Meek Models Shall Inherit The Earth

Hans Gundlach [1] **Jayson Lynch** [1] **Neil Thompson** [1]

Abstract

The past decade has seen incredible scaling of AI systems by a few companies, leading to inequality in AI model performance. However, we develop a model which illustrates that under a fixed-distribution next-token objective, the marginal capability returns to raw compute shrink substantially. Under the current scaling paradigm, we argue that these diminishing returns are strong enough that even companies that can scale their models exponentially faster than other organizations will eventually have little advantage in capabilities. As part of our argument, we give several interpretations of what proxies like training loss difference mean in terms of empirical benchmark data and theoretical performance models. Finally, we present some of the policy implications of this result for the governance of AI systems.

1. Introduction

Artificial intelligence systems have grown considerably in the last decade (Sevilla et al., 2022). This trend has drastically changed the landscape of machine learning. Large corporations now dominate the training of many state of the art (SOTA) models, including systems such as GPT (Brown et al., 2020), Llama (Touvron et al., 2023), and Gemini (Research, 2023). Further, it is getting harder to run inference on these models, which now involves the use of multiple GPUs for the largest systems. What do these trends mean for the effects of AI on society?

If model investment growth continues in this direction, only centralized entities such as the government and corporations can train and use these AI systems (Cottier et al., 2024). At the same time, other sources warn about the diminishing returns to AI scaling (Lohn, 2023; Lu, 2025; Thompson et al., 2021). These have raised speculation that AI is ”hitting a wall” (Caputo, 2025). In this paper, we want to outline models of how performance inequality develops between deep learning models. This model leads us to the counterintuitive conclusion that relevant AI performance levels could converge under the current AI scaling paradigm. Hence, models trained or run with limited resources, “meek” models, will have more comparable performance to state-of-the-art models. We argue that this could imply greater democratization of AI systems and lead to a world where **meek models shall inherit the earth**.

2. Modeling Training Inequality

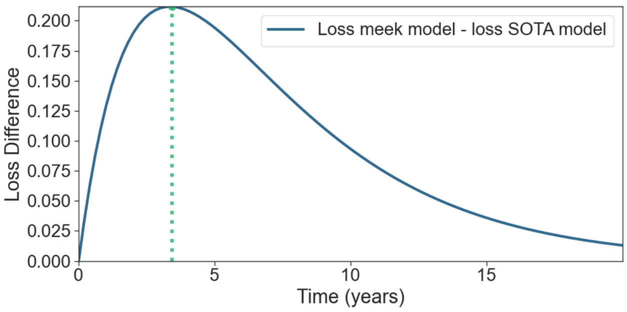


Figure 1: Graph of loss Difference between a model with 3.6x yearly compute scaling and a meek model with a constant compute budget ($1000 budget). Both models start with an initial compute budget of $1000. Initially, the model with an exponentially growing compute budget is able to surpass a model creating with constant training budget. However, this gap eventually declines as the top model faces decreasing returns to compute scaling.

The first model we construct focuses on the difference in training loss between SOTA models and a ”meek” model trained with a fixed capital training budget on the same data (we assume $\$1000$ dollars at $\approx 10^{17}$ GPU Flops per dollar (Li, 2021) training budget). We assume that scaling performance is governed by chinchilla-like scaling laws (Hoffmann et al., 2022). Chinchilla laws give a relationship between optimal compute usage C and the log-likelihood loss L. We will refer to this as simply the loss for the rest of the paper.

The compute-only form can be derived using Pearce & Song (2024)’s approach. This formulation of the original analysis gives a compute-only formulation of Hoffmann et al.

[1] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. Correspondence to: Hans Gundlach <hansgund@mit.edu>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada. Copyright 2025 by the author(s).

(2022)'s scaling laws given by the equations below.

$$L_{opt}(C) = 1070\, C^{-0.154} + 1.7 \quad (1)$$

$$L_{opt}(C) = A\, C^{-\alpha} + L_0 \quad (2)$$

Over time, the amount of effective Flops per dollar increases. This change is due to two effects.

The first effect is an increase in hardware capability. This effect is based on trends like Moore's law where the number of transistors on integrated circuits increases by approximately a factor of 2 every 2 years. Our models assume hardware growth $g_h = 1.4$ consistent with Moore's law and trends in GPU price performance for GPUs used in ML research (Hobbhahn & Besiroglu, 2022; Rupp, 2015).

The second effect is due to algorithmic progress, the fact that better algorithms make it possible to learn more effectively with less computation. For example, the discovery of transformers made it possible to train AI models much more effectively in parallel. Ho et al. (2024) discovered that algorithmic progress in language models is remarkably rapid consistent over time. Due to algorithmic progress, effective computational resources double approximately every 8 months. We label the growth rate in the effective computation in one year due to algorithmic progress $g_{alg} = 2.8$.

We label the compute budget in Flops at time $t = 0$ as C_0. Therefore, the effective computation resources over time at a given budget C_0 is $(g_{alg}g_h)^t C_0$. To account for the compute budget of large corporations and the progress of SOTA models we must consider a third factor – the growth rate in compute investment. Compute usage in language models has also grown at a steady exponential rate for the largest models. Between 2010 and 2022 the compute used for training deep learning model grew by a factor of 5 yearly (Sevilla et al., 2022). We then divide the compute growth rate by the hardware growth rate to get the growth rate in compute investment $g_i = 5/1.4 = 3.57$.

Now, that we have an expression for compute over time we can find the difference in loss between theoretical SOTA models with exponential growth in investment and "meek" individuals with a constant ($1000 compute budget).

$$\begin{aligned}\text{Training Loss Difference} &= \text{Loss Meek} - \text{Loss SOTA}\\ &= A((g_{alg}g_h)^t C_0)^{-\alpha} - A((g_{alg}g_h g_i)^t C_0)^{-\alpha} \quad (3)\end{aligned}$$

Figure 1 shows a graph of this relationship over time.

The Inflection Point in Training Loss Advantage An important point to note about Figure 1 is the inflection point in the training advantage curve. At a critical point in time, the diminishing returns to compute scale in addition to the exponential growth in the shared factors of algorithmic and hardware progress, overwhelm the large model provider's exponentially growing compute budget. At this point in time in our model, increasing investment is only able to create a narrow loss advantage over models trained with a very modest budget. Setting the derivative of the loss difference equal to zero lets us solve for the inflection time as:

$$\text{Training Inflection Time} = \frac{1}{\alpha \ln g_i}\left[\ln\left(\frac{\ln(g_h g_{alg} g_i)}{\ln(g_h g_{alg})}\right)\right]$$

It is important to note that this inflection time is the time since scaling has begun in our model. GPT-2 was trained in 2019 for around $\$25,000 - \$50,000$ (UMATechnology, 2025), with this baseline, our model would predict a peak advantage in the early 2020s. Using Pre-chinchilla compute scaling where $L - L_0 \propto C_T^{-0.057}$ (Kaplan et al., 2020), the inflection time is significantly longer at about 10 years. Trends in AI resource scaling may change significantly. We also believe there are other reasonable parameter choices. Our model conclusions do not depend significantly on these variations (see Appendix D).

2.1. Inference Time Scaling

We are in the middle of a transition from scaling training compute to scaling inference computation (You, 2025), which poses a challenge to the pre-training centered analysis used in our model. We might not care if we cannot run state of the art models if we get the same result leveraging cheap models with significant inference compute. Will this new paradigm eventually yield the same diminishing returns as pretraining compute? Under the popular model of inference scaling where inference compute can substitute for and multiply training computeVillalobos & Atkinson (2023), our results remain valid. This means doubling training compute while halving inference compute leads to the same level of performance. In this case, exponentially increasing inference compute would lead to similar diminishing returns in terms of loss and benchmark performance. You (2025) already projects progress in reasoning models to slow significantly. However, significant inference compute might yield new types of pretraining and inference compute might be qualitatively different (see Section 4).

3. Does Loss Difference Actually Capture Something Important?

Language models loss is traditionally given as the average negative log-likelihood loss per token on a given test set. Another common metric is Perplexity, which is 2 (or e if measured in nats) to the power of the negative log-likelihood loss. These metrics measure how well a language model is able to mimic textual data and have served as crucial milestones in the development of language models. The most straightforward interpretation of loss difference is as a measure of how much better one model is able to predict text than another. More formally, given two models with loss L_1 and L_2, the difference $L_1 - L_2 = \Delta L$ gives the number of extra nats/bits per token necessary for one model needs to encode text over the other. However, this perspective does not yield intuitive measures of model performance and does not directly measure useful capabilities, such as the ability to perform economically valuable tasks.

Another reasonable perspective is to look at historical trends in loss as a measure. The loss of GPT-3 (davinci) is 4.36 while the loss of GPT-2 (large) is 5.16 (Ho et al., 2024). Since the beginning of deep learning, loss has decreased steadily and tracked AI progress (Ho et al., 2024). However, this loss might not correspond to general model capabilities or intelligence. Yet, more tractable metrics like image classification accuracy for vision transformers have a remarkably similar power-law form to this log-likelihood loss (see equation 4) (Zhai et al., 2022). Our analysis holds with this power law formulation as well.

$$\min_{N,D} L = 0.09 + \frac{0.26}{(C + 0.01)^{0.35}} \tag{4}$$

We speculate that most model capabilities are monotonic functions of loss. Some models' capabilities are accurately captured by such scaling laws while other capabilities are better modeled as abrupt discontinuities. We extend our analysis to these capabilities in Section 3.1. However in less circumscribed or competitive setting loss difference might not capture the relative performance difference between models, see Section 4.

3.1. Loss to Benchmark Performance

How does loss measure actual capabilities? Here we provide evidence that loss can be strictly monotonically translated to benchmark performance. For instance, MMLU (Massive Multitask Language Understanding), benchmark performance can be modeled as a sigmoid of training compute (Owen, 2024). Since loss is a monotonically decreasing function of compute, it can be used as a proxy. In Appendix B, we make a sigmoid fit of MMLU scores to loss. Using this sigmoid-translates loss has similar dynamics to those we have previously outlined between large and small models. At first there is little difference in capabilities, then a large difference emerges, followed by an eventual convergence. We can also consider the case where an overall task requires p-steps where each step requires correct benchmark performance, which can be modeled as individual benchmark performance to the power of p (see Appendix B). In this case we get an interesting relationship as seen in Figure 2. As the number of necessary tasks increases, so does the length of time large model builders have an advantage. The maximum loss difference decreases as well. This is due to our fit and the nature of the MMLU benchmark. No model in our dataset have MMLU performance above 80%. Therefore, the maximum performance for a p-level task is 0.8^p. With high accuracy tasks, this effect would be less pronounced.

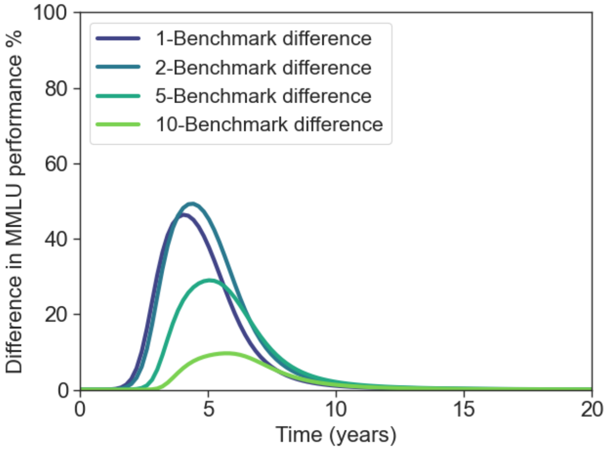


Figure 2: Difference in MMLU performance between SOTA and meek model. 2 and 5 benchmark performance identify difference in capability in tasks which involve multiple (2 to 5) correct MMLU answers. The difference in height is a direct result of the 80% sigmoid fit ceiling, a higher ceiling would yield a different height relationship.

3.2. Hypothesis Test View

Benchmark performance is a reasonable proxy for some capabilities. However, many AI benchmarks are close to saturation (Ott et al., 2022). Benchmarks have a performance threshold that they cannot exceed. Convergence on these tasks might not reflect general intelligence differentials but rather the circumscribed nature of the task. Here we present a more theoretical information theory-based perspective using the assumptions of Barnett & Besiroglu (2023), which parallels our main conclusion. This approach considers two models, A and B, which try to model some base text sampled from a distribution p_0, and we ask how many tokens N it takes an ideal observer to distinguish which model is better. In this case, p_0 corresponds to the distribution of human text, which we model using common assumptions as stationary and ergodic (Jurafsky & Martin, 2025). The expected number of tokens needed to differentiate the two models is given by equation 5, where α is the probability that we reject the true hypothesis (i.e, the test concludes that model A predicts the distribution better than model B, while the reverse is true and vice versa). We set this at 5%

$$E_{p_0}[N] = \frac{(1-\alpha)\log\left(\frac{1-\alpha}{\alpha}\right) + \alpha\log\left(\frac{\alpha}{1-\alpha}\right)}{\Delta L} \tag{5}$$

Equation 5 shows that as the loss difference decreases, the number of tokens necessary for discrimination increases. An exponential decrease leads to an exponential increase in necessary discriminator tokens. Figure 4 shows the growth in the number of discrimination tokens necessary, increasing over time, which supports the conclusion using our other approaches. [1]

4. Limitations: New Training Paradigms and Adversarial Settings

We explain this convergence as models growing increasingly close to the fixed distribution of human text. There are only so many abilities necessary to predict human text, and as models grow larger they master narrower less common abilities (see (Michaud et al., 2024)). This is likely true under naive inference scaling as well. However, powerful AIs can do more than just human imitation. They can far exceed humans and learn new kinds of skills. RL and synthetic data techniques promise to change drastically the distributions learned by AIs. It is no longer a question of how well AIs are learning but what they are learning. We are uncertain about our model in these new cases where AIs are trained on adaptively chosen data or their own synthetic data. Further, there are instances where exponentially small differences in overall loss may correspond to large capability differentials. For instance, in order to model the small set of tokens dealing with elementary math the model has to internalize the rules of arithmetic (Michaud et al., 2024). This is particularly the case in adversarial settings where agents are incentivized to win by finding situations where the competitor is unfamiliar. In competitive games, adversaries with exponentially increasing compute continuously diverge with their competitor (Jones, 2021). Yet, diminishing returns may return at high levels of compute as agents approach perfect play (Neumann & Gros, 2022).

5. AI Governance Discussion

Our model points to several conclusions for AI governance. First, there might exist a crucial **"governance window"** where large entities have a large advantage over ubiquitous AI models. During this period, regulations can be more targeted. This period is particularly advantageous as trusted organizations can gain experience with powerful models and design safety procedures for these models before they become ubiquitous. However, it is a double-edged sword, as it allows small groups to accumulate power and influence during this period.

Can increased AI investment help AI safety? If the benefits of centralization and using the "governance window" are strong enough, accelerating AI might have a positive effect on safety. Appendix D Fig 7 depicts the loss advantage with different SOTA investment rates. If a safe organization makes a very large investment in AI it will have access to highly capable systems much earlier than consumers. In this case the large AI loss advantage gives an organization much more time to do safety research before these system becomes ubiquitous.

Is money a long-term moat? Companies and countries may care about having a competitive advantage by maintaining private/proprietary foundation models. We've seen that even drastic differences in compute investment in these models create only a modest difference in important capability measures. It is possible that in the long run, no entity can hope to keep a large advantage under the current paradigm simply by having more capital.

AI for all If these trends continue then it seems access to high performance deep learning models will become as ubiquitous as computer ownership. Given that computers are also becoming less expensive and more widely used, it seems reasonably likely that a large fraction of the world's population could have access to powerful deep learning models. This suggests many people will be able to share in the benefits and productivity increases that might come with improved AI and suggests a likely dispersion of power to individuals. However, this poses a risk to safety if individuals can access dangerous capabilities like bioweapons design.

5.1. We Need to Rethink AI Regulation

Much of current AI governance is focused on monitoring and limiting access to large frontier systems. These include US export controls on GPU hardware. The US and EU focus on models trained with above 10^{26} and 10^{25} Flops, respectively (Caputo, 2025). Our work shows that simply restricting total compute may not suffice to keep frontier AI capability from becoming ubiquitous. Future AI governance would either need to drastically increase capacity for monitoring and safeguarding systems or find new targets to be able to effectively limit access to powerful models. These could include regulating data, new research breakthroughs, and algorithms (Caputo, 2025).

6. Conclusion

AI training is stretching the limits of data, computation, and energy and continued scaling may slowdown in the near future (Sevilla et al., 2024). However, even if continued scaling is possible there are deeper limits to its progress. We've developed a strategic model based on training loss

[1] We must note that this is the expected number of discrimination tokens using a random natural language sample. Fewer discrimination tokens would be needed if we narrow the focus to specialized knowledge.

and have developed several interpretations of its importance. We hope that this discussion develops a conversation about the nature of current capability benchmarks. Our model points to multiple stages of AI development and eventually a world where powerful AI is more ubiquitous than it is now. However, we emphasize that the future of AI is uncertain and new technical methods are already in development which could address these issues. Yet, bottlenecks in AI training are not only hard technical problem but pose unique challenges for AI governance. We need to develop AI governance methods that are effective in a world where AI development is less centralized. In this vein, we hope our investigation raises awareness for research into the strategy and governance of AI to prepare for a world where meek models inherit the earth.

7. Acknowledgments

We would like to thank Zachary Brown for his feedback on this manuscript.

Impact Statement

Our paper makes claims about ways to govern AI systems. AI will have a varied impact across a wide range of groups. We caution AI researchers and policymakers that significant policy AI regulation needs to take into account the details of the groups involved as well as the state of current technology. We have outlined a model and its implications, but significantly more research is needed before changing policy affecting society at large.

References

Barnett, M. and Besiroglu, T. The direct approach, 2023. URL https://epochai.org/blog/the-direct-approach. Accessed: 2024-09-20.

Besiroglu, T., Erdil, E., Barnett, M., and You, J. Chinchilla scaling: A replication attempt, 2024. URL https://arxiv.org/abs/2404.10102.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Caputo, N. A. Governing ai beyond the pretraining frontier. *arXiv preprint arXiv:2502.15719*, 2025.

Chen, C., Borgeaud, S., Irving, G., Lespiau, J.-B., Sifre, L., and Jumper, J. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., and Owen, D. The rising costs of training frontier ai models, 2024.

Franken, P. and Lisek, B. On wald's identity for dependent variables. *Z. Wahrscheinlichkeitstheorie verw Gebiete*, 60:143–150, June 1982. doi: 10.1007/BF00531818. URL https://doi.org/10.1007/BF00531818.

Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., and Sevilla, J. Algorithmic progress in language models, 2024. URL https://arxiv.org/abs/2403.05812.

Hobbhahn, M. and Besiroglu, T. Trends in gpu price-performance, 2022. URL https://epoch.ai/blog/trends-in-gpu-price-performance. Accessed: 2025-05-12.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Jones, A. L. Scaling scaling laws with board games, 2021. URL https://arxiv.org/abs/2104.03113.

Jurafsky, D. and Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition, Jan 2025. URL https://web.stanford.edu/~jurafsky/slp3/ed3book_Jan25.pdf. Online manuscript draft released January 25, 2025.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Li, C. Tesla A100 server total cost of ownership analysis, September 2021. URL https://lambda.ai/blog/tesla-a100-server-total-cost-of-ownership. Lambda Deep Learning Blog.

Lohn, A. Scaling ai. Technical report, Technical report, Center for Security and Emerging Technology, 2023.

Lu, C.-P. The race to efficiency: A new perspective on ai scaling laws. *arXiv preprint arXiv:2501.02156*, 2025.

Michaud, E. J., Liu, Z., Girit, U., and Tegmark, M. The quantization model of neural scaling, 2024. URL https://arxiv.org/abs/2303.13506.

Neumann, O. and Gros, C. Scaling laws for a multi-agent reinforcement learning model. *arXiv preprint arXiv:2210.00849*, 2022.

Nowak, R. D. Lecture 9: Sequential testing. https://nowak.ece.wisc.edu/ece830/ece830_fall11_lecture9.pdf, 2011. ECE 830: Statistical Signal Processing, Fall 2011, University of Wisconsin–Madison.

Ott, S., Barbosa-Silva, A., Blagec, K., Brauner, J., and Samwald, M. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793, 2022.

Owen, D. How predictable is language model benchmark performance?, 2024. URL https://arxiv.org/abs/2401.04757.

Pearce, T. and Song, J. Reconciling kaplan and chinchilla scaling laws. *arXiv preprint arXiv:2406.12907*, 2024.

Research, G. Introducing gemini: Google's next-generation language model. https://research.google.com/gemini, 2023.

Rupp, K. 40 years of microprocessor trend data. https://www.karlrupp.net/2015/06/40-years-of-microprocessor-trend-data/, June 2015. Accessed: 2025-05-09.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2022. doi: 10.1109/ijcnn55064.2022.9891914. URL http://dx.doi.org/10.1109/IJCNN55064.2022.9891914.

Sevilla, J., Besiroglu, T., Cottier, B., You, J., Roldán, E., Villalobos, P., and Erdil, E. Can ai scaling continue through 2030?, 2024. URL https://epoch.ai/blog/can-ai-scaling-continue-through-2030. Accessed: 2025-05-12.

Thompson, N. C., Greenewald, K., Lee, K., and Manso, G. F. Deep learning's diminishing returns, September 2021. URL https://spectrum.ieee.org/deep-learning-computational-cost. IEEE Spectrum, "The Great AI Reckoning" special report, accessed 11 May 2025.

Touvron, H., Martin, T., Narang, S., Yao, L., Zhang, Z., and Stenetorp, P. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

UMATechnology. How much did gpt-2 cost? https://umatechnology.org/how-much-did-gpt-2-cost/, January 2025. Accessed: 2025-05-09.

Villalobos, P. and Atkinson, D. Trading off compute in training and inference, 2023. URL https://epochai.org/blog/trading-off-compute-in-training-and-inference. Accessed: 2024-07-24.

Wu, Y., Sun, Z., Li, S., Welleck, S., and Yang, Y. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*, 2024.

You, J. How far can reasoning models scale?, May 2025. URL https://epoch.ai/gradient-updates/how-far-can-reasoning-models-scale. Newsletter article.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers, 2022. URL https://arxiv.org/abs/2106.04560.

A. Full Hypothesis Testing Framework

Here we outline the hypothesis testing framework introduced in Section 3.2, which is based on a model developed by Barnett & Besiroglu (2023). This approach asks how many tokens it takes to declare one model better than another using the Sequential Probability Ratio Test (SPRT) (Nowak, 2011).

A.1. Sequential Probability Ratio Test (SPRT)

Our discrimination model is based on the SPRT test (Nowak, 2011) and the framework developed in Barnett & Besiroglu (2023). We use the same assumptions they make here to show how our results can incorporate their model; however, assessing when all of these conditions fully hold is out of scope for our analysis and deserves future research. Consider two models, A and B, with predictive probabilities p_A and p_B. We continue measuring the model's probability on given tokens until a likelihood threshold is reached, in which case we declare that one model predicts the text better than the other. Here, we derive the test threshold assuming text samples are generated iid from a true distribution p_0 and then extend our proof to the stationary and ergodic case, which is a common modeling assumption for language (Jurafsky & Martin, 2025). This is a good approximation; however, natural language can depend on arbitrarily far words, which breaks these assumptions.

$$X_1, X_2, \ldots, X_n \sim p_0$$

Here, we use an information-theoretic interpretation of the loss as the cross entropy between the model's probability distribution and the true distribution p_0.

$$H(p_0, p_A) = H(p_0) + D_{KL}(p_0 \parallel p_A) \tag{6}$$

(Barnett & Besiroglu, 2023) identifies $D_{KL}(p_0 \parallel p_A) = C^{-\alpha}$ with the non-irreducible part of the loss function.

H_A is the hypothesis that A is the better model. Likewise, H_B represents the hypothesis that B is the better model. Let Z_i represent the log-likelihood ratio per token given by: $Z_i = \log \frac{p_A(x_i)}{p_B(x_i)}$. The cumulative log likelihood ratio is given by:

$$Z_N = \sum_{i=1}^{N} Z_i = \log\left(\prod_{i=1}^{N} \frac{p_A(X_i)}{p_B(X_i)}\right), i = 1, 2, \ldots \tag{7}$$

Next, we choose two threshold values, A_{th} and B_{th}. We choose to accept H_A if $Z_N \geq log(A_{th})$ and we accept H_B if $Z_N \leq log(B_{th})$. These thresholds can be chosen such that the probability of falsely rejecting H_A is α while the probability of falsely rejecting H_B is β. The threshold values that have such properties are (Nowak, 2011):

$$A_{th} = \frac{1-\beta}{\alpha}, B_{th} = \frac{\beta}{1-\alpha} \tag{8}$$

Now, we can use Wald's stopping theorem to find N, the expected number of tokens necessary to differentiate the better model. Wald's stopping theorem holds under stationary and ergodic distributions (Franken & Lisek, 1982). See also (Barnett & Besiroglu, 2023).

$$E_{p_0}[Z_N] = E_{p_0}[N] E_{p_0}[Z_i] \tag{9}$$

$$E_{p_0}[Z_i] = E_{p_0}[\log \frac{p_A(x_i)}{p_B(x_i)}] = \\ D_{KL}(p_0 \parallel p_B) - D_{KL}(p_0 \parallel p_A) = \Delta L \tag{10}$$

The proof above also works for stationary and ergodic processes as well, due to ergodicity:

$$\frac{1}{n} Z_n \xrightarrow[n\to\infty]{\text{a.s.}} E_{p_0}[Z_i] \tag{11}$$

We can evaluate the expected value of Z_N.

$$E_{p_0}[Z_N \mid H_A] = (1-\beta) \log A_{th} + \beta \log B_{th} \tag{12}$$

Under H_B it is:

$$E_{p_0}[Z_N \mid H_B] = (1-\alpha) \log B_{th} + \alpha \log A_{th} \tag{13}$$

This means the expected number of tokens necessary to distinguish the two models under each hypothesis is given by:

$$E_{p_0}[N \mid H_A] = \frac{(1-\beta) \log A_{th} + \beta \log B_{th}}{\Delta L} \tag{14}$$

$$E_{p_0}[N \mid H_B] = \frac{(1-\alpha) \log B_{th} + \alpha \log A_{th}}{|\Delta L|}. \tag{15}$$

Finally, we set $\alpha = \beta$ because we want symmetrical considerations. Let us consider the case where H_A corresponds to the larger model (ie, it predicts the text better) and $\Delta L > 0$ as the default. We can then write the expected number of discrimination tokens as:

$$E_{p_0}[N] = \frac{(1-\alpha) \log\left(\frac{1-\alpha}{\alpha}\right) + \alpha \log\left(\frac{\alpha}{1-\alpha}\right)}{\Delta L} \tag{16}$$

A.2. Bayes Slowdown Factor

Similar to (Barnett & Besiroglu, 2023) we experiment with a Bayesian slowdown factor to more closely measure practical distinguishability rather than ideal distinguishability.

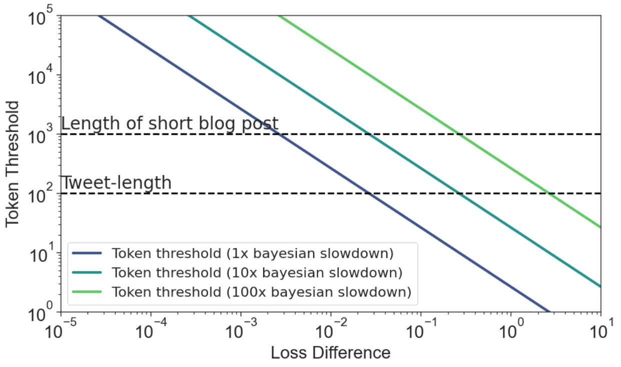


Figure 3: Expected number of tokens needed to differentiate text based on loss using SPRT. Bayesian Slowdown= 1 corresponds to an ideal discriminator.

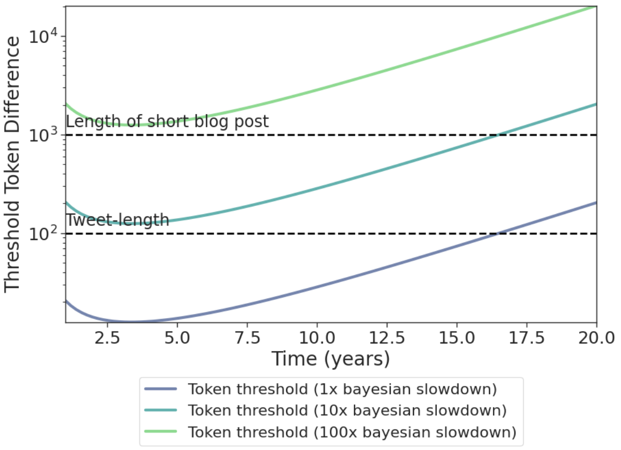


Figure 4: Expected number of tokens needed to differentiate SOTA-model vs meek model using SPRT. Bayesian Slowdown= 1 represents the number of tokens needed by a perfect discriminator.

$$\log(\text{Posterior odds}) = \log(\text{Prior odds}) + \frac{\log(\text{Bayes factor})}{\text{Slowdown}}. \tag{17}$$

Accounting for such a factor scales the number or expected tokens proportionally by the size of the slowdown (see Fig 3 and Fig 4).

A.3. Speculative Sampling Implications

We also think such a model is of independent interest in modeling the gains from speculative sampling. (Chen et al., 2023). Speculative sampling uses a small, cheap draft model to generate most tokens, while a more expensive target model is used for tokens the draft model generates incorrectly. The rise in similarity between models of different sizes motivates the use of techniques like speculative sampling. Our model explains why the target model and draft model are indistinguishable on most tokens. We therefore predict increase usage of techniques like speculative sampling in the future.

B. Benchmark Fit

In order to determine the relationship between loss, we used a framework similar to (Owen, 2024). We fit a sigmoid function to the loss L. We determined the loss for each model using Chinchilla scaling laws from data on parameters, and data for each model from Owen (2024).

$$\text{Benchmark-Performance} = \frac{A}{1 + e^{-k(L-x_0)}} + b \tag{18}$$

For tasks that are composed of multiple steps, each of which depends on benchmark performance, we exponentiate the benchmark performance (which represents the one-shot success rate).

$$\text{p-Benchmark-Performance} = \left(\frac{A}{1 + e^{-k(L-x_0)}} + b\right)^p \tag{19}$$

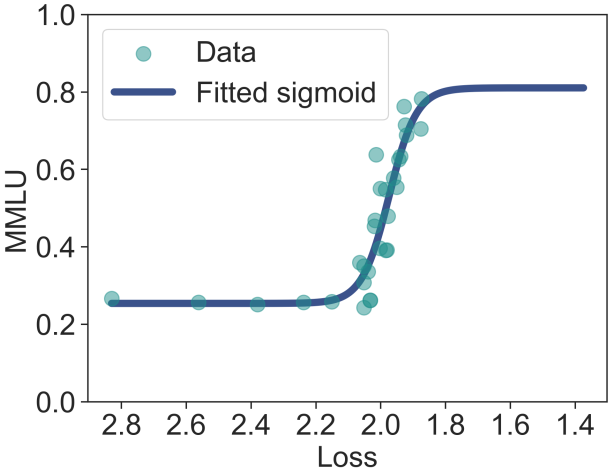


Figure 5: Sigmoid fit of MMLU benchmark performance vs inferred loss.

C. Modeling Zero-Shot Inference Inequality

In many cases, individuals do not want to train their own models but simply want to run inference on a state of the art model. Therefore, we want to model the difference between the performance of models run with a fixed inference budget vs the performance of state of the art model. Here, we present a rough model to do this. We compare against the same SOTA models as before (we do not assume inference compute budget scales exponentially). We use the useful heuristic that the compute used to perform inference is approximately the square root of the training compute (Villalobos & Atkinson, 2023). Therefore, we can take the square of the inference compute available at a given cost to arrive at a rough estimate of the training compute. This is a very rough approximation, but will help us illustrate some of the different dynamics for inference than in the training compute case. Next, we scale the training compute by the growth due to algorithmic progress to get the effective training compute. We label the computer price of inference C_{inf}. The price of a single inference is many orders of magnitude lower than the cost of training an ML model.

The loss for a given level of inference compute is $L_0 + A(g_{alg}^t(g_h^t C_{inf})^2)^{-\alpha}$. However, this loss does not account for algorithmic progress in inference computation, which is separate from algorithmic progress in training. Inference performance benefits from some algorithmic advances in training (i.e., better model architectures), while there are algorithmic advances in training that don't help in inference (i.e., better data processing) and vice versa. We label algorithmic growth in inference g_{inf}. Since we have no known estimates of this value we assume that it is the same as the algorithmic growth in training compute for our analysis. Therefore, our best estimate of inference loss over time for a fixed inference budget is:

$$\text{Meek Inference Loss} = L_0 + A((g_h^t g_{inf}^t C_{inf})^2 g_{alg}^t)^{-\alpha} \tag{20}$$

If we make the assumption that $g_{alg} = g_{inf}$, then inference computation has a larger growth rate than investment in the SOTA model. Therefore, we use the modified loss difference formula equation 21. We choose an initial inference budget of $10^{-8}$$ so that we have some initial SOTA-model builder advantage. Fig 6 illustrates the much faster convergence of loss-difference. Therefore, it may be much harder to monitor inference runs of SOTA models vs monitoring training of SOTA models. However, this only captures the difference in generating individual tokens. This does not account for larger players investing more in inference scaling, which allows them to increase performance with majority voting or longer reasoning (Wu et al., 2024).

$$\max(\text{Meek Inference Loss-SOTA Loss}, 0) = \max(A(\left((g_h^t g_{inf}^t C_{inf})^2\, g_{alg}^t\right)^{-\alpha} - ((g_{alg}\, g_h\, g_i)^t C_0)^{-\alpha}), 0) \tag{21}$$

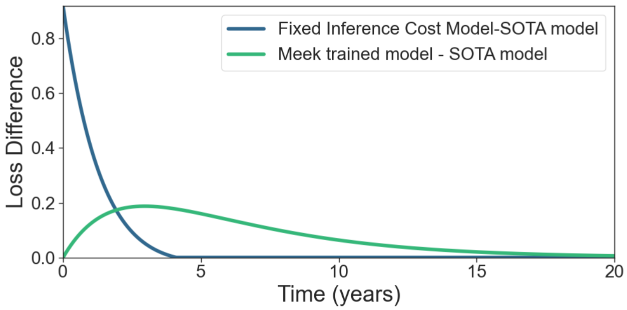


Figure 6: Graph of loss difference in inference vs training performance. The inference difference is between a SOTA model and a model that can be run with an example fixed inference budget. For comparison, we have the training loss difference between the SOTA model and the meek model with a fixed training budget.

D. Robustness and Variation Section

D.1. Variations in Model Investment Trends

The graphs we have presented are based on several key assumptions. These assumptions are that growth in compute investment, algorithms, and hardware will consistently continue. In this section, we want to highlight what will happen to measures of AI inequality if growth in Hardware, Algorithms, or Investment decreases, stalls, or increases. Fig 7 illustrates the greater and longer loss-advantage/capability-differential possible with larger growth in compute investment. This has both positive and negative effects on AI safety. Further, this centralization effect/advantage-duration has diminishing returns as the growth rate increases.

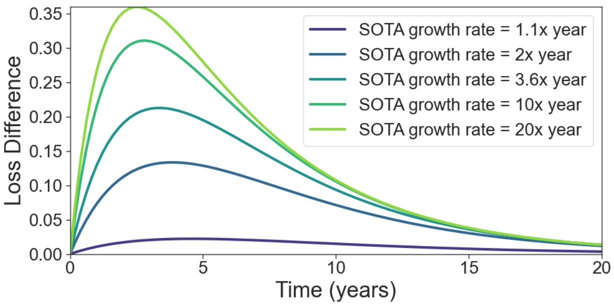


Figure 7: Loss difference between SOTA and meek models with different levels of SOTA compute investment growth g_i.

Fig 8 illustrates variation in our SOTA model builder's initial compute capital. Variation in initial capital has little the effect in the long-run in our model.

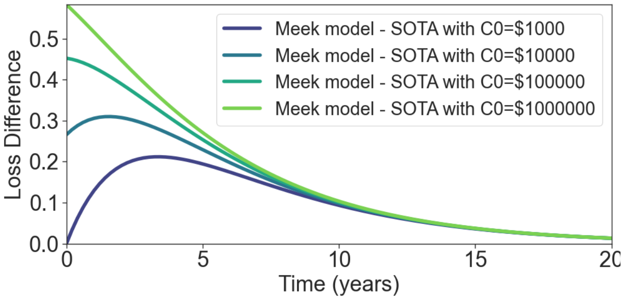
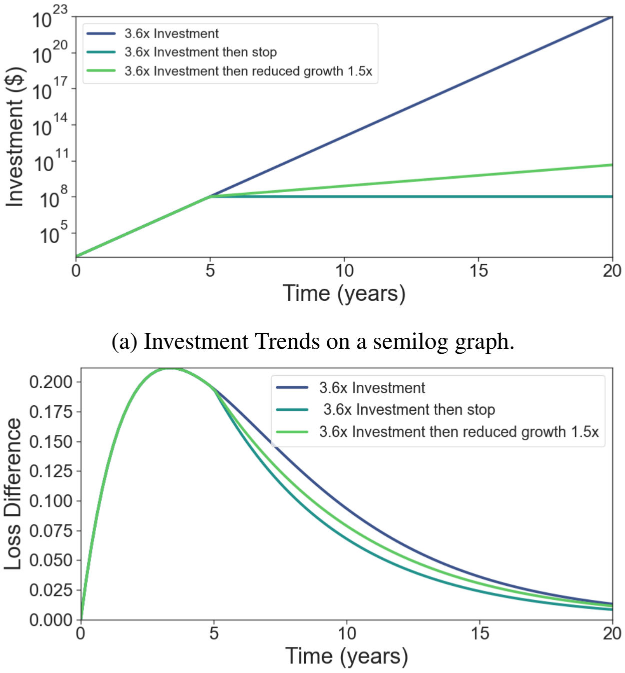


Figure 8: Initial compute capital advantage makes little difference in loss over time. Meek model training budget is kept constant at 1000$.

We might be interested in how our model changes in response to AI training growth slowing or coming to a halt after five years. This could be due to energy-limits or lack of remaining training data (Sevilla et al., 2024). Fig 9 illustrates that such stagnation has little effect on the loss difference in our framework, as AI builders are already in a regime with steep diminishing returns to loss.

Finally, there is the possibility that scaling laws could have significant variations. For instance, the difference between the scaling exponent estimated by Besiroglu et al. (2024) and the exponent estimated in the original chinchilla paper (Hoffmann et al., 2022) as described in Pearce & Song (2024). The difference between these two situations is represented in Fig 10.



(a) Investment Trends on a semilog graph.

(b) Loss trends based on investment schedules in Fig 9a

Figure 9: Model investment growth trajectories vs loss-difference. Surprisingly, large exponential variation in investment trajectory leads to little change in loss.

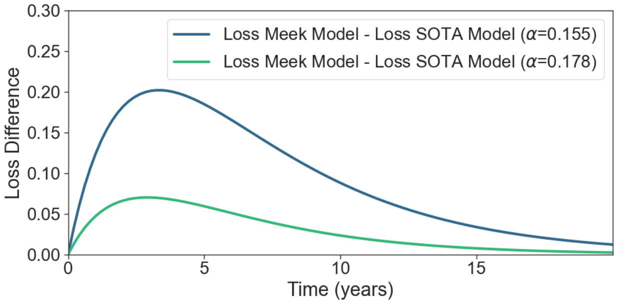


Figure 10: Differences in effect between original chinchilla exponent $\alpha = 0.154$ and the reanalyzed exponent proposed by Besiroglu et al. (2024).