AI: WHY "MEEK," LOW-BUDGET MODELS COULD SOON MATCH HIGH-BUDGET PERFORMANCE

IN THIS BRIEF

- *Large organizations that can afford to grow their artificial intelligence model investments would seem to be the primary controllers of powerful AI, both now and in the future. But a new analysis shows the rise of a capability convergence in which "meek" AI models—those trained or run with limited resources— can approach the performance of models with greater investment.*
- *To explore this issue, a team of three researchers created a scheme for what's known as Model Training Inequality. The researchers focused on the training-loss difference between state-of-the-art (SOTA) models and a "meek" model trained with a fixed capital training budget on the same data distribution. Their main finding: As the top model faces decreasing returns to compute scaling, the gap between high- and low-budget AI systems eventually shrinks.*
- *In addition, the researchers find that another measure—Inference Time Scaling, which focuses on how efficiently a model can allocate compute during inference—shows that, at least for now, the same holds true. That is, exponentially increasing inference compute leads to diminishing returns.*
- *Further, the researchers find that the benchmark performance advantage held by top-tier AI models due to large training budgets compared to more modestly-resourced models has narrowed. That's due to ongoing improvements in hardware and algorithms.*
- *The researchers suggest that future AI governance will need to move beyond simply restricting access to powerful computing hardware. Instead, this governance may also need to regulate data sources, algorithmic innovations and emerging research. Only in this way can we address the widespread and relatively inexpensive availability of advanced AI capabilities.*

RESEARCH OVERVIEW

Greater democratization of artificial intelligence systems may be on tap for the future, even if current performance differences between costly and modest AI systems would seem to argue otherwise.

It's true that large corporations currently dominate the training of state-of-the-art (SOTA) AI models, including GPT, Llama and Gemini (Brown et al., 2020; Touvron et al., 2023; Pichai & Hasabis, 2023). In part, that's because running inference on these models often requires multiple graphical processing units (GPUs), a costly endeavor. One logical conclusion is that in the future, only centralized entities such as government agencies and large businesses will be able to afford to train and use these AI systems (Cottier et al., 2024). If this occurs, then powerful AI would be controlled by just a small number of organizations. Conversely, organizations with more limited computation budgets would suffer from a lack of access to equivalent AI performance.

However, this assumption fails to consider a possible inflection point in which further AI investments yield only narrow advantages for SOTA models. In fact, AI performance gains could even decrease substantially. That's because models being trained and run with far more modest resources are now approaching the capabilities of today's best models.

That's the case made by a team of three MIT researchers—Hans Gundlach, Jayson Lynch and Neil Thompson—in their recent working paper, **Meek Models Shall Inherit the Earth**.

To support their argument, the researchers developed a model showing that the marginal capability returns to raw compute over time will shrink substantially. Indeed, even the companies that scale their AI models the fastest could gain little to no capability advantage, at least as measured by prominent capability metrics.

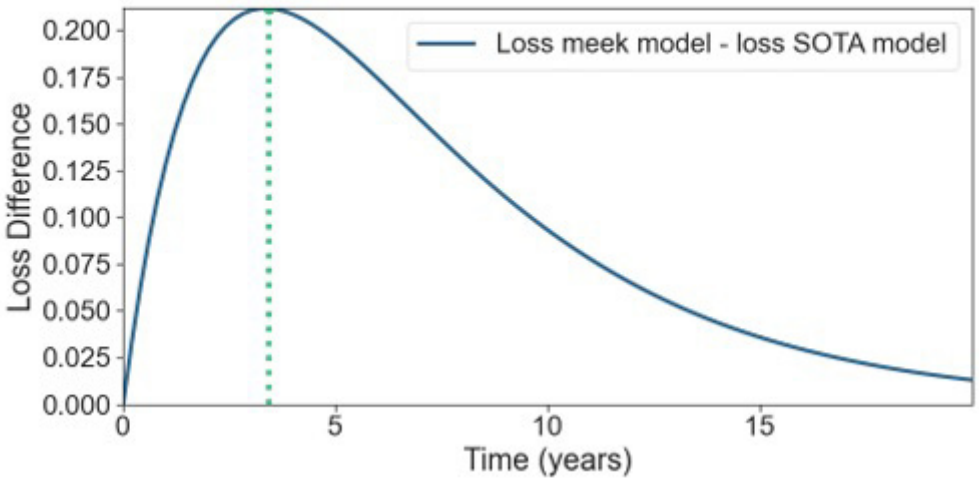

Figure 1: Initially, the AI model with a growing compute budget surpasses the model with a constant training budget. But over time, the top model faces decreasing returns to compute scaling, and the gap declines.

The researchers' models also explore the factors behind what's known as "capability convergence"—that is, "meek" models approaching the performance of the best models overall.

The researchers also highlight possible AI strategy and policy requirements that could arise in the future:

- Diminishing returns to compute: As more compute is invested in training AI models, the marginal gains in performance decrease substantially. In effect, each additional dollar spent ultimately provides less and less improvement.
- Convergence of capabilities: Over time, "meek" models trained with modest resources will approach the capabilities of SOTA models.
- Implications for AI governance: Current strategies for AI regulation and safety focus on restricting access to large-scale compute. But in the future, that approach may be too limited because robust capabilities will also be available from small-scale AI systems.

THE INVESTIGATION

To test their hypotheses, the researchers developed mathematical models and analyzed empirical data. For Model Training Inequality, they compared a SOTA model with an exponentially growing compute investment (3.6x per year) to a "meek" model on the same data distribution. Both started with an investment of $1,000 at the rate of approximately 1,017 GPU floating-point operations per second (FLOPS) per dollar (Li, 2021).

But while the SOTA model received exponentially more funding over time, the "meek" model's budget remained constant. Scaling performance is governed by what are known as chinchilla-like scaling laws (Hoffmann et al., 2022), which give a relationship between optimal compute usage and log likelihood loss. Loss measures how well a model predicts the correct answer by penalizing the model when it assigns a low probability to the true outcome.

The researchers also used the Massive Multitask Language Understanding (MMLU) benchmark. It affects log loss performance by providing a standardized way to evaluate how well a large language model (LLM) predicts correct answers across a wide range of tasks.

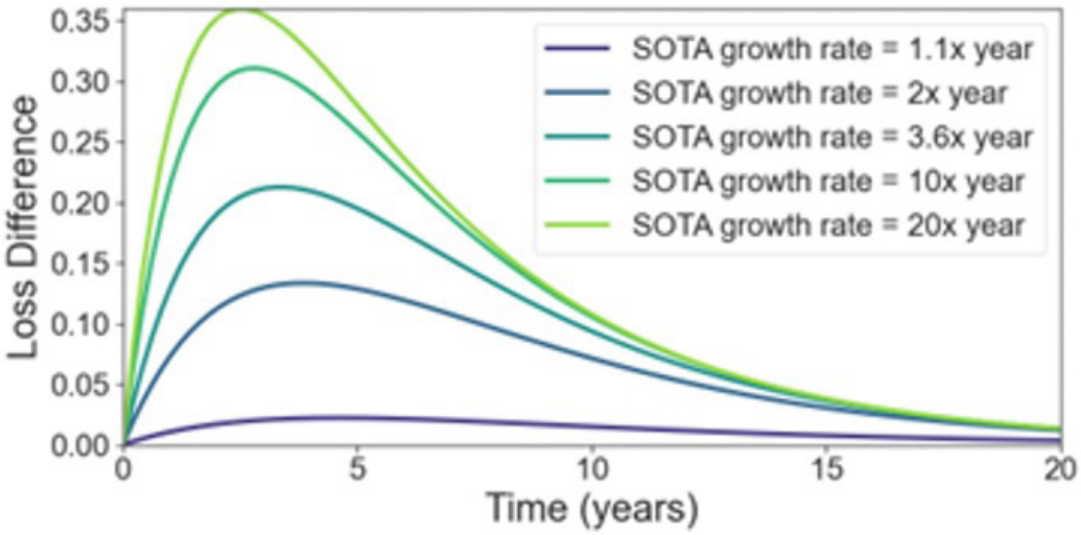

Figure. 2. The loss difference is the change in loss values between two models that shows how much better or worse one model is at predicting the correct answers compared to another. This figure shows the loss difference between SOTA vs. meek models with different levels of SOTA compute investment growth.

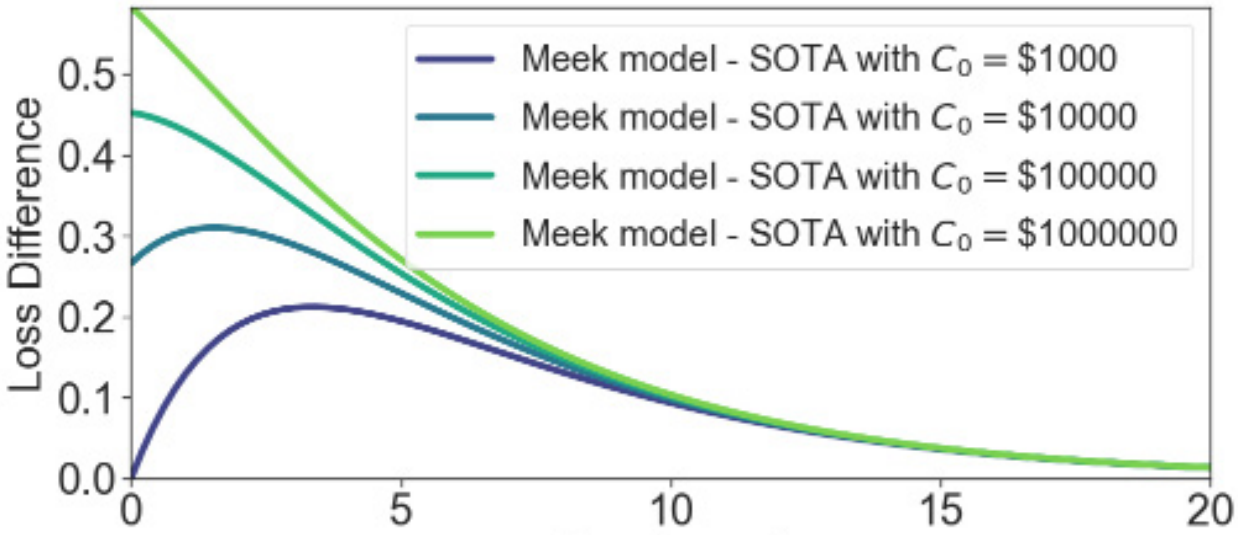

Figure 3. This chart shows that the initial compute capital advantage makes little difference in loss over time. The "meek model" training budget is kept constant at $1,000.

Initially, the researchers found that an AI model with an exponentially growing compute budget will surpass another model created with a constant training budget.

But over time, the effects of both Moore's Law—which observes that compute power doubles about every two years—and better algorithms begin to reverse the situation. Even though the budget may keep growing, returns eventually diminish. And because smaller improvements are obtained for each extra dollar spent, the gap narrows (Figure 1).

This surprising outcome is due, in part, to the fact that algorithmic progress in LLMs is remarkably consistent over time (Ho et al., 2024). Effective computational resources double about every eight months. What's more, better hardware and smarter algorithms help everyone, regardless of their budget size.

The researchers also found that the inflection point occurs when the training model size hits sizes with particularly diminishing returns, and exponential growth in compute budgets can no longer keep up with the combined progress in algorithms and hardware. At that point, investing more creates only a narrow advantage over models trained with far more modest budgets. The researchers also found that for SOTA models with huge training budgets, the biggest advantages may have already arrived.

Looking ahead, the researchers explored an important question: Given that we're in the middle of a transition from AI training to AI inference—that is, from teaching AI models to running them—will this new paradigm eventually yield the same diminishing returns as did AI pre-training? They conclude the answer is yes, at least under current inference compute conditions. That is, the more an organization scales inference compute, the less extra performance it ultimately receives.

The researchers also explored the performance on the MMLU benchmark, which can be used as a proxy for measuring advanced knowledge in a wide variety of domains from international law to religion. The researchers found that when using the MMLU benchmark, there's a plateau where performance flattens out, regardless of more computer or data being added.

What this means, the researchers say, is that for simple tasks, there will at first be little difference in capabilities between SOTA and "meek" AI models. But over time, a large difference will emerge, eventually followed by a convergence.

Convergence on these tasks might reflect the circumscribed nature of the task's performance, rather than general intelligence. That's because benchmarks have a performance threshold which they cannot exceed.

For this reason, the researchers present a more theoretical information theory-based perspective (Ho et al., 2023). This parallels their conclusion that, due to the shrinking loss difference, the number of tokens necessary to differentiate the meek model from the SOTA model increases gradually as the two models become more indistinguishable. In practice, users will find it increasingly difficult to tell the difference between "meek" and SOTA models.

The researchers also highlight empirical trends using data sourced from the Artificial Analysis LLM Leaderboard (Artificial Analysis, 2025), which tracks metrics on commercially available LLMs. Specifically, the data show a rapid convergence in capabilities among commercially available AI models with fixed inference budgets. The caveat here is that the data is suggestive, but still sparse.

THE RESULTS

The researchers show that over time, even drastic differences in compute investment in AI models create only modest differences in important capabilities (Figure 2).

Over the long term, they speculate, it's possible that no entity can maintain a large advantage by simply investing more (Figure 3).

For this reason, the researchers say that progress in both AI training and inference will face diminishing returns.

Organizations will gain smaller and smaller improvements even as their compute budgets grow. Eventually, they'll reach an inflection point where additional spending no longer delivers meaningful gains.

Of course, the future remains uncertain. The researchers point to the possibility of new paradigms and adversarial settings. And they're uncertain about their model's validity if powerful AI models are trained on either adaptively chosen or synthetic data. In these cases, AI systems could go beyond imitating humans, learning new, extra-human skills. If this happens, then the question would be not how well AIs are learning, but what they're learning.

The researchers also discuss adversarial contexts, such as competitive games, where adversaries with exponentially increasing compute might continuously diverge from their competitors (Jones, 2021). In these settings, small performance differences can make large impacts. Still, diminishing returns may return at high levels of compute as agents approach perfect play (Neumann & Gros, 2022), which refers to how artificial agents—like those trained with reinforcement learning—get closer to making optimal decisions.

CONCLUSIONS

The researchers' model has important implications for AI governance. Regulations, they say, can be targeted at large companies for as long as those companies are close to the inflection point, before they hit large diminishing returns.

As the companies gain experience with powerful models, they can design safety procedures for these models before they become ubiquitous. At the same time, small groups not subject to these restrictions can accumulate power and influence.

Given existing trends, we all could soon gain access to high-performance deep learning models. These capabilities could become as ubiquitous as computer ownership via PCs and smartphones. And again, the effects could be both positive and negative. On the plus side, many people could enjoy the productivity advantages of enhanced AI. But on the other, some people would undoubtedly use the systems for nefarious purposes.

With this in mind, governments may need to expand their AI governance beyond "frontier AI," a term for highly capable, general-purpose AI models that either match or exceed today's capabilities. When it comes to frontier AI, simply restricting total compute, using measures such as U.S. export controls on GPU hardware, may not be enough. Safeguards for a world where frontier AI capabilities are easily accessible, even to those with limited capital and compute, could come in the form of regulating data, algorithms and new research breakthroughs. Whatever may come, the researchers say we must prepare for a world in which "meek" AI models inherit the earth.

REPORT

Read **the full working paper**.

ABOUT THE AUTHORS

Hans Gundlach is a Researcher at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), where he studies the technological and economic future of deep learning and quantum computing.

Jayson Lynch is a Research Scientist at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL). Jayson works on theoretical computer science problems and researches trends in algorithmic innovation and its impact on the world.

Neil Thompson is a Principal Research Scientist at the MIT Initiative on the Digital Economy (IDE) and the Director of the MIT FutureTech program.

REFERENCES

Artificial Analysis (2025). **LLM leaderboard: Comparison of over 100 AI models from OpenAI, Google, DeepSeek & others**.

Atkinson, D., et al. (2023). **Direct approach interactive model**. Epoch AI report.

Brown, T., et al. (2020). Brown, T., et al. (2020). **Language models are few-shot learners**. Advances in Neural Information Processing Systems, vol. 33, pp.1877–1901.

Cottier, B., et al. (2024). **The rising costs of training frontier AI models**.

Ho, A., et al. (2024). **Algorithmic progress in language models**.

Hoffman, J., et al. (2022). **Training compute-optimal large language models**.

Jones, A.L. (2021). **Scaling scaling laws with board games**.

Li, C. (2021). **Tesla A100 server total cost of ownership analysis**. Lambda Deep Learning blog.

Neumann, O.; Gros, C. (2022). **Scaling laws for a multi-agent reinforcement learning model**.

Pichai, S.; Hasabis, D. (2023). **Introducing Gemini: Our largest and most capable AI model**.

Touvron, H., et al. (2023). **Llama: Open and efficient foundation language models**.

MIT Initiative on the Digital Economy
MIT Sloan School of Management
245 First St, Room E94-1521
Cambridge, MA 02142-1347
ide.mit.edu

Albert Scerbo,
Associate Director
267-980-2616
ascerbo@mit.edu

Our Mission: The MIT Initiative on the Digital Economy (IDE) is shaping a brighter digital future. We conduct groundbreaking research on the promise—and peril—of new digital techonologies including generative artificial intelligence (GenAI), quantum computing, data analytics, and distributed marketplaces. We also investigate the rise of fake news and misinformation and the development of a digital culture. Through research and the convening of leaders from academia, industry and government, the IDE provides critical, actionable insight for people, businesses, and government to understand and benefit from new technologies and how they're repidly changing the ways we live, work and communicate.

Contact Us: David Verrill, Executive Director,
MIT Initiative on the Digital Economy
617-452-3216
dverrill@mit.edu

Become a Sponsor: The generous support of individuals, foundations and corporations help to fuel cutting-edge research by MIT faculty and graduate students. It also enables new faculty hiring, curriculum development, events and fellowships.