CSET CENTER for SECURITY and EMERGING TECHNOLOGY

# Summary of *Scaling AI: Cost and Performance of AI at the Leading Edge*

While simply scaling up the sizes of models, datasets, and compute budgets has driven much of the recent progress in artificial intelligence (AI), various pressures are acting against this trend. Training larger models is expensive as is operating them when deployed. And there can be diminishing returns to increasing scale for many AI capabilities.

In the face of these pressures, it seems that demand for the largest models is slowing. The computing power to train the largest models grew steeply throughout most of the 2010's but appears to be climbing at a reduced rate through the start of the present decade. There also appears to be limited demand for the very largest openly available models. Open source developers overwhelmingly download smaller models than the largest and most powerful ones available from the public model repository Hugging Face.

Scaling up models still provides a viable path to further progress. Even with diminishing returns, small jumps in technical performance may unlock valuable or risky capabilities that either justify further investment or precautionary intervention. But there are alternative paths to improved performance in addition to simply scaling up investments. Innovative developers can, in some cases, make smaller models with similar or better performance. They can also be more inventive about how they use the models that already exist.

A detailed understanding of these trends and alternative possible trajectories allows decision-makers to understand the available policy levers. It allows them to allocate budgets, determine computing's relative importance in AI progress, and perhaps even anticipate some of the risks and opportunities from continued progress in AI.

## For more information:

- Download the report: https://cset.georgetown.edu/publication/scaling-ai
- Contact: John Bansemer (john.bansemer@georgetown.edu).