CPSC 573 Final Project: Data Cleaning Report

AI Models Dataset from Epoch AI

Zoya Malik

CPSC 573 – Data Analysis, University of Calgary

March 2026

Contents

1 Introduction

1.1 Dataset Source

This report examines the **AI Models** dataset published by Epoch AI, a leading research institute dedicated to tracking and forecasting developments in artificial intelligence. The dataset is publicly available at https://epoch.ai/data/ai-models and is licensed under Creative Commons Attribution 4.0 (CC-BY 4.0).

Citation: Epoch AI, "Data on AI Models," 2026, available at https://epoch.ai/data/ai-models. Licensed under CC-BY 4.0.

1.2 Why This Dataset

As a Computer Science student, it is impossible to ignore the transformation happening around us. Large language models, multimodal systems, and autonomous agents have gone from research curiosities to tools that reshape how software is built, how research is conducted, and how entire industries operate. Yet much of the public conversation about AI is driven by anecdote and hype rather than careful measurement. This dataset offers something different: a rigorous, structured record of over three thousand AI models spanning decades of development, compiled by researchers whose explicit mission is to bring empiricism to AI forecasting.

What draws me to this dataset is the breadth of questions it can answer. The raw numbers on training compute and parameter counts let us trace the exponential scaling laws that have defined the deep learning era, but the dataset goes further. It records which organizations built each model, in which countries, using what hardware, at what cost, and whether the resulting weights were released to the public. These dimensions open the door to questions that sit at the intersection of technology, economics, and policy. These are questions that feel urgent to anyone who will spend their career working alongside (and competing with) these systems.

I am particularly curious about the dynamics that the summary statistics alone cannot reveal. Has the relationship between model size and training compute changed as researchers have pursued efficiency gains? Is the frontier of AI development becoming more geographically concentrated or more distributed? Are open-weight releases gaining ground, or is the trend toward closed, proprietary systems? This project is an opportunity to move beyond impressions and let the data speak, applying the analytical techniques from CPSC 573 to a domain that is both personally fascinating and professionally consequential.

2 Dataset Description

2.1 Overview

The Epoch AI Models dataset is a comprehensive catalogue of artificial intelligence systems, assembled and maintained by researchers at Epoch AI. It is a **primary source**: the data is collected directly by Epoch AI's research team through systematic review of publications, technical reports, and direct communication with model developers. This distinguishes it from aggregated or secondary datasets, as the provenance and methodology behind each data point are documented by the collectors themselves.

The dataset is distributed across four CSV files, each representing a different scope or curation criteria. Together, they provide both breadth (thousands of models across all domains) and depth (detailed resource and performance metrics for the most significant systems).

2.2 File Structure

Table 1: Dataset file structure

File	Rows	Cols	Description
`all_ai_models.csv`	∼3,237	57	The master dataset containing every model tracked by Epoch AI. The most comprehensive file and the superset from which the other files are drawn.
`frontier_ai_models.csv`	∼137	61	A curated subset of frontier (state-of-the-art) models that pushed the boundaries of capability at release. Contains additional columns for detailed cost and hardware utilization metrics.
`large_scale_ai_models.csv`	∼491	34	Models meeting a threshold for scale in training compute, parameters, or dataset size. Uses a focused column set emphasizing resource characteristics.
`notable_ai_models.csv`	∼981	47	Models deemed notable by Epoch AI based on citation impact, historical significance, or influence on subsequent research. Balances breadth with depth of annotation.

The relationship among these files is hierarchical. The `all_ai_models.csv` file is the master record, and the other three files are curated subsets applying different selection criteria (frontier significance, scale, or notability) to highlight models of particular analytical interest. A single model may appear in multiple subsets if it satisfies more than one criterion.

2.3 Key Column Groups

The 57 columns in the master dataset can be organized into four thematic groups:

Model Identity and Context. These columns describe what each model is and where it comes from: `Model` (name), `Domain` (e.g. Language, Vision, Multimodal), `Task`, `Organization`, `Country (of organization)`, `Publication date`, `Authors`, `Reference`, and `Link`. They provide the categorical and temporal dimensions needed for cross-sectional and time-series analysis.

Training Resources. These columns quantify the inputs required to build each model: `Parameters`, `Training compute (FLOP)`, `Training dataset size (total)`, `Training time (hours)`, `Epochs`, `Hardware quantity`, `Training hardware`, `Training compute cost (2023 USD)`, `Numerical format`, `Training power draw (W)`, and hardware utilization metrics (`Hardware utilization (MFU)`, `Hardware utilization (HFU)`). This group is central to studying scaling laws and the economics of AI development.

Performance and Impact. Columns such as `Citations`, `Notability criteria`, `Frontier model`, and `Confidence` capture how each model was received and how it compares to its contemporaries. These allow analysis of what distinguishes impactful models from the broader population.

Accessibility and Openness. Columns including `Model accessibility`, `Open model weights?`, `Training code accessibility`, and `Base model` record how open or closed each model is. These are essential for studying trends in open-source AI and the diffusion of model capabilities.

2.4 Suitability for Analysis

This dataset is well suited for the goals of CPSC 573 for several reasons:

1. **Primary source.** The data is collected and curated by domain experts, providing a high degree of reliability and a clear methodology to evaluate.
2. **Richness.** With 57 columns spanning numeric, categorical, temporal, and boolean types, the dataset supports a wide range of statistical and machine learning techniques, from regression and clustering to classification and time-series analysis.
3. **Temporal and cross-sectional dimensions.** Models span from the earliest neural networks to systems released in 2025, enabling both longitudinal trend analysis and point-in-time comparisons.
4. **Real-world messiness.** The dataset contains missing values, mixed units, free-text fields, and inconsistencies that make data cleaning a substantive exercise rather than a formality.
5. **Relevance.** For a Computer Science student, there is no more directly relevant domain than the one that is actively reshaping the discipline itself.

3 Data Cleaning and Transformation

This section describes the cleaning and transformation steps applied to the Epoch AI Models dataset to produce a unified, analysis-ready dataset.

3.1 Raw Data Overview

The primary dataset (`all_ai_models.csv`) contained **3,237** rows and **57** columns. Three supplementary datasets were also used: `notable_ai_models.csv` (981 rows), `frontier_ai_models.csv` (137 rows), and `large_scale_ai_models.csv` (491 rows).

3.2 Cleaning Steps

Step 1: Whitespace and newline removal. Cleaned 41 string columns; fixed 5,538 cells with embedded newlines. Model names and other text fields contained embedded `\n` characters and excess whitespace that were normalized to single spaces.

Step 2: Deduplication. Removed 15 duplicate entries (kept most complete record). Duplicate model names were identified, and for each set of duplicates, the record with the fewest missing values was retained.

Step 3: Country deduplication. Deduplicated country values from 369 to 211 unique values. Values such as "Australia,Australia" were collapsed to "Australia" while preserving genuinely multi-country entries like "China,United States of America".

Step 4: Primary value extraction. Created `primary_domain` (3,152 non-null), `primary_task` (3,118 non-null), `primary_country` (3,142 non-null). For multi-valued fields (Domain, Task, Country), the first value was extracted into `primary_domain`, `primary_task`, and `primary_country` columns to facilitate grouping and visualization.

Note: these non-null counts reflect the dataset at this intermediate step (3,222 rows), prior to the auxiliary merge in Step 7. After the merge, the final counts in the cleaned file are: ***`primary_domain`*** *3,235 non-null;* ***`primary_task`*** *3,201 non-null;* ***`primary_country`*** *3,225 non-null.*

Step 5: Date parsing. Parsed 3,209 of 3,222 dates successfully; created `year` (3,209 non-null) and `decade` columns. The `Publication date` column was converted to datetime format, and `year` and `decade` columns were derived for temporal analysis.

Step 6: Numeric type conversion. Converted 11 columns to numeric: Parameters: 2,083 non-null; Training compute (FLOP): 1,379 non-null; Training dataset size (total): 1,299 non-null; Training time (hours): 547 non-null; Training compute cost (2023 USD): 224 non-null; Epochs: 791 non-null; Hardware quantity: 837 non-null; Citations: 1,270 non-null; Hardware utilization (MFU): 61 non-null; Hardware utilization (HFU): 25 non-null; Training power draw (W): 765 non-null.

Step 7: Auxiliary dataset merging. Flagged 137 frontier, 906 notable models; filled 0 missing Organization categorizations; appended 83 models from notable/large_scale CSVs. Boolean indicators `is_frontier` and `is_notable` were added. Organization categorization data was backfilled from the notable models dataset where missing in the primary dataset. Models present in the supplementary datasets but absent from the primary dataset were appended.

Step 8: Derived columns. Created: `log_parameters` (2,151 non-null), `log_compute` (1,385 non-null), `is_open_weights` (1,272 True), `org_type` (3,205 non-null). `log_parameters` and `log_compute` provide $\log_{10}$-transformed scales suitable for visualization. `is_open_weights` is a boolean indicator derived from the "Open model weights?" field. `org_type` captures the primary organization type (Industry, Academia, Government, or Research collective) using a majority-vote approach on multi-valued categorizations.

Missing value treatment. No imputation was applied to any column; missing values in numeric and categorical columns are retained as `NaN` throughout. Analyses that rely on sparse fields (e.g. training cost, training time) will therefore operate on the available subset of records rather than the full dataset.

3.3 Cleaning Summary

Final dataset has 3,305 rows and 76 columns (started with 3,237 rows and 57 columns). The cleaned dataset is saved as `all_ai_models_cleaned.csv` and a machine-readable log of all cleaning steps is available in `cleaning_log.json`.

Table 2: Step-by-step cleaning summary

Step	Before	After	Details
strip_whitespace	3,237	3,237	Cleaned 41 string columns; fixed 5,538 cells with embedded newlines
deduplicate	3,237	3,222	Removed 15 duplicate entries (kept most complete record)
clean_country_duplicates	369	211	Deduplicated country values from 369 to 211 unique values
split_csv_fields	3,222	3,222	Created primary_domain (3,152), primary_task (3,118), primary_country (3,142); pre-merge counts
parse_dates	3,222	3,222	Parsed 3,209 of 3,222 dates; created year and decade columns
type_conversion	3,222	3,222	Converted 11 columns to numeric
merge_auxiliary	3,222	3,305	Flagged 137 frontier, 906 notable; appended 83 models from auxiliary CSVs
derived_columns	3,305	3,305	Created log_parameters, log_compute, is_open_weights, org_type
final_summary	3,237	3,305	Final dataset: 3,305 rows, 76 columns

4 Data Profiling

This section presents a comprehensive profile of the *raw* AI Models dataset prior to cleaning. The goal is to characterize the structure, completeness, distributions, and quality of the data that feeds into subsequent analysis.

4.1 Dataset Dimensions

Table 3: Raw dataset dimensions

File	Rows	Columns
all_ai_models	3,237	57
frontier_ai_models	137	61
large_scale_ai_models	491	34
notable_ai_models	981	47

The primary dataset (`all_ai_models.csv`) contains 3,237 models across 57 columns, making it the most comprehensive of the four files. The frontier, large-scale, and notable subsets are curated views with fewer rows but partially overlapping columns.

4.2 Missing-Value Analysis

Out of 3,237 rows, per-column completeness varies dramatically:

Table 4: Per-column completeness (raw dataset)

Column	Non-Null	% Missing
Post-training compute (FLOP)	1	100.0%
Post-training compute notes	1	100.0%
Archived links	22	99.3%
Hardware utilization (HFU)	25	99.2%
Training compute lower bound	26	99.2%
Training compute upper bound	41	98.7%
Training cloud compute vendor	53	98.4%
Hardware utilization (MFU)	61	98.1%
Training data center	66	98.0%
Foundation model	76	97.7%
Utilization notes	88	97.3%
Frontier model	137	95.8%
Batch size notes	197	93.9%
Training chip-hours	215	93.4%
Training compute cost (2023 USD)	224	93.1%
Batch size	251	92.2%
Finetune compute (FLOP)	259	92.0%
Approach	301	90.7%
Finetune compute notes	302	90.7%
Numerical format	339	89.5%
WikiText and Penn Treebank data	350	89.2%
Possibly over 1e23 FLOP	501	84.5%
Training time (hours)	547	83.1%
Training time notes	599	81.5%
Hugging Face developer id	609	81.2%
Base model	672	79.2%
Training power draw (W)	765	76.4%
Epochs	791	75.6%
Notability criteria notes	793	75.5%
Hardware quantity	837	74.1%
Notability criteria	932	71.2%
Training hardware	1,163	64.1%
Citations	1,270	60.8%
Training dataset size (total)	1,323	59.1%
Training compute (FLOP)	1,379	57.4%
Training compute estimation method	1,438	55.6%
Training compute notes	1,613	50.2%
Dataset size notes	1,634	49.5%
Accessibility notes	1,659	48.8%
Parameters notes	1,730	46.6%
Parameters	2,091	35.4%
Training code accessibility	2,296	29.1%
Authors	2,466	23.8%

continued on next page

Column	Non-Null	% Missing
Model accessibility	2,488	23.1%
Open model weights?	2,488	23.1%
Abstract	2,833	12.5%
Reference	3,076	5.0%
Task	3,119	3.6%
Organization categorization	3,137	3.1%
Country (of organization)	3,149	2.7%
Domain	3,153	2.6%
Organization	3,157	2.5%
Link	3,202	1.1%
Publication date	3,219	0.6%
Model	3,237	0.0%
Confidence	3,237	0.0%
Last modified	3,237	0.0%

The high missingness in training cost (∼93%), training time (∼83%), and dataset size (∼59%) reflects the reality that these details are frequently unreported in AI research publications. This has important implications for any analysis relying on these fields.

4.3 Column Classification

Dimensions (categorical): Domain, Task, Organization, Country (of organization), Confidence, Model accessibility, Organization categorization, Frontier model, Open model weights?, Approach, Training hardware, Numerical format.

Measures (numeric): Parameters, Training compute (FLOP), Training dataset size (total), Training time (hours), Training compute cost (2023 USD), Epochs, Hardware quantity, Citations, Hardware utilization (MFU), Hardware utilization (HFU), Training power draw (W).

Other columns: 34 columns including model name, authors, abstract, notes fields, and metadata.

4.4 Descriptive Statistics for Numeric Measures

All numeric measures exhibit extreme right skew, consistent with log-normal distributions spanning many orders of magnitude.

Table 5: Descriptive statistics for numeric measures (raw dataset)

Measure	Count	Mean	Median	Std	Min	Max	Skew	Kurt
Parameters	2,091	1.06×10^{11}	3.00×10^{9}	2.22×10^{12}	10	1.00×10^{14}	43.75	1,964.6
Training compute (FLOP)	1,379	1.64×10^{24}	3.46×10^{21}	1.97×10^{25}	40	5.00×10^{26}	21.08	468.3
Training dataset size	1,299	2.65×10^{12}	3.30×10^{9}	2.42×10^{13}	0	7.99×10^{14}	28.12	903.8
Training time (hours)	547	457.4	144.0	878.9	0.1	9,022.8	4.72	31.37
Training compute cost ($)	224	6,015,290	27,006.5	37,626,531	3.90	387,842,678	8.72	79.04
Epochs	791	561.7	15.00	7,265.3	0	191,400.0	23.82	610.3
Hardware quantity	837	975.2	32.00	7,990.3	1	200,000.0	20.27	475.7
Citations	1,270	4,147.4	409.0	14,666.2	0	215,518.0	8.71	98.30
HW utilization (MFU)	61	0.3715	0.3649	0.1110	0.171	0.560	0.075	-0.905
HW utilization (HFU)	25	0.4909	0.4935	0.1488	0.236	0.927	0.783	1.87
Training power draw (W)	765	771,924	25,518	4,887,626	75.83	109,948,656	16.55	339.6

The extreme skewness and kurtosis values confirm that raw-scale statistics are dominated by outliers. Log-scale analysis is essential for meaningful comparisons.

4.5 Frequency Distributions

4.5.1 Top 15 Organizations by Model Count

Table 6: Top 15 organizations by model count

Organization	Models
OpenAI	108
Google DeepMind	93
Alibaba	93
Google	79
Meta AI	67
NVIDIA	63
Microsoft	45
DeepMind	43
ByteDance	36
DeepSeek	31
Mistral AI	30
Stanford University	29
Facebook AI Research	26
Stability AI	26
IBM	24

4.5.2 Domain Breakdown

Table 7: Model count by domain (raw `Domain` values)

Domain	Models
Language	1415 (44.9%)
Biology	375 (11.9%)
Vision	270 (8.6%)

Domain	Models
Image generation	161 (5.1%)
Speech	121 (3.8%)
Multimodal,Language,Vision	68 (2.2%)
Video	55 (1.7%)
Games	50 (1.6%)
Language,Vision,Multimodal	45 (1.4%)
Robotics	39 (1.2%)
Video,Vision	36 (1.1%)
Other	29 (0.9%)
Multimodal,Vision,Language	25 (0.8%)
Audio	22 (0.7%)
Recommendation	21 (0.7%)
Materials science	16 (0.5%)
Multimodal,Language,Vision,Video	16 (0.5%)
Earth science	15 (0.5%)
Vision,Image generation	15 (0.5%)
3D modeling	14 (0.4%)
Image generation,Vision	13 (0.4%)
Language,Multimodal,Vision	12 (0.4%)
Vision,Language	11 (0.3%)
Mathematics,Language	10 (0.3%)
Mathematics	10 (0.3%)
110 further multi-label combinations each account for <0.5% of models.	

4.5.3 Country Breakdown (Top 15)

Table 8: Top 15 countries by model count (raw values)

Country	Models
United States of America	1,041
China	538
United States of America,United States of America	199
United Kingdom of Great Britain and Northern Ireland	127
China,China	67
Korea (Republic of)	59
Canada	56
United States of America,United States of America,United States of America	55
France	43
China,China,China	40
Germany	37
United States of America,France	35
Japan	33
Russia	27
Hong Kong	20

4.6 Temporal Patterns

4.6.1 Models Published Per Year

Table 9: Models published per year

Year	Models
1950–2009: fewer than 20 models per year; 204 total.	
2010	16
2011	13
2012	18
2013	24
2014	35
2015	36
2016	83
2017	79
2018	87
2019	145
2020	112
2021	198
2022	217
2023	520
2024	946
2025	538

The dataset shows exponential growth in AI model publication, with a sharp acceleration from 2017 onward coinciding with the transformer revolution.

4.7 Cross-File Comparison

- **Shared columns across all 4 files:** 29
- **Total unique columns (union):** 74

Unique columns per file (not found in other files):

- `all_ai_models`: Approach, Archived links, Foundation model, Hugging Face developer id, Last modified, Possibly over 1e23 FLOP, Training cloud compute vendor, Training compute lower bound, Training compute upper bound, WikiText and Penn Treebank data
- `frontier_ai_models`: API prices, Base model compute, Estimated over 1e25 FLOP, FLOP/$, Hardware release date, Maybe over 1e25 FLOP, Model versions (benchmarks), Power per GPU, Training dataset size
- `large_scale_ai_models`: (DEPRECATED) Training dataset size (datapoints)
- `notable_ai_models`: Hardware acquisition cost

Model overlap between files:

Table 10: Model overlap between dataset files

File Pair	Shared Models
all_ai_models & frontier_ai_models	137
all_ai_models & large_scale_ai_models	483
all_ai_models & notable_ai_models	906
frontier_ai_models & large_scale_ai_models	37
frontier_ai_models & notable_ai_models	123
large_scale_ai_models & notable_ai_models	157

4.8 Correlation Analysis (Log-Transformed)

Pearson correlations computed on $\log_{10}$-transformed values for positive entries:

Table 11: Pearson correlations among log-transformed numeric measures

	Params	Comp.	Data	Time	Cost	Epochs	HW qty	Cit.	MFU	HFU	Power
Parameters	1.00	0.89	0.79	0.39	0.74	−0.68	0.65	−0.03	−0.22	−0.11	0.66
Compute	0.89	1.00	0.86	0.65	0.93	−0.61	0.78	0.00	−0.03	0.13	0.81
Dataset	0.79	0.86	1.00	0.64	0.64	−0.71	0.74	0.04	0.04	0.49	0.74
Train time	0.39	0.65	0.64	1.00	0.67	−0.27	0.53	0.20	0.20	0.40	0.51
Cost	0.74	0.93	0.64	0.67	1.00	−0.46	0.87	−0.00	−0.00	0.28	0.87
Epochs	−0.68	−0.61	−0.71	−0.27	−0.46	1.00	−0.50	−0.18	−0.04	−0.47	−0.51
HW quantity	0.65	0.78	0.74	0.53	0.87	−0.50	1.00	0.38	0.38	−0.14	0.99
Citations	−0.03	0.00	0.04	0.20	−0.00	−0.18	0.38	1.00	−0.10	−0.18	0.38
MFU	−0.22	−0.03	0.19	−0.04	−0.10	−0.04	−0.00	−0.10	1.00	0.24	−0.03
HFU	−0.11	0.13	0.49	0.40	0.28	−0.47	−0.14	−0.18	0.24	1.00	−0.30
Train power	0.66	0.81	0.74	0.51	0.87	−0.51	0.99	0.38	−0.03	−0.30	1.00

Parameters, training compute, and dataset size show strong positive correlations on log scale, confirming well-known scaling relationships in AI research.

4.9 Key Findings and Data Quality Issues

1. **Severe missingness in cost and resource columns.** Training cost is missing for ∼93% of models, training time for ∼83%, and dataset size for ∼59%. These gaps limit quantitative analysis of resource requirements.
2. **Extreme value ranges.** Numeric measures span many orders of magnitude (e.g. parameters range from single digits to hundreds of billions). Log-transformation is essential for meaningful statistical analysis.
3. **US dominance.** The United States is the leading country of origin for AI models, followed by China and the United Kingdom.
4. **Language models dominate.** The Language domain accounts for the largest share of models, reflecting the NLP focus of recent AI research.
5. **Exponential growth.** Model publication rates have grown exponentially, with the post-2017 transformer era seeing the steepest increase.
6. **Column schema differences.** The four CSV files share a common core of columns but each has unique fields, requiring careful alignment for cross-file analysis.

4.10 Raw vs. Cleaned Data Comparison

- **Raw:** 3,237 rows, 57 columns
- **Cleaned:** 3,305 rows, 76 columns
- **Net rows added during cleaning:** **+68** (15 removed via deduplication, 83 appended from auxiliary datasets)

Columns with improved completeness after cleaning:

Table 12: Completeness improvements after cleaning

Column	Raw % Missing	Cleaned % Missing	Improvement
Notability criteria	71.2%	69.5%	1.7 pp
Citations	60.8%	59.4%	1.4 pp
Authors	23.8%	23.1%	0.7 pp
Reference	5.0%	4.5%	0.5 pp

5 Visualizations

This section presents 10 annotated charts that characterize the AI Models dataset, covering data quality, temporal trends, scaling patterns, and demographic distributions.

Figure 1: **Missing Data Pattern Across AI Models Dataset.** A heatmap of the null indicator matrix (rows = sampled records, columns = all 57 features). Red cells indicate missing values. Columns such as Training time (83% missing) and Training compute cost (93% missing) show near-total sparsity, while core identifiers (Model, Confidence, Last modified) are fully populated. This pattern motivates imputation and column-drop decisions in the cleaning pipeline.

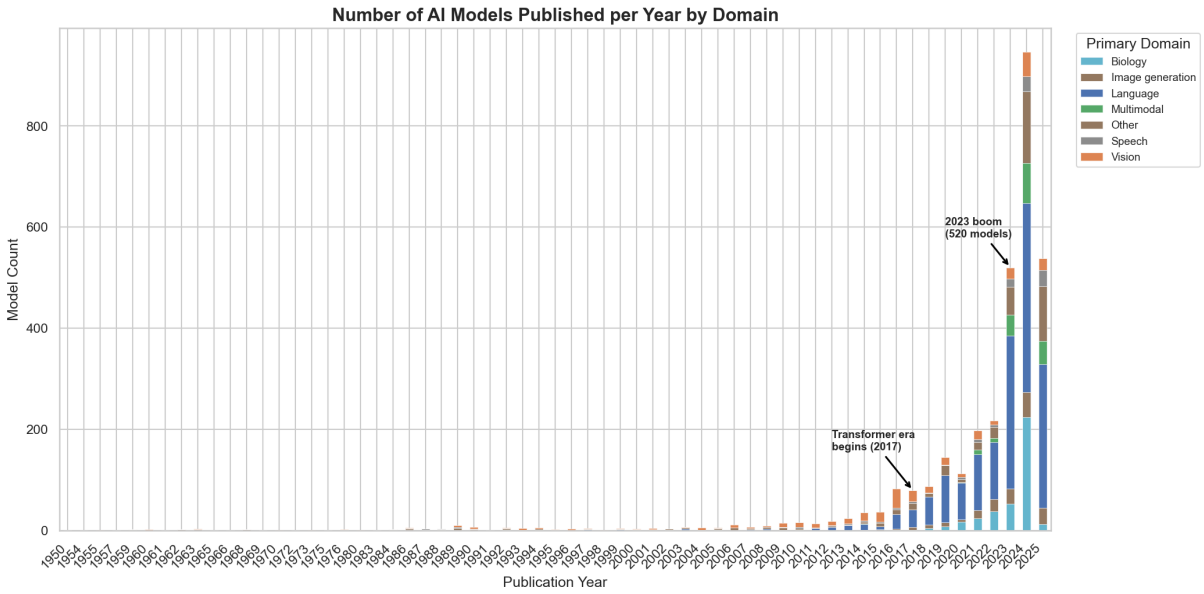

Figure 2: **Number of AI Models Published per Year by Domain.** A stacked bar chart colored by primary domain. Key inflection points are annotated: the post-2017 Transformer era marks a shift toward Language-dominant publications, and the 2023 boom reflects the surge in generative AI releases. Language models consistently dominate, followed by Biology and Vision.

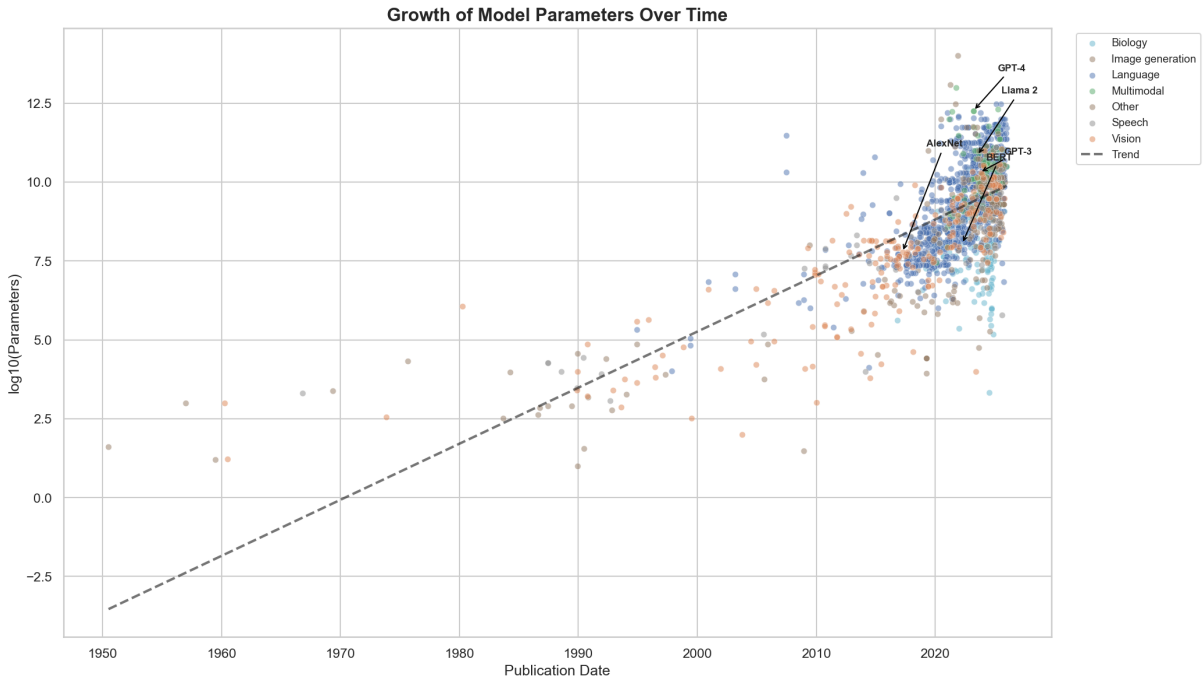

Figure 3: **Growth of Model Parameters Over Time.** A scatter plot of $\log_{10}$(Parameters) versus publication date, colored by domain, with a linear trend line. Landmark models (GPT-3, GPT-4, Llama, BERT, AlexNet) are annotated. The trend demonstrates roughly exponential parameter scaling, with Language and Vision models leading the frontier. A noticeable spread after 2020 reflects the diversification of both massive and efficient models.

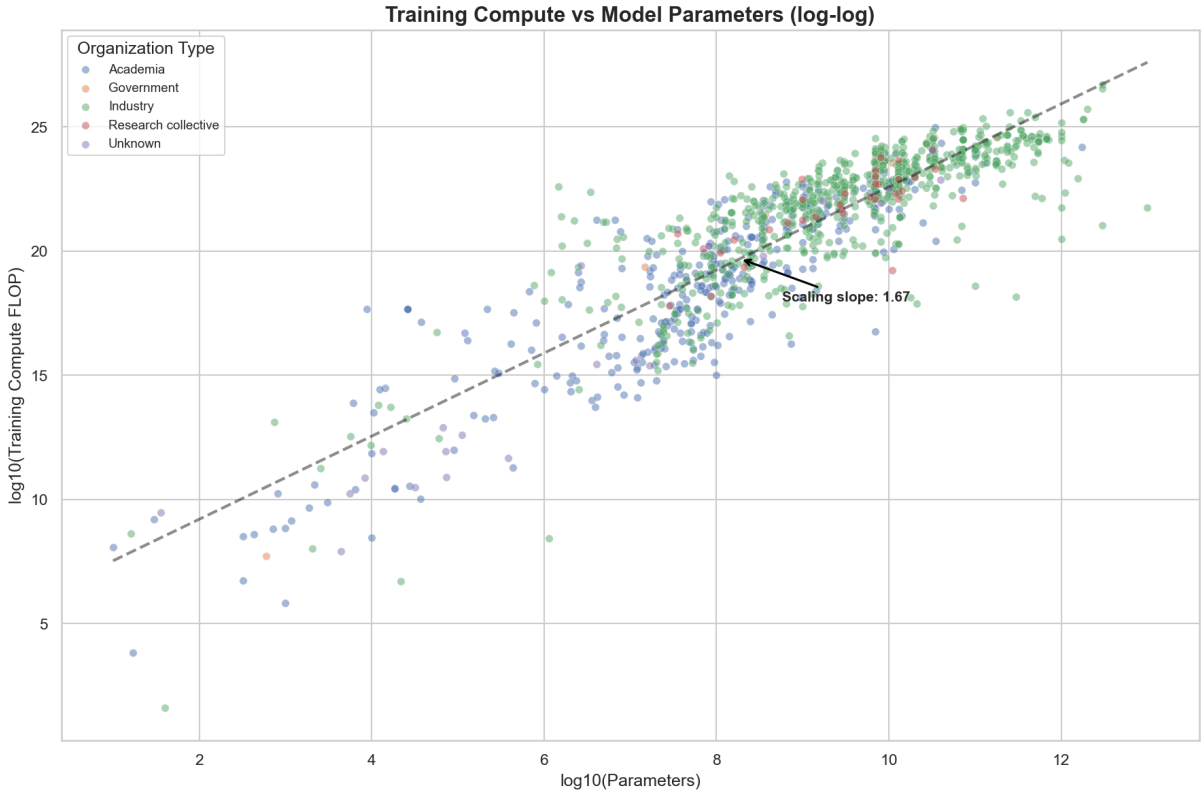

Figure 4: **Training Compute vs. Model Parameters (log–log scale).** Colored by organization type (Industry, Academia, etc.). A fitted trend line reveals the scaling slope between compute and parameters. Industry models cluster at higher compute budgets, while academic models tend toward smaller-scale experiments. The annotated slope quantifies the compute–parameter scaling law observed in practice.

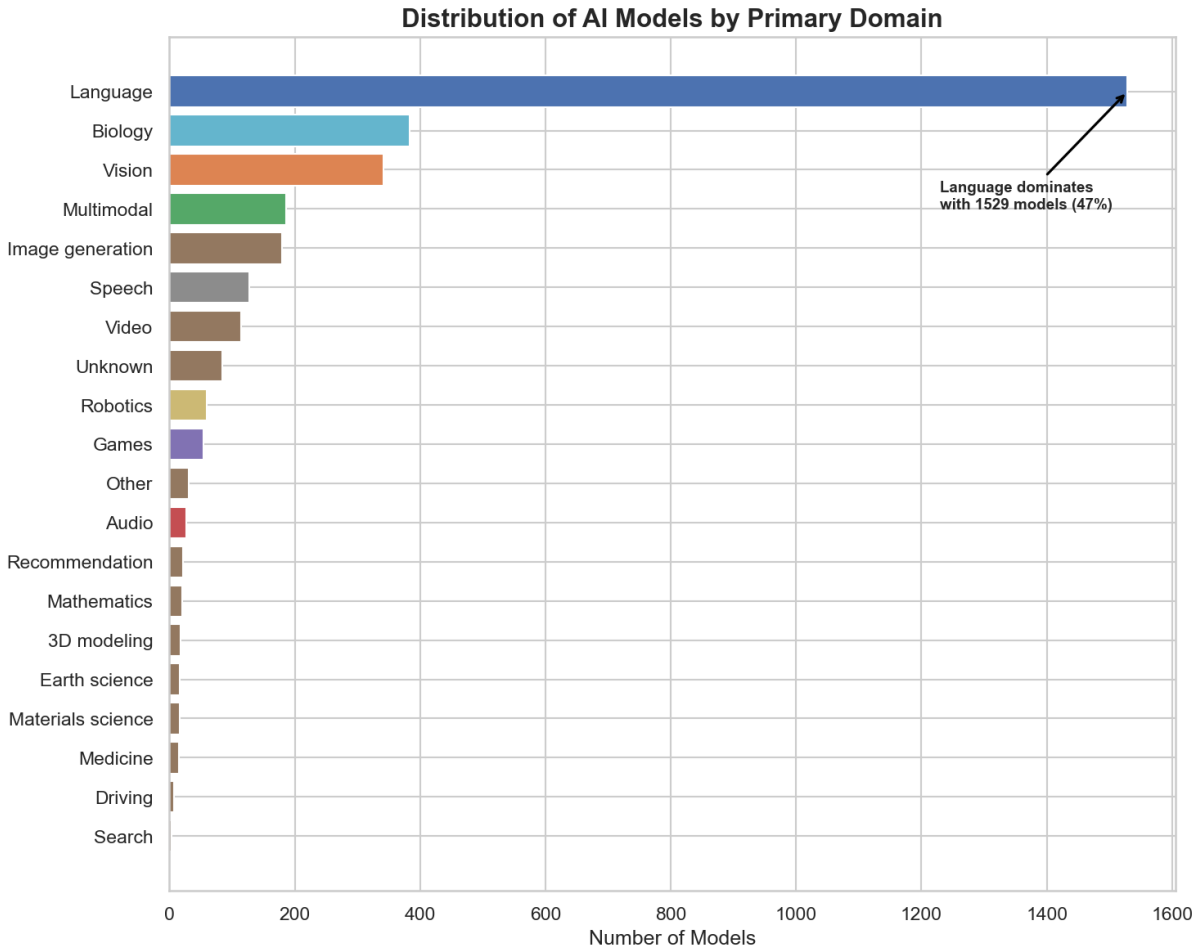

Figure 5: **Distribution of AI Models by Primary Domain.** A horizontal bar chart of the top 20 domains (using the first domain from comma-separated values). Language dominates with over 1,400 models (∼44% of the dataset), followed by Biology, Vision, and Image Generation. This skew is important context for any domain-level analysis.

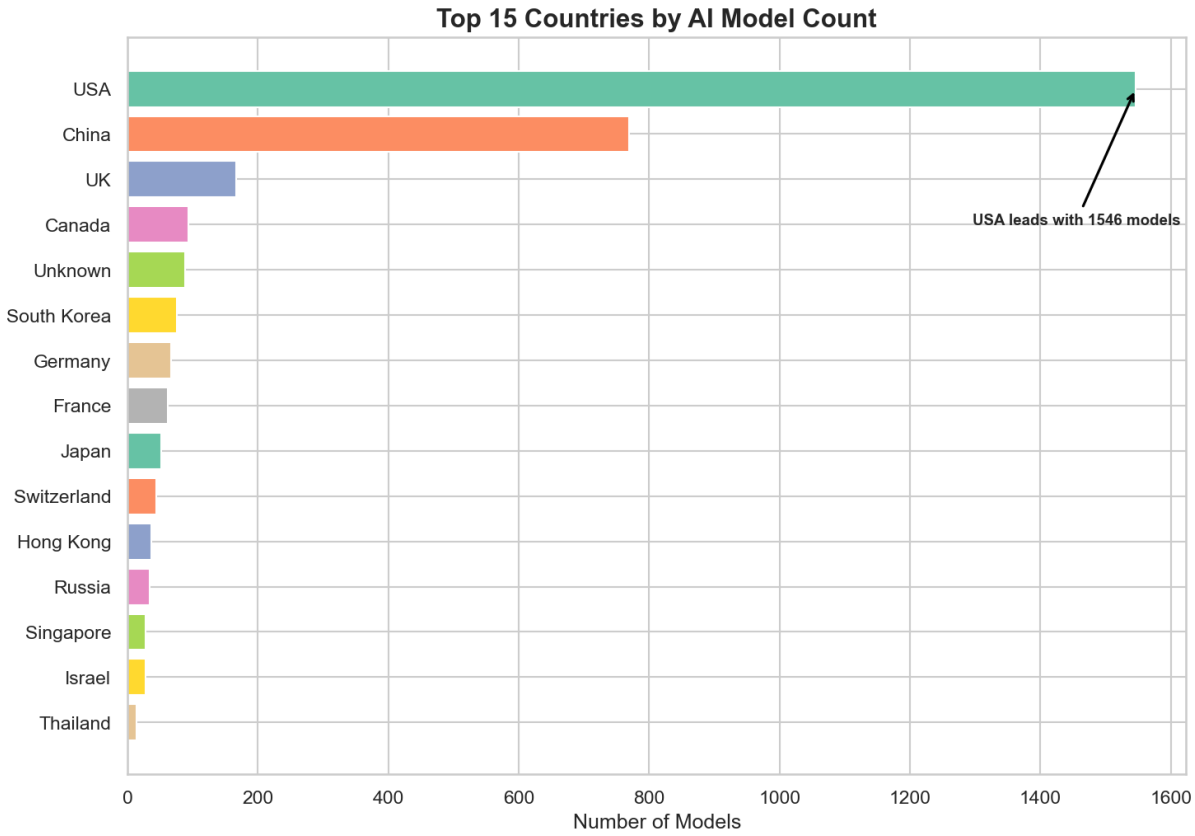

Figure 6: **Top 15 Countries by AI Model Count.** The USA leads with over 1,000 models, followed by China, the UK, and South Korea. Country was extracted as the first value from the comma-separated "Country (of organization)" field. The US–China dominance mirrors global AI investment patterns.

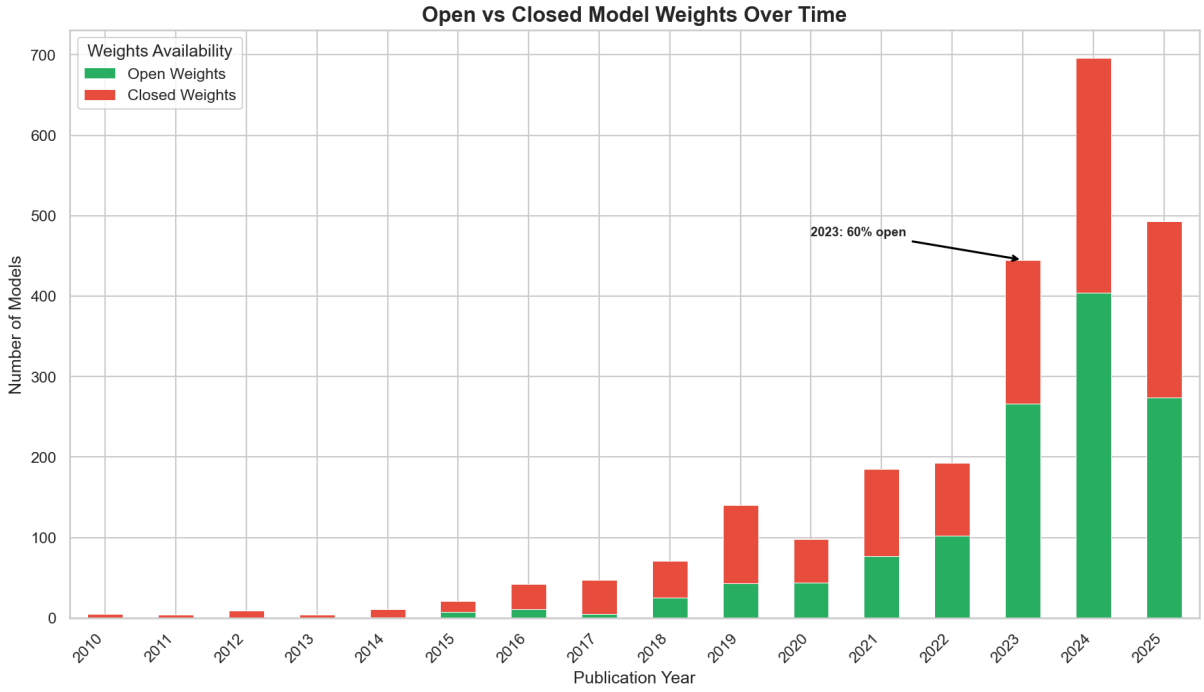

Figure 7: **Open vs. Closed Model Weights Over Time.** A stacked bar chart showing the proportion of open-weight and closed models per year (2010–2026). The proportion of open-weight models has shifted over time, with the annotation highlighting the 2023 open-source ratio. The rapid growth of both categories in recent years reflects the broader acceleration of AI development.

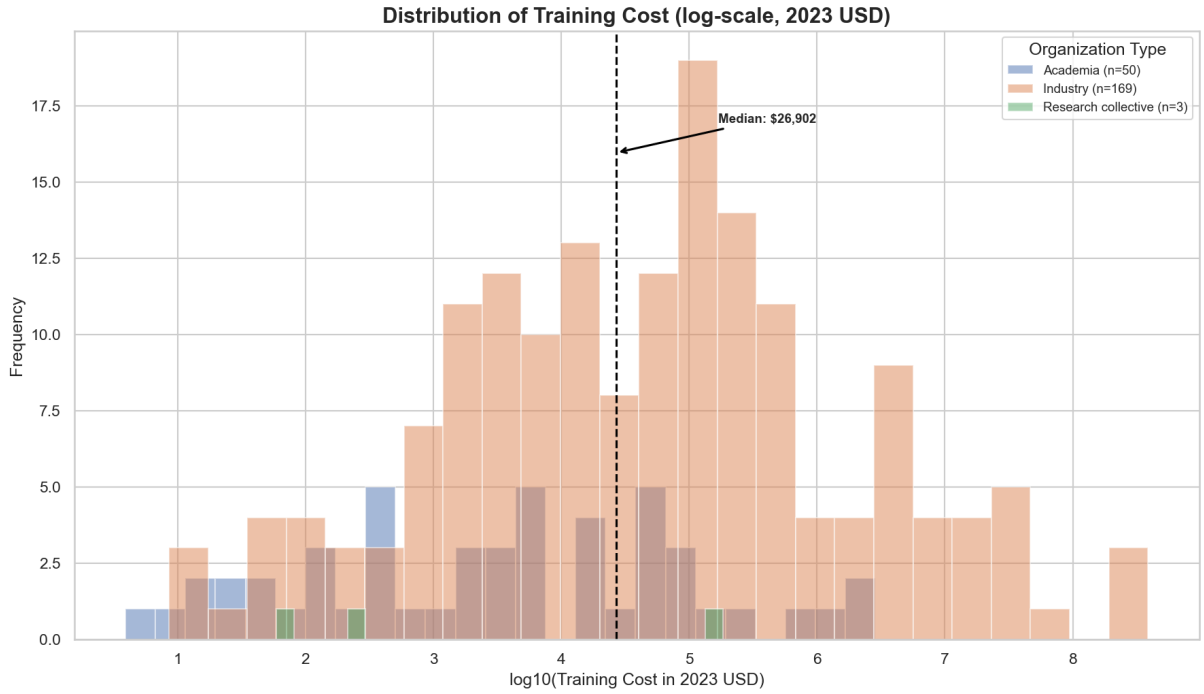

Figure 8: **Distribution of Training Cost (log-scale, 2023 USD).** A histogram of $\log_{10}$(training cost) split by organization type. The median cost is annotated with a dashed vertical line. Only ∼7% of models have cost data, so this represents a subset of primarily large-scale, well-documented systems. Industry models skew toward higher costs.

Figure 9: **Data Quality: Missing Values by Column.** A horizontal bar chart showing null counts for key columns in the raw dataset. Training cost has the highest missingness (∼93%), followed by Training time (∼83%). Core metadata columns (Domain, Organization, Country) have relatively low missingness (<5%). This chart motivates the selection of imputation strategies and the decision to exclude extremely sparse columns from certain analyses.

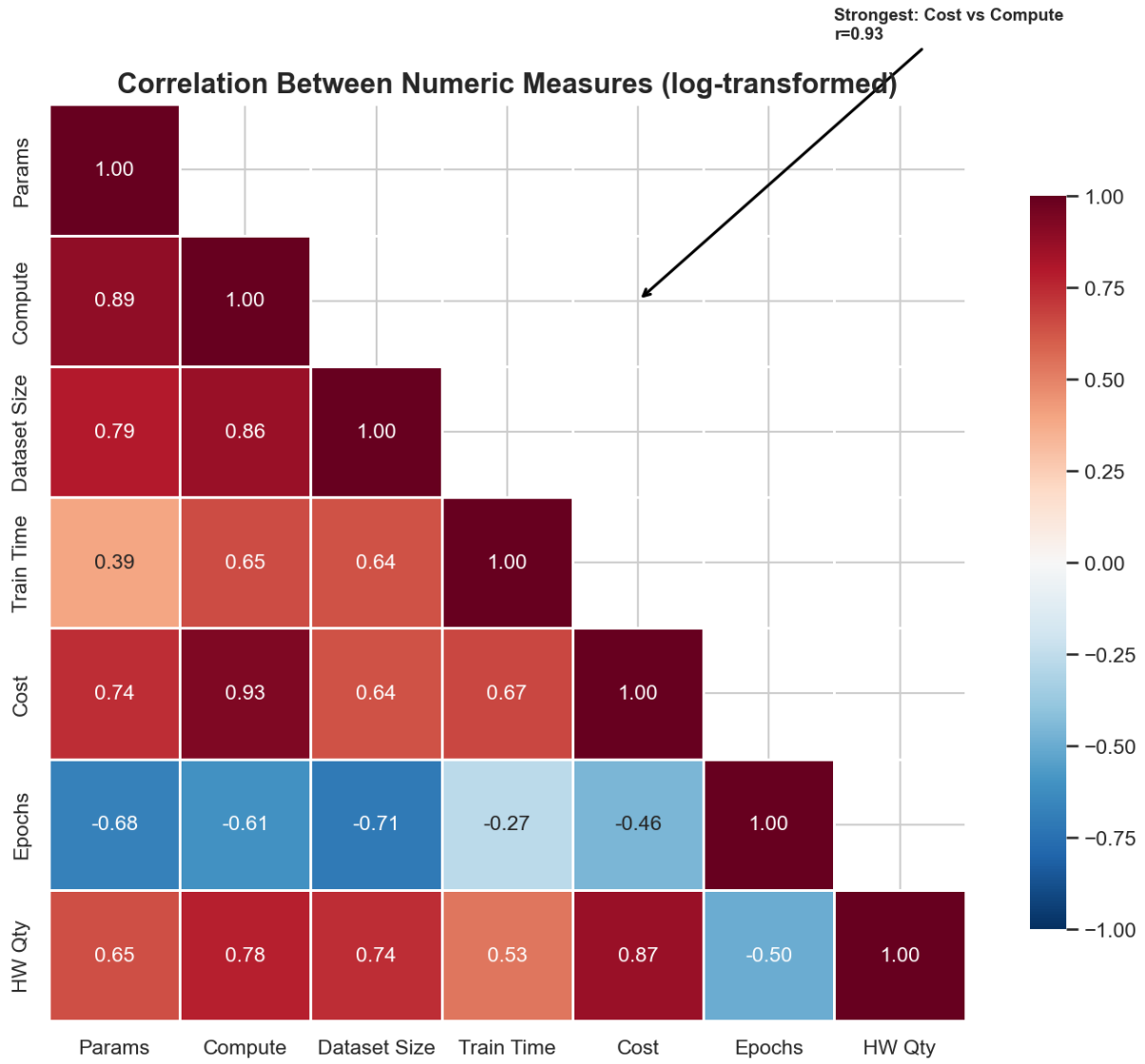

Figure 10: **Correlation Between Numeric Measures (log-transformed).** A lower-triangular heatmap of Pearson correlations among log-transformed numeric columns (Parameters, Compute, Dataset Size, Training Time, Cost, Epochs, Hardware Quantity). The strongest correlation is annotated. Compute and Parameters show the expected strong positive relationship, consistent with neural scaling laws. Epochs show weaker correlations with other measures, reflecting diverse training regimes.

All figures generated by `scripts/visualizations.py` *from the raw dataset (*`data/all_ai_models.csv`*, $n = 3{,}237$ models).*

6 Analysis Questions

The following seven questions are designed to go beyond surface-level description and investigate the relationships, trends, and mechanisms underlying the AI Models dataset. Each question targets a specific analytical technique and is motivated by a substantive concern about the trajectory of AI development.

1. Why has the relationship between model size (parameters) and training compute diverged since 2020, and what does this reveal about changing efficiency strategies across Industry vs. Academia?

Early scaling laws suggested a relatively stable power-law relationship between parameter count and the floating-point operations needed to train a model. However, recent work on techniques such as mixture-of-experts, sparsity, and improved training recipes suggests that this relationship may be shifting, and shifting differently depending on whether the model originates in an industrial lab with large compute budgets or an academic group optimising under constraints. Investigating this divergence through **regression analysis** (fitting separate scaling relationships for pre- and post-2020 models, stratified by organization type) would reveal whether efficiency gains are real and who is benefiting from them.

2. How has the geographic concentration of frontier AI development shifted over time, and why do certain countries dominate specific domains?

The geography of AI research has significant implications for policy, talent flows, and technological sovereignty. While the United States and China are often cited as the two dominant players, this dataset allows us to test whether frontier-level contributions are becoming more or less concentrated and whether specific countries have carved out specializations in particular domains (e.g. Language models vs. Vision or Robotics). **Time-series analysis** of country-level model counts, combined with **chi-squared tests** for independence between country and domain, would quantify these patterns and identify statistically significant associations.

3. What factors predict whether an AI model will be released with open weights, and has the open-source movement in AI accelerated or decelerated since 2023?

The debate over open vs. closed AI models is one of the most consequential in the field, touching on safety, competition, and scientific reproducibility. This dataset contains both the outcome variable (`Open model weights?`) and a rich set of potential predictors: organization type, domain, model size, country, and whether the model is classified as frontier. A **logistic regression or classification model** trained on these features would identify which factors most strongly predict openness, while tracking the open-weight rate over time would reveal whether the trend is accelerating, plateauing, or reversing in the wake of recent policy discussions.

4. Is there a meaningful clustering of AI models based on their resource profiles (compute, parameters, dataset size, training time), and do these clusters align with domain or organizational boundaries?

It is tempting to assume that AI models fall into natural tiers (small research models, mid-scale production models, and frontier behemoths), but the actual structure of the

resource landscape may be more complex. Applying **unsupervised clustering methods** such as K-means or DBSCAN to the log-transformed resource columns (training compute, parameters, dataset size, training time) would reveal whether discrete clusters exist in the data. Overlaying domain and organization labels on the resulting clusters would then test whether the resource groupings correspond to meaningful real-world categories or cut across them in unexpected ways.

5. Why has the cost of training frontier models grown super-linearly, and what is the relationship between training cost, compute, and the hardware used?

Media reports frequently cite escalating training costs, ranging from millions to hundreds of millions of dollars, but the underlying drivers are not always clear. Is cost growth simply a consequence of using more compute, or do hardware choices, utilization rates, and cloud pricing play independent roles? **Regression analysis** on the `Training compute cost (2023 USD)` column, using compute, hardware type, hardware quantity, and utilization as predictors, would decompose the cost curve and identify which factors contribute most to the headline figures. This analysis is especially relevant given that cost may eventually become a binding constraint on further scaling.

6. How do the publication patterns and scaling trajectories differ between Language models and Multimodal models, and what does this suggest about the future convergence of AI modalities?

Language models have received the most public attention, but multimodal systems that integrate text, image, audio, and video are increasingly prominent. The dataset provides an opportunity to compare these two domains along multiple axes: publication frequency over time, parameter growth rates, compute requirements, and the organizations producing them. A **comparative time-series analysis**, plotting and statistically testing the scaling trajectories of Language vs. Multimodal models, would reveal whether multimodal systems are following the same exponential curve with a time lag, converging toward a common frontier, or charting an entirely different path.

7. What is the relationship between a model's notability and its measurable characteristics: are notable models simply bigger, or do other factors like accessibility and novel architecture play a role?

The `notable_ai_models.csv` subset is curated based on criteria that go beyond raw scale, including citation impact and historical significance. This raises the question of what actually distinguishes a notable model from a merely large one. A **feature importance analysis**, using a classification model (e.g. random forest or gradient boosting) to predict whether a model appears in the notable subset based on its numeric and categorical features, which would quantify the relative contribution of size, compute, accessibility, domain, organization type, and architecture. If notability is driven by factors beyond scale, that finding would challenge the prevailing narrative that bigger is always better.