The Scaling Era: An Oral History of AI, 2019–2025

Dwarkesh Patel with Gavin Leech

Chapter 1
Scaling

About *The Scaling Era*

How did we build large language models? How do they think, if they think? What will the world look like if we have billions of AIs that are as smart as humans, or even smarter? In a series of in-depth interviews with leading AI researchers and company founders—including Anthropic CEO Dario Amodei, DeepMind cofounder Demis Hassabis, OpenAI cofounder Ilya Sutskever, MIRI cofounder Eliezer Yudkowsky, and Meta CEO Mark Zuckerberg—Dwarkesh Patel provides the first comprehensive and contemporary portrait of the technology that is transforming our world. Drawn from his interviews on the Dwarkesh Podcast, these curated excerpts range from the technical details of how LLMs work to the possibility of an AI takeover or explosive economic growth. It also includes 170+ definitions and visualizations, classic essays on the theme, and previously unpublished interviews. The Scaling Era offers readers unprecedented insight into a transformative moment in the AI's development—and a vision of what comes next.

Stripe Press
Ideas for progress
South San Francisco, California
press.stripe.com

Primer

If you're new to machine learning, here are some key terms you'll need to know and how they fit together.

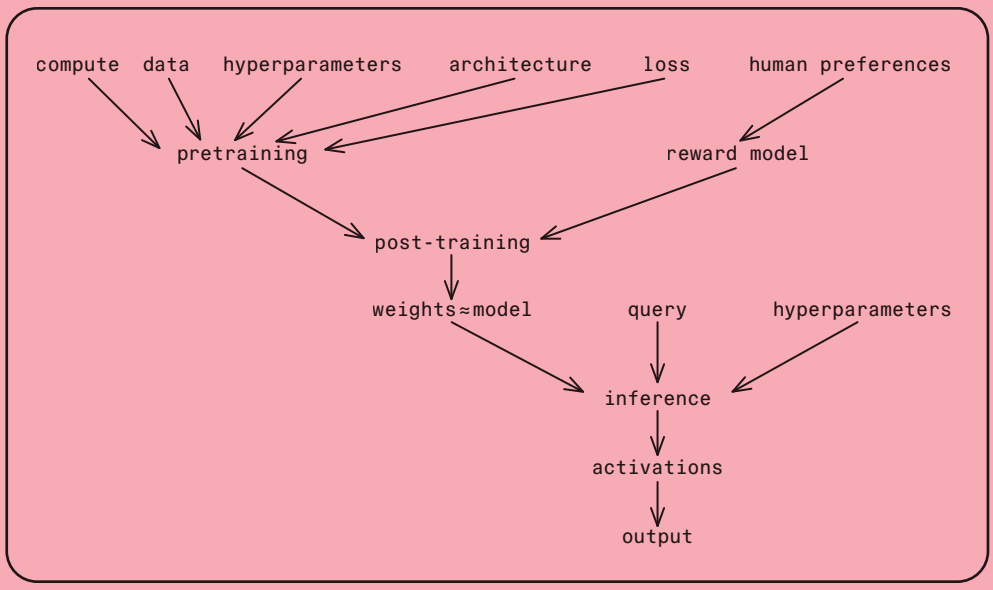

Figure 1. A bird's-eye view of machine learning.

PARAMETER
A variable that helps define a system or a transformation applied to input data; a dimension in model space. In machine learning, a numerical value that is adjusted iteratively during model training to encode patterns learned from the data.

(ARTIFICIAL) NEURAL NETWORK
A type of computer separated into three parts: the input layer, where data enters; the hidden layers, where most computation occurs; and the output layer, where predictions are made. Each layer contains many units (10,000, for example), interconnected by many weights. Unlike traditional computers, neural networks can learn programs by automatically adjusting these weights. The concept dates back to the 1940s, and was rebranded in the 21st century as deep learning.

WEIGHT
A parameter that defines the strength of the connection between two units in a neural network; where the algorithms performed on inputs to produce outputs are defined. Metaphorically, weights are like the synapses in the brain.

ACTIVATION
The value a model produces when processing a specific query, which depends on the weights it has learned during training and the inputs provided by the user; what gets input into the next layer of neurons in the model. Metaphorically, activations are like the electrical and neurotransmitter activity in the brain, or the model's active thoughts, associations, and goals.

LEARNING
The process of adjusting weights in a model after it processes data, enabling improved predictions based on past performance.

ARCHITECTURE
The structure of a model, including how its components connect to one another and how it is trained.

MODEL
The AI system produced by training an architecture on data. A program that has learned to perform specific tasks.

TRANSFORMER
A modern neural network architecture notable for its parallel design and ability to learn context and relationships using a mechanism called self-attention. This attention mechanism dynamically assigns varying importance to different parts of the input data.

LARGE LANGUAGE MODEL (LLM)
A neural network trained on text data to produce a probabilistic model of language. The term has become a misnomer in recent years, as LLMs are now also trained on audio, images, and other modalities, such as amino-acid sequences. The leading LLMs are based on the Transformer architecture.

TOKEN
The basic unit of data in an LLM, typically representing roughly one word. However, Transformers can be trained to emit more than just text tokens. Models can also output actions (such as searching the web) and pixels (as in image generators), among many other data types.

PROMPT
The input provided by the user, typically a query or instruction that the model responds to.

PRETRAINING
The process of creating an initial LLM, setting the values of its weights. During pretraining, the model is exposed to vast amounts of data and learns through trial and error by predicting the next token in a document.

POST-TRAINING
The process of adapting the pretrained model to be more of an assistant (instruction tuning), or to make it more professional, to make it less toxic, or to satisfy some other criteria (reinforcement learning from human preferences) by training further on data from chat sessions or rankings.

LOSS
A measure of how far a prediction is from the truth. In LLMs, "loss" is typically shorthand for the average autoregressive loss: the average error the model makes when predicting the next word in previously unseen documents.

HYPERPARAMETER
A parameter that governs how a model is trained or operates. It's "hyper" because it governs the parameters (weights) of the model.

FLOATING-POINT NUMBER
A computer representation of a real number, for example 0.00000024361.

FLOATING-POINT OPERATION (FLOP)
An arithmetic operation performed on a floating-point number. Updating a large model on a single data point might require billions of FLOPs. This measurement is often confused with FLOP/S, which measures the rate of floating-point operations per second.

BASE LLM
A model trained on a vast corpus of human text (as well as audio and images) in a semi-supervised manner by predicting the next word in a document and being updated in proportion to the magnitude of its error in predicting the next token. Technically, this refers to a pretrained, decoder-only Transformer.

INSTRUCTION-TUNED LLM
A base LLM that has been fine-tuned on examples of chat sessions so that it can respond in dialogue form as an assistant.

RLHF'D LLM
An instruction-tuned LLM that has been further refined using reinforcement learning from human feedback (RLHF). This process involves optimizing the model based on a learned representation of human preferences to reduce harmful, offensive, commercially sensitive, or inhuman responses.

SCAFFOLDED LLM
An RLHF'd LLM equipped with tools like chain-of-thought prompting, web search, vector databases, symbolic solvers, code interpreters, episodic memory, and search and self-criticism over possible responses. Also known as an augmented language model. Most systems available for public use are scaffolded LLMs.

LLM AGENT
A system that can solve open-ended or long-horizon tasks that require planning and executing sequences of actions and perceptions. A simple example is placing an LLM inside a prompt loop that continues until the task is completed. An LLM agent is sometimes also referred to as a scaffolded LLM, confusingly.

ARTIFICIAL GENERAL INTELLIGENCE (AGI)
An AI system capable of performing any task a human can perform, any task a group of humans can perform, or any task the average human can perform. Example tasks are boundless, but imagine an AGI and its copies performing every role in a large corporation, including strategy, design, management, production, and distribution; performing Nobel-level scientific research, including the experiments and breakthrough mathematical insights; or executing a coup on a major world government. The term "AGI" is sometimes used to refer specifically to human-level AI, while "ASI" (artificial superintelligence) denotes AI systems that surpass human-level capabilities.

SCALING
Massively increasing a model architecture's size (measured in parameters), the optimization used to train it (measured in FLOPs), the data used for training it (measured in bytes), or the computation required for each query (measured in tokens).

THE SCALING HYPOTHESIS
The idea that increasing the size, training data, and computational inputs of LLMs will be sufficient to achieve AGI.

Chapter 1
Scaling

Why is bigger better?

> As the rising flood reaches more populated heights, machines will begin to do well in areas a greater number can appreciate... When the highest peaks are covered, there will be machines that can interact as intelligently as any human on any subject. The presence of minds in machines will then become self-evident.
> —Hans Moravec, 1997•

When OpenAI released GPT-2 in 2019, it was barely discussed outside of AI circles. Three years later, GPT-3.5 took the world by storm, with perhaps the fastest recorded user growth of any software in history.••

There are a few reasons for this, not least the friendlier user interface of ChatGPT. But a key reason is that GPT-3.5 was much smarter than its predecessors. The main reason it got smarter is scaling: The researchers used roughly the same design as GPT-2 but created a much larger version trained on much more data.•••

It's hard to overstate the magnitudes. The compute needed to train a leading model is now 10 billion times higher than it was in 2010.[19] If the compute used for a 2010 AI model was the size of a laptop, the compute used for Google DeepMind's Gemini Ultra, released in 2023, would be the size of New York City.•••• In this period, the compute used to train each frontier model doubled every six months—four times faster than Moore's law predicts.[20]

OpenAI took a $4 million risk in training GPT-3 because it had stumbled upon so-called neural scaling laws: curves that predict how much models will improve as we increase the resources used to create them.[21] So far, these laws

• While Moravec's predictions were incredibly prescient, he did not fully account for the need for huge amounts of training data and computational resources. Without the internet, there would have been no massive free dataset; without training data, there would be no LLMs.

•• Reportedly, ChatGPT had more than 100 million monthly active users within two months of launching. By way of comparison, it took TikTok nine months to achieve this milestone. Milmo, "ChatGPT Reaches 100 Million Users."

••• In a 2024 paper, researchers at Epoch AI estimated the contribution from scaling data and compute compared to that of improving the training algorithm. Their findings suggest that two-thirds of the gains come from increasing data and compute. Ho et al., "Algorithmic Progress."

•••• A laptop is approximately 0.09 m^2, so 10 billion laptops tiling the plane would cover 900 km^2. For comparison, New York City's area is 778.2 km^2.

have accurately predicted the improvement resulting from exponential increases in investment.• This has inspired confidence in $100 million training runs, which would otherwise have been perceived as wildly unreasonable business risks. The scaling hypothesis is the idea that this resource-intensive strategy is all it will take to build a human-level AI system, and possibly systems that surpass human-level intelligence.••

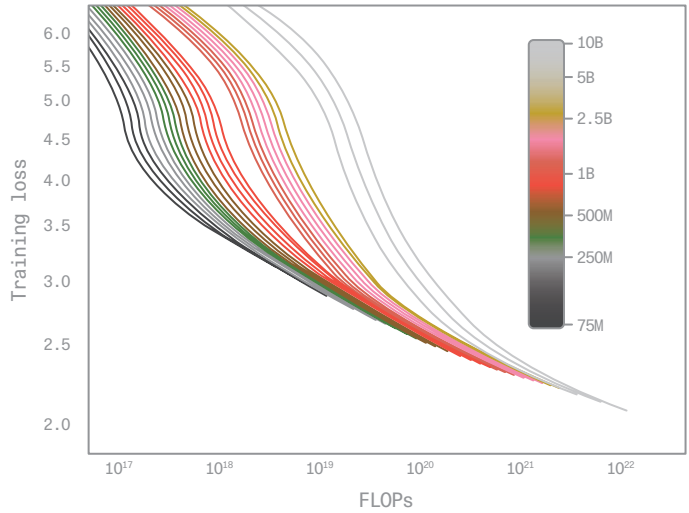

Figure 2. An example of a scaling law. The model improves (that is, its loss decreases) smoothly over a 100,000x increase in training compute (measured in FLOPs), and this smooth curve remains consistent across a wide range of model sizes. Each colored line represents a single model training run; the heat map separately encodes the number of parameters in the trained model.[22] Hoffman, "Training Compute-Optimal Large Language Models."

• Though the predictions are vague: "It will get better by this much on one general metric (the loss)," not "It will be able to prove novel theorems or do this particular thing." In July 2024, on the *In Good Company* podcast, Anthropic cofounder and CEO Dario Amodei said, "Right now, [it costs] $100 million [to train a model]. There are models in training today that are more like $1 billion. I think [we will get to] $10 or $100 billion… in 2025, 2026, maybe 2027." In the Appendix, you'll find the blogger Nostalgebraist's account of this predictability revolution in AI progress.

•• Note, however, that the term "scaling hypothesis" is used inconsistently. The original meaning focused on increasing the number of (dense) parameters. Later, the meaning shifted to refer to increasing training compute, which reflected the amount of human training data used, since data is only used once in a training run. As of this writing, the active area being scaled is the combined compute used in generating synthetic data, pretraining, post-training, and inference. Amodei and Hernandez, "AI and Compute"; Branwen, "Scaling Hypothesis."

More recently, a second form of scaling has emerged: inference scaling.• This involves increasing the compute used to answer each question by training the model to think longer (by using more tokens in its response) or by applying explicit algorithms on top of an LLM to explore multiple paths.•• This strategy yields significant improvements on tasks that require chains of reasoning. Because inference is bottlenecked differently than training,••• and because these methods might generate highly useful reasoning trace training data, this approach might well drive further AI progress.

TOKEN
The basic unit of data in an LLM, typically representing roughly one word. However, Transformers can be trained to emit more than just text tokens. Models can also output actions (such as searching the web) and pixels (as in image generators), among many other data types.

REASONING TRACE
The text output of a full step-by-step reasoning process. These outputs enable process supervision, a training method that gives the model feedback multiple times per response.

But inference scaling is just an elaboration of the general principle: So far, the (exponentially) more compute and data you put in, the more intelligence you get out. This effect is so clear and so important that I call the period since 2016 the scaling era of AI.••••

In this chapter, we hear from some pioneers of scaling about why it works, discuss the evolutionary neuroscience of human and artificial intelligence, and speculate about whether scaling will continue to create increasingly impressive systems.

I. DWARKESH PATEL

Fundamentally, what is the explanation for why scaling works? Why is the universe organized such that if you throw big blobs of compute at a wide enough distribution of data, the thing becomes intelligent?

• Also known as test-time compute scaling.

•• There's a one-to-one ratio between outputting and processing. The more tokens output, the more time the model has to think about a given query. The simplest approach to letting an LLM search is majority voting: Have it generate many completions per question (hundreds, for example) and take the most common answer. See Lewkowycz, "Solving Quantitative Reasoning Problems."

••• For example, inference requires much less RAM, and therefore fewer GPUs, than training. It also doesn't face the same data wall (the shortage of new training data) as training.

•••• We usually think of OpenAI's billion-parameter GPT-2 model from 2019 as the beginning of the scaling era, but it wasn't the first big model. In 2007, researchers trained a limited type of language model with 300 billion parameters—roughly the same size as GPT-3. And the winners of the 2008 Netflix Prize trained 10 billion-parameter models. Researchers at Epoch AI used a simple regression to date the year scaling really got going and concluded that "a separate trend of models breaks off the main trend between 2015 and 2016." Brants et al., "Large Language Models in Machine Translation"; Bell et al., "BellKor 2008 Solution"; Sevilla et al., "Compute Trends."

FEATURE
A variable used by a model to make predictions or decisions; a dimension in the space the model thinks in. For example, when classifying the species of a flower, a useful feature is the width of its petals. Traditionally, a developer had to do feature engineering: handing the model features relevant to the task. Instead, deep learning models learn features, developing their own representations of the important parts of the training data. In some cases, these features reflect recognizable concepts, like "straight line" or "malevolent AI."

PARAMETER
A variable that helps define a system or a transformation applied to input data; a dimension in model space. In machine learning, a numerical value that is adjusted iteratively during model training to encode patterns learned from the data.

LOSS
A measure of how far a prediction is from the truth. In LLMs, "loss" is typically shorthand for the average autoregressive loss: the average error the model makes when predicting the next word in previously unseen documents.

CIRCUIT
A collection of neurons in a model that form a stable pattern of activation in response to certain inputs, enabling the model to perform simple tasks like detecting straight lines in an image or determining whether one quantity is greater than another. Circuit-level interpretability would represent an understanding of every such circuit—a complete explanation for the LLM's behavior. It may not be possible.

DARIO AMODEI
CEO of Anthropic

The truth is that we still don't know. It's almost entirely just a [contingent] empirical fact. It's a fact that you could sense from the data, but we still don't have a satisfying explanation for it.

If I were to try to give one—and I'm just waving my hands when I say this—there are these ideas in physics around long-tail or power-law correlations or effects. When you have a bunch of features, you get a lot of [the total information] in the early part of the distribution, before the tails. For language, that would be big things like figuring out that there are parts of speech or that nouns follow verbs. Afterwards, you learn more and more subtle correlations.

It makes sense why every order of magnitude added captures more of the distribution. What's not clear at all is why it scales so smoothly with the number of model parameters, and why it scales so smoothly with the amount of data.

DWARKESH PATEL

By "scaling law," we're referring to the fact that when you go from Claude 1 to Claude 2, there's a smooth improvement in how well the model predicts the next token. We may not know why it's happening. But can you at least predict empirically, here is the loss at which this ability will emerge, here is the place where this circuit will emerge? Is that at all predictable, or are you just looking at the loss number?•

DARIO AMODEI

That is much less predictable. What's predictable is this statistical average, this loss, this entropy. It's sometimes predictable even to several significant figures, which you don't see outside of physics. You don't expect to see it in this messy empirical field.

Specific abilities are very hard to predict. Back when I was working on GPT-2 and GPT-3, we were asking, "When does arithmetic come into place? When do models learn to code?" Sometimes it's very abrupt. It's like how you can

• Here, I'm alluding to the difference between predicting the average autoregressive loss after pretraining—how well the LLM predicts the next token—and predicting the system's downstream task performance—how well it does on real tasks people want it to do, like writing code or doing homework—also known as its emergent capabilities. An up-to-date discussion is in Schaeffer et al., "Downstream Capabilities of Frontier AI Models."

predict statistical averages of the weather, but the weather on one particular day is very hard to predict.

One of the first things [OpenAI cofounder and former chief scientist] Ilya Sutskever said to me was, "Look. The models just want to learn. You have to understand this. The models just want to learn." It was a bit like a Zen koan. I listened to this and I became enlightened.

II.

DWARKESH PATEL

Scaling is the main way these models are getting better. Why does that work? Why is the universe this way?

JARED KAPLAN
Cofounder of Anthropic

I have a few hypotheses. But maybe first we should talk about what scaling is. Scaling is this relation we've noticed: As you make AI systems larger—increasing the number of parameters they have, training on more data, or increasing the total amount of compute used for training—you get really, really predictable trends for the performance of these systems as you scale up.

This holds true over many, many orders of magnitude. Although at first we were only looking at language models in the GPT-1, GPT-2 era, it also seems true of multimodal models and all sorts of other AI systems. This universality is really striking and really important. If you have a phenomenon that only occurs in some really niche situation, then maybe there's a niche explanation for what's going on there. Whereas scaling seems so universal that you'd expect there to be some kind of simple general explanation.

I can talk a little bit about theories we've developed to explain this. They're the kinds of theories physicists like—you just have to make the right assumptions and the result follows. But we don't really know all of the details of how neural networks work. We're still very confused. There's a lot left to understand, even to just validate some of our hypotheses.

DWARKESH PATEL

What is that simple explanation? I understand that the full theory might not be clear, but what's the general heuristic?

JARED KAPLAN

You can think of neural networks as mapping their data to some kind of data manifold that has some dimensionality. All neural networks are really doing, then, is basically fitting a curve to that data manifold.

MULTIMODAL MODEL
A model that can simultaneously process multiple data types (modalities), such as text, images, and audio. Figuratively, it's like having multiple senses and being able to correlate and reason about them.

ARTIFICIAL NEURAL NETWORK
A type of computer separated into three parts: the input layer, where data enters; the hidden layers, where most computation occurs; and the output layer, where predictions are made. Each layer contains many units (10,000, for example), interconnected by many weights. Unlike traditional computers, neural networks can learn programs by automatically adjusting these weights. The concept dates back to the 1940s, and was rebranded in the 21st century as deep learning.

DATA MANIFOLD
A structure representing all possible data points, often conceptualized as a surface with a complex shape in a high-dimensional space. Notably, "data" manifold is a misnomer, as the manifold has a far lower dimension than the original data. The word "surface" is also somewhat misleading, as it suggests three dimensions, whereas the manifold of large LLMs is estimated at more than 90 dimensions.

This is all very abstract, but probably everyone who's done a little bit of science has done an experiment, gotten a bunch of data points on x versus y, and fit some kind of curve to that. The idea is that maybe neural networks are doing something, abstractly, as simple as fitting some multi-dimensional curve. In general, what's the simplest way you can fit a curve? You can just chop up your x-axis or your data into little bins and then model each bin separately.

So if you make that assumption—which is a huge assumption; we don't really know where this data manifold lives or if it really exists—then you can argue that as

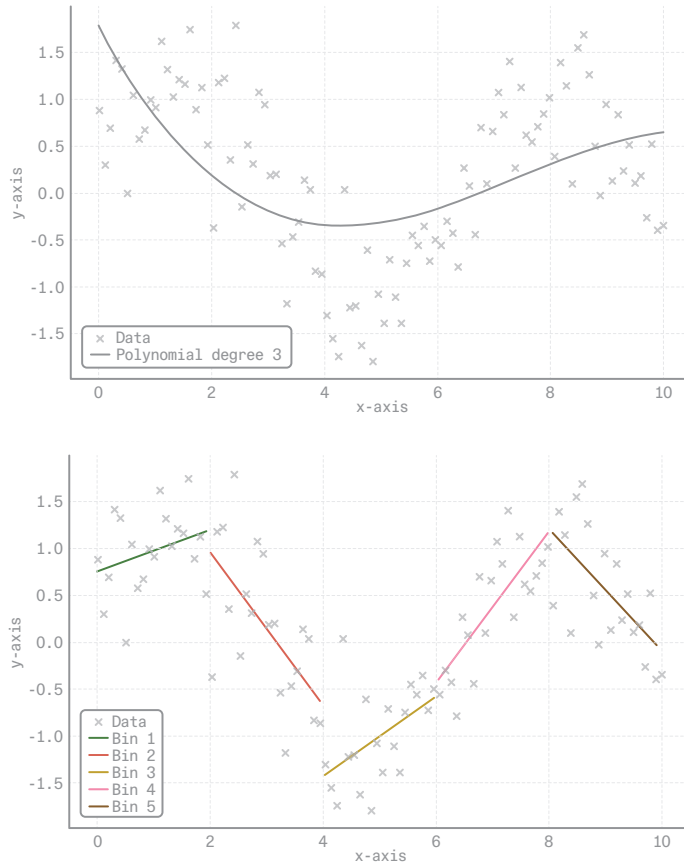

Figure 3. Two ways to fit the same data. On the top is a relatively sophisticated model: a polynomial of degree 3 (that is, it has four parameters). On the bottom is a piecewise linear model obtained by binning (splitting) the data into intervals of width 2 and fitting a simple line to each bin (using 10 parameters). As Kaplan notes, the linear models fit the data better. This data is just a sine wave with noise; the underlying generator has only three parameters. The optimal model for this particular data in terms of both accuracy and parameter efficiency is a sine wave, but the piecewise approach is applicable to data in general.

you scale up the number of parameters, all you're really doing is cutting up your data manifold into more and more high-resolution pieces. You can then ask, how will the error that you get scale as you chop it up into more and more pieces? Because we've hypothesized that the data manifold has some intrinsic dimensionality, you can run some numbers and you'll find that you get a power-law scaling in the number of parameters as you scale up.

INTRINSIC DIMENSION
The minimum number of parameters needed to represent the data as simply as possible, or to solve a given optimization problem

POWER LAW
A relationship between two variables, x and y, where y scales as a power of x ($y = x^k$) and the relationship stays the same at any scale. A simple example is the area of a square (area = length^2).

III. DWARKESH PATEL

How seriously do you take these scaling laws? There's a paper that says you need such-and-such many more orders of magnitude of training compute to get all the reasoning out.[23] Do you take that seriously, or do you think it breaks down at some point?

ILYA SUTSKEVER
Cofounder of Safe Superintelligence Inc.

The thing is, the scaling law tells you what happens to the *log* of your next-word prediction accuracy. There's a whole separate challenge of linking this next-word prediction accuracy to actual reasoning capability. I do believe there is a link, but it's complicated. We may find that there are other things that can give us more reasoning per unit of effort. I think reasoning tokens can be helpful.

REASONING TOKEN
1. Output: An LLM token that uses more test-time compute per query, enabling step-by-step reasoning through multiple chains of thought. The current central example is OpenAI's o1, which hides its reasoning tokens from users. This approach has its own scaling laws with different constraints.
2. Input and output: Special symbols that denote the role of the sentence within a broader argument. A simple example would be labeling text with "Premise:" and "Conclusion:." Hypothetically, more sophisticated tokens like "<Make a plan>" or "<Go back and check your work>" could improve LLMs' reasoning. Recent Anthropic models are reported to use such tokens.

IV. DWARKESH PATEL

If the current scale-up works, we're going to get to AGI really fast, like within the next 10 years. If the current scale-up doesn't work, we're left with the baseline—the economy growing at only 2 percent a year, so we have only 2 percent more resources a year to spend on AI. You're talking about decades, then, before you can train a $10 trillion model.•

Let's talk about your thesis that the current AI scale-up would work. What's the evidence from AI itself, or from the evolution of primates and other animals?

CARL SHULMAN
Independent adviser to Open Philanthropy

The best way to think about this might be: In the 2000s, before the deep learning revolution, how did I think about AGI timelines? How have I updated since then based on what has happened with deep learning?

AGI TIMELINE
A prediction about when AGI will arrive, often expressed as the first year where there is a 50 percent likelihood of AGI existing.

• OpenAI's Sam Altman recently estimated that it would take a total of $7 trillion to build the necessary new AI compute clusters. Carchidi, "Is OpenAI's Sam Altman's Future Worth $7 Trillion?"

Back then I would have said, we know the brain is an information-processing device. Human intelligence works. Intelligence is possible. Not only is it possible, it was created by evolution on Earth. That gives us something of an upper bound [on the size of the search necessary to produce intelligence], in that brute force [that is, evolutionary trial and error] was sufficient.

There are some complexities. What if it was a freak accident and it didn't happen on any of the other planets? I have a paper with [philosopher] Nick Bostrom about this.[24] Basically, it's not that important. There's convergent evolution. Octopi are also quite sophisticated. If a special event was required at the level of forming cells at all, or forming brains at all, we get to skip that because we already exist and we're choosing to build computers. We have that advantage. So evolution gives something of an upper bound. Really intensive, massive brute-force search and things like evolutionary algorithms can produce intelligence.

DWARKESH PATEL

Isn't the fact that octopi and other mammals got to the point of being pretty intelligent but not human-level intelligent evidence that there's a hard step between a cephalopod and a human?

CARL SHULMAN

It doesn't seem particularly compelling. One source of evidence is work by Suzana Herculano-Houzel, a neuroscientist who has dissolved the brains of many creatures to determine how many neurons are present. She's found a lot of interesting scaling laws. She has a paper discussing the human brain as a scaled-up primate brain.[25] Across a wide variety of animals, mammals in particular, there are certain characteristic changes in the number of neurons and the size of different brain regions as things scale up. There's a lot of structural similarity.

You can explain a lot of what is different about us with a brute-force story. You expend resources on having a bigger brain, keeping it in good order, and giving it time to learn. We have an unusually long childhood. We spend more compute by having a larger brain than other animals—more than three times as large as chimpanzees—and by having a longer childhood. We're spending more compute in a way that is analogous to having a bigger model and training it for longer.

With AI models, we see these large, consistent benefits from increasing compute spent in those ways. We see

qualitatively new capabilities showing up over and over again, particularly in the areas that AI skeptics call out. In my experience over the last 15 years, people call out things like, "Ah, but the AI can't do that, and it's because of a fundamental limitation." We've gone through a lot of them. There were Winograd schemas, catastrophic forgetting, quite a number of these, and they have repeatedly gone away through scaling.

Most creatures wind up with small brains because they can save that biological energy and that time to reproduce and so on. Humans seem to have wound up in a self-reinforcing niche where we greatly increase the returns to having large brains. Language and technology are the obvious candidates. You have humans around you who know a lot of things, and they can teach you. Compared to almost any other species, we have vastly more instruction from parents and society. You're getting way more from your brain per minute because you can learn a lot more useful skills. You can then provide the energy to feed that brain [through ingenuity], such as by hunting and gathering, and by having fire, which makes digestion easier.

Humans play a lot, and we keep playing as adults, which is very weird compared to other animals. We're more motivated to copy those around us than other primates. These motivational changes keep more of our attention and effort on learning, and that pays off more when you have a bigger brain and a longer lifespan in which to learn.

A mayfly or a mouse that tried to invest in a giant brain and a very long childhood would be quite likely to be killed by some predator or some disease before they were able to use it. That means you actually have exponentially increasing costs in a given niche. If I have a 50 percent chance of dying every few months as a little mammal or lizard, the cost of going from three months of learning and childhood development to 30 months is not 10x less benefit; it's a 1,024x reduction in the benefit I get from what I learn, because 99.9 percent of such animals will have been killed before that point.•

We're in a specific niche. We're large, long-lived animals

WINOGRAD SCHEMA
A type of grammatical puzzle that requires common-sense reasoning to solve. The task involves identifying the meaning of a pronoun in a sentence with multiple possible subjects. The canonical example, from Terry Winograd, is the following pair of contrasting sentences:

A: The city councilmen refused the demonstrators a permit because they feared violence.
B: The city councilmen refused the demonstrators a permit because they advocated violence.

In sentence A, "they" refers to the councilmen, while in B, it refers to the demonstrators. The Winograd Schema Challenge, a benchmark for this task, was declared solved in 2019 after an LLM achieved 90 percent accuracy. Kocijan et al., "Winograd Schema Challenge."

CATASTROPHIC FORGETTING
A phenomenon in which an AI's performance declines on learned tasks it has not interacted with for an extended period of time. As Shulman mentions, this limitation can be overcome.

• Shulman obtains this result as follows: If you have a 50 percent chance of dying in the next three months and a 50 percent chance of dying in the three months after that, your probability of surviving to six months (and, therefore, of reaping the benefits of the investment into intelligence) is calculated as 50% × 50% = 25%. Over 30 months, the probability is thus 0.510 or 1/1,024, so the expected value of investing in your own development is reduced accordingly.

with language and technology, so we can learn a lot from our groups. That means it pays off to expand our investment into intelligence.

DWARKESH PATEL

Other species also live in flocks or packs. They play with each other. Why isn't that a hill they could have climbed to human-level intelligence? If it's because of something like language or technology, humans were getting smarter before we got language.• Especially given how valuable it is and the fact that we've dominated the world as a result, there should be other species that had the beginnings of a cognitive revolution. You'd think there would be selective pressure for it.

CARL SHULMAN

Evolution doesn't have foresight. What gets more surviving offspring and grandchildren in *this* generation is the thing that becomes more common. Evolution doesn't think, "If you do this, then in a million years, you'll have a lot of descendants." It's about what survives and reproduces *now*.

In fact, on average, social animals do have larger brains. Part of that is probably due to the social applications of bigger brains: keeping track of which group members have helped you before so that you can reciprocate, or remembering who's dangerous within the group. So there's some correlation there, but it seems that it's enough to just invest a bit more [in intelligence], but not to the point where a mind can easily develop language and technology and pass it on.

You see bits of tool use in some other primates. They have an advantage compared to whales, who don't have hands, which rules out a bunch of ways brains can pay off. Primates use sticks to extract termites. Capuchin monkeys open clams by smashing them with a rock. What they don't have is the ability to sustain culture. Maybe a particular primate will discover one of these tactics and it'll be copied by their immediate group, but they're not holding onto the tactic that well. It's easy to forget things, easy to lose information. So they remained technologically stagnant for hundreds of thousands of years.

We can look at some comparable human situations. There's an old paper by [economist] Michael Kremer that talks about the technological growth in human societies

• See, for example, what we infer about our distant ancestors' ability to handle figurative grammars. Watson et al., "Nonadjacent Dependency Processing."

on different continents.[26] Eurasia is the largest integrated connected area. Africa is partly connected to it, but the Sahara desert restricts the flow of information and technology. Then you have the Americas, which, after colonization from the land bridge, were largely separated and are smaller than Eurasia. Then you have Australia, and then smaller islands like Tasmania. The paper finds that technological progress seems to have been faster with larger, connected groups of people. In the smallest groups, like in Tasmania, they actually lost technology, like some fishing techniques.

If you have fewer people, there's less innovation. Moreover, you can easily get an imbalance between the rate at which you lose technologies to local disturbances and the rate at which you create new technologies. The great change brought by hominids and humanity is that we wound up accumulating tech faster than we lost it. Accumulating those technologies allowed us to expand our population, which then reinforced all of this. The tech also created additional demand for intelligence, so our brains became three times as large as those of chimpanzees and our ancestors.

DWARKESH PATEL

The crucial point for AI is that the selective pressures against intelligence in other animals are not acting against neural networks. The model isn't going to get eaten by a predator if it spends too much time becoming more intelligent. Unlike evolution, we're explicitly training them to become more intelligent. So we have a good first-principles reason to think that if scaling made our minds this powerful, and if the things that prevented other animals from scaling don't impinge on AI, then AI should just continue to become very smart.

CARL SHULMAN

Yeah. We are also growing them in a technological culture, with jobs like software engineering, which depend much more on cognitive output and less on things like metabolic resources devoted to the immune system or big muscles to throw spears with.

V.

DWARKESH PATEL

[AI researcher] Richard Sutton's "Bitter Lesson" essay• says that there are two things you can scale: search and learning. LLMs are about the learning aspect. You've worked on search throughout your career, where you have an agent interacting with an environment.•• Is that the direction

SEARCH
An area of computer science focused on finding solutions that satisfy a given specification when no explicit algorithm is known. Many of AI's scientific successes, such as protein structure prediction and theorem proving, have resulted from using deep reinforcement learning (non-LLM neural networks) to solve complex search problems.

AGENT
An autonomous system that perceives and acts in pursuit of a goal; a system capable of working out what it needs to learn and do in order to achieve an objective.

that needs to be explored again? Or is that something that needs to be added to LLMs, so they can interact with their data or the world or in some way?

SHANE LEGG
Cofounder and chief AGI scientist at Google DeepMind

Yeah, that's on the right track. These foundation models are world models of a kind, and to do really creative problem solving, you need to start searching. Think about something like AlphaGo and the famous Move 37. Where did that come from? Did it come from data it had seen of human games? No. It came from the model identifying a move as being unlikely but plausible and then, via a process of search, coming to understand that it was actually a very good move.

WORLD MODEL
A low-dimensional, stable representation of reality that captures essential structures and relationships, as opposed to a complex web of millions of statistical associations.

ALPHAGO
DeepMind's most famous game-playing AI and the first computer system to surpass human-level performance at Go.

MOVE 37
An extremely surprising move the AlphaGo system played against a world-class human player. To observers, the move initially seemed like a bizarre error, but it was eventually recognized as part of an unprecedented strategy. Although they also involve neural networks, the Alpha systems come from a different lineage of AI than LLMs, namely reinforcement learning and tree search. Between 2010 and 2022, these lineages formed DeepMind's distinctive effort toward AGI.

To get real creativity, you need to search through spaces of possibilities and find these hidden gems. That's what creativity is. Current language models don't really do that. They're mimicking the data. They're mimicking all the human ingenuity they've seen from all these internet data, which are originally derived from humans.

These models can blend things. They can do Harry Potter in the style of Kanye West, even though that's never been done before. But a system that goes beyond that—generalizing in novel ways and doing something truly creative, not just blending existing things—requires searching through a space of possibilities for these hidden gems. That requires search. So I don't think we'll see systems that truly step beyond their training data until we have powerful search in the process.•••

- "The Bitter Lesson" is computer scientist Richard Sutton's very brief summary of 70 years of AI research, published in 2019. (It's included in the Appendix.) He writes that sophisticated methods using limited compute will always lose out to "[simple] methods that continue to scale with increased computation." To an AI scientist of the old guard, this lesson is bitter because it involves relatively little insight, theory, or human intervention. Instead, the improved performance comes from a sheer increase in resources. Halevy et al. made much the same point in 2009 in "The Unreasonable Effectiveness of Data."

•• Like AlphaGo's use of RL policies and tree search.

••• After this interview, we began to see a trend toward inference scaling (also known as test-time compute scaling), effectively allowing an LLM search over many possible responses.

VI. DWARKESH PATEL

How linked are longer context windows to the ability to do long-horizon tasks, ones that require you to engage with an assignment for many hours? Or is it unrelated?

SHOLTO DOUGLAS

Reinforcement learning infrastructure lead at Anthropic

I would take issue with the idea that context length is the reason that agents haven't taken off. I think that's more about nines of reliability and the model successfully doing composite things. If you can't chain tasks successively with high enough probability, then you won't get something that looks like an agent. GPT-4 or Gemini Ultra-class models aren't enough. But maybe the next increment on model scale means that you get that extra nine. Even though the loss isn't going down that dramatically, that small amount of extra ability gives you the extra reliability. Obviously, you need some amount of context to fit long-horizon tasks, but I don't think that's been the limiting factor up to now.

Over a couple of orders of magnitude, we've seen models go from being unable to do anything to being able to do huge amounts. It feels to me that each incremental order of magnitude gives more nines of reliability, which unlocks things like agents. But at least at the moment, it doesn't feel like reasoning improves linearly but rather somewhat sublinearly.

CONTEXT WINDOW
The space within a model for usable information, measured in tokens, during a single pass. The context window includes the developer's prompt, the user's input, the model's output, and the resulting conversation history. Figuratively, it's like the model's working memory. Modern context windows can now be book-length: 100,000 tokens or more.

NINES OF RELIABILITY
A measure of reliability expressed as the number of nines in an uptime percentage. For example, three nines represents 99.9 percent reliability; six nines indicates 99.9999 percent reliability. The intuition here is that if a long-horizon task consists of 10 subtasks, having one nine of reliability at each subtask results in a 34 percent success rate at the overall task (0.9^{10})—effectively useless. Having two nines results in a 90 percent success rate. So small improvements to the model could have large effects on its ability to perform complex tasks.

DWARKESH PATEL

A friend made the point that if you look at new applications unlocked by GPT-4 relative to what GPT-3.5 unlocked, it's not clear that it's that much more impressive. GPT-3.5 could run Perplexity, or whatever. So if there's a diminishing increase in capabilities that cost exponentially more, that's actually a bearish sign of what GPT-5 will unlock in terms of economic impact.

SHOLTO DOUGLAS

For me, the jump between 3.5 and 4 is pretty huge, so another jump of that size is ridiculously good, if GPT-5 is a 3.5-to-4-sized jump in terms of its ability to do SATs and this kind of stuff. It doesn't feel like we're going to jump to utter genius in the next generation of models. But it does feel like we'll get to very smart models, plus lots of reliability. It's unclear what that looks like.

I don't want people to come away thinking that models aren't going to get much better. The jumps we've seen so far are huge. Even if those continue on a smaller scale, we're still

SAMPLE EFFICIENCY
A measure of how much the model's performance improves per training example. Here, we're talking about the model learning tasks it wasn't necessarily pretrained to perform. In-context learning is much more sample efficient—the model can learn to perform complicated tasks like linear regression from just a handful of examples.

in for extremely smart, very reliable agents over the next couple of orders of magnitude. We have a lot more jumps coming. Even if those jumps are smaller, relatively speaking, that's still a pretty stark improvement in capability.

TRENTON BRICKEN
Interpretability researcher at Anthropic

Not only that, but if you believe the claims that GPT-4 has around 1 trillion parameters... The human brain has 30 to 300 trillion synapses. It's obviously not a 1-to-1 mapping between machine parameters and animal synapses, and we can debate these numbers, but it seems pretty plausible that we're still below the scale of the human brain.

DWARKESH PATEL

Crucially, the counterpoint is that the algorithmic overhead is really high. Even if you can't keep dumping more compute beyond models that cost $1 trillion, the fact that the brain is so much more data-efficient implies that if we have the compute, and if we have the brain's algorithm to train, and if we could train [a model] as sample efficient as humans are from birth, then we could make AGI.

TRENTON BRICKEN

I never know exactly how to think about the sample efficiency stuff, because a lot of things are hardwired in certain ways in humans, like the coevolution of language and the brain's structure. So it's hard to say. There are also some results that indicate that if you make your model bigger, it becomes more sample efficient.[27]

A STRATEGIC COMPUTE OVERHANG is the situation where sufficient resources are available to run multiple instances of a powerful AI as soon as one is trained. This is one source of AI takeover risk, since the sudden availability of many AI instances is key to many plausible takeover scenarios, and because acquiring existing compute is a relatively fast process, meaning that once one AI is capable of exfiltrating itself, it could rapidly proliferate.

A TACTICAL COMPUTE OVERHANG is when significant algorithmic advancements suddenly enable us to train AGI on a much smaller budget than previous training runs required.

The TEST-TIME COMPUTE OVERHANG is Aschenbrenner's term for the idea that allowing models to think longer (that is, expend more compute answering a given query) could significantly improve their performance without much further training. The simplest version of this is best-of-n sampling: just query the model n times and take the best answer. A 2024 paper found that in small models, more sophisticated allocation of test-time compute improves performance as much as increasing the models' parameters fourteenfold. The OpenAI o1 family of models operates along these lines, but it likely required substantial retraining with reinforcement learning to produce its own iterative chain-of-thought skills.

CHAIN-OF-THOUGHT PROMPTING
A prompting technique that improves a model's ability to reason by making it think step by step (that is, generate intermediate reasoning steps). This simple and cheap change expands the class of problems a trained model can handle.

VII. LEOPOLD ASCHENBRENNER
Cofounder of Situational Awareness LP

A key question for AI progress in the next few years is how hard it is to unlock the test-time compute overhang. Right now, GPT-4 can do a few hundred tokens of chain-of-thought prompting. That's already a huge improvement. Before, answering a math question was shotgun—and if you tried to answer a math question by saying the first thing that came to mind, you wouldn't be very good. GPT-4 instead thinks for a few hundred tokens. It's equivalent to me thinking for three minutes.

Now suppose GPT-4 could think for millions of tokens. That's [10,000x] more test-time compute spent on one problem. It can't do it now. It writes some code and can do a little bit of iterative debugging, but it eventually gets stuck and can't correct its errors. There's a big overhang.

In another area of ML, there's a great paper on AlphaGo that shows you can trade off train-time and test-time compute.[28] If you use four orders of magnitude (OOM) more test-time compute, that's almost like a 3.5x OOM bigger model. A few million tokens might be a few months of human working time. There's a lot more you can do in a few months of working time than just getting an answer *right now*. How hard is it to unlock that?

The reason it might not be that hard is that there are only a few extra tokens to learn to use. You need to learn things like error-correction tokens: "Ah, I made a mistake, let me think about that again." You need to learn planning tokens: "I'm going to start by making a plan. I'm going to write a draft, and now I'm going to critique my draft and think about it." These aren't things the models can do now,• but the question is, how hard is it to get there?

There are two paths to agents. When Sholto Douglas was on your podcast, he talked about scaling leading to more nines of reliability. That's one path. The other path is the unhobbling path. The model needs to learn this System 2 process I described earlier. If it can learn that, it can use millions of tokens per query and think coherently.

Here's an analogy. When you drive, you're on autopilot most of the time. Sometimes you hit a construction zone or an intersection. Sometimes my girlfriend is in the passenger seat and I'm like, "Ah, be quiet for a moment, I need to figure out what's going on." You go from autopilot to System 2. Scaling improves that System 1 autopilot. The brute-force way to get to agents is improving that system. But if, instead, you can get a System 2 working, you can quickly jump to something more agentified, and test-time compute overhang is unlocked.

DWARKESH PATEL

Is there some loss function that easily enables System 2 thinking? There aren't many animals with System 2 thinking. It took a long time for evolution to give it to us. Pretraining uses trillions of tokens of internet text and gets you all of these capabilities, but not much of a System 2. What's the reason to think this will be an easy unhobbling?

UNHOBBLING
A term coined by Aschenbrenner to describe techniques that make an LLM more consistent, autonomous, and strategic. These include chain-of-thought prompting, RLHF, and the use of scaffolds like calculators and search engines—essentially, any method other than scaling. Another word for it is schlep.

SYSTEM 2 THINKING
A mode of explicit, effortful, and sequential reasoning, exemplified by activities like mathematical derivation. It contrasts with System 1 thinking, which is fast, automatic, and intuitive. The terms originate from Keith Stanovich and Richard West's theory of human reasoning. Instilling System 2 thinking in an LLM might be as simple as having it learn a new higher-order algorithm using its existing representations. A 2024 paper by Piantadosi et al. summarizing decades of evidence suggests that human concepts and reasoning are also vector-based, just like a neural network.

LOSS FUNCTION
A mathematical expression that specifies the quality of a prediction or decision, defining the training objective of an AI system. During pretraining, the loss function guides how LLMs are trained to produce more accurate predictions. Post-training is guided by different losses, such as human preferences or predicted human preferences.

• True at the time of the interview, but OpenAI's o1 model, released in September 2024, supports the general claim that LLMs will be able to search and perform chain-of-thought reasoning independently.

REINFORCEMENT LEARNING FROM HUMAN FEEDBACK (RLHF)
An LLM post-training technique that uses a proxy for human preferences to guide the model toward producing more human-like and socially desirable outputs. Originally developed as an AI alignment technique, RLHF has also been crucial to making LLMs more capable and commercially viable.

BOOTSTRAPPING
A form of self-supervised learning; training a language model on raw data without requiring a human to provide tags or answers. This increases the amount of available training data by a factor of millions. There is currently no full equivalent of this for robotics, although some exciting work uses a pretrained LLM to help a robot plan the necessary sequence of actions.

SELF-PLAY
A training method in which an AI system is trained on data generated by a copy of the system (also known as synthetic data). This approach offers two key advantages: it provides virtually unlimited training data at a lower cost, and it naturally scales in difficulty as the system improves, because the improved system can be swapped in as the data generator. So far, however, self-play has only been effective for so-called closed-world problems, such as games.

DATA WALL
(or DATA BOTTLENECK)
A looming challenge for training better LLMs posed by the need for more high-quality data. Since models like GPT-4 were likely trained on much of the material available on the internet, the low-hanging-fruit—trillions of tokens of free, human-generated content online—is sometimes thought to be exhausted. However, most of the people interviewed in this book disagree. Notably, Meta's Llama 3.1 model used some amount of synthetic data.

LEOPOLD ASCHENBRENNER

First of all, pretraining is magical. It gave us a huge advantage for models of general intelligence because you can predict the next token. However, there's a common misconception. Predicting the next token lets the model learn incredibly rich representations. Representation learning is the magic of deep learning. Rather than just learning statistical artifacts, the models learn models of the world. That's why they can generalize, because they learned the right representations.

When you pretrain a model, you get this raw bundle of capabilities. That's useful. The unhobbling from GPT-2 to GPT-4 took this raw mass and RLHF'd it into a good chatbot. That was a huge win. Look at the original InstructGPT paper.[29] When comparing RLHF versus non-RLHF models, RLHF is equivalent to increasing the model size 100 times in terms of the resulting increase in human evaluators' preference ratings. InstructGPT also started to do simple chain of thought. You still have the advantage of all these raw capabilities, and there's still a huge amount you're not doing with them.

This pretraining advantage is also the difference between LLMs and robotics. People used to say the slow progress in robotics was a hardware problem. The hardware issue is getting solved, but you still don't have this huge advantage of bootstrapping with pretraining. You don't have all this unsupervised learning you can do. You have to start right away with RL self-play.

The question is why RL and unhobbling might work. Bootstrapping is an advantage. You [as a human] are not being pretrained anymore. You were pretrained in grade school and high school. At some point, you transition to being able to learn by yourself. You weren't able to do it in elementary school. High school is probably where it started. By college, if you're smart, you can teach yourself. Models are just starting to enter that regime.

This requires a little bit more scaling, and then you figure out what goes on top. It won't be trivial. A lot of deep learning seems obvious in retrospect. There's some obvious cluster of ideas. There are some ideas that seem a little dumb but work. There are a lot of details you have to get right. We're not going to get this next month. It'll take a while to figure out.

DWARKESH PATEL

For you, a while is, like, half a year.

LEOPOLD ASCHENBRENNER

Between six months and three years. But it's possible. It's also very related to the data wall issue.

Pretraining is kind of like the teacher lecturing to you. The words are flying by. You're just getting a little bit from it. That's not what you do when you learn by yourself. When you learn by yourself—say, you're reading a dense math textbook—you're not just skimming through it once. You read a page, think about it, have some internal monologue going on, and have a conversation with a study buddy. You try a practice problem and fail a bunch of times. At some point it clicks and you're like, "This made sense." Then you read a few more pages.

We've bootstrapped our way to just starting to be able to do that with models. The question is, can you use all this self-play, synthetic data, and RL to make that thing work? Right now, there's in-context learning, which is super sample efficient. Gemini just learns a language in context.[30] Pretraining, on the other hand, is not at all sample efficient.

SYNTHETIC DATA
Training examples generated by computer programs or AIs instead of humans. Use of synthetic data is standard practice in science, where it is called simulation. In AI, synthetic data has struggled to capture the tails—rare but crucial thoughts that humans can generate easily. One exception is when the ground truth is known or verifiable, as in the case of games and mathematics. However, labs have made progress on data synthesis, and training on some synthetic data is now helpful and standard. A reported 20 percent of the training data for the Hunyuan-Large model was synthetic.

What humans do is a kind of in-context learning. You try a practice problem, fail, and at some point you figure it out in a way that makes sense to you. That's the best possible data for you because it's the way *you* would have solved the problem, rather than reading how somebody else solved the problem, which doesn't initially click.

DWARKESH PATEL

Suppose this is the way things go, and we get these unhobblings...

LEOPOLD ASCHENBRENNER

And scaling. Scaling provides this baseline enormous force of improvement. GPT-2 was amazing [for its time]. It could string together plausible sentences, but it could barely do anything. It was kind of like a preschooler. GPT-4, on the other hand, could write code and do hard math, like a smart high schooler. This big jump in capability is explored in my essay series.[31] I count the orders of magnitude of compute and algorithmic progress.

IN-CONTEXT LEARNING (ICL)
The ability of a model to learn or improve on tasks using the instructions and examples provided in the prompt, without requiring any further gradient updates. Sufficiently large models can perform this type of dynamic learning within their activations. ICL is essentially a learning algorithm inside of the learning algorithm—meta-learning. The simplest version of this, where a user gives the model examples of the task, is called few-shot prompting.

Scaling alone, by 2027 or 2028, is going to do another preschool-to-high-school-sized jump on top of GPT-4. At a per-token level, the models will be incredibly smart. They'll gain more reliability. With unhobblings, they'll look less like chatbots and more like agents or drop-in remote workers.• That's when things really get going.

VIII. DWARKESH PATEL

How do you make sense of the fact that when you give LLMs a lot of data in any specific domain, they tend to get better in just that domain? Wouldn't we expect a general improvement across all of the different areas?

DEMIS HASSABIS
Cofounder and CEO of Google DeepMind

You do sometimes get surprising improvement in other domains. For example, when these large models improve at coding, that can actually improve their general reasoning.[32] There is evidence of some transfer, although we would like a lot more evidence of that. But that's how the human brain learns, too. If we experience and practice a lot of things, like chess, creative writing, et cetera, we also tend to specialize and get better at that specific thing, even though we're using general learning techniques and systems in order to get better in that domain.••

TRANSFER
The ability to apply acquired knowledge effectively in different contexts, particularly to solve real-world problems beyond the original learning environment.

DWARKESH PATEL

As somebody who's been in this field for a long time and seen different trends come and go, what do you think the strong version of the scaling hypothesis gets right? What does it get wrong?

DEMIS HASSABIS

This is an empirical question right now. It was pretty surprising to almost everyone, including the people who first worked on the hypothesis, how far we've gotten. The models clearly have some form of concepts and abstraction. Five years ago, I would have said that we needed an additional algorithmic breakthrough to get that—maybe one more like how the brain works. I think that's still true if we want *explicit* abstract concepts, neat concepts, but it seems that these systems can already implicitly learn them.

STRONG SCALING HYPOTHESIS
A current prevailing hypothesis in AI that holds that LLMs can achieve human-level intelligence with sufficient data and compute, with costs potentially in the range of trillions of dollars.

We've got to push scaling as hard as we can. It's an empirical question whether we will hit a brick wall. No one knows. In the meantime, we should also double down on innovation. You can think of half of our effort as having to do with scaling. The other half has to do with inventing the next architectures and algorithms that will be needed,

ARCHITECTURE
The structure of a model, including how its components connect to one another and how it is trained. As of this writing, leading model architectures are still designed by humans.

• That is, a substitute for a human performing a laptop job remotely from home.

•• Psychologists call improving in general by training in specific far transfer. It's the holy grail of education, in the sense that it is elusive. Niplav, "Transfer Learning in Humans."

knowing that larger and larger scaled models are coming down the line. My bet is that you need both.

I also think it's interesting and unexpected that these systems have some sort of grounding, even though they don't experience the world multimodally. I think we get some grounding through the RLHF feedback systems, because the human raters are, by definition, grounded, so their feedback is grounded too. Also, maybe language contains more grounding than we previously thought.

SYMBOL GROUNDING PROBLEM
A fundamental requirement and challenge for any general AI system: the ability to translate between sensory data and abstract representations (for example, between a set of written or spoken instructions and the corresponding objects and sequences of actions in the real world). To be effective, the system must ground symbols in the appropriate real-world objects or events. Hassabis is, in my view, correct that this problem has more or less been sidestepped. For example, OpenAI's 2021 CLIP system learned how to translate between images and text descriptions.

LABEL
In supervised learning, a label is the correct or desired answer, which is applied to each input in the training data. It is used to guide the model's learning process. Here, I'm referring to the notion that at some point humans won't be able to produce useful labels because we won't be able to understand the outputs of superhuman models.

DWARKESH PATEL

Two things might change that would make grounding more difficult. One is that as these models get smarter, they're going to be able to operate in domains where we just can't generate enough human labels because we're not smart enough. If the model makes a million-line pull request, for example, how do we tell it whether this is within the constraints of our morality and the end goal we wanted or not?

The other thing has to do with what you were saying about compute. So far, we've been doing next-token prediction, and in some sense, that's a guardrail. You have to talk as a human would talk, and maybe think as a human would think. Now, additional compute might be spent on reinforcement learning, which just *somehow* gets to the objective. We can't really trace how it got there. When you combine those two, how worried are you that the grounding goes away?

DEMIS HASSABIS

You have to have some grounding for a system to achieve goals in the real world. But these systems are becoming more multimodal [as of February 2024], ingesting things like video and audiovisual data as well as text data. The system correlates those things together. That is a form of proper grounding. So I do think our systems are going to start to understand the physics of the real world better.

IX. TRENTON BRICKEN

Machine learning research is just so empirical. This is honestly one reason why I think our solutions might end up looking more brain-like than otherwise. Even though we wouldn't want to admit it, the whole community is doing a kind of greedy evolutionary optimization over the landscape of possible AI architectures. It's no better than evolution.

X. DWARKESH PATEL

You're one of the only people outside of OpenAI who noticed the way AI was progressing in 2020, and you're maybe the only one who had a detailed model of scaling. What sort of process made you able to develop this model of what was happening with LLMs in your "Scaling Hypothesis" post? •

GWERN BRANWEN
Freelance writer and researcher

I was just a patient reader of everything, noting anomalies and then going back once in a while and checking again.

If I had to give an intellectual history, it would start in the mid-2000s, when I was reading [computer scientists Hans] Moravec and [Ray] Kurzweil. They were making the fundamental connectionist argument that getting enough computing power will result in a neural network that matches the human brain, and that until that computing power is available, trying to build AI is basically futile.

I found this argument very unlikely. It's very much a "build it and they will come" view of progress, which I did not think was correct. I thought it was ludicrous to suggest that simply because there's some supercomputer that matches the human brain in compute, that would summon the correct algorithm out of nonexistence. I thought, "You can't just buy a bunch of computers and expect to get an AI out. That's magical thinking." So, because I was super skeptical of the argument, I didn't pay too much attention to it.

But as part of my interest in transhumanism and AI risk, I was paying close attention to Shane Legg's blog posts, where he extrapolates the connectionist argument out with updated numbers, giving very precise predictions, like, "We're going to get the first general assistant around 2019, and then around 2025 we'll get agents and generalist capabilities, and by 2030 we should have AGI." I was, again, very skeptical. But along the way, [the semantic network] DanNet and [the image classification model] AlexNet came out—a very impressive success story of connectionism.[33] I thought, "Is it an isolated success story, or is it what Kurzweil and Moravec and Legg were predicting?" I started thinking that maybe scaling was not quite as stupid as I'd first thought.

It was this gradual trickle of drops hitting me as I went along. The dataset sizes kept getting bigger. The models kept getting bigger. The training runs crept up from using

CONNECTIONISM
A school of thought in cognitive science and AI that seeks to explain cognition in terms of neural networks. By the 2000s, the decades-long philosophical debate between the connectionists and the symbolists had largely subsided, just in time for deep learning to vindicate the connectionist argument. See, for example, Hans Moravec's "When Will Computer Hardware Match the Human Brain?" and Ray Kurzweil's *The Age of Spiritual Machines*.

• Excerpted in the Appendix.

one cheap consumer GPU to two, and then to training on eight. The system's abilities kept getting broader and broader. Every few weeks, every few months, another drop. Finally, I went, "Maybe intelligence really is just a lot of compute applied to a lot of data. Huh. If that was true, it would have a lot of implications."

So there was no real eureka moment. I was just continually watching this trend that no one else seemed to see, except a handful of people like Ilya Sutskever or [computer scientist] Jürgen Schmidhuber. I just paid attention, noticing that the world looked more like [the connectionists'] world than like my world, where algorithms are super important and you need deep insight.

Then GPT-1 comes out. I was like, "Wow, this unsupervised sentiment neuron is learning on its own. That's pretty amazing." And then GPT-2 came out, and I was like, "Holy shit!" I looked at the prompting and the summarization, like, "Do we live in their world? *Can* we just keep scaling Transformers?"

Then GPT-3 comes down—the crucial test. Going from GPT-2 to GPT-3 is one of the biggest scale-ups in all of neural network history. If scaling was bogus, then the GPT-3 paper would be super unimpressive. Whereas if scaling was true, you would automatically get much more impressive results than GPT-2, guaranteed. I opened up the second page [of the paper] and I saw the few-shot learning chart, and I'm like, "Holy shit, we live in the scaling world. Legg and Moravec and Kurzweil were right!"

And then I turned to Twitter, and everyone else was like, "This shows scaling doesn't work! Why is GPT-3 not state of the art at everything?"[34] I was so angry at them that I had to write all this up.[35]

DWARKESH PATEL

In 2020, AI was already a thing. People were writing best-selling books about it. But none of those books were about scaling. What were people failing to account for?

GWERN BRANWEN

For the most part, they were suffering from two issues. First, they had not paid attention to all of the scaling results before 2020. They had not appreciated the fact that AlphaZero was discovered in part by doing Bayesian optimization on the hyperparameters and noticing that you can get rid of more and more of the tree search and get a better model. That was a critical insight that could only have been gained by having so much compute that you could train many,

TRANSFORMER
A modern neural network architecture notable for its parallel design and ability to learn context and relationships using a mechanism called self-attention. This attention mechanism dynamically assigns varying importance to different parts of the input data.

ALPHAZERO
Another game-playing AI developed by DeepMind that superseded AlphaGo. Unlike its predecessor, its training was pure self-play, using no human data. The system was also able to learn multiple games.

BAYESIAN OPTIMIZATION
A method of searching a space that is expensive to sample, such as the space of possible hyperparameter settings for a training run. Roughly, this involves creating a second model to predict how good a given setting will be, using this proxy to decide which settings to try next, and updating the model based on how good the setting actually was.

HYPERPARAMETER
A parameter that governs how a model is trained or operates. It's "hyper" because it governs the parameters (weights) of the model.

MONTE CARLO TREE SEARCH
A method for identifying an appropriate sequence of actions by searching over an abstract decision tree. As an example, a game of chess can be represented as a branching tree of all possible sequences of moves. It is a powerful instance of symbolic AI, a rival of statistical machine learning that uses rules, logic, human representations, and explicit algorithms.

many versions and see the difference. And they simply did not know about the 2017 Baidu paper on scaling laws.[36] It should have been the paper of the year, but it didn't have any immediate impact. People were too busy discussing Transformers or AlphaZero.

Another issue is that they made the basic error I had made, thinking that algorithms are more important than compute. That's partly due to a systematic falsification of the actual origins of ideas in the research literature. Papers do not tell you where ideas come from; they just tell you a nice-sounding story about how something was discovered.[37] So even if you appreciate the role of trial and error and compute in your own experiments, you probably think, "I got lucky. Over in the next lab, they do things with the power of thought and deep insight." But it turns out that everywhere you go, compute, trial and error, and serendipity play enormous roles in how things actually happen.

Once you understand that, you understand why compute comes first. You can't do trial and error or serendipity at scale without it. You can write down all these beautiful ideas but you can't test them, or you can only test a few instances of it, so you typically find that it doesn't work, and you give up and do something else. Reading the old deep learning literature, you see all sorts of ideas that were completely correct but that no one could prove, like ResNets being first published way back in 1988 instead of 2015. The researchers didn't have the compute to train a version that would have worked.

RESNET
Residual neural network. An important precursor to the Transformer architecture that uses residual connections between units (also known as skip connections) to let information flow between nonconsecutive layers.

Why believe that scaling was not going to work? Because you didn't notice the results that were key in retrospect. Another was BigGAN scaling to 300 million images.[38] There are still people today who will tell you with a straight face that GANs cannot scale past millions of images. If you don't know [otherwise], you could easily think, "GANs are broken. [We need a better algorithm.]" But if you do know that, then you think to yourself, "How can algorithms be so important when all these different generative architectures work, as long as you have lots and lots of GPUs?"

That's the common ingredient: lots and lots of GPUs. That's probably the root cause of not seeing scaling as a coherent paradigm and always [underrating it]. Even in 2020, you would still have AI people saying, "We'll get AGI in 2050." You could still think, very reasonably, that we still need lots and lots more incredible algorithmic breakthroughs [before we get AGI].

XI. DWARKESH PATEL

Regarding your original point about LLMs needing episodic memory, you mentioned that these are problems that we can solve, not fundamental impediments. When you say that, do you think they will be solved through scaling, or do each of these require a specific fine-grained architectural solution?

SHANE LEGG

I think it'll be architectural, because current architectures don't really have what you need. They basically have a context window, which is very fluid, and they have weights, which [knowledge] gets baked into very slowly. To my mind, the model's activations are like working memory in your brain, and the weights are like the synapses in your cortex.

ACTIVATION
The value a model produces when processing a specific query, which depends on the weights it has learned during training and the inputs provided by the user; what gets input into the next layer of neurons in the model. Metaphorically, activations are like the electrical and neurotransmitter activity in the brain, or the model's active thoughts, associations, and goals.

Now, the brain separates these things out. It has a separate mechanism for rapidly learning specific information. That's a different type of optimization problem compared to slowly learning deep generalities.[39] There's a tension between the two. But you want both. You want to be able to hear someone's name and remember it the next day. You also want to be able to integrate information over a lifetime to see deeper patterns in the world.

These are quite different optimization targets, different processes, but a comprehensive system should be able to do both. So it's conceivable that you could build one system that does both. You can also see that because they're quite different things, it makes sense for them to be done differently. I think that's why the brain does it separately.

XII. DWARKESH PATEL

A big open question is whether reinforcement learning will allow these models to use self-play or synthetic data to get over data bottlenecks. It sounds like you're optimistic about this.

DEMIS HASSABIS

I'm very optimistic. First of all, there's still a lot more data that can be used, especially if one considers multimodal data. Obviously, society is adding more data to the internet all the time. There's a lot of scope for creating synthetic data. We're looking at that in different ways, partly through simulation—for example, using very realistic game environments to generate realistic data—but also through self-play. That's where systems interact with each other or converse with each other. It worked very well for us with AlphaGo and AlphaZero. We got the systems to play

against each other, learn from each other's mistakes, and build up a knowledge base that way. There are some good analogies for that. It's a little bit more complicated to build a general kind of world data.

DWARKESH PATEL

How do you get to the point where the synthetic data the models are outputting on self-play is not just more of what's already in their dataset but something they haven't seen before? Something that would actually improve their abilities.

DEMIS HASSABIS

There's a whole new science needed there. This is important for things like fairness and trying to remove bias from the system and making sure that the dataset is representative of the distribution you're trying to learn. We're still in the nascent stage of [optimal] data curation and data analysis and analyzing the holes in our data distribution. There are many tricks one can use, like overweighting or replaying certain parts of the data. Or, if you identify some gap in your dataset, that's where you put synthetic generation to work.

XIII.

DWARKESH PATEL

What should we make of the fact that these models require so much training and the entire corpus of internet data in order to become merely *sub*human? Should we be worried about how inefficient these models seem to be?

DARIO AMODEI

That's one of the remaining mysteries. One way you could phrase it is that the models are maybe two to three orders of magnitude [100x to 1,000x] smaller than the human brain, while at the same time being trained on three to four orders of magnitude [1,000x to 10,000x] more data. Compare the number of words a human sees as it's developing until age 18. I think it's in the hundreds of millions.[40] Whereas for the models, we're talking about trillions.

What explains this? The models are smaller than brains, so they need a lot more data.• Or perhaps the analogy to the brain is not quite right or is breaking down. There's some missing factor. This is just like in physics, when we

• A lesser-known scaling law is that larger models are actually more sample efficient—they learn more from each data point than smaller models. See Kaplan et al., "Scaling Laws."

couldn't explain the Michelson-Morley experiment• or one of the other 19th-century physics paradoxes. It's something we don't quite understand. Humans see so little data and they still do fine.

One theory could be that it's our other modalities that do it. How do we get 10^{14} bits into the human brain? Maybe most of it is images.•• Maybe a lot of what's going on inside the human brain is that our mental workspace involves simulated images, or something like that.

Honestly, we have to admit it's weird. It doesn't match up. This is one reason I'm a bit skeptical of biological analogies. I thought in those terms five or six years ago. Now that we have these models in front of us, it feels like the evidence from these analogies has been screened off by what we've actually seen. What we've seen are models that are much smaller than the human brain, and yet they can do a lot of the things that humans can do. And yet, paradoxically, they require a lot more data to do those things.

Maybe we'll discover something that makes it all efficient. Maybe we'll understand why the discrepancy is present. At the end of the day, I don't think it matters if we keep scaling the way we are. What's more relevant at this point is just measuring the abilities of the model and seeing how far they are from humans' abilities. They don't seem terribly far to me.

• A crucial 1887 experiment that disproved the aether theory of light propagation, paving the way for the discovery of special relativity.

•• Here's a rough comparison: The eyes receive around 10 million bits of information per second. An ordinary person can read five words per second, which, in English, is about 50 bits per second—thousands of times slower. Spoken language tends to be about 39 bits per second. The skin is another high-bandwidth channel, processing perhaps 1 million bits per second, though it still doesn't approach the optic information rate. One caveat is that most optic information doesn't reach the brain. Markowsky, "Information Theory"; Coupé et al., "Different Languages, Similar Encoding Efficiency"; Koch et al., "How Much the Eye Tells the Brain."

About the authors

Dwarkesh Patel is the host of *Dwarkesh Podcast*, where he produces deeply researched interviews with both obscure intellectuals and the most influential figures of our time.

Gavin Leech is the cofounder of Arb Research, a consultancy that does empirical and conceptual work in various sciences. He is miscellaneous at gleech.org.

Stripe Press
Ideas for progress
South San Francisco, California
press.stripe.com