# Trends in Landmark Machine Learning Model Training

Christopher Denq, Robi Rahman, Lauren Yu

## Background

Since the 1970s, there has been a trend of biennial doubling of the computing power of integrated circuit chips, colloquially known as Moore's law. With the invention of the internet in 1983, the world's data began to accumulate at an exponential pace[1], beginning with text and tabular data and expanding to encompass images, audio, video, and more, from billions of different users and sources.

Artificial neural networks were proposed as early as 1943, but the paradigm was largely ineffective and stagnant due to computational bottlenecks, until the invention of the backpropagation algorithm in 1975 and the advent of parallel distributed processing in the 1980s allowed ANNs to be trained efficiently on useful tasks.[2] In recent decades, developers have sought to create increasingly sophisticated models, aided by increases in available data and processing power, and new AI capabilities have emerged as the scale of these models has increased.[3] For example, emergent abilities in language models include detecting figures of speech, transliterating sounds between languages, and unscrambling words whose letters have been shuffled.[4]

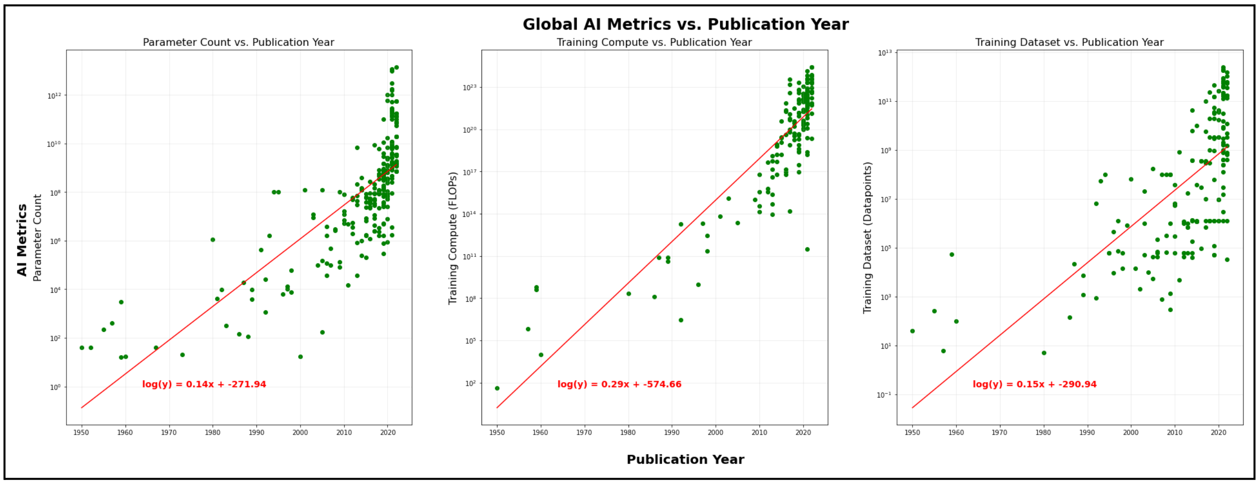
The AI research organization [Epoch] has created a [database of landmark machine learning systems] developed since 1950, selected for meeting notability criteria[5] and indicating their number of parameters, training dataset sizes, and amount of compute used to train them.[6] By examining the associated publications, we find the number of authors of each nationality who contributed to the development of each model. This allows statistical analysis of the trends in model training that are happening in each country. The United States has historically led the development of machine learning models based on neural networks, but other countries are rapidly expanding their technology sectors and investing in R&D, so the gap is quickly closing.

## Historical Trends

Using the whole Epoch machine learning systems dataset, we can determine the historical worldwide average growth rates in our three metrics of choice: parameter count, training compute (in floating-point operations), and training dataset size (in datapoints). As seen below, the compute used in training of landmark ML models echoes the global trend in processor manufacturing (increasing by a factor of 10 every 3.45 years), with parameter count

1 https://www.statista.com/statistics/871513/worldwide-data-created/
2 https://en.wikipedia.org/wiki/Artificial_neural_network#History
3 https://openreview.net/pdf?id=yzkSU5zdwD
4 https://arxiv.org/pdf/2206.04615.pdf
5 Uniqueness and relevance to ML, as well as major historical influence, improvement on the state of the art, and/or having over 1000 citations
6 https://epochai.org/blog/compute-trends

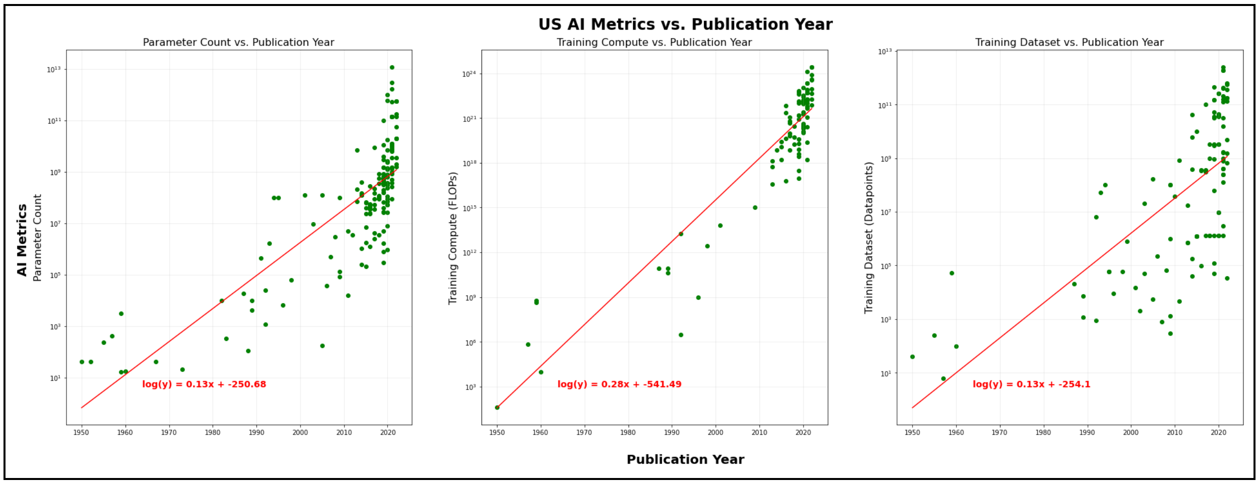
and dataset size following a slower but similar exponential trajectory (x10 every 7.14 and 6.67 years, respectively).




| Country | Metric | Equation Format | Growth Rate | Intercept | Correlation (R^2) | P-Value | Standard Error | Num of Papers | Earliest Publication | Latest Publication |
|---|---|---|---|---|---|---|---|---|---|---|
| World | Parameter Count | log(y) = mx + b | 0.139006 | -2.719407e+02 | 0.592666 | 2.175640e-54 | 0.007026 | 271 | 1950-07-02 | 2022-11-08 |
| | | y = 10^(mx) * 10^(b) | 1.377229 | 1.146199e-272 | 0.592666 | 2.175640e-54 | 0.007026 | 271 | 1950-07-02 | 2022-11-08 |
| | Training Compute (FLOPs) | log(y) = mx + b | 0.294649 | -5.742015e+02 | 0.784516 | 7.821675e-47 | 0.013291 | 137 | 1950-07-02 | 2022-11-08 |
| | | y = 10^(mx) * 10^(b) | 1.970830 | 0.000000e+00 | 0.784516 | 7.821675e-47 | 0.013291 | 137 | 1950-07-02 | 2022-11-08 |
| | Training Dataset (Datapoints) | log(y) = mx + b | 0.152855 | -2.994864e+02 | 0.562124 | 5.338152e-20 | 0.013358 | 104 | 1950-07-02 | 2022-09-21 |
| | | y = 10^(mx) * 10^(b) | 1.421852 | 3.262983e-300 | 0.562124 | 5.338152e-20 | 0.013358 | 104 | 1950-07-02 | 2022-09-21 |

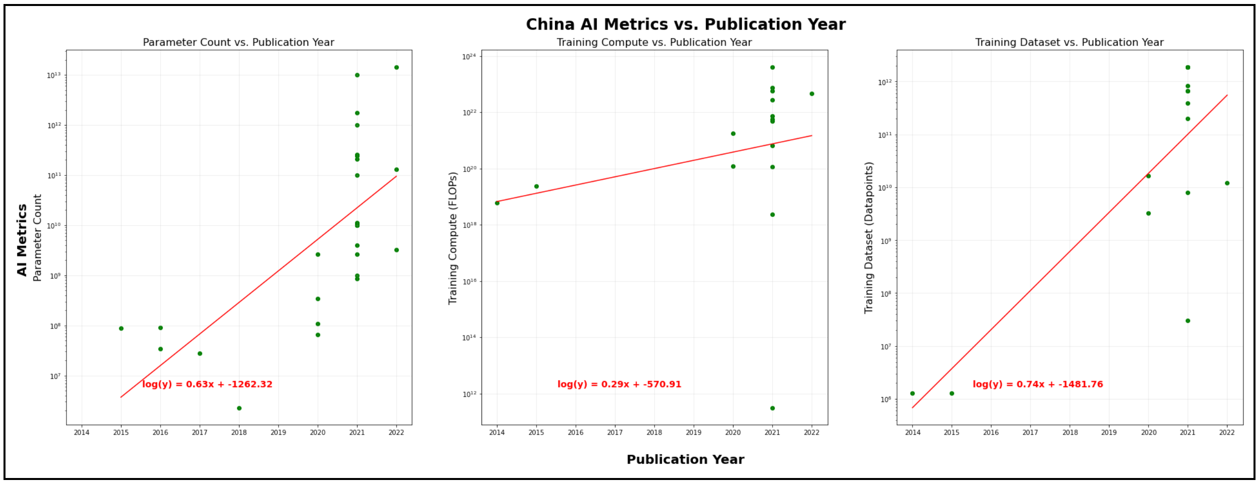
We can then examine subsets of the data corresponding to the machine learning systems which had contributing authors from particular countries, to determine the trends in ML development in those countries and compare them to the global trend. While our analysis encompassed all 31 countries home to authors of publications in the Epoch dataset, we will highlight the results from the six with the most publications: the United States, United Kingdom, Canada, China, Germany, and Switzerland.

Starting with results for the United States, its growth rates are not far off from the world average. This is unsurprising as the bulk of historical AI publications have come from US institutions, and many of the international collaborations have American coauthors. Notably, the variance in residuals was substantially smaller within the trend in compute than in the trends in parameter counts and dataset sizes; of all three, compute had the strongest correlation with publication date. This implies that compute usage more closely follows the steady historical development of technology over the past several decades than do the other factors.




| Country | Metric | Equation Format | Growth Rate | Intercept | Correlation (R^2) | P-Value | Standard Error | Num of Papers | Earliest Publication | Latest Publication |
|---|---|---|---|---|---|---|---|---|---|---|
| United States | Parameter Count | log(y) = mx + b | 0.128462 | -2.506810e+02 | 0.635534 | 2.532260e-39 | 0.007439 | 173 | 1950-07-02 | 2022-09-21 |
| | | y = 10^(mx) * 10^(b) | 1.344195 | 2.084728e-251 | 0.635534 | 2.532260e-39 | 0.007439 | 173 | 1950-07-02 | 2022-09-21 |
| | Training Compute (FLOPs) | log(y) = mx + b | 0.280669 | -5.457238e+02 | 0.852551 | 1.402388e-34 | 0.013132 | 81 | 1950-07-02 | 2022-09-21 |
| | | y = 10^(mx) * 10^(b) | 1.908400 | 0.000000e+00 | 0.852551 | 1.402388e-34 | 0.013132 | 81 | 1950-07-02 | 2022-09-21 |
| | Training Dataset (Datapoints) | log(y) = mx + b | 0.136233 | -2.657964e+02 | 0.608430 | 1.895528e-14 | 0.013769 | 65 | 1950-07-02 | 2022-09-21 |
| | | y = 10^(mx) * 10^(b) | 1.368463 | 1.598192e-266 | 0.608430 | 1.895528e-14 | 0.013769 | 65 | 1950-07-02 | 2022-09-21 |

When comparing US trends with those of China, we see a stark contrast in historical parameter count and training dataset size growth rates. Specifically, China's parameter counts and dataset sizes have been growing faster than those of the US by factors of 4.85 and 5.69 respectively. On the other hand, Chinese compute growth is on par with the American and global rates. Out of all 31 countries represented in the dataset, China has had the fastest historical growth rate in parameter count (x10 every 1.59 years) and in training dataset size (x10 every 1.36 years). In compute growth rates, China ranks 8th out of 31 – behind the UK, Germany, and Switzerland but ahead of Canada and the US. (Complete results are available in the hyperlinked spreadsheet.) This asymmetric level of growth makes sense in light of the ongoing geopolitical circumstances: China has been focusing on increasing its own domestic AI output using ever-larger datasets and models with billions of parameters, but has been limited by its dependency on US-sourced semiconductor chips. With the recent CHIPS Act soon coming into effect for the US semiconductor industry, and new export controls expected to reduce Chinese semiconductor production, it is not clear how long the Chinese technology industry can continue with growth rates at or near the highest in the world.

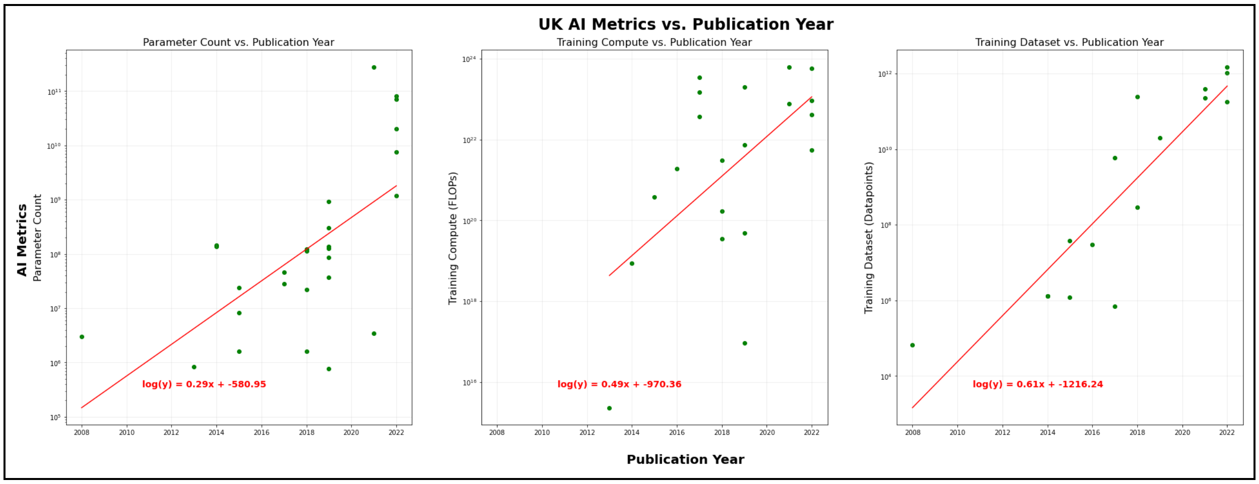



| Country | Metric | Equation Format | Growth Rate | Intercept | Correlation (R^2) | P-Value | Standard Error | Num of Papers | Earliest Publication | Latest Publication |
|---|---|---|---|---|---|---|---|---|---|---|
| China | Parameter Count | log(y) = mx + b | 0.629719 | -1262.315079 | 0.462259 | 0.000096 | 0.135838 | 27 | 2015-01-09 | 2022-08-04 |
| | | y = 10^(mx) * 10^(b) | 4.263034 | 0.000000 | 0.462259 | 0.000096 | 0.135838 | 27 | 2015-01-09 | 2022-08-04 |
| | Training Compute (FLOPs) | log(y) = mx + b | 0.285829 | -556.786508 | 0.024204 | 0.579812 | 0.503351 | 15 | 2015-01-09 | 2022-08-04 |
| | | y = 10^(mx) * 10^(b) | 1.931206 | 0.000000 | 0.024204 | 0.579812 | 0.503351 | 15 | 2015-01-09 | 2022-08-04 |
| | Training Dataset (Datapoints) | log(y) = mx + b | 0.790246 | -1586.284608 | 0.546501 | 0.014557 | 0.254513 | 10 | 2015-01-09 | 2021-12-23 |
| | | y = 10^(mx) * 10^(b) | 6.169438 | 0.000000 | 0.546501 | 0.014557 | 0.254513 | 10 | 2015-01-09 | 2021-12-23 |

Turning to the United Kingdom, we find that in every metric, its historical growth rates have been faster than that of the global average. Most notably, the UK's exponential growth in compute and dataset size has been outpacing the global averages by factors of approximately 1.69 and 4.07, respectively. Out of the 31 countries sampled, the UK has had the 2nd-fastest historical growth rates in dataset size (x10 every 1.65 years), 5th-fastest in compute (x10 every 2.03 years), and 3rd-fastest in parameter count (x10 every 3.43 years). This suggests that the UK may become a serious contender for the international AI race in the near future. London is home to DeepMind, currently one of the leading artificial intelligence research labs, so a reasonable expectation is for the UK to continue producing a large share of landmark ML systems in the near future. The British government has emphasized this as a national priority, with recent major investments into AI, large language models, and supercomputing.[7][8]

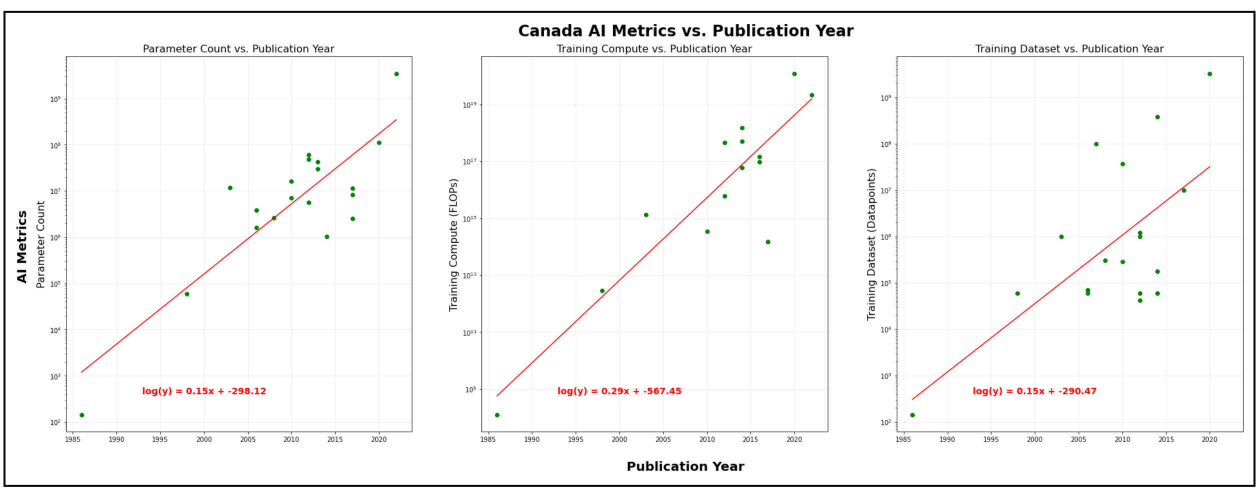
7 https://www.theguardian.com/technology/2023/mar/15/uk-to-invest-900m-in-supercomputer-in-bid-to-build-own-britgpt

8 https://www.gov.uk/government/news/government-commits-up-to-35-billion-to-future-of-tech-and-science



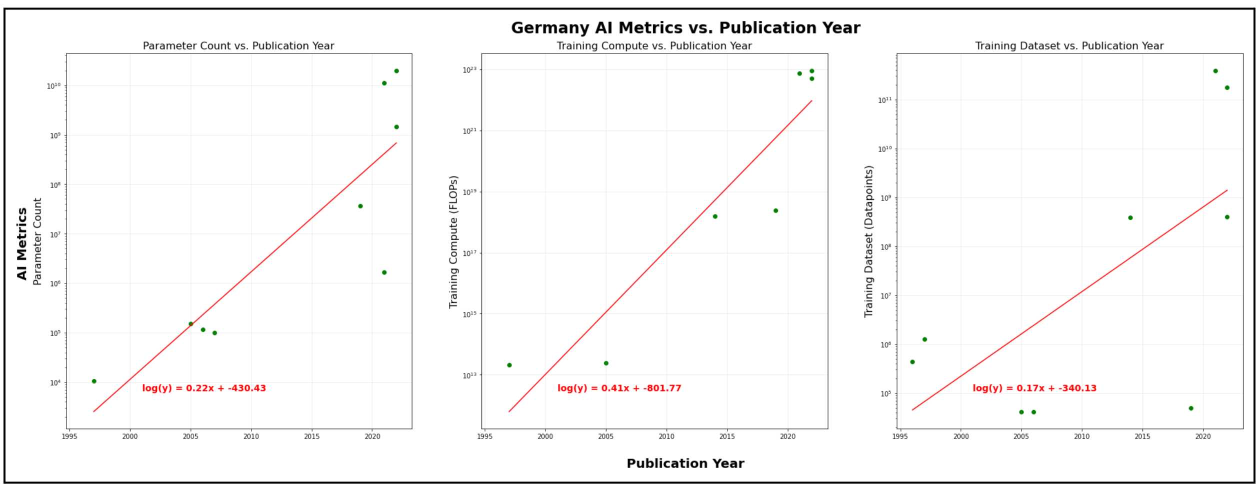

| Country | Metric | Equation Format | Growth Rate | Intercept | Correlation (R^2) | P-Value | Standard Error | Num of Papers | Earliest Publication | Latest Publication |
|---|---|---|---|---|---|---|---|---|---|---|
| United Kingdom | Parameter Count | log(y) = mx + b | 0.291892 | -580.951824 | 0.401517 | 0.000388 | 0.071273 | 27 | 2008-07-15 | 2022-05-12 |
| | | y = 10^(mx) * 10^(b) | 1.958359 | 0.000000 | 0.401517 | 0.000388 | 0.071273 | 27 | 2008-07-15 | 2022-05-12 |
| | Training Compute (FLOPs) | log(y) = mx + b | 0.591668 | -1173.279820 | 0.435702 | 0.005392 | 0.179959 | 16 | 2013-01-01 | 2022-05-12 |
| | | y = 10^(mx) * 10^(b) | 3.905420 | 0.000000 | 0.435702 | 0.005392 | 0.179959 | 16 | 2013-01-01 | 2022-05-12 |
| | Training Dataset (Datapoints) | log(y) = mx + b | 0.624103 | -1249.884250 | 0.765254 | 0.002019 | 0.130648 | 9 | 2014-09-04 | 2022-03-29 |
| | | y = 10^(mx) * 10^(b) | 4.208262 | 0.000000 | 0.765254 | 0.002019 | 0.130648 | 9 | 2014-09-04 | 2022-03-29 |

Next, examining Canada, we see that it has similar growth rates to the US, albeit with proportionately fewer publications (i.e. roughly the same quantity per capita). Canada has, on average between our three metrics, some of the strongest correlations between time and exponential growth out of all countries, and fewer publications early in the recorded period but more of them in the past two decades. This suggests that Canada has potential to become a more active player in the international AI race.



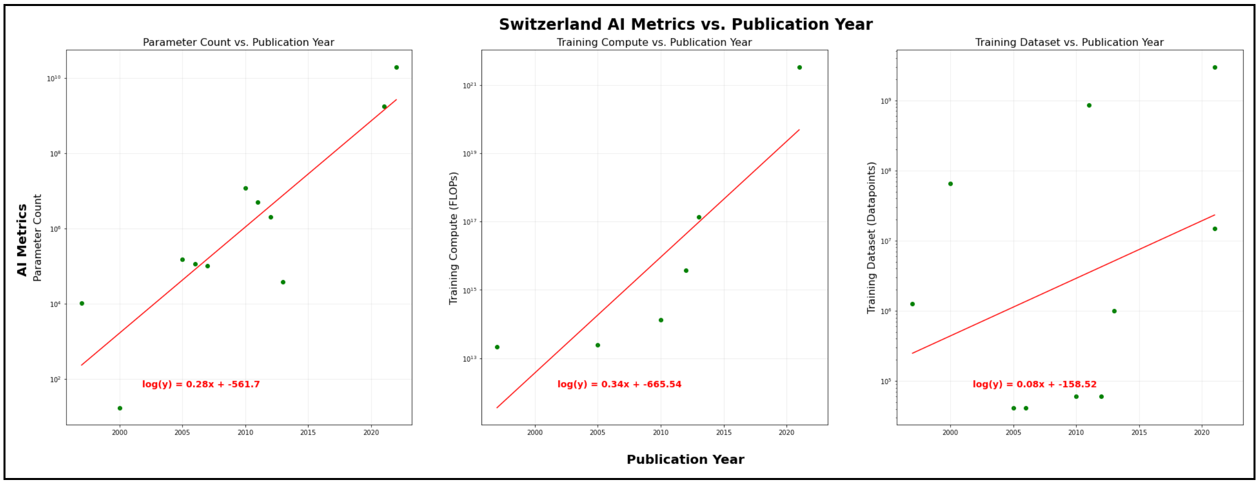
| Country | Metric | Equation Format | Growth Rate | Intercept | Correlation (R^2) | P-Value | Standard Error | Num of Papers | Earliest Publication | Latest Publication |
|---|---|---|---|---|---|---|---|---|---|---|
| Canada | Parameter Count | log(y) = mx + b | 0.151661 | -2.981199e+02 | 0.726794 | 0.000004 | 0.022552 | 19 | 1986-10-01 | 2022-04-14 |
| | | y = 10^(mx) * 10^(b) | 1.417949 | 7.587753e-299 | 0.726794 | 0.000004 | 0.022552 | 19 | 1986-10-01 | 2022-04-14 |
| | Training Compute (FLOPs) | log(y) = mx + b | 0.285356 | -5.579979e+02 | 0.805657 | 0.001023 | 0.052972 | 9 | 1986-10-01 | 2022-04-14 |
| | | y = 10^(mx) * 10^(b) | 1.929107 | 0.000000e+00 | 0.805657 | 0.001023 | 0.052972 | 9 | 1986-10-01 | 2022-04-14 |
| | Training Dataset (Datapoints) | log(y) = mx + b | 0.165089 | -3.255502e+02 | 0.760105 | 0.010528 | 0.041477 | 7 | 1986-10-01 | 2020-11-23 |
| | | y = 10^(mx) * 10^(b) | 1.462475 | 0.000000e+00 | 0.760105 | 0.010528 | 0.041477 | 7 | 1986-10-01 | 2020-11-23 |

Lastly, looking at Germany and Switzerland, we find that their growth trends in parameter count and compute power are similar to those in the UK, but their growth in training dataset sizes is much slower. All of the publications from these countries are in English and none involve language modeling of German text, so this is not an artifact of slower growth in German-language web content. This may suggest that UK researchers have been atypically successful in scaling up recent models to train on large datasets, consistent with the recent finding[9] by DeepMind (a British AI lab) that optimizing models for performance requires higher ratios of compute and data to parameters than most leading developers have been using.




| Country | Metric | Equation Format | Growth Rate | Intercept | Correlation (R^2) | P-Value | Standard Error | Num of Papers | Earliest Publication | Latest Publication |
|---|---|---|---|---|---|---|---|---|---|---|
| Germany | Parameter Count | log(y) = mx + b | 0.217242 | -430.429191 | 0.754861 | 0.002362 | 0.046792 | 9 | 1997-11-15 | 2022-04-13 |
| | | y = 10^(mx) * 10^(b) | 1.649080 | 0.000000 | 0.754861 | 0.002362 | 0.046792 | 9 | 1997-11-15 | 2022-04-13 |
| | Training Compute (FLOPs) | log(y) = mx + b | 0.407085 | -801.065332 | 0.872363 | 0.006390 | 0.077857 | 6 | 1997-11-15 | 2022-04-13 |
| | | y = 10^(mx) * 10^(b) | 2.553199 | 0.000000 | 0.872363 | 0.006390 | 0.077857 | 6 | 1997-11-15 | 2022-04-13 |
| | Training Dataset (Datapoints) | log(y) = mx + b | 0.179371 | -353.501908 | 0.371069 | 0.199287 | 0.116761 | 6 | 1997-11-15 | 2022-04-13 |
| | | y = 10^(mx) * 10^(b) | 1.511371 | 0.000000 | 0.371069 | 0.199287 | 0.116761 | 6 | 1997-11-15 | 2022-04-13 |

[9] https://arxiv.org/pdf/2203.15556.pdf




| Country | Metric | Equation Format | Growth Rate | Intercept | Correlation (R^2) | P-Value | Standard Error | Num of Papers | Earliest Publication | Latest Publication |
|---|---|---|---|---|---|---|---|---|---|---|
| Switzerland | Parameter Count | log(y) = mx + b | 0.282456 | -5.617010e+02 | 0.765947 | 0.000418 | 0.052046 | 11 | 1997-11-15 | 2022-05-10 |
| | | y = 10^(mx) * 10^(b) | 1.916269 | 0.000000e+00 | 0.765947 | 0.000418 | 0.052046 | 11 | 1997-11-15 | 2022-05-10 |
| | Training Compute (FLOPs) | log(y) = mx + b | 0.339051 | -6.655369e+02 | 0.760262 | 0.023553 | 0.095197 | 6 | 1997-11-15 | 2021-06-08 |
| | | y = 10^(mx) * 10^(b) | 2.182988 | 0.000000e+00 | 0.760262 | 0.023553 | 0.095197 | 6 | 1997-11-15 | 2021-06-08 |
| | Training Dataset (Datapoints) | log(y) = mx + b | 0.126063 | -2.473845e+02 | 0.306289 | 0.254605 | 0.094859 | 6 | 1997-11-15 | 2021-06-08 |
| | | y = 10^(mx) * 10^(b) | 1.336789 | 4.125514e-248 | 0.306289 | 0.254605 | 0.094859 | 6 | 1997-11-15 | 2021-06-08 |

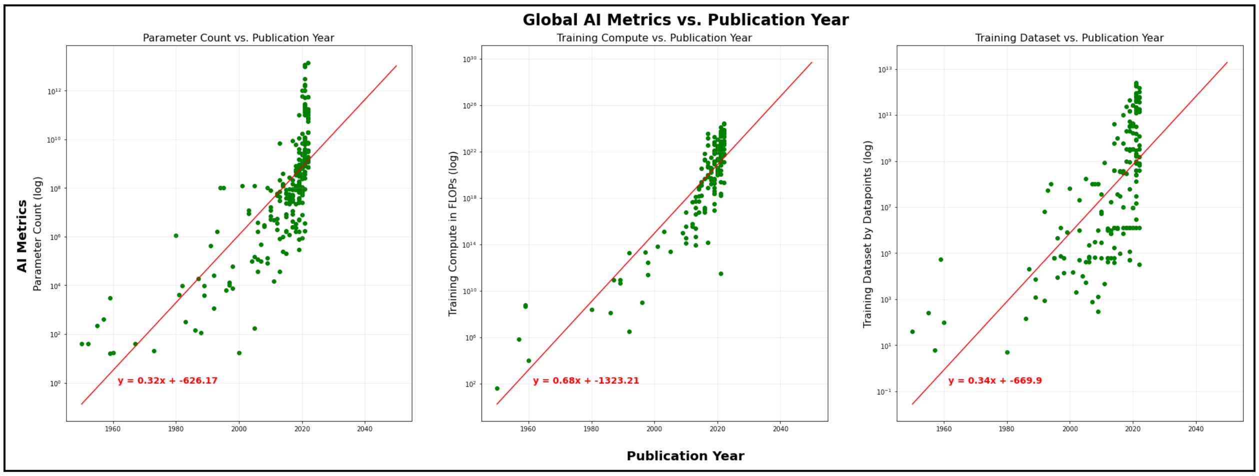
Regression coefficients for all 31 countries in the dataset are provided in a table[10] in exponential and logarithmic formats, but when interpreting these, consider the following caveat. Diagnostic plots[11] of the linear regressions demonstrate that the residuals exhibit heteroskedasticity, with older data points having more variance and more outliers before 2010. The landmark ML systems are following the exponential growth trend much more closely in recent years than before 2010, in a pattern consistent with Epoch's finding that there have been different "eras" of machine learning development with different rates of growth. Therefore, it should be possible to improve prediction accuracy by partitioning the dataset by publication year and extrapolating different trend lines for different time periods of research.

---

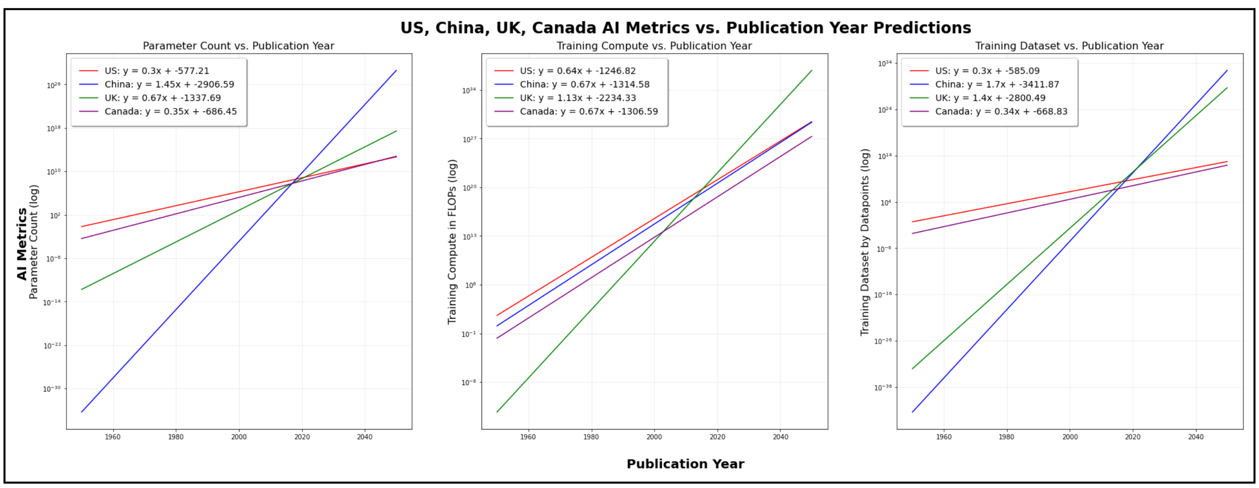
10 https://docs.google.com/spreadsheets/d/1FR6h_e4NexHjtt1GXw-YKYbLGsMBqpy-x0CUE_u9Yuk/edit?usp=sharing

11 Regression diagnostic base code from Python package statsmodel: https://www.statsmodels.org/dev/examples/notebooks/generated/linear_regression_diagnostics_plots.html

# Projections and Conclusion




Using the semi-log graphs derived from Epoch's dataset, we compared the predictions for future machine learning growth between the 7 countries with the most data points from past ML publications conducted within each country. The trend was graphed from 1950 to 2050 to estimate the growth of each respective country over this 100 year period.




China has outpaced other countries in the parameter count and training dataset metrics of growth, and has a relatively higher rate of growth in training compute per year, only outmatched by the UK. If current growth rates continue, China is projected to have higher training compute metrics than the US by 2045, but as mentioned before, the recent export controls may throw a wrench in China's rapid expansion of ML infrastructure, and the CHIPS Act may provide a one-off bump in US hardware or permanently increase its growth rate, either of which would delay this catch-up point.

The UK also demonstrates a larger growth in metrics than the US and Canada in all departments. The US and Canada appear to have similar development rates, though Canada's growth rates are slightly higher, so its trends intersect the US in compute by 2045 and training data by 2055 if using a linear projection. However, Canada has about 1/10th the population of the US, so it is more realistic to expect them to catch up in per-capita ML development and taper off at the US growth rate, rather than for Canada to train models with more compute than the US despite having 1/10th as many researchers.

Projections beyond the near future need to take into consideration other factors besides past growth rates. For instance, the cost of AI development cannot exceed national research expenditures, so taking into account the size of each country's technology industry may provide a more realistic estimate for progress made, as well as when the rate of model scaling will taper off. There are several limiting factors that prevent these trends from continuing forever: Moore's law is based on ever-shrinking transistors, but these cannot be made smaller than silicon atoms, and the stock of text and image data is not growing fast enough to keep up with the expansion of ML training datasets[12]. Additionally, as noted earlier, there have been at least three different eras of ML growth in the past 15 years, so it may be fundamentally impractical to project anything farther than a decade away while the whole industry goes through worldwide changes on a timescale shorter than decades. Future investigation should account for the number of researchers working in the academic and technology sectors, availability of training data, and trends in processor performance to determine the bounds of the current regime of growth in these machine learning models.

[12] https://epochai.org/blog/will-we-run-out-of-ml-data-evidence-from-projecting-dataset