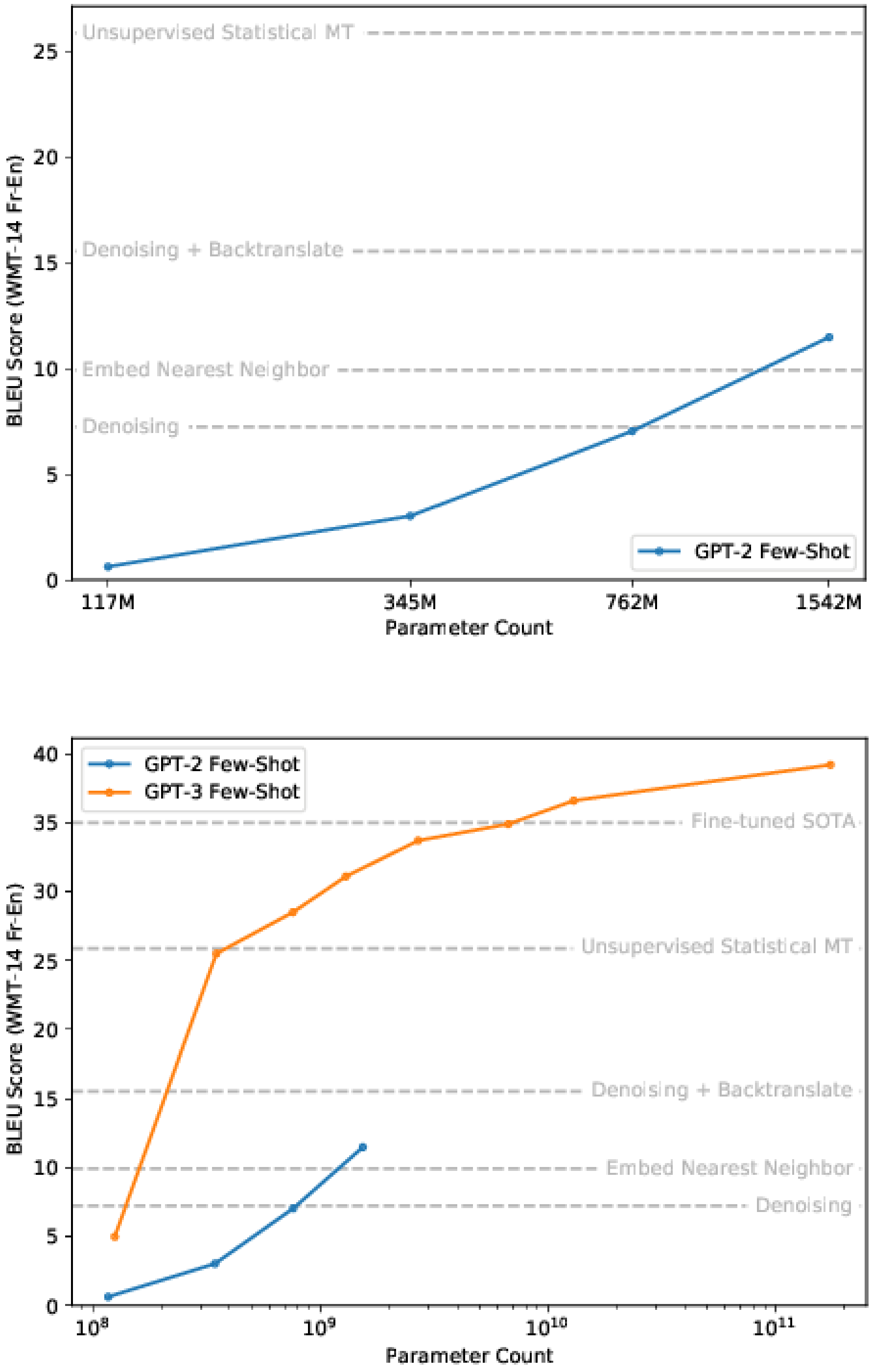
There have already been several examples of quantitative differences leading to important qualitative changes in machine learning.

Storage and Learning. The emergence of machine learning as a viable approach to AI is itself an example of More Is Different. While learning had been discussed since the 1950s, it wasn't until the 80s-90s that it became a dominant paradigm: for instance, IBM's first statistical translation model was published in 1988, even though the idea was proposed in 1949[1]. Not coincidentally, 1GB of storage cost over $100k in 1981 but only around $9k in 1990 (adjusted to 2021 dollars). The Hansard corpus used to train IBM's model comprised 2.87 million sentences and would have been difficult to use before the 80s. Even the simple MNIST dataset would have required $4000 in hardware just to store in 1981, but that had fallen to a few dollars by 1998 when it was published. Cheaper hardware thus allowed for a qualitatively new approach to AI: in other words, More storage enabled Different approaches.

Compute, Data, and Neural Networks. As hardware improved, it became possible to train neural networks that were very deep for the first time. Better compute enabled bigger models trained for longer, and better storage enabled learning from more data; AlexNet-sized models and ImageNet-sized datasets wouldn't have been feasible for researchers to experiment with in 1990.

Deep learning performs well with lots of data and compute, but struggles at smaller scales. Without many resources, simpler algorithms tend to outperform it, but with sufficient resources it pulls far ahead of the pack. This reversal of fortune led to qualitative changes in the field. As one example, the field of machine translation moved from phrase-based models (hand-coded features, complex systems engineering) to neural sequence-to-sequence models (learned features, specialized architecture and initialization) to simply fine-tuning a foundation model such as BERT or GPT-3. Most work on phrase-based models was obviated by neural translation, and the same pattern held across many other language tasks, where hard-won domain-specific engineering effort was simply replaced by a general algorithm.

Few-shot Learning. More recently, GPT-2 and GPT-3 revealed the emergence of strong few-shot and zero-shot capabilities, via well-chosen natural language prompting.

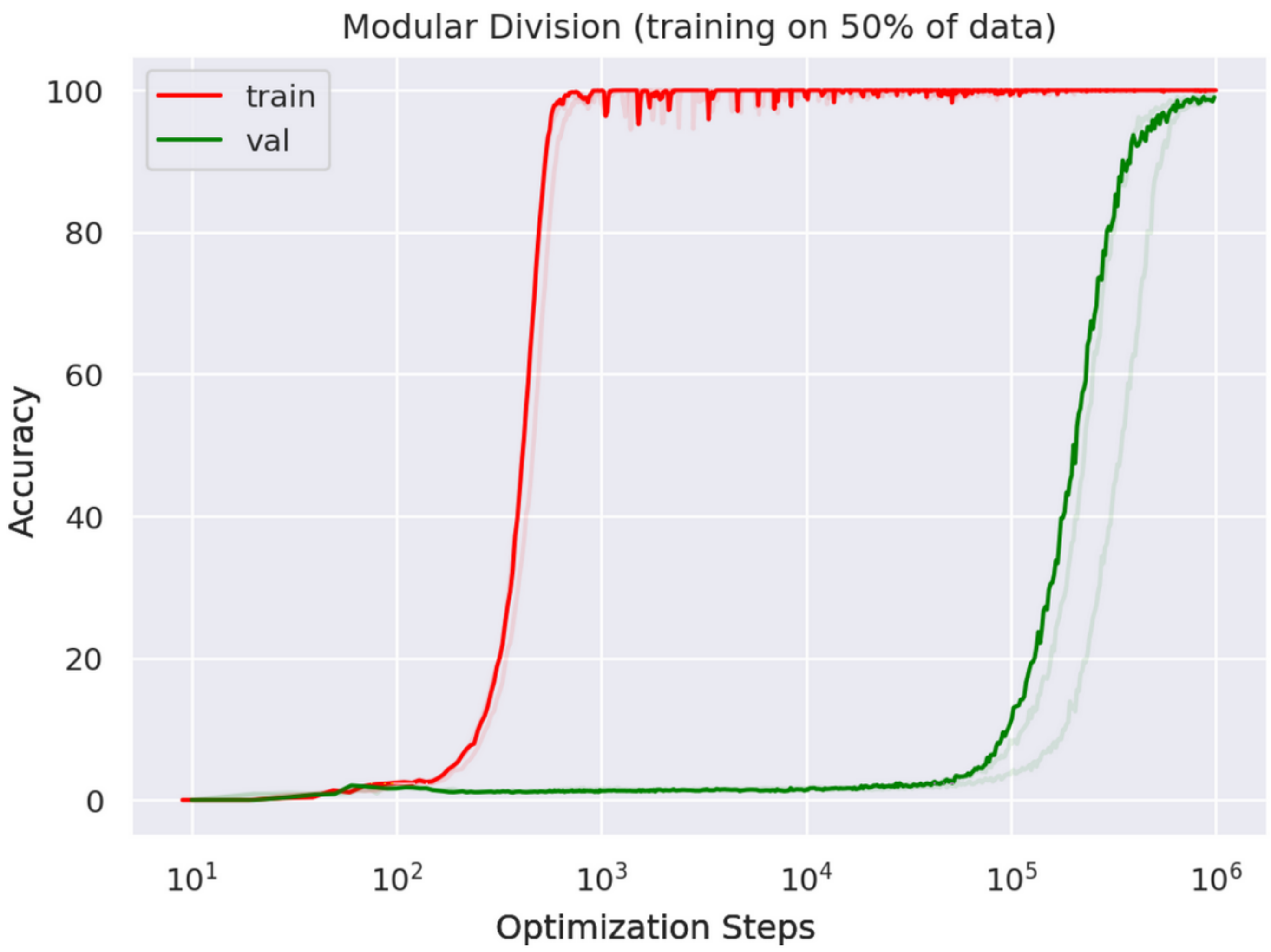


Top: Few-shot machine translation performance (BLEU score) for GPT-2. Bottom: GPT-3 *(trained on more data) has an even starker curve, going from 5 to 25 BLEU between 100M and 400M parameters. Unsupervised baselines, as well as fine-tuned state-of-the-art, are indicated for reference.*

This was an unexpected and qualitatively new phenomenon that only appeared at large scales, and it emerged without ever explicitly training models to have these few-shot capabilities. Comparing GPT-2 to GPT-3 shows that the exact model size needed can vary due to the training distribution or other factors, but this doesn't affect the basic point that new capabilities can appear without designing or training for them.

Grokking. In 2021, Power et al. identified a phenomenon they call "grokking", where a network's generalization behavior improves qualitatively when training it for longer (even though the training loss is already small).

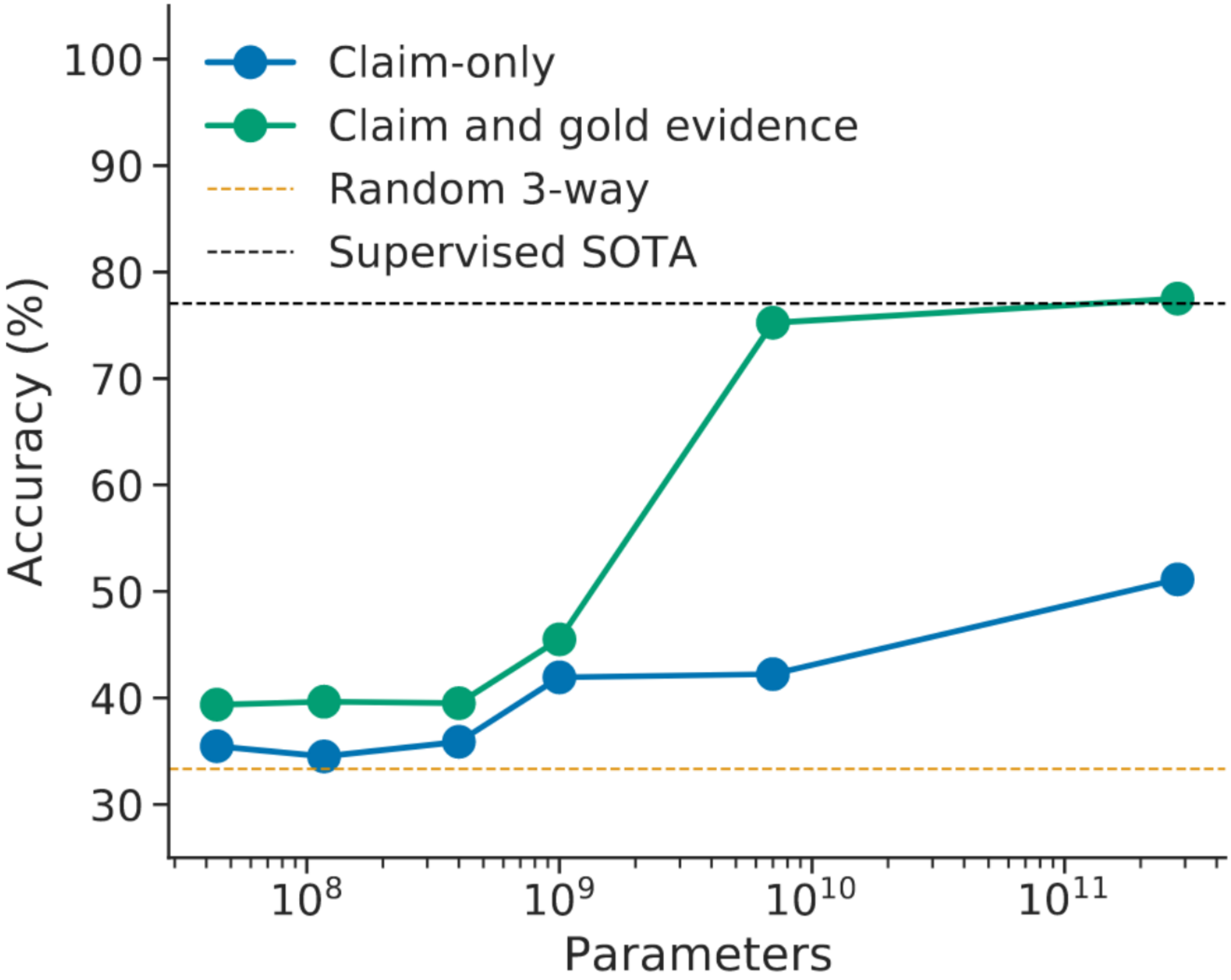
Specifically, for certain algorithmically generated logic/math datasets, neural networks trained for 1,000 steps achieve perfect train accuracy but near-zero test accuracy. However, after around 100,000 steps the test accuracy suddenly increases, achieving near-perfect generalization by 1 million steps.



This shows that even for a single model, we might encounter qualitative phase transitions as we train for longer.

Other potential examples. I'll briefly list other examples from recent papers. I don't think these examples are as individually clear-cut, but they collectively paint an interesting picture:

- McGrath et al. (2021) show that AlphaZero acquires many chess concepts at a phase transition near 32,000 training steps.
- Pan et al. (2021) show that reward hacking sometimes occurs via qualitative phase transitions as model size increases.
- DeepMind's recent Gopher model exhibits a phase transition on the FEVER task, acquiring the ability to evaluate evidence provided as side information (Figure 3):



- Wei et al. (2021) show that instruction-tuning hurts small models but helps large models (see Figure 6).
- Some few-shot tasks such as arithmetic show phase transitions with model size (see Brown et al. (2020), Figure 3.10).
- This researcher shares an anecdote similar to the “grokking” paper.

What This Implies for the Engineering Worldview

In the introduction post to this series, I contrasted two worldviews called Philosophy and Engineering. The Engineering worldview, which is favored by most ML researchers, tends to predict the future by looking at empirical trends and extrapolating them forward. I myself am quite sympathetic to this view, and for this reason I find emergent behavior to be troubling and disorienting. Rather than expecting empirical trends to continue, emergence suggests we should often expect new qualitative behaviors that are not extrapolations of previous trends.

Indeed, in this sense Engineering (or at least pure trend extrapolation) is self-defeating as a tool for predicting the future[2]. The Engineering worldview wants to extrapolate trends, but one trend is that emergent behavior is becoming more and more common. Of the four phase transitions I gave above, the first (storage) occurred around 1995, and the second (compute) occurred around 2015. The last two occurred in 2020 and 2021. Based on past trends, we should expect future trends to break more and more often.[3]

How can we orient ourselves when thinking about the future of AI despite the probability of frequent deviations from past experience? I'll have a lot more to say about this in the next few posts, but to put some of my cards on the table:

- Confronting emergence will require adopting mindsets that are less familiar to most ML researchers and utilizing more of the Philosophy worldview (in tandem with Engineering and other worldviews).
- Future ML systems will have weird failure modes that don't manifest today, and we should start thinking about and addressing them in advance.
- On the other hand, I don't think that Engineering as a tool for predicting the future is entirely self-defeating. Despite emergent behavior, empirical findings often generalize surprisingly far, at least if we're careful in interpreting them. Utilizing this fact will be crucial to making concrete research progress.

• • •

[1] From the IBM model authors: “In 1949 Warren Weaver suggested that the problem be attacked with statistical methods and ideas from information theory, an area which he, Claude Shannon, and others were developing at the time (Weaver 1949). Although researchers quickly abandoned this approach, advancing numerous theoretical objections, we believe that the true obstacles lay in the relative impotence of the available computers and the dearth of machine-readable text from which to gather the statistics vital to such an attack. Today, computers are five orders of magnitude faster

than they were in 1950 and have hundreds of millions of bytes of storage. Large, machine-readable corpora are readily available.” ↩

[2] This is in contrast to using Engineering to *build capable and impressive systems* today. If anything, recent developments have strongly solidified Engineering’s dominance for this task. ↩

[3] This list is probably subject to selection bias and recency effects, although I predict that my point would still hold up for a carefully curated list (for instance, I didn’t include the several ambiguous examples in my count). I would be happy to bet on more phase transitions in the future if any readers wish to take the other side. ↩

Jacob Steinhardt ▾

1 Comments

Sign in to join the conversation.

Mordechai Rorvig 1 year ago

Good post. I'll be interested to read the next ones. I particularly like the list of concrete studies or observations of sudden phase shift behavior. I wonder though if these phase shifts aren't still ultimately limited and defined by the architecture. I'm not sure that we have architectures that can phase shift as far as we might want, or as far as we're afraid they would go.

♡ 0

Powered by Cove

← **Previous Post**

Next Post →