Data in the Wild: Giants vs. Specialists – A Comparative Analysis of Scientific Output (2023)

ANTÓNIO SANTOS, IT University of Copenhagen, Denmark
FRANCISCO OLIVEIRA, IT University of Copenhagen, Denmark
HENRIQUE ALEIXO, IT University of Copenhagen, Denmark
TIBOR SZOLOMAIER, IT University of Copenhagen, Denmark
VINCENZO SABINO, IT University of Copenhagen, Denmark

1 Introduction

Scientific research drives global innovation, but it is not distributed equally around the world. Although large economies like the United States and China are well known to produce the highest volume of research, the factors that drive research efficiency are less clear. This report examines the relationship between a nation's development and its scientific output in three key fields: Artificial Intelligence (AI), Environmental Science, and Medicine.

The goal of this analysis is to look beyond simple quantity. By focusing solely on the total number of publications can overlook smaller nations that utilize their resources effectively. We aim to separate scientific success from population size and economic power to understand true research intensity. Specifically, we investigate whether success depends solely on national wealth, or if different fields require specific conditions, such as digital infrastructure for AI or strong public health systems for Medicine.

To achieve this, we combined data from multiple open sources to map the global research landscape in 2023. By comparing socioeconomic indicators with scientific output, we distinguished between two types of leaders: the "Giants," who dominate through sheer size, and the "Specialists," who achieve high efficiency through focused development. All code and data used for this analysis are publicly available in our GitHub repository: https://github.com/Francisco772/Data-in-the-wild.

2 Data Collection

To build the dataset for this analysis, we combined information from two open sources. We needed one source for the economic and social factors, and another for the scientific research itself.

For the socioeconomic data, we used the World Bank Open Data platform [2]. We selected eight indicators that provide a broad picture of a country's status, including economic measures like GDP and GDP per capita, but also social metrics like life expectancy, internet access, and tertiary school enrollment. The data was downloaded as a raw CSV file.

For the research data, we used OpenAlex [1], an open index of the global research system. Unlike the static World Bank files, this data had to be retrieved programmatically. We wrote a Python script to query their API, specifically looking for works published in 2023. We retrieved the total publication count for every country. Then, we filtered these queries by specific fields to isolate counts for Artificial Intelligence, Environmental Science, and Medicine.

Authors' Contact Information: António Santos, IT University of Copenhagen, Copenhagen, Denmark, anms@itu.dk; Francisco Oliveira, IT University of Copenhagen, Copenhagen, Denmark, faol@itu.dk; Henrique Aleixo, IT University of Copenhagen, Copenhagen, Denmark, hale@itu.dk; Tibor Szolomaier, IT University of Copenhagen, Copenhagen, Denmark, tisz@itu.dk; Vincenzo Sabino, IT University of Copenhagen, Copenhagen, Denmark, vins@itu.dk.

3 Data Processing and EDA

Our raw datasets, particularly from the World Bank, were not ready for immediate integration. We implemented a multi step processing method to ensure data quality and consistency.

(1) Cleaning Aggregates
Our first challenge with the World Bank dataset was that it mixed individual countries with broad regional aggregates. In our first observation, we identified 31 rows that did not represent sovereign nations but rather groupings such as "Arab World," "High income," and "Sub-Saharan Africa". Including these would have severely skewed our analysis, essentially double-counting populations and GDP.
We systematically filtered these out by identifying their unique ISO-3 codes (e.g., ARB for Arab World, HIC for High Income) and removing them from the dataset.

(2) Geographic Standardization
To enable regional analysis, we needed to map every country to its respective continent, so we constructed a dictionary mapping ISO-3 codes to their continental regions (Africa, Asia, Europe, North America, Oceania, South America).
By creating a continent mapping for all countries,we were able to later color code some of our visualizations by continent and identify regional trends in research output.

(3) Feature Correlation and Imputation Strategy
Before addressing the missing values in our dataset, we performed a preliminary correlation analysis to understand the relationships between our development indicators. We hypothesized that a country's infrastructure metrics (like internet access) would be strongly predictive of its social metrics (like education and health), which would allow us to accurately estimate missing data points.(1)

To ensure the accuracy of our analysis, we used a systematic, column-by-column inspection method. Instead of relying on broad automated filters, we examined each development indicator individually to identify data irregularities

Although it is standard practice to use multivariate methods to detect complex errors, this approach would be a risk in our geopolitical analysis. In our context, statistical outliers are rarely mistakes, they represent sovereign nations with distinct characteristics. For instance, the United States and China are outliers in almost every dimension due to their immense GDP and populations.

Automated algorithms would likely detect these unique profiles as anomalies and maybe remove them. However, deleting these data points would exclude the very "Giants" of research that are central to our study. Consequently, we deliberately avoided multivariate exclusion. Instead, we prioritized correcting specific data points within individual columns, ensuring that no major nation was removed from the analysis due to its scale.

1. Handling Distributional Skew (Transformation): The column-by-column inspection highlighted extreme skew in our "volume" metrics (GDP, Population), where a few nations dominated the scale. Rather than treating these extreme values as noise to be deleted, we treated them as valid signals that needed scaling. We applied a logarithmic transformation (np.log1p) to these specific columns. This approach "pulled" the extreme outliers back into a comparable range without losing the valuable information they represent, normalizing the distributions for our correlation analysis while preserving the rank order of the nations.

2. Handling Missing Values : Consistent with our idea of preventing the meaning of the data, we rejected the standard practice of deleting rows with missing values. This would have biased our results and failed to show a true picture of the global research.

Instead, we chose to fill missing data, selecting the method best suited to each specific variable. Rather than using simple global averages, we prioritized context-aware techniques. For indicators where our initial analysis revealed strong correlations, we used those highly correlated features to define similarity groups (such as wealth bins) and imputed values based on those group medians. For metrics driven by location, we utilized regional

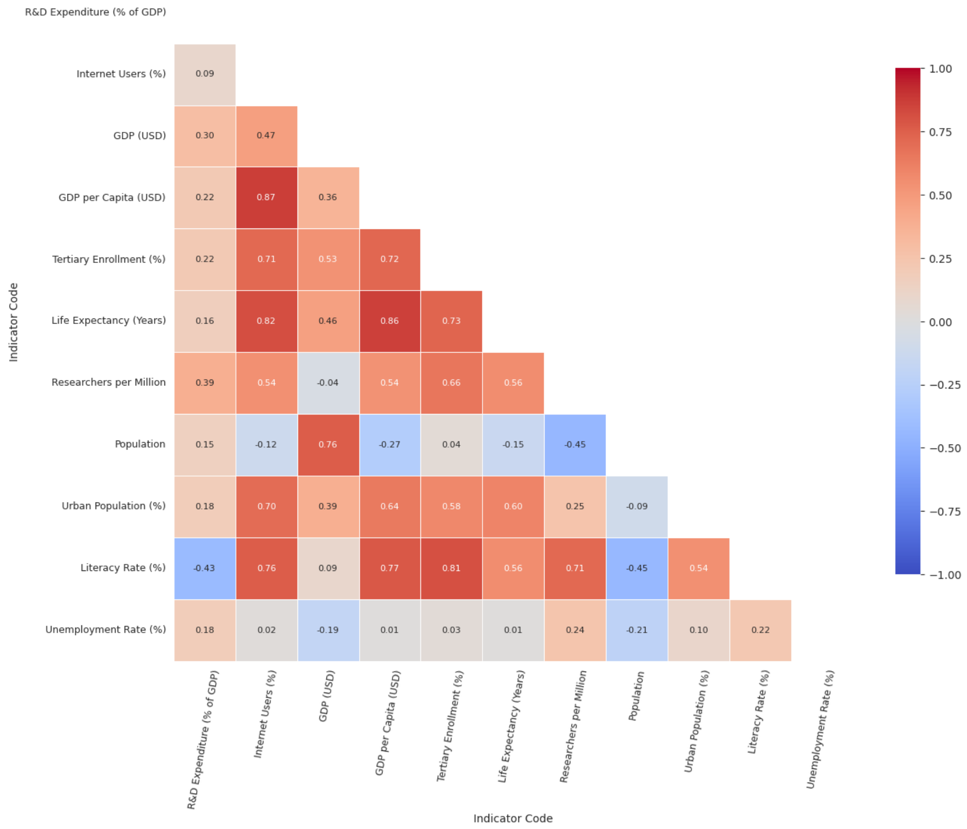

Fig. 1. Heatmap Correlation between the development indicators

imputation, filling gaps using continental medians. Finally, where possible, we avoided estimation entirely by calculating values directly from their available underlying components (e.g. deriving per-capita metrics from total population data).

The most significant challenge was Tertiary Enrollment (%), which had the highest proportion of missing values. Before imputing, we investigated the nature of this missingness to determine if it was systematic. By plotting the distributions of other key variables (like Life Expectancy and GDP) for countries with and without enrollment data, we observed clear patterns, identifying the data as Missing At Random (MAR).

This confirmation allowed us to implement Multivariate Imputation by Chained Equations (MICE). Because the missingness was related to observed data, MICE could effectively model the missing values as a function of the other highly correlated features (Life Expectancy, Internet, GDP). We ran the algorithm for 10 iterations, ensuring that the imputed education data was statistically consistent with the country's entire socioeconomic profile.

Finally, for indicators where the missingness was too extensive to be reliably imputed,(Research and Development Expenditure,Literacy Rate (%),Researchers per Million) which was missing for the majority of nations, we made the decision to drop the column entirely.

4 Methods and Results

Before performing our main analysis, we ran a preliminary correlation matrix to test the relationship between the raw paper count and our development indicators. This step was critical to determine if we could simply compare total output across nations.

The results revealed an overwhelming bias toward scale. The total number of papers was nearly perfectly correlated with GDP (0.95) and strongly correlated with Population (0.70). This confirmed that richer and more populous countries produce more papers simply due to capacity. To identify which countries were actually "performing better" relative to their resources, we needed to normalize the data,so we designed a custom Research Efficiency Metric to penalize "size" and reward "intensity." We defined the final metric as a weighted combination of Z-scores, using the correlations we just found as the weights:

$$\text{Efficiency Metric} = W_{pop} \cdot Z\left(\frac{\text{Papers}}{\text{Population}}\right) + W_{gdp} \cdot Z\left(\frac{\text{Papers}}{\text{GDP}}\right)$$

Where $W_{pop} \approx 0.70$ and $W_{gdp} \approx 0.95$. This formula ensures that a country is ranked based on how much it outperforms the expected output for an economy and population of its size.

To validate the effectiveness of our Efficiency Metric, we compared the global rankings under the traditional "Volume" view against our new "Efficiency" view. The divergence between these two lists confirms that our metric successfully separated scientific success from economic or population mass.

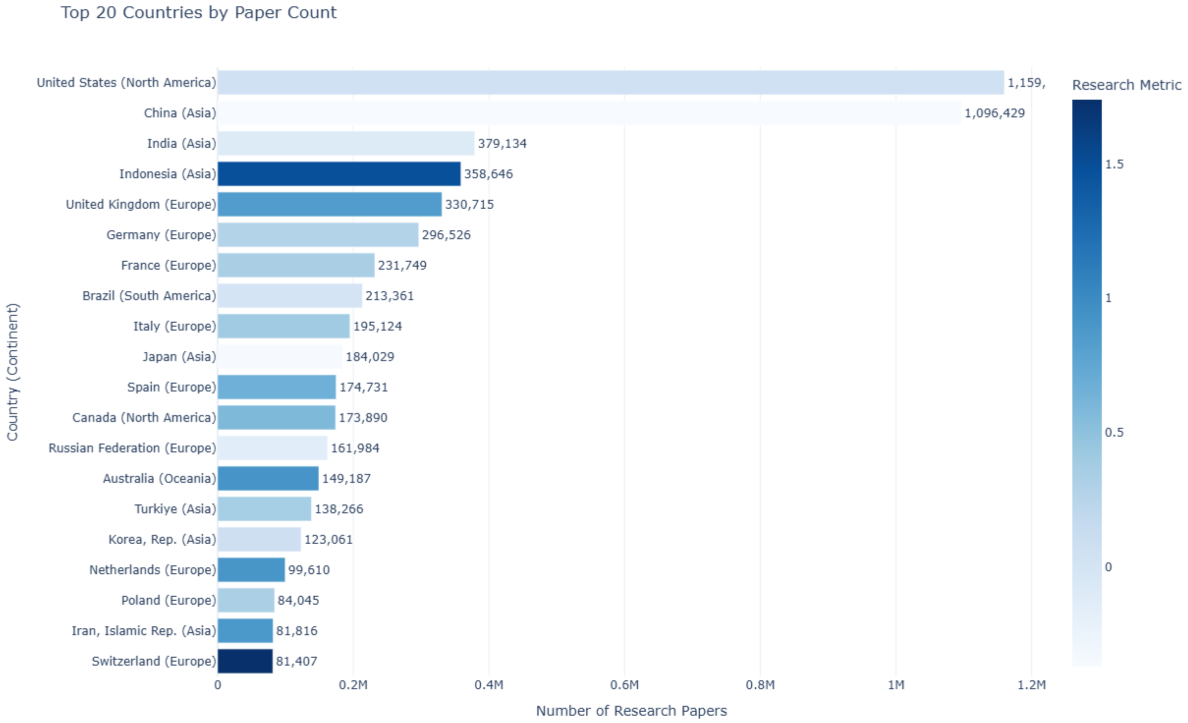

Fig. 2. Top 20 countries ranked by number of papers

When ranked by raw publication count, the list is predictable.(2) The world's largest economies (China, the United States, and India) dominate the field, producing over 2 million papers combined. However, the color scale (representing our Efficiency Metric) reveals a critical insight: Scale does not equal efficiency. The bars for China and the US are the longest, but their lighter color indicates that their output, while massive, is roughly what we would expect for nations of their colossal size. Their efficiency scores are near zero (average) or even negative.

In contrast, countries like the United Kingdom, Indonesia and Switzerland stand out as they appear in the top tier for volume, yet their darker blue shading shows they also maintain a high efficiency score, outperforming their expected output.

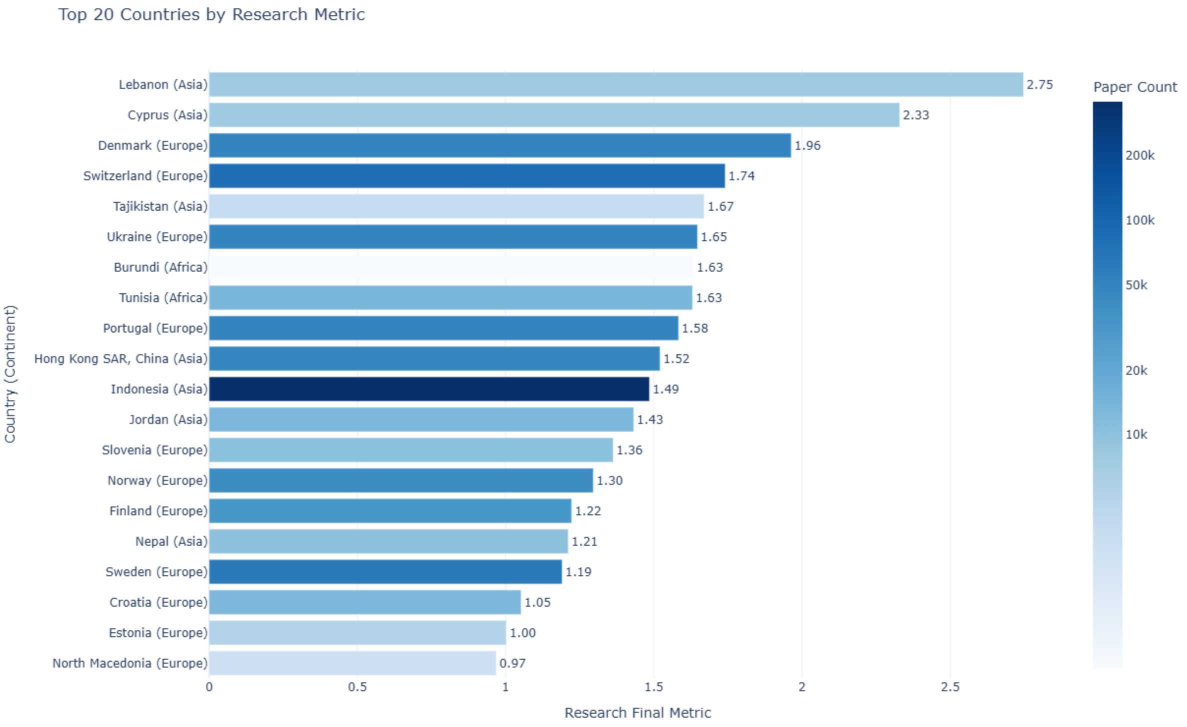

Fig. 3. Top 20 countries ranked by research metric

When we rank countries by our Research Efficiency Metric(3), the list changes completely. The big countries like the US and China disappear, and smaller nations take the top spots. Lebanon is the number one country in this index, followed by Cyprus, Denmark, and Switzerland. The top of the list is very diverse. It includes wealthy European nations like Denmark and Switzerland, but it also includes developing nations like Tajikistan and Burundi. This proves that our formula works as intended. By penalizing GDP and Population, we are not just finding the richest countries. We are finding the countries that produce the most research relative to their specific situation, whether they are wealthy or not.

Topic Specialization: Who Leads in What?

Finally, we looked at each of our three research fields individually to see if the "Efficiency Leaders" change depending on the topic. We compared the Top 20 countries by Volume (Total Papers) against the Top 20 by Efficiency (Our metric) for each field.

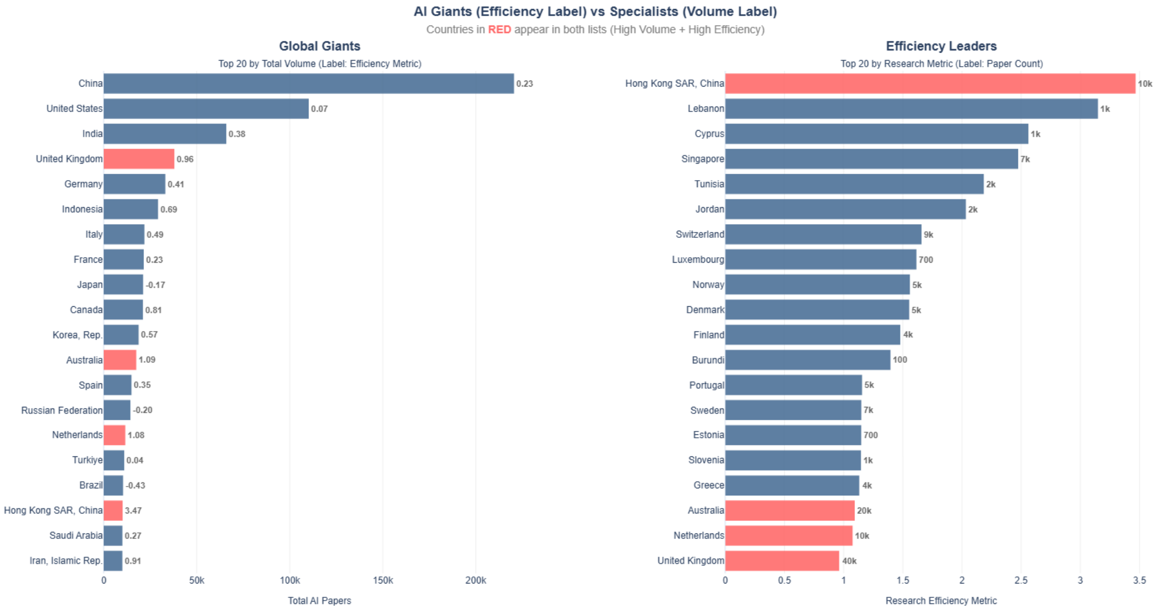

Fig. 4. Artificial Intelligence: Giants vs Specialists

For AI, the volume list on the left is dominated by the usual countries like China and the United States (4). However, the labels on these bars reveal their efficiency scores are often very low (e.g. 0.07 for the US), indicating that their dominance is largely a function of their massive size rather than per-capita intensity. Conversely, the efficiency list on the right is populated by "Tech Hubs" like Singapore and Switzerland. While the labels show their total paper counts are a fraction of the giants, their high ranking proves they are achieving outsized results relative to their resources.

The most impressive finding is the group of countries highlighted in red. These nations: Hong Kong, the United Kingdom, Australia, and the Netherlands, appear in the Top 20 for both volume and efficiency. This proves that is possible to scale up research production to a global level without sacrificing the high intensity typically associated with smaller specialists.

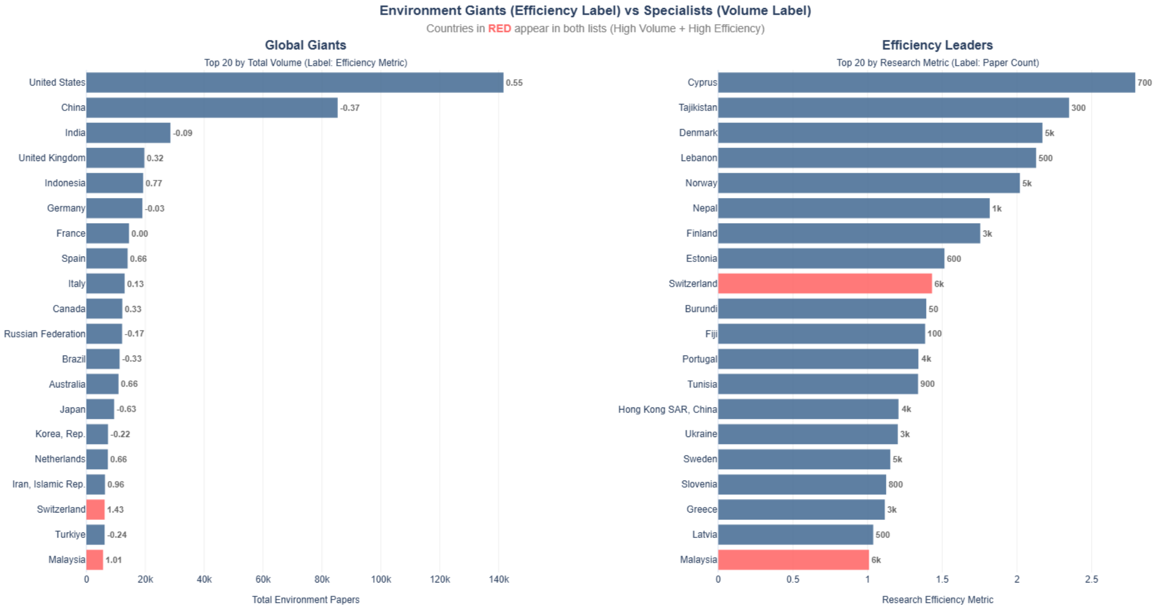

Fig. 5. Environment Research: Giants vs Specialists

The analysis of Environmental Science reveals a distinct landscape compared to other fields (5). When we compare the Top 20 countries by volume against those by efficiency, we see almost no overlap, suggesting a strong divide between the producers of mass research and the most intensive specialists.

The volume list remains consistent with other fields, dominated by massive economies like China and the United States. However, their efficiency scores are relatively low, confirming that their leadership is primarily based on scale rather than per-capita intensity. On the other side, the efficiency ranking reveals a unique mix of leaders. Cyprus takes the top spot, followed by a diverse group that includes Tajikistan, Denmark, Lebanon, Norway, and Nepal. This creates a blend of wealthy Nordic or European nations and developing nations where environmental resource management may be a critical national priority.

Unlike other fields, it is rare for a country to be a leader in both volume and efficiency in Environmental Science.(5) Only two nations appear in both Top 20 lists: Switzerland and Malaysia. These are the only that have managed to achieve massive global scale while maintaining the high intensity of a specialist.

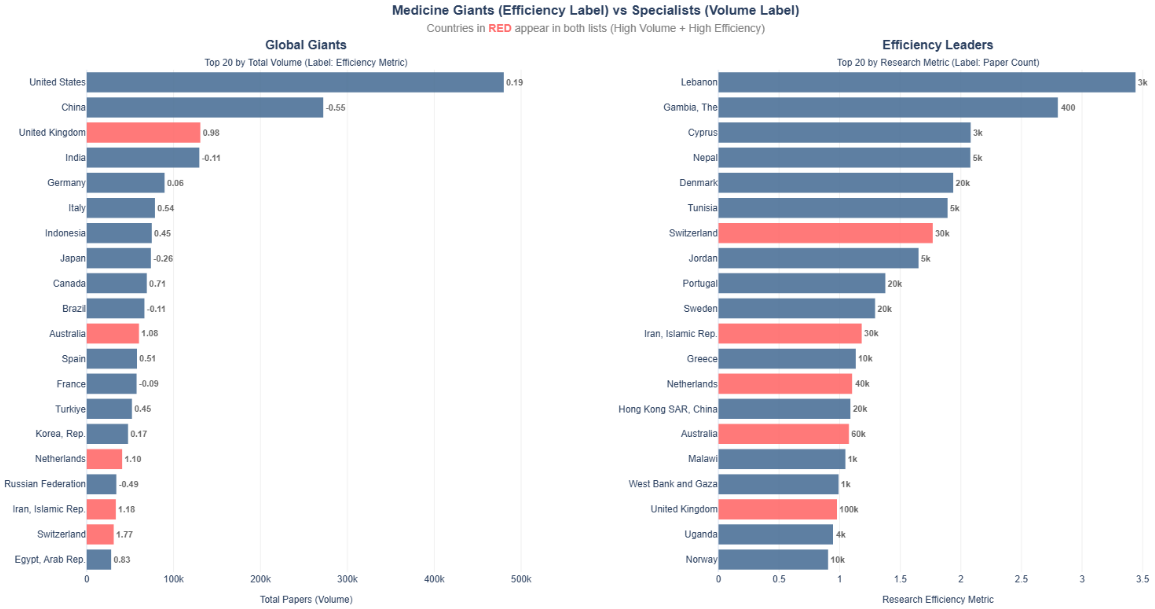

Fig. 6. Medicine: Giants vs Specialists

The results for Medical research (6) are very different from AI or Environment. Here, we see a mix of wealthy European nations and smaller developing countries leading the way in efficiency.

The volume list is standard, with the United States and China producing the most medical research by far. However, their efficiency scores are low, meaning their high output is mostly due to their large size. On the other hand,on the efficiency list Lebanon takes the number one spot, followed by Gambia and Cyprus. This top 20 includes wealthy countries with strong public health systems like Denmark and Switzerland, but also developing nations like Nepal and Tunisia. Five nations: the United Kingdom, Australia, the Netherlands, Switzerland, and Iran appear in both the Top 20 for volume and the Top 20 for efficiency. These nations prove that high volume doesn't have to mean low efficiency.

Dimensionality Reduction Strategy

The rankings above showed us clearly who the "Giants" and the "Specialists" are. However, we also wanted to see if these high-performing nations share common traits, like similar levels of wealth, health, or education, that separate them from the rest of the world. To find these patterns, we couldn't just look at one number at a time. We needed a way to look at all our development indicators at once to see which countries are truly similar to each other. To analyze relationship between national development and research output, we used a two-part strategy to simplify our multi-dimensional data.

First, to capture subtle, non linear patterns, we utilized an Uniform Manifold Approximation and Projection (UMAP). Unlike linear methods that focus on the overall spread of data, UMAP is designed to keep similar data points close together. This capability was essential for identifying distinct "neighborhoods" or clusters of nations. It allowed us to group countries based on their shared development characteristics, rather than simply looking at their raw output numbers.

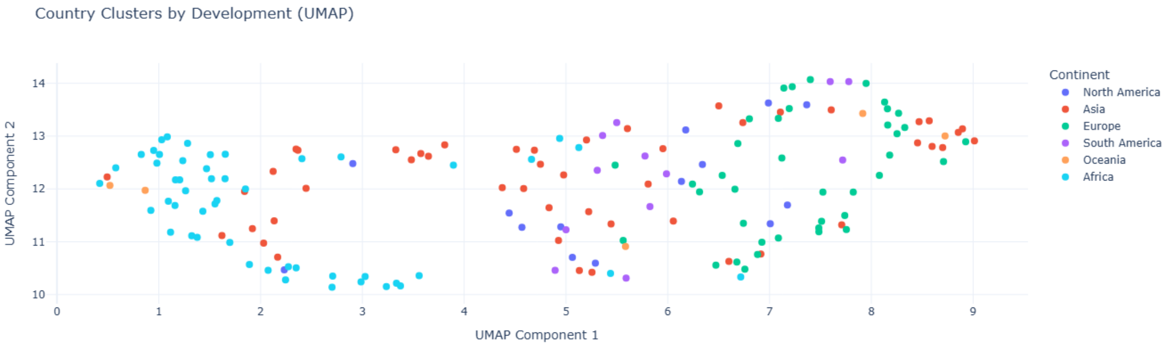

Fig. 7. Country Clusters by Development (UMAP)

After recreating the UMAP visualization (7), we noticed a clear pattern based on geography. Even though the algorithm was not given the continent labels, it organized the countries into a structure that closely resembles a real-world map.

The most obvious finding is that African nations (shown in light blue) form a separate, isolated group. This suggests that the development profile of these nations is distinct from the rest of the world. In contrast, European countries (green) are packed closely together, indicating that they share very similar levels of wealth, health, and technology.

Asian nations (red) act as a bridge between these two groups. Unlike the tight cluster of Europe, the Asian countries stretch across the middle of the graph. This reflects the huge variety within the continent, which includes both wealthy technological leaders and emerging economies.

This clustering proves that national development is not random and it is strongly dependent on the location. This geographic split helps explain the unequal distribution of scientific research we observed earlier, wealthy regions like Europe share a high capacity for science, while other regions face a distinct set of developmental hurdles.

In parallel, we applied Principal Component Analysis (PCA) across all three research domains: Artificial Intelligence, Environmental Science, and Medicine. We used PCA to combine our eight indicators into a single Development Index. This one feature captures 57.31% of the information (variance) from the original dataset.This approach provided the "big picture" view.

To analyze the global distribution of science, we plotted the derived "Development Index" against Research Efficiency for each domain. Across all three fields AI, Medicine, and Environmental Science.

A consistent finding across all plots is the position of African countries, these countries cluster in the bottom-left quadrant, falling below the global average for both development and research efficiency. This confirms the structural gap identified in our UMAP analysis, indicating that lower socioeconomic development creates a significant barrier to scientific intensity.

While the general trend is similar, distinct regional leaders appear in different fields:

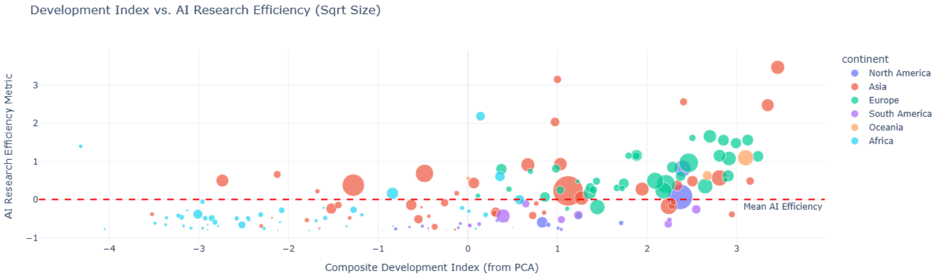

Fig. 8. Development Index vs. AI Research Efficiency

Artificial Intelligence (8): This field shows a unique trend where Asian nations are particularly successful. We see a cluster of Asian countries that have achieved very high scores in both national development and research efficiency, performing as well as or even better than traditional Western leaders.

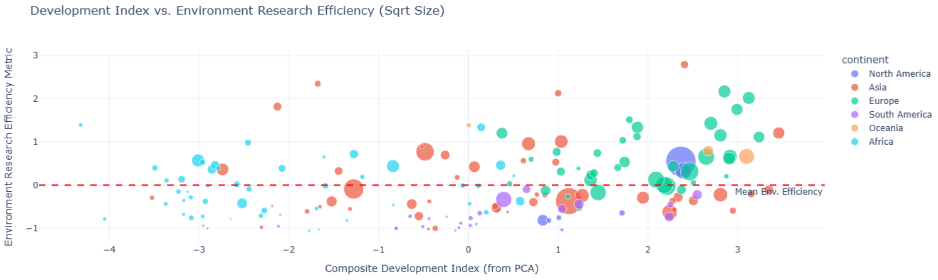

Fig. 9. Development Index vs. Environment Research Efficiency

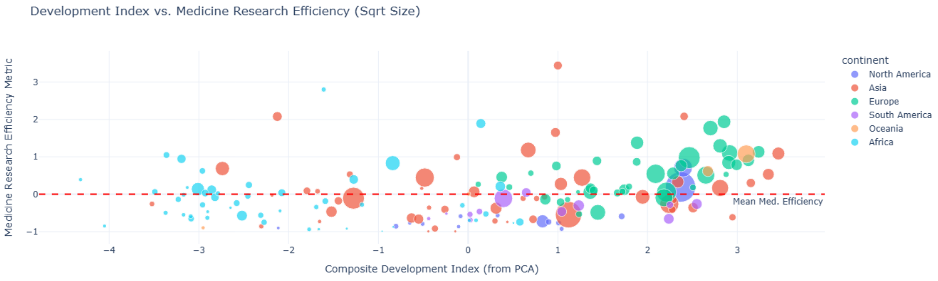

Fig. 10. Development Index vs. Medicine Research Efficiency

Environmental Science and Medicine (9)(10): In these two areas, Europe is the clear leader. On average, European nations consistently rank high in both development and efficiency. This indicates that while Asia is a major driver of new digital technologies, Europe remains the central hub for research focused on public health and environmental protection.

5 Conclusion

Lastly, to visualize global scientific strategies, we separated nations into distinct categories based on their performance. We assigned each country a specific profile label, such as:

- "All-Round Powerhouse" (Countries that rank in top 20 in all research fields)
- "AI Specialist", "Environment Specialist", or "Medicine Specialist" (Countries that rank in the top 20 in only one specific field.)
- "AI / Env Specialist", etc. (Countries that rank in the top 20 in two different fields.)
- "Not in Top 20" (Does not belong in the top 20 in any category)

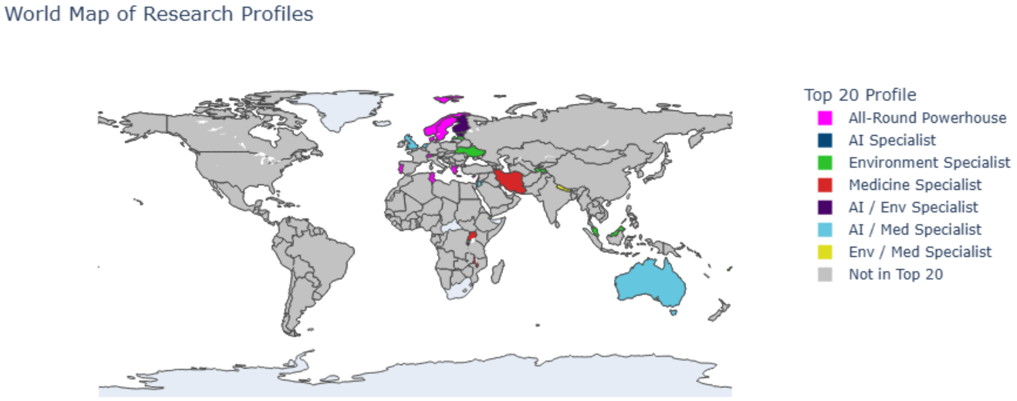

Fig. 11. World Map of Research Profiles

The "All-Round Powerhouse" group, countries that are highly efficient in all three fields, is mostly made up of wealthy nations like Switzerland, Sweden, and Denmark. This makes sense, as they have the resources to fund a broad range of science. However, it is interesting to see that smaller economies like Tunisia and Lebanon also appear in this top tier, proving that you don't need to be a global superpower to maintain a balanced, high-quality research output.

The "Specialists" tell a different story. For AI, the leaders are Singapore and Luxembourg, small, wealthy, high-tech hubs. But for Medicine and Environment, the list changes completely. The most efficient nations here are not the usual wealthy giants, but developing countries like Uganda, Malawi, Fiji, and Tajikistan.

This confirms a major difference between the fields: high efficiency in AI seems to require expensive digital infrastructure and capital, whereas efficiency in Medicine and Environment is often driven by urgent national priorities or specific geographic challenges, regardless of a country's wealth.

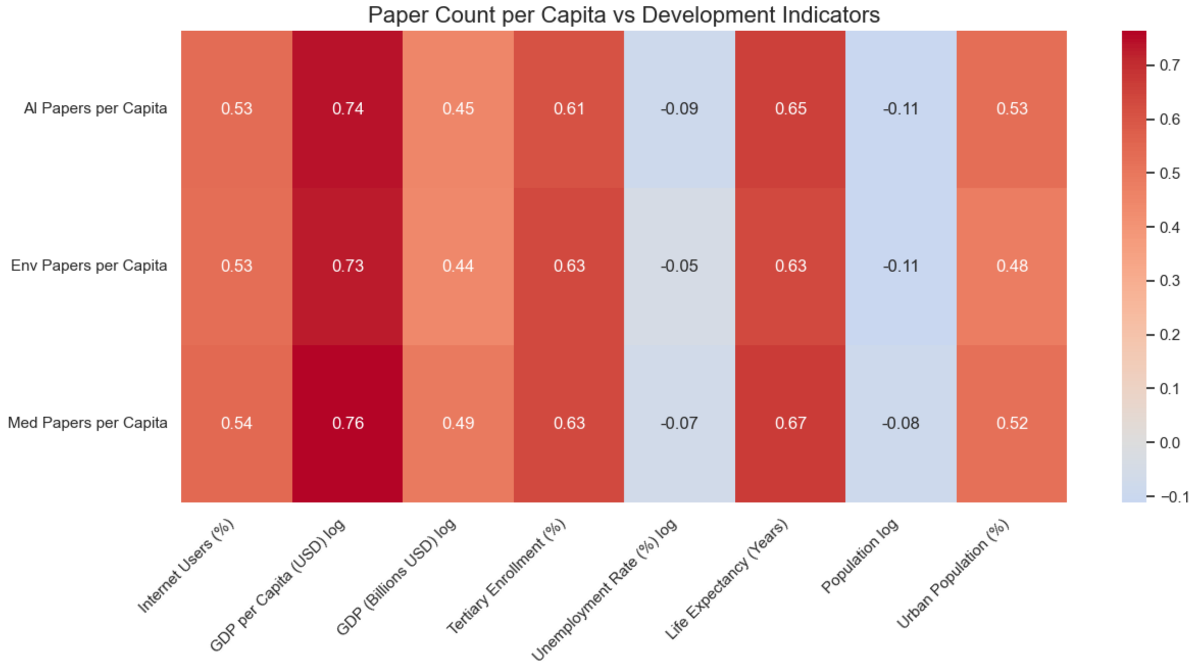

Fig. 12. Correlation: Paper count per capita vs. Development Indicators

To conclude, we decided to make a correlation heatmap to provides a final insight into our core research question: "Does a specific research area correlate more strongly with a specific development indicator?"

To find this, we correlated our normalized research intensity (papers per capita) with 8 of our cleaned development indicators.

While Economic Wealth (‘GDP per Capita‘) is, as expected, the strongest overall predictor across all fields (correlations 0.73–0.76), the secondary drivers reveal distinct patterns for each research domain.

For Medicine & AI, Health is Key: After wealth, the strongest predictor for both Medicine and AI research is "Life Expectancy (Years)" (0.67 for Medicine, 0.65 for AI). This suggests that a society’s overall health outcomes are a critical indicator of its capacity to perform advanced research in these high-tech and biological fields.

For Environment, Education takes the Lead: Interestingly, for Environment research, "Tertiary Enrollment (%)" (0.63) matches or slightly edges out Life Expectancy as the second most important factor. This suggests that environmental research intensity may be more uniquely driven by the sheer availability of university-level talent and education infrastructure.

The "Small Country" Advantage: The correlation with "Population_log" is consistently negative across all fields (-0.08 to -0.11). This confirms that massive scale is not an advantage for efficiency. Smaller nations are often more effective at achieving high per-capita research output than global giants.

Important Limitations

It is critical to state the limitations of this analysis.

Correlation is not causation, so these numbers only show an association. We cannot conclude that "Life Expectancy" causes more AI research. This analysis is descriptive, not explanatory. These variables are likely all related to a country’s overall development ecosystem in a complex, bidirectional way. Besides this, we are only looking at 8 specific indicators. Other factors we did not study (such as national research funding, specific educational policies, or levels of international collaboration) could be far more important.

Therefore, these findings don't provide a "final answer" but rather point to interesting associations within our specific dataset.

6 Statement on Generative AI Usage

In accordance with course guidelines, we acknowledge the use of Large Language Models, specifically Google Gemini, as a supplementary tool during the preparation of this project.It was utilized primarily to assist with the debugging and optimization of the code, as well as to refine the grammatical structure and clarity of the report.

References

[1] Jason Priem, Heather Piwowar, and Richard Orr. 2025. OpenAlex: A fully open index of scholarly works, authors, venues, institutions, and concepts. https://openalex.org

[2] The World Bank. 2023. World Development Indicators. https://datatopics.worldbank.org/world-development-indicators/