# History of Scaling Laws in AI Research

← HOME

JUMP TO: 2026 2025 2024 2023 2022 2020 2019 2017 2016 2015 2014 2012 2009 2001 HOME

RESEARCH PAPERS & FINDINGS | INDUSTRY, COMPUTE & COMMENTARY

Newest first ↓

## 2026

Nov. 10, 2026
CROWN »
**Ben Buchanan & Tantum Collins, *The Bitter Struggle: Superintelligence, Superpowers, and the Fate of the World***
"The scaling laws have lent a predictability to AI progress that other scientific fields lack. Even if why scaling works remains unknown, the fact of its success has instilled in developers the confidence needed to underwrite the construction of larger and larger infrastructure, trusting each generation of models will be better than the last."

"Scaling increases not just the domain-specific capability of AI systems but also their *generality*. . . . Scale does not just make systems better at individual tasks; counterintuitively, it makes them better at almost everything."

*Ben Buchanan & Tantum Collins, The Bitter Struggle 24, 27 (2026).*

Oct. 6, 2026
FARRAR, STRAUS & GIROUX »
**Kevin Roose, *The AGI Chronicles***
OpenAI's scaling laws, used to develop GPT-3, were "clearly a major breakthrough."

*Kevin Roose, The AGI Chronicles 69 (2026).*

## 2025

2025
ACM SIGKDD EXPLORATIONS »
**Chen et al., "A Survey of Scaling in Large Language Model Reasoning"**
"Scaling has played a pivotal role in the development of state-of-the-art LLMs such as GPT-4, ... enabling [it] to generalize across a broad range of tasks with unprecedented accuracy and fluency."

*Zihan Chen et al., 28 ACM SIGKDD Explorations Newsletter 63, 63 (2025).*

Nov. 25, 2025
DWARKESH PODCAST (YOUTUBE) »
**Ilya Sutskever on the Dwarkesh Podcast**
"Pre-training was the thing to scale. It was a particular scaling recipe. The big breakthrough of pre-training is the realization that this recipe is good."

*Interview with Ilya Sutskever by Dwarkesh Patel.*

▶ Watch from 18:58 on YouTube ↗

2025
STRIPE PRESS »
**Dwarkesh Patel & Gavin Leech, *The Scaling Era: An Oral History of AI, 2019–2025***
"So far, the (exponentially) more compute and data you put in, the more intelligence you get out. The effect is so clear and so important that I call the period since 2016 the scaling era of AI."

*Dwarkesh Patel & Gavin Leech, The Scaling Era 22 (2026).*

## 2024

Nov. 12, 2024
YOUTUBE »
**Dario Amodei explains scaling laws**
"We very much got the informal sense that the more data and the more compute and the more training you put into these models the better they perform."

*Remarks by Dario Amodei.*

▶ Watch on YouTube ↗

## 2023

Aug. 8, 2023
DWARKESH PODCAST (YOUTUBE) »
**Dwarkesh Podcast, "Dario Amodei (Anthropic CEO) - Scaling, Alignment, & AI Progress"**
Amodei explains that scaling laws are an empirical fact, but on why scaling works so smoothly, "the truth is that we still don't know."

*Interview with Dario Amodei by Dwarkesh Patel.*

▶ Watch from 1:01 on YouTube ↗

## 2022

**NOV. 30, 2022 · CHATGPT LAUNCHES (GPT-3.5) ↗**

May. 29, 2022
DEEPMIND (ARXIV) »
**Hoffmann et al., "Training Compute-Optimal Large Language Models" (Chinchilla)**
Model size and training tokens should be scaled in equal proportion, so most large models were undertrained; 70B-parameter Chinchilla, trained on 1.4 trillion tokens, beats 280B-parameter Gopher on the same compute.

*Jordan Hoffmann et al., arXiv 1, 8, 12 (2022).*
Cited 5,979 times (Google Scholar, Nov. 2026)

## 2020

Oct. 28, 2020
OPENAI (ARXIV) »
**Henighan et al., "Scaling Laws for Autoregressive Generative Modeling"**
"We identify empirical scaling laws for the cross-entropy loss in four domains: generative image modeling, video modeling, multimodal image↔text models, and mathematical problem solving. In all cases autoregressive Transformers smoothly improve in performance as model size and compute budgets increase, following a power-law plus constant scaling law."

*Tom Henighan et al., arXiv 1 (2020). Co-authors include Jared Kaplan and Dario Amodei.*
Cited 783 times (Google Scholar, Nov. 2026)

May 28, 2020
OPENAI (ARXIV) »
**Brown et al., "Language Models Are Few-Shot Learners" (GPT-3)**
OpenAI trains eight models from 125 million to 175 billion parameters to test how few-shot performance changes with scale, sizing them using the Kaplan scaling laws.

*Tom B. Brown et al., arXiv 6-9 (2020). Co-authors include Ilya Sutskever and Dario Amodei.*
Cited 86,091 times (Google Scholar, Nov. 2026)

Jan. 23, 2020
OPENAI (ARXIV) »
**Kaplan et al., "Scaling Laws for Neural Language Models"**
Language-model loss falls as a smooth power law in model size, dataset size and training compute, across more than seven orders of magnitude.

*Jared Kaplan et al., arXiv 3-8 (2020). Co-authors include Dario Amodei.*
Cited 9,769 times (Google Scholar, Nov. 2026)

May 28, 2020
GWERN.NET (ESSAY) »
**Gwern Branwen, "The Scaling Hypothesis"**
"As increasing computational resources permit running such algorithms at the necessary scale, the neural networks will get ever more intelligent."

*Gwern Branwen's essay on GPT-3, begun the day the GPT-3 paper appeared; later revised and followed by "Scaling Hypothesis Revisited" (2022).*

## 2019

Feb. 14, 2019
OPENAI »
**Radford et al., "Language Models are Unsupervised Multitask Learners" (GPT-2)**
"Our approach motivates building as large and diverse a dataset as possible in order to collect natural language demonstrations of tasks in as varied of domains and contexts as possible."

"When a large language model is trained on a sufficiently large and diverse dataset it is able to perform well across many domains and datasets."

*Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei & Ilya Sutskever, OpenAI 1, 10 (2019).*
Cited 25,649 times (Google Scholar, Nov. 2026)

Mar. 13, 2019
RICH SUTTON (ESSAY) »
**Rich Sutton, "The Bitter Lesson"**
"[B]reakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning."

*Rich Sutton argues 70 years of AI research show general methods that leverage computation beat approaches built on human knowledge.*

## 2017

Dec. 1, 2017
BAIDU RESEARCH (ARXIV) »
**Hestness et al., "Deep Learning Scaling is Predictable, Empirically"**
"We empirically validate that DL model accuracy improves as a power-law as we grow training sets for state-of-the-art (SOTA) model architectures in four machine learning domains: machine translation, language modeling, image processing, and speech recognition. These power-law learning curves exists [*sic*] across all tested domains, model architectures, optimizers, and loss functions."

*Joel Hestness et al., arXiv 13 (2017).*
Cited 1,665 times (Google Scholar, Nov. 2026)

2017
DARIO AMODEI (OPENAI MEMO) »
**Dario Amodei, "The Big Blob of Compute Hypothesis"**
"[N]eural nets generalize better when you simply expose them to a wider distribution of data that forces them to find a general solution rather than just memorizing a few special cases."

*Dario Amodei, The Big Blob of Compute Hypothesis 7 (2017). Internal memo written while Amodei was at OpenAI; excerpt published by Kevin Roose ("The Essay That Started the AI Race") Oct. 6, 2026.*

July 30, 2017
GOOGLE RESEARCH & CMU (ARXIV) »
**Sun, Shrivastava, Singh & Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era"**
"Performance [of computer vision] increases logarithmically based on volume of training data. We find there is a logarithmic relationship between performance on vision tasks and the amount of training data used for representation learning."

*Chen Sun, Abhinav Shrivastava, Saurabh Singh & Abhinav Gupta, arXiv 2 (2017) (v2, Aug. 4, 2017). Trained on Google's JFT-300M dataset of 300 million images.*
Cited 4,206 times (Google Scholar, Nov. 2026)

## 2016

Nov. 18, 2016
MIT PRESS »
**Goodfellow, Bengio & Courville, *Deep Learning***
Textbook says scaling "completely changed what was possible with deep learning."

*Ian Goodfellow, Yoshua Bengio & Aaron Courville, Deep Learning 19 (2016).*
Cited 506,738 times (Google Scholar, Nov. 2026)

## 2015

Dec. 8, 2015
BAIDU RESEARCH (ARXIV) »
**Amodei et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin"**
"Scaling Data[:] Our English and Mandarin corpora are substantially larger than those commonly reported in speech recognition literature. . . . This implies that a speech system [for speech recognition] will continue to improve with more labeled training data."

*Dario Amodei et al., arXiv 16 (2015).*
Cited 4,545 times (Google Scholar, Nov. 2026)

May 27, 2015
NATURE »
**LeCun, Bengio & Hinton, "Deep Learning"**
Review predicts deep learning will "take advantage of increases in the amount of available computation and data," and that new methods "will only accelerate this progress."

*Yann LeCun, Yoshua Bengio & Geoffrey Hinton, 521 Nature 436 (2015). Published online May 27, 2015; print issue May 28.*
Cited 123,744 times (Google Scholar, Nov. 2026)

## 2014

Dec. 19, 2014
BAIDU RESEARCH (ARXIV) »
**Hannun et al., "Deep Speech: Scaling up end-to-end speech recognition"**
"In this paper, we describe an end-to-end speech system, called 'Deep Speech', where deep learning supersedes these processing stages. Combined with a language model, this approach achieves higher performance than traditional methods on hard speech recognition tasks while also being much simpler. These results are made possible by training a large recurrent neural network (RNN) using multiple GPUs and thousands of hours of data."

*Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates & Andrew Y. Ng, arXiv 1 (2014).*
Cited 3,230 times (Google Scholar, Nov. 2026)

2014
ILYA SUTSKEVER (YOUTUBE) »
**Ilya Sutskever, "The Deep Learning Hypothesis"**
"If you have a hard problem, we just need to take a big deep neural network and we'll solve it if you have a lot of data as well."

*Talk by Ilya Sutskever.*

▶ Watch from 2:36 on YouTube ↗

## 2012

Dec., 2012
UNIV. OF TORONTO (NEURIPS 2012) »
**Krizhevsky, Sutskever & Hinton, "ImageNet Classification with Deep Convolutional Neural Networks" (AlexNet)**
"All of our experiments suggest that our results can be improved simply by waiting for faster GPUs and bigger datasets to become available."

*Alex Krizhevsky, Ilya Sutskever & Geoffrey E. Hinton, 25 Advances in Neural Information Processing Systems 1097 (2012). AlexNet won the ImageNet challenge with a top-5 error of 15.3%, vs. 26.2% for the runner-up.*
Cited 160,752 times (Google Scholar, Nov. 2026)

## 2009

June 2009
IEEE CVPR (PRINCETON) »
**Deng, Dong, Socher, Li, Li & Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database"**
"We hope that the scale, accuracy, diversity and hierarchical structure of ImageNet can offer unparalleled opportunities to researchers in the computer vision community and beyond."

*Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li & Li Fei-Fei, 2009 IEEE Conf. on Computer Vision & Pattern Recognition 248, 249 (2009). Fei-Fei Li led the project; ImageNet later supplied the training data for AlexNet (2012).*
Cited 105,018 times (Google Scholar, Nov. 2026)

June 2009
STANFORD (ICML 2009) »
**Raina, Madhavan & Ng, "Large-scale Deep Unsupervised Learning Using Graphics Processors"**
"[S]caling up existing DBN [deep belief network] and sparse coding models to use more parameters, or more training data, might produce very significant performance benefits."

*Rajat Raina, Anand Madhavan & Andrew Y. Ng, Proceedings of the 26th Int'l Conf. on Machine Learning, Montreal, Canada, 2008, at 1.*
Cited 1,377 times (Google Scholar, Nov. 2026)

Mar. 2009
GOOGLE (IEEE INTELLIGENT SYSTEMS) »
**Halevy, Norvig & Pereira, "The Unreasonable Effectiveness of Data"**
"[I]nvariably, simple models [for speech recognition] and a lot of data trump more elaborate models based on less data."

*Alon Halevy, Peter Norvig & Fernando Pereira, The Unreasonable Effectiveness of Data, 24 IEEE Intelligent Systems 8, 8 (2009).*
Cited 2,995 times (Google Scholar, Nov. 2026)

## 2001

July 2001
MICROSOFT RESEARCH (ACL 2001) »
**Banko & Brill, "Scaling to Very Very Large Corpora for Natural Language Disambiguation"**
"We have shown that for a prototypical natural language classification task, the performance of learners can benefit significantly from much larger training sets."

*Michele Banko & Eric Brill, Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, at 7 (2001).*
Cited 1,164 times (Google Scholar, Nov. 2026)