# Anton Korinek

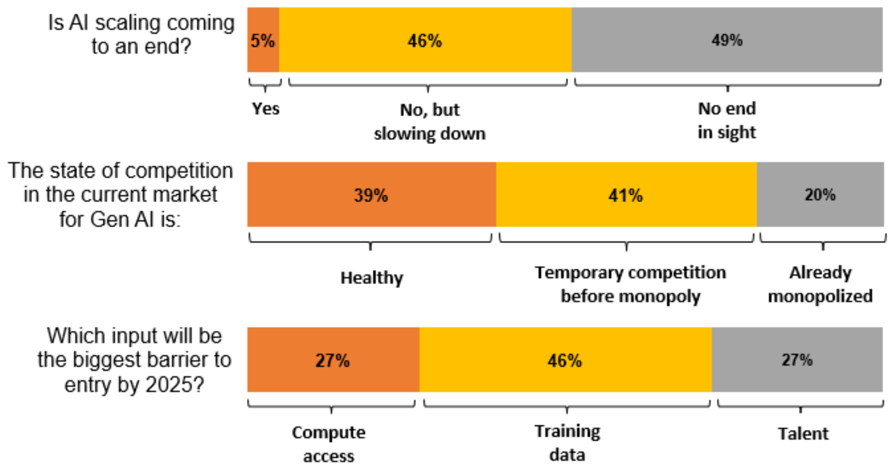
# AI Scaling Laws and Market Structure

On Tuesday, November 26, Anton Korinek joined Markus' Academy for a conversation on AI Scaling Laws and Market Structure. Anton Korinek is a Professor of Economics at the University of Virginia and a Visiting Fellow at Brookings.

A few highlights from the discussion.

- **A summary in three bullets**
  - The effective computing power employed to train AI models is growing by a factor of 10x per year. Investing has been enabled by AI's scaling laws, whereby performance improves predictably from additional computing power
  - There are concerns that AI models are hitting a "data wall," but further model improvements could come from synthetic data, more efficient training algorithms, and scaling outside of the training stage
  - The LLM market is highly competitive, but high training costs, network effects, and vertical integration between LLM developers and chips manufacturers raise concerns about future market concentration

- **[0:00] Markus' introduction and poll questions**
  - There is much debate about AI's impact on TFP growth: Acemoglu (2024) estimates a boost of less than 0.71% over the next 10 years in total, while Goldman Sachs (2023) projects it will bring 1.5% in productivity growth annually
  - We can bring insights from the economics of platforms to understand the IO of AI. The sector also has a layered structure, with base LLMs and under them a more competitive layer of bespoke technology
  - As investors rush to pick winners, winner-take-all economies may lead to extreme investment booms, similar to the dot-com bubble. Mispricing could be more severe if most of the funding is private (as companies cannot be shorted)



- [9:47] **The AI Scaling laws**
    - The primary driver of AI progress over the past 15 years has been the exponential growth in computational power ("compute"), with total compute used to train models increasing 4x per year.
    - On top of the computational power we have also seen exponential growth in algorithmic efficiency (2.5x per year)
    - These two effects entail that the effective amount of computing power employed is growing by a factor of 10x per year
    - Two AI scaling “laws” have established predictable relationships in the benefits of scaling, enabling labs to anticipate the returns from investments in computational power
    - The first law governs how computing power improves model intelligence during the training stage. Model intelligence (what labs produce) is a function of the number of parameters in the model and the amount of training data
    - Recent experience suggests that the production function is close to Cobb Douglas in its inputs, with an optimal ratio of 1 parameter per 20 tokens of training data (for example with GPT-3.5 inefficiently deviating from this)
    - Consistently, doubling compute leads to proportional improvements in model accuracy; however, due to diminishing marginal returns, the absolute gains in performance decline as compute scales
    - The second law governs the ability of models to do inference and reason. If models are given more computing power (and thus more time) during usage, the quality of their output improves
    - By some estimates, the most expensive models last year cost hundreds of millions of dollars to train. An extrapolation of the current trend in training costs suggests that they could reach $10bn per model in 2026 and $1tr per model by 2030 (Korinek and Vipra, 2024)
    - If scaling laws stop holding and output begins to disappoint, required investments may no longer justify returns, potentially leading to an AI winter
    - There are concerns that AI models are hitting a “data wall,” where improvements in quality are constrained by the lack of more training data. However, this is unlikely to be a game changer because of several countervailing forces

        (1) The use of synthetic data<br>
        (2) More efficient training algorithms (the 2.5x growth mentioned before), for example algorithms that prioritize higher-quality data during training<br>
        (3) New scaling paradigms: labs could scale other things besides the training compute, for example using more computing power when the model produces output

- [27:48] **Snapshot of the AI market**
    - The market is characterized by fierce competition, between AI labs, between the big tech companies backing them, and between countries
    - The LMSYS leaderboard, which ranks AI models based on users’ preferences, provides a good measure of competition among top models.

    Google DeepMind is tied with OpenAI for the top spot, closely followed by xAI, China's 01.AI, and Anthropic (Korinek, 2024)
  - Despite the export controls on chips, China's models have advanced rapidly in the last 6 months. In part this is because the controls have not been watertight, but also because their labs excel at efficiently using their compute
  - Three Chinese models are in the top 10, though it is possible they get a leg up in the ratings because they are optimized for chat interactions
  - Frontier models are improving rapidly, with LMSYS scores increasing and token pricing dropping significantly—by more than 80% in just over a year for GPT-4
  - OpenAI is the clear leader in fundraising, recently raising $6bn at a $157bn valuation. It is followed by xAI and Anthropic, each valued around $40-$50 billion, while other companies are valued in the single billions
  - The market for computing power is highly concentrated, with NVIDIA holding a 92% market share and achieving 75% gross profit margins. Although Google's production of its own chips for internal use may reduce NVIDIA's effective market share, NVIDIA's chips are still considered superior

- [39:43] **Where is this all going?**
  - Scaling alone will lead to significant progress. However, the declared goal of frontier AI labs is to develop an Artificial General Intelligence (AGI), which could perform any cognitive task done by humans
  - Progress in scaling and AGI could enable recursive self-improvement, where AI models enhance their own capabilities, triggering an intelligence explosion
  - One key lesson from the past year is that combining AI with robotics can significantly enhance robots' capabilities. The scope for physical automation would grow rapidly if an AGI is achieved
  - Geogg Hinton recently said: "I have suddenly switched my views on whether these things are going to be more intelligent than us. I now predict 5 to 10 years." On the other hand Yann LeCun argues that our pets are still smarter than frontier AI
  - Acemoglu's (2024) estimates for AI's impact on GDP growth are only likely if we consider models like GPT-3.5, but overall, there is Knightian uncertainty around the impact of AI
  - Amidst this uncertainty the best approach is scenario planning (Korinek, 2023), with three possible scenarios: (1) Business as usual: with a productivity boom akin to the internet boom, (2) gradual advancement towards AGI within 10-20 years, (3) AGI within 2-5 years
  - Rapid advances in AGI would turbo-charge growth but eventually depress wages due to labor substitution. The AI labs may not capture much of the gains but there would be large rents in the market for computing power

- [47:54] **Concentration concerns**
  - Concentration in the AI industry raises concerns about income concentration, the political influence of AI companies, their societal impact, and their potential to gain geopolitical power

  - As a result there has been a lot of antitrust action, for example with a a joint statement by the FTC, DOJ, CMA and the EU
  - Several factors may drive concentration in the sector. Firstly, the high costs of training foundation models may lead to natural monopolies
  - Secondly, as we saw with social media, fierce competition in the early stages may primarily serve to consolidate market shares, with network effects eventually leading to market tipping. However, in AI network externalities are weaker and are mainly driven by the ability to collect user feedback
  - Further market tipping may come from an "intelligence feedback loop" (Korinek and Vipra, 2024); for example, 25% of all the code used by Google is already AI generated
  - Lastly there is potential for upstream vertical integration between labs and computing power producers: we are already seeing NVIDIA strategically allocating chips, while Amazon and Microsoft are working on proprietary chips
  - We could also see downstream integration in products, with Google and Microsoft already incorporating AI into existing services
  - The tradeoff between concentration and safety risks is evident with open-source models. While they lower entry barriers (by removing the need for market participants to train foundation models), they can be exploited for malicious purposes
  - Before publishing their open-source Lamma, Meta tries to address safety concerns through additional instruction fine-tuning and reinforcement learning. However anyone can undo this layer at a relatively small cost (< million dollars)

- [1:04:58] **Regulation and conclusion**
  - Foundation models may soon have many of the characteristics of utilities. Regulation should ensure appropriate data governance, prevent entrenched monopolies, and ensure interoperability
  - Excessive reliance on a few models could exacerbate systemic risks, where cyberattacks or errors could propagate throughout the economy
  - Korinek would be in favor of slightly slowing down AI progress, but it is hard to do because of the competition. Even if Google and OpenAI agreed to slow down, competitors abroad would push ahead
  - A global agreement would be required. To make it enforceable countries would have to be able to keep track of the compute (Anderljung et al., 2023); in a second step countries could jointly promote research on AI alignment and responsible development
  - Should we pull the plug on AI when "AI agents" develop ways to communicate with each other beyond human understanding, as Eric Schmidt has argued? Already with embeddings we are at the stage of black boxes. As AI becomes more integrated in the economy it will become impossible to unplug, in the same way it is impossible to stop electricity without major social & economic turmoil

**Timestamps:**
[9:47] **The AI Scaling laws**
[27:48] **Snapshot of the AI market**