Is the Scaling Hypothesis Falsifiable?

Bruce Rushing and Javier Gomez-Lavin

June 24, 2024

Abstract

The scaling hypothesis in artificial intelligence claims that a model's cognitive ability scales with increased compute. This hypothesis has two interpretations: a weak version where model error rates decrease as a power law function of compute, and a strong version where as error rates decrease new cognitive abilities unexpectedly emerge. We argue that the first is falsifiable but the second is not because it fails to make exact predictions about which abilities emerge and when. This points to the difficulty of measuring cognitive abilities in algorithms since we lack good ecologically valid measurements of those abilities.

Word Count: 4421

1 Introduction

The scaling hypothesis in artificial intelligence (AI) predicts that cognitive ability scales as computational resources increase for training general learning algorithms. In contrast, critics argue that cognitive ability depends more on algorithmic design than brute-force

learning. Proponents respond that larger models reliably reduce hold-out test error, which yields breakthroughs in vision, audio, and language processing. Skeptics retort that these breakthroughs are illusory—the result of flawed testing methodology and that scale cannot overcome the sins of "dumb" learning techniques.

We argue that distinguishing between the weak and strong interpretations of the scaling hypothesis can help resolve this debate. The weak version holds that test error follows power law scaling, while the strong version posits that new cognitive capabilities emerge as test error decreases. We find that the weak interpretation is falsifiable yet uninteresting, although the strong one is more important but harder to evaluate empirically–and not falsifiable in its current guise.

Importantly, the debate over the scaling hypothesis revolves around a tricky theoretical proposition—that general learning algorithms make better predictions because they learn physical simulations of the data-generating process—and the twin problems of predicting when cognitive abilities emerge and how to measure them. Proponents of scaling emphasize the unpredictability of new cognitive abilities as models improve, but this is an empty check that yields no empirical cash; predictions should be made about *exact cognitive abilities* that will emerge in advance of scaling. Finally, this last point of exactness is a hard problem because we lack good theories about the ecology and developmental trajectory of AI intelligence, *pace* our knowledge of human analogues. We simply do not know what are the targets of AI cognition to be able to construct tests whose measurements are correlated enough for us to know we are measuring some underlying AI psychological faculty.

Here is how our argument proceeds. First, we introduce the scaling hypothesis and discuss the close connection scientists have posited between lower test loss and model cognitive ability. Second, we split the scaling hypothesis into the weak and strong versions,

and we argue that the weaker version easily can be falsified while the stronger version cannot given the current problems with testing models. Third, we discuss the issue of ecological validity of tests for AI cognitive ability.

2 The Scaling Hypothesis

The scaling hypothesis originally arose from two observations. First, there is what Sutton calls the bitter lesson: applying large amounts of compute has been a more successful method of solving tasks in domain after domain rather than leveraging human knowledge to design algorithms specific to these tasks (Sutton (2019)). Sutton catalogs how seventy years of AI research show that special algorithms that utilize human knowledge may work in the short-term, but they are ultimately outclassed by more general learning and search methods that utilize the increasing compute provided by Moore's law. Chess, Go, speech recognition, and computer vision all show this same pattern. Second, deep learning algorithms, which utilize relatively simple learning methods such as gradient descent, have consumed the field of machine learning and artificial intelligence only because of the advent of large datasets and specialized compute hardware; the fundamental algorithms and methods existed in the 1980s but grew effective in the 2010s thanks to bigger datasets and the better parallel processing of graphics processing units (Goodfellow et al. (2016), 17). The general experience for machine learning engineers and scientists who use these algorithms is that the quantity of compute has a quality of its own—often architecture and hyperparameter choice are irrelevant for effectiveness so long as the proper compute budget is used. Together, these two observations have led to the popular thesis that intelligence is easy: fix the right learning and search algorithms and just apply more data

and more parameters to get better results.

"Better results", however, is a vague claim. To be more precise, advocates for scaling often mean two distinct claims. The first claim is a technical one about how machine learning models are judged and an observed regularity found across the field of deep learning. The second claim is a more practical one about what those models can do in the world and the behavioral and psychological theories applied to deep learning models. We discuss each in turn.

2.1 The Varieties of Scaling

Prior to diving into our interpretations of the scaling hypothesis in machine learning, it's worthwhile to disambiguate our claims from other mentions of scaling that exist in related fields. In high performance computing (HPC) contexts, "scaling-up," "scaling-out," refer to architectural choices made when enlarging an HPC system, with scaling-up yielding improved computing elements while scaling-out involves adding more elements to the array (Hwang et al. (2014)). These in turn differ from strong and weak scaling in HPC, where strong scaling demonstrates a proportional decrease in time-to-solution as a function of an increased number of compute nodes, while applications that increase a problem-size as the number of compute nodes increase (thereby keeping the per-node-workload constant) typify weak scaling (Shoukourian et al. (2014)).

These uses of scaling differ from how similar terms have been applied in the mind and brain sciences. In neuroscience and cognitive science, "scaling laws" pick out several phenomena. Kello et al. (2010) use it to refer to the repeated patterns that occur at various levels, or scales, of organisms. In a similar vein, van Hemmen (2014) articulates the scaling

hypothesis as the claim that various scales of description require distinct mathematical and conceptual resources if our goal is to generate adequate explanations of phenomena located at said levels. Neither of these applications of scaling laws or hypotheses captures how the term is used in machine learning, although its use in HPC and cloud computing is arguably related. In the following section we'll focus on the novel use of these terms in machine learning.

2.2 Scaling Hypothesis and Loss

The weak scaling hypothesis concerns the appearance of scaling laws in supervised and self-supervised learning. We discuss each type of learning in turn before talking about scaling laws.

Supervised learning develops models through a loop: a model with initially random parameters predicts a target feature from training samples, and a loss function judges the predictions' accuracy. The loss informs a learning algorithm, updating the model's parameters. The process repeats until a stopping criterion is met; the model is then evaluated on a hold-out test set to assess generalization and avoid memorization. This procedure is supervised because datasets have human-selected target features.

Unsupervised learning requires no human-identified targets. Normally this requires a different training regime from the one discussed above. However, in self-supervised learning, we leverage the inductive features of datasets like sequences to avoid the need for a teacher while preserving the supervised learning feedback loop. In sequences, earlier entries inform later ones, providing a target feature: the next item; for example, the phrase "the cat jumped on the" greatly informs the predictor about what word comes next.

Models can be trained unsupervised with a supervised loop by using previous sequence items as features and judging predictions of the next item. Evaluation occurs on a hold-out test set, with better models minimizing loss.

Scaling laws appear in both supervised and self-supervised learning cases. These scaling laws all exhibit the same behavior: they show that loss, as a function of overall compute spent on training a model, fits a decreasing power law. A power law between two variables x and y is a function of the form roughly $y = f(x) = ax^k$ where a change in x leads to a proportional change in y and a and k are the constants that determine the degree of that change. For example, if y is the area of a square and x is the side length, then there is a power law relating y and x, i.e. $y = f(x) = x^2$. The first scaling laws were discovered in supervised learning, and they on average show that across different tasks such as vision, natural language processing, and audio processing, loss and compute have a power law relationship where the exponent, k, varies between -0.07 and -0.35 (Hestness et al. (2017), 2). Here compute is a function of training data size—how many samples are shown to the model—and the number of model parameters. This trend is robust and occurs across nearly all model types and variations on the common supervised training script. While initially discovered in the supervised learning context, scaling laws have been found in language models (LM), which use self-supervised learning to become proficient at natural language processing. Most famously, scaling laws have been shown to apply even when datasets grow to encompass significant parts of the internet and as model parameter counts climb into the trillions. OpenAI demonstrated that the same decreasing power laws hold first with LMs in the low billions of parameters (Radford et al. (2019)) and then with LMs in the hundreds of billions of parameters (Kaplan et al. (2020)) (see figure 1); similar trends were observed by Deepmind on LMs, with loss evolving predictably as a

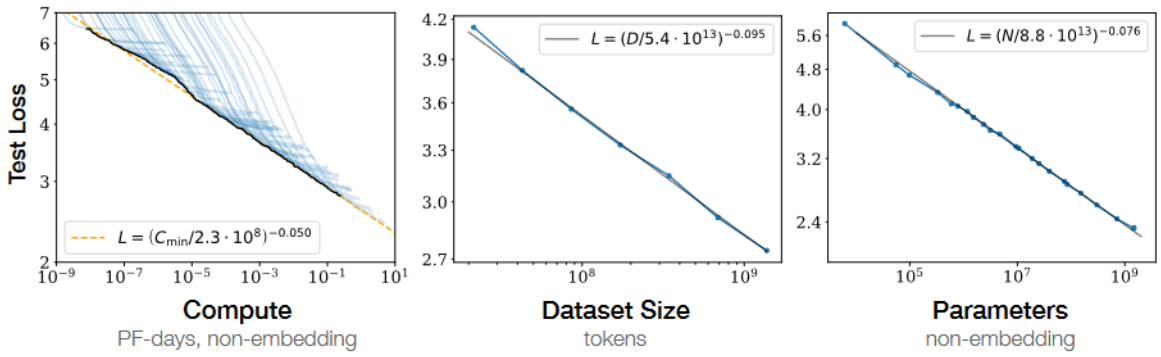

Figure 1: An example of scaling laws across multiple model architectures for LMs. Scaling laws hold for compute, training data size, and parameter counts robustly across a wide variety of models. The black line indicates the fitted power law. Originally figure 1 one from (Kaplan et al. (2020)).

function of scale (Rae et al. (2021)), and those trends have continued as compute spent in training increases (see figure 1 in Achiam et al. (2023)). So across both supervised and self-supervised learning contexts, across different tasks ranging from image classification to natural language processing, and across a large zoo of model architectures and varied training regimes, scaling laws appear again and again.

Importantly, machine learning engineers use scaling laws to guide the design and training of models. These are not just curious regularities. Engineers and researchers know that model selection and design can occur at lower parameter counts and on smaller datasets because they expect model performance to scale accordingly as parameters and data increase. This means that resources need not be wasted on large training runs to find the optimal model architecture for a given problem. And because the scaling laws have a predictable decreasing power law form, researchers can compute beforehand the optimal parameter count and data set size for LMs. Deepmind demonstrated that there

are decreasing marginal returns to compute being spent on parameters versus data or data versus parameters (Hoffmann et al. (2022)). These "Chinchilla" scaling laws help guide engineers on LM model selection when they know there is a hardware limit on the number of parameters or a collection problem on the size of the dataset. Thus scaling laws fulfill an important research and engineering role in model selection and design.

We term the claim that test loss is a decreasing function of compute spent while training the *weak* scaling hypothesis. The weak scaling hypothesis just asserts that empirically, model test loss is related to compute by a decreasing power law. Naturally, one might ask the question of what this means in terms of a model's intelligence or cognitive ability. It is on this question that scientists link the scaling laws with the emergence of new abilities, which we turn to discuss now.

2.3 Scaling Hypothesis and Emergent Abilities

Machine learning scientists expect lower loss on hold-out test sets to be evidence for greater cognitive ability because they believe that better prediction and generalization on test data is driven by the compression of inductive regularities in the training data. Instead of memorizing patterns in the data, models have to learn statistical summaries that are more "dense" in an information-theoretic sense to be better at predicting data they have never seen before. Researchers believe those statistical summaries to contain physical world modeling, i.e. simulations of the underlying data generating process (Sutskever (2023)).[1] These ideas are connected to statistical learning theory and, allegedly, an appeal to Ockham's Razor; models that empirically do well on hold-out test data are theoreti-

[1]In an interview, Sutskever, the chief scientist at OpenAI, argued that next token prediction—predicting the next item in a sequence—would require modeling the underly-

cally guaranteed to uniformly converge to the "simplest" model capable of predicting the data.[2] Researchers then make the inference that "simplest" here must mean something like a physical world model or simulation. So the expectation is that models that minimize hold-out test loss will learn those physical world models. Call this the compression hypothesis. From this hypothesis, researchers infer that better models gradually would

ing process that generated the sequence:

> Because if you think about it, what does it mean to predict the next token well enough? It's actually a much deeper question than it seems. Predicting the next token well means that you understand the underlying reality that led to the creation of that token. It's not statistics. Like it is statistics but what is statistics? In order to understand those statistics to compress them, you need to understand what is it about the world that creates this set of statistics? (Patel (2023)

He expanded on this argument in a conference presentation at the Simons Institute by making the connection between compression and Kolmogorov Complexity explicit when arguing for why neural networks generalize to off-training data (Sutskever (2023)).

[2]More precisely, "simplest" here is a type of cardinality concerning making decisions (predictions) about the data, the Vapnik-Chervonenkis (VC) dimension (Vapnik et al. (1994)). A related idea is that the models with the lowest VC dimension are also the models that minimize the length of the shortest computer program necessary to produce the data set (Kolmogorov Complexity). Both notions of simplicity, however, have been attacked as irrelevant for the statistical guarantees. See Herrmann (2020) for a critique of the underlying mathematical framework and Sterkenburg (2023) for a defense.

have new capabilities emerge as the total compute on them increases.

From compression arise purported cognitive abilities, including recently reported breakthroughs in LM performance on logical argumentation, sports understanding, and figure-of-speech parsing (consult Figure 2 below). Supposing that this view is on the right track, we're still confronted with two important questions: First, what makes an emergent ability emergent *per se*, and what analogs do we have for such abilities? As some ML researchers note, an ability is considered emergent just in case, "it is not present in smaller models but is present in larger models" (Wei et al. (2022), 2). At first blush this characterization seems far too broad, after all an ability might be latent at an earlier stage or have a characteristic developmental trajectory. Consider that language—certainly a cognitive achievement—only "emerges" in children after several years of concerted effort beginning with one-syllable utterances and single-word phrases until they can begin to grasp grammatical structure and engage in more meaningful conversations. In effect, every cognitive ability in humans (or other animals) could be said to be emergent under this description. Instead, we can turn to Schaeffer et al. (2023) who propose two criteria that emergent abilities ought to have: *sharpness* and *unpredictability*. That is, to be emergent, an ability should display something like a phase transition, where evidence for an ability goes from essentially undetectable in smaller or earlier models to near maximum performance in a saltatory fashion. Second, we should not be able to easily predict when these phase-like transitions in a model's ability occur. Arguably, these un-predicted, phase-like transitions in cognitive abilities are just exactly what we see as we scale models ever upwards (consult Figure 2).

Do we have robust analogues for emergent cognitive abilities in other creatures? Some cognitive scientists from a dynamical systems bent argue that many core cognitive abilities–

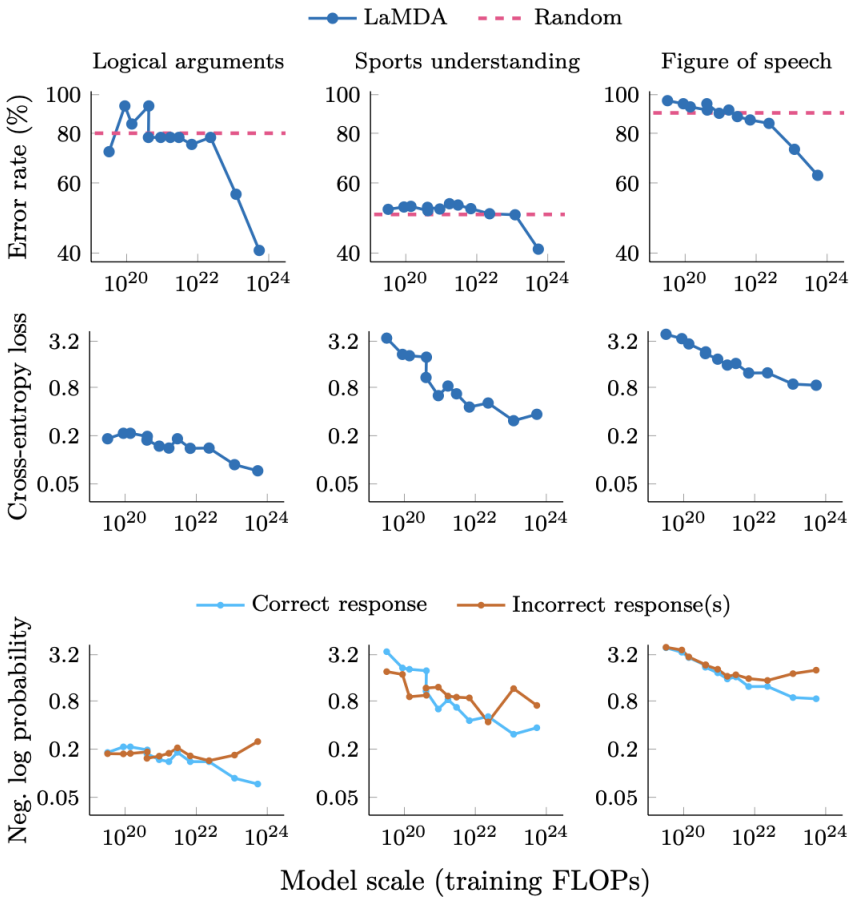

Figure 2: Emergence on abilities from LaMDA related to loss metrics. Abilities that allegedly emerge include logical argumentation, sports understanding, and figure-of-speech understanding. Originally figure 6 from Wei et al. (2022).

ranging from lexical entries and rules (a representational system *par excellence*) to decision making–are emergent phenomena (McClelland (2010)). [3] In this context, "emergence" is a part-whole mereological relationship, where a property of the whole (e.g., intelligence) isn't found in any single component (McClelland (2010), 752). In this way, dynamical systems proponents provide a narrative that explains how agents comprised of simple processing systems can behave as if they have vast stores of innate knowledge and domain-specific capacities. While similar in spirit to the kind of emergence brought about by the compression hypothesis and strong scaling in ML, there is a crucial discrepancy; namely, while both groups see emergent abilities as arising from sharp, phase-transition-like jumps from the capacities simpler predecessors, cognitive scientists don't suppose that these breaks are "unpredictable." Indeed, we *know* that there are organisms that have a range of cognitive abilities (us), the question is how to model the causal basis of these achievements. Representationalists and fans of modularity take these abilities to be products of symbol manipulation (e.g. Fodor (2000)), while dynamical systems theorists rely on emergence to explain how we arrive at such abilities from simpler processing structures. However, both these views are in principle falsifiable and arguably that's a principle task cognitive science these past three decades. By contrast, and as we describe in section 3.2, the strong scaling hypothesis in ML with its inherent commitment to a kind of unpredictability makes it a bad candidate to build a falsifiable scientific framework around.

[3] In fact, McClelland goes even further to argue that many elements of cognitive architecture including attention and declarative memory are also emergent phenomena. Postle (2006) raises a similar point about working memory.

3 Is the Scaling Hypothesis Falsifiable?

The key question around the falsifiability of the scaling hypothesis is what is being measured? We have two options: first, the power laws of the weak scaling hypothesis, and second, the emergent capabilities of models as test loss decreases. We address each in turn.

3.1 Falsifiable Loss Curves

Recall that the weak scaling hypothesis is the claim that hold-out test set loss for deep learning models fits a decreasing power law as a function of compute. More model parameters and more training data proportionally decrease the test loss function in a manner congruent with a power law whose exponent is between negative 0.1 and 0.35. This is falsifiable; either the test loss fits such a power law or it does not. If it does not fit such a power law, then one would see some other pattern in the loss. For example, the scaling law would be falsified if test loss fails to continue to decrease as a function of compute—this is the "flattening" scenario that LM skeptics often believe may happen soon. So the weak scaling hypothesis is easily falsifiable.

Unfortunately, while the weak scaling hypothesis is falsifiable, it is empirically uninteresting. Both machine learning scientists and philosophers have invested in scaling laws because of their connection with the alleged improvement in model capability. Were there no connection between lower test loss and model abilities, then the falsification of the scaling laws would be of little consequence. As a result, the question of whether the strong scaling hypothesis is falsifiable is the one worth answering.

3.2 Non-falsifiable Emergent Capabilities

By contrast, the strong scaling hypothesis in ML with its twin criteria of *sharpness* and *unpredictability* make it a poor candidate for falsifiability. Baking in the requirement that emergent cognitive abilities arise unpredictability as a model scales results in an impoverished suite of hypotheses for researchers to test—that is, aside from the ever-present (and naturally un-falsifiable) premise that ever larger models *may* lead to new abilities. As previously mentioned, this is distinct from the pattern of explanation that proponents of emergent abilities deploy in the cognitive sciences (McClelland (2010)), as their project runs in the other direction as they begin with a description of our already present suite of cognitive abilities and attempt to explain how they emerge from groups of simpler (non-representational) processing units.

There are two final criticisms of the strong scaling hypothesis: that purportedly "emergent" properties may be a mirage caused by metric choice, and that beyond worries of unpredictability, there are serious concerns about the ecological validity of target abilities in the first place. As Schaeffer et al. (2023) points out, the twin signatures of emergence: sharpness and unpredictability, may fall out of a researcher's metric of choice. That is, "emergent abilities are a mirage caused primarily by [...] choosing a metric that nonlinearly or discontinuous deforms per-token error rates," (2023, 2). In their paper, they show that purportedly emergent abilities (arithmetic task performance by the GPT family of models) "evaporate" when switching from a nonlinear to a linear metric (in this case from accuracy to token edit distance). They further show a double dissociation by inducing emergent-like abilities in vision tasks via metric choice (8). This doesn't negate all prior claims of emergence in ML, but it cautions us to choose (and ideally preregister our pre-

dictions about) a diverse set of metrics when hunting for evidence of emergent abilities in ML models.

Finally, there is a worry about the ecological validity of the entire project; namely, what is a good in principle way to identify target emergent cognitive abilities in ML models? Our human cognitive achievements—language use, problem-solving, social coordination—are products of a rich evolutionary story of adaptation and of a complex developmental trajectory. It's not clear to us that these achievements are the right targets for the kinds of abilities that LMs might come to, or already, possess. That's in part because LMs haven't faced the same evolutionary and developmental hurdles that we humans have managed for hundreds of thousands of years. While these achievements might serve as good placeholders for genuine targets of emergent abilities in ML models, we ought to justify these choices and little of that work has been done. Consider claims surrounding IQ scores for large LMs. Journalists and commentators have noted that recent LMs like GPT-4 and Claude 3 have IQs at 150 and 100 respectively based on their performance on IQ tests; however, a closer examination shows that LM performance on these tests lacks the correlation with human training-relative performance on the test, which signifies the test on LMs fails to measure g (Recueil (2024)). The upshot is that such scores are effectively meaningless given their lack of construct validity as applied to LMs.

All told, the strong scaling hypothesis with its reliance on emergent abilities yields an impoverished framework from which to make and test tractable, falsifiable hypotheses about cognitive-like abilities in ML models. The reliance on unpredictability, *pace* similar efforts in cognitive science, the impact of metric choice, and the very ecological validity of the enterprise of comparing machine and human intelligence present serious challenges for proponents of this version of the hypothesis.

4 Discussion

When confirming or falsifying a hypothesis, it is a good heuristic to make that hypothesis exact. We have argued that while there are no issues in making the weak scaling hypothesis exact, there are more serious problems with the strong scaling hypothesis. Those problems come in two varieties: first, we have to evaluate better the compression hypothesis that more cognitive ability is downstream of physical world modeling, which is a result of "simpler" models being learned during training, and second, we need better measures of cognitive abilities concerning AIs.

Progress has already been made in evaluating part of the compression hypothesis. For example, labs have shown that LMs can build linear representations of an Othello game board that are causally relevant to the model's decisions when playing Othello (Li et al. (2023)), LMs learn compositional representations from next-word prediction (Lepori et al. (2023)), and LMs apply vector arithmetic on representations of color spaces (Merullo et al. (2023)). This suggests that at least some of the time, LMs can build something that looks like a physical simulation when one squints at it. However, there has been little connection to the complexity of the models learned during training; there has been little work done on analyzing how the functional capacity of models changes—if at all—during training. Most customary analyses look at the complexity of the model before training by the use of heuristic measures such as parameter counts, though these measures happen to be irrelevant to the actual theoretical guarantees provided by statistical learning theory.[4]

[4]Parameter counts need not directly translate into VC dimension. For example, the set of models with a single parameter ω defined by $f(x) = \sin \omega x$ where $\omega \geq 0$ has an infinite VC dimension. A more general point is made by Romeijn (2017) about what counts as

So there is additional work that needs to be done in quantifying the complexity of trained models that do well on hold-out test data.

Unfortunately, very little work has been done on the problem of how to properly measure AI cognitive abilities. This is important for assessing whether there are emergent abilities as test loss decreases. It is not because people haven't tried—the proliferation of benchmarks like BIG-bench (Srivastava et al. (2022)) and MMLU (Hendrycks et al. (2021)) show that people are attempting to quantify model capabilities; instead, the problem is just genuinely hard. The principal difficulty is that when measuring human intelligence, we can construct tests that take advantage of the many ecologically valid reasoning tasks humans can execute. Those tests then turn out to be highly correlated—meaning that we can be confident that they measure the same psychological construct in humans. However, in non-human animals and even more so for machines, we find it difficult to identify the right sort of tests that are sufficiently correlated to allow us to infer they measure something like animal or machine cognition. Some progress has been made on this in comparative psychology by exploiting the ecological niches that animals have adapted into; but with machines trained via gradient descent, it is unclear whether it is even correct to ascribe an adapted niche that can then be leveraged to construct good tests. We lack a good theoretical understanding of what machine intelligence is for to be able to measure the smartness of machines.

We should emphasize that this is not to say that machines are not intelligent or cognitively capable. Far from it. Just because we lack good measures does not mean there is no there there; after all, the fact that thermometers have not existed for most of human history does not imply things were not hot or cold until the invention of temperature.

"complex" and "simple" here.

Ultimately, cognitive ability is reflected in reliable success in action, and from an empirical perspective, models are now capable of actions above and beyond predicting the next word. The difficulty is that we do not know what those actions are for—what is the success condition relative to the models' behavior?—to build valid tests. For example, with humans and other great apes, we expect—and find—theory of mind tests to be sufficiently correlated with one another because we have a story that theory of mind makes sense given the importance of cooperation and competition between conspecifics in apes' evolution. However, there is no corresponding story with LMs making the theory of mind tests we use for great apes uncorrelated and constructively invalid for LMs. We might say that by better predicting text involving human interactions, an LM learns to physically simulate those humans' theory of mind. But this just backs the measurement issue up to the truth of the compression hypothesis. If that hypothesis turns out to be false or if we cannot confirm it, then we need some way of identifying the targets of machine cognition to better evaluate machine intelligence.

So what steps can we offer moving forward? There are at least two general lessons that we can draw from our discussion of the scaling hypothesis in ML: a conceptual point about target choice and a practical empirical recommendation. First, while it might be tempting for ML researchers to devise benchmarks that seek to simulate the properties and limits of human intelligence, we think that such moves should be justified by a theoretical framework rather than being taken merely for granted. Such a framework should make clear predictions not only about the kinds of abilities that LM models at various scales can achieve, but that give an explanatorily plausible story as to *why* the models can achieve such performance. Perhaps this requires a broader interdisciplinary effort to determine just what an ecological narrative about machine intelligence might look like, but any

well-thought-out justification will be better than the story we have at present.

Finally, there are at least two practical empirical considerations that we can distill from the current state-of-the-art in ML research. Metrics matter, and deploying a diverse range of metrics that include both linear and non-linear evaluations of LM abilities will be key to providing stronger evidence for any future claims of emergent abilities. Second, and this is borne from our experience in the social sciences, is a suggestion that ML practitioners develop norms and accessible protocols around the preregistration of experimental designs, hypotheses, and analyses. We hope that these recommendations will enable us to more effectively predict task performance as LM models continue to scale, and in effect conclusively answer the question that we set out at the start of this paper.

References

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Fodor, J. (2000). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. MIT Press.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*. MIT press.

Hendrycks, D., C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt (2021). Measuring massive multitask language understanding.

Herrmann, D. A. (2020). Pac learning and occam's razor: Probably approximately incorrect. *Philosophy of Science 87*(4), 685–703.

Hestness, J., S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou (2017). Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

Hoffmann, J., S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Hwang, K., Y. Shi, and X. Bai (2014). Scale-out vs. scale-up techniques for cloud performance and productivity. In *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, pp. 763–768.

Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kello, C. T., G. D. Brown, R. Ferrer-i Cancho, J. G. Holden, K. Linkenkaer-Hansen, T. Rhodes, and G. C. Van Orden (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences 14*(5), 223–232.

Lepori, M., T. Serre, and E. Pavlick (2023). Break it down: Evidence for structural compositionality in neural networks. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), *Advances in Neural Information Processing Systems*, Volume 36, pp. 42623–42660. Curran Associates, Inc.

Li, K., A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*.

McClelland, J. L. (2010). Emergence in cognitive science. *Topics in Cognitive Science 2*(4), 751–770.

Merullo, J., C. Eickhoff, and E. Pavlick (2023). A mechanism for solving relational tasks in transformer language models.

Patel, D. (2023). Ilya sutskever (openai chief scientist) - building agi, alignment, future models, spies, microsoft, taiwan, & enlightenment.

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience 139*(1), 23–38.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog 1*(8), 9.

Rae, J. W., S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Recueil, C. (2024). Nonhuman intelligence.

Romeijn, J.-W. (2017). Inherent complexity: A problem for statistical model evaluation. *Philosophy of Science 84*(5), 797–809.

Schaeffer, R., B. Miranda, and S. Koyejo (2023). Are emergent abilities of large language models a mirage?

Shoukourian, H., T. Wilde, A. Auweter, and A. Bode (2014, Jun.). Predicting the energy and power consumption of strong and weak scaling hpc applications. *Supercomputing Frontiers and Innovations 1*(2), 20–41.

Srivastava, A., A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Sterkenburg, T. F. (2023). Statistical learning theory and occam's razor: The argument from empirical risk minimization.

Sutskever, I. (2023). An observation on generalization. Simons Institute Workshop: Large Language Models and Transformers.

Sutton, R. (2019). The bitter lesson. *Incomplete Ideas (blog) 13*(1), 1–2.

van Hemmen, J. L. (2014). Neuroscience from a mathematical perspective: key concepts, scales and scaling hypothesis, universality. *Biological Cybernetics 108*(5), 701–712.

Vapnik, V., E. Levin, and Y. Le Cun (1994). Measuring the vc-dimension of a learning machine. *Neural computation 6*(5), 851–876.

Wei, J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.