



Global Spatial Distribution of Open-Source Artificial Intelligence

Anqi Wei, Baihui Wang, Charlotte Li



Guidelines

O1 Introduction

O2 Data Collection

O3 Data Cleaning

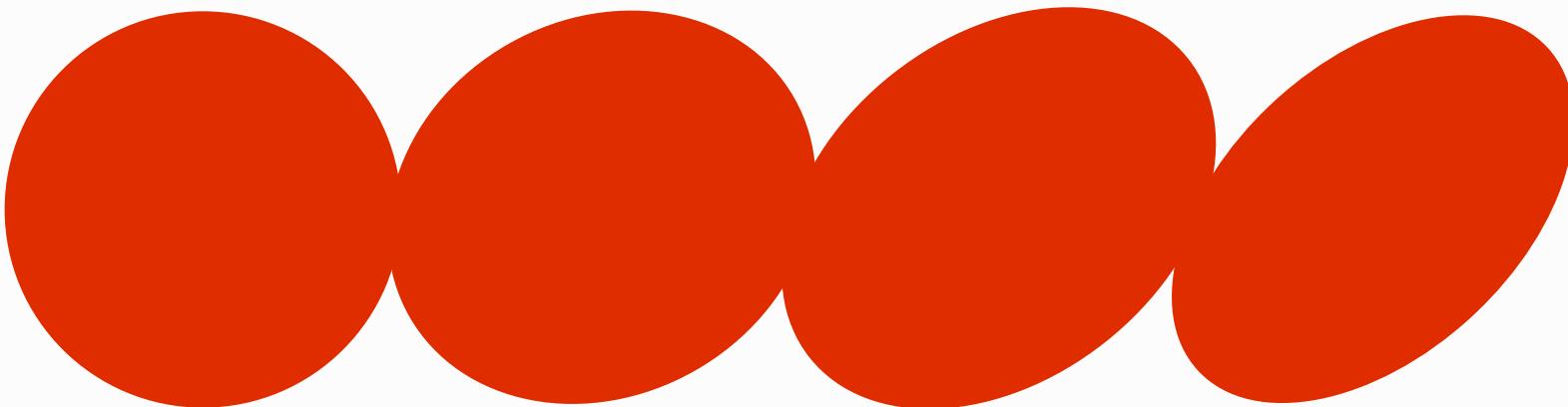
O4 Data Visualization

O5 Data Wrangling

O6 Data Analysis

O7 Conclusion

Introduction



Topic Description, Research Question & Hypothesis

Introduction

Key Focus:

- Analyze global spatial patterns of open-source AI development.
- Explore socioeconomic drivers (GDP, education, scholar publication).

Description of Topic

Significance:

- Democratizing AI innovation vs. persistent geographic inequality.
- Policy implications for equitable technological access.

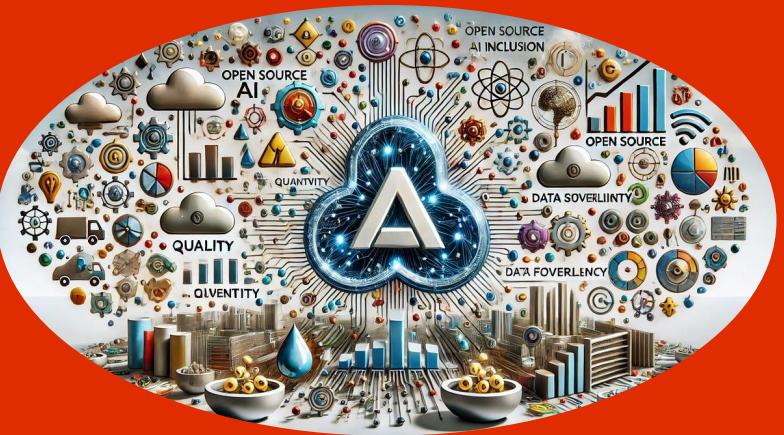


RQ1: Which cities/countries dominate AI-related open-source projects?

RQ2: How do economic development and higher education explain these patterns?

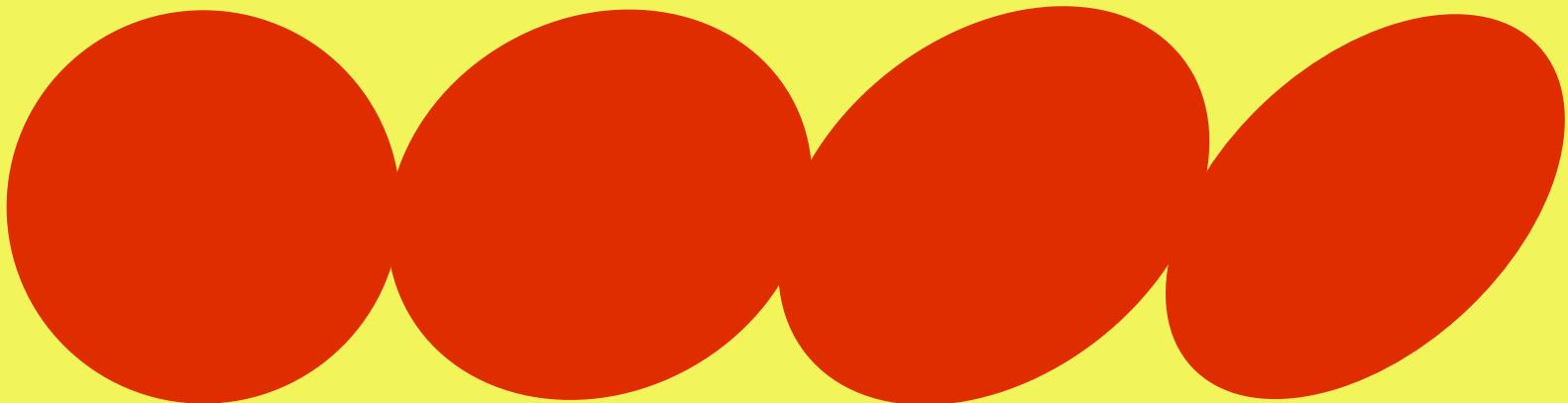
Introduction -
Research Question

Hypothesis



- **H1:** Higher GDP density → Higher open-source AI activity.
- **H2:** Strong AI publication presence → Reinforces open-source contributions.

Data Collection



Data Sources, Data Collection & Scraping

Data Sources

Source	Description
GitHub	Top 270 AI repositories (star-based ranking); 11,256 contributors' locations.
World Bank	GDP & GDP per capita (2023)
UNESCO GERD	R&D Investment as % of GDP (proxy for innovation capacity).
UN Population Data	City/country population metrics for per capita calculations.
Our World in Data	Annual AI Publications for tracking AI research output.

Data Collection

Data Collection & Scraping

GitHub Workflow:

- Scrapped AI repositories using Python (pagination + token rotation to bypass API limits).
- Extracted metadata: stars, contributors, commit frequency, geolocations.

Geocoding:

- OpenStreetMap + Google Maps APIs to standardize city/country names.
- Handled multilingual entries via Google Translate API (with manual validation).



Data Cleaning n Visualization

Baihui Wang

Rank	Repository	Stars	Contributor	Location
1	AutoGPT	171829	Auto-GPT-Bot	
2	AutoGPT	171829	Pwuts	Delft, NL
3	AutoGPT	171829	waynehamadi	
4	AutoGPT	171829	Torantulinc	UK
5	AutoGPT	171829	Swiftyos	
6	AutoGPT	171829	hunteraraujo	
7	AutoGPT	171829	richbeales	Kent, UK
8	AutoGPT	171829	majdyz	Amsterdam
9	AutoGPT	171829	aarushik93	Berlin
10	AutoGPT	171829	ntindle	Texas
11	AutoGPT	171829	BillSchumacher	
12	AutoGPT	171829	SilenNaihin	
13	AutoGPT	171829	dependabot[bot]	
14	AutoGPT	171829	Bentlybro	Some Where In Space I Think?
15	AutoGPT	171829	kcze	Scotland
16	AutoGPT	171829	lc0rp	
17	AutoGPT	171829	collijk	United States
18	AutoGPT	171829	AndresCdk	Colombia
19	AutoGPT	171829	p-i-	Oxford, England
20	AutoGPT	171829	sadmuphir	Somewhere
21	AutoGPT	171829	maiko	Serris
22	AutoGPT	171829	drikusroor	Utrecht, Netherlands
23	AutoGPT	171829	Abhi19920	India
24	AutoGPT	171829	erik-megai	Denver
25	AutoGPT	171829	Wladastic	Germany
26	AutoGPT	171829	coditamar	
27	AutoGPT	171829	k-boikov	Sofia, Bulgaria
28	AutoGPT	171829	sweetlilmre	
29	AutoGPT	171829	csolar	
30	AutoGPT	171829	rihp	Schengen
31	AutoGPT	171829	Tymec	
32	AutoGPT	171829	kinance	
33	AutoGPT	171829	dschonhol	Boston, MA
34	AutoGPT	171829		

Initial Data



Data Cleaning



Extract countries from locations:

Example:

Location	Extracted Country
"San Francisco, USA"	"USA"
"Berlin, Germany"	"Germany"
"Japan"	"Japan"

Extract cities from locations:

Example:

Location	Extracted City
"San Francisco, USA"	"San Francisco"
"Berlin, Germany"	"Berlin"
"Japan"	NaN

AutoGPT	171829	Auto-GPT-Bot			
AutoGPT	171829	Pwuts	Delft, NL		
AutoGPT	171829	waynehamadi			
AutoGPT	171829	Torantulinc	UK		
AutoGPT	171829	Swiftyos			
AutoGPT	171829	hunteraraujo			
AutoGPT	171829	richbeales	Kent, UK		
AutoGPT	171829	majdyz	Amsterdam		
AutoGPT	171829	aarushik93	Berlin		
AutoGPT	171829	ntindle	Texas		



AutoGPT	171829	Pwuts	Delft, NL	Delft	Netherlands		
AutoGPT	171829	Torantulinc	UK	Ouaka	Ködörösêse tî Bêafîka / République centrafricaine		
AutoGPT	171829	richbeales	Kent, UK	Kent	United Kingdom		
AutoGPT	171829	majdyz	Amsterdar	Amsterdar	Nederland		
AutoGPT	171829	aarushik93	Berlin	Berlin	Deutschland		
AutoGPT	171829	ntindle	Texas	Texas	United States		

Geocoding—handle partial location input

The image displays three distinct map interfaces side-by-side, illustrating different approaches to handling partial location input:

- OpenStreetMap:** Shows a detailed map of North America with state/province boundaries and major cities. A sidebar on the left provides an introduction to the platform, including links to "Learn More" and "Start Mapping".
- Google Maps Platform:** Shows a similar map of North America, but with a prominent blue banner at the bottom stating "Build awesome apps with Google's knowledge of the real world". Below the banner, text reads "Create real-world, real-time experiences with the latest Maps, Routes, and Places features from Google Maps Platform. Built by the Google team for developers everywhere." Three buttons are present: "Get Started", "See our new pricing structure", and "Visit documentation →".
- Search Bar:** A simple search bar containing the text "The University of Tokyo".

First-time geocoded data

Repository	Stars	Contributor	Location	City	Country
AutoGPT	171829	Pwuts	Delft, NL	Delft	Netherlands
AutoGPT	171829	Torantulinc	UK	Ouaka	Ködörösé tî Béafrîka / République centrafricaine
AutoGPT	171829	richbeales	Kent, UK	Kent	United Kingdom
AutoGPT	171829	majdyz	Amsterdar	Amsterdar	Nederland
AutoGPT	171829	aarushik93	Berlin	Berlin	Deutschland
AutoGPT	171829	ntindle	Texas	Texas	United States
AutoGPT	171829	Bentlybro	Some Wh	Unknown	Unknown
AutoGPT	171829	kcze	Scotland	Glasgow	Scotland
AutoGPT	171829	collijk	United Sta	United Sta	Unknown
AutoGPT	171829	AndresCde	Colombia	Colombia	Unknown
AutoGPT	171829	p-i-	Oxford, En	Oxford	England
AutoGPT	171829	sadmuphir	Somewher	Somewher	United Kingdom
AutoGPT	171829	maiko	Serris	Serris	France
AutoGPT	171829	drikusroor	Utrecht, N	Utrecht	Netherlands
AutoGPT	171829	Abhi19920	India	India	Unknown
AutoGPT	171829	erik-megai	Denver	Denver	United States
AutoGPT	171829	Wladastic	Germany	Deutschlar	Unknown
AutoGPT	171829	k-boikov	Sofia, Bul	Sofia	Bulgaria
AutoGPT	171829	rhp	Schengen	Schengen	Lëtzebuerg
AutoGPT	171829	dschonhol	Boston, M.	Boston	MA
AutoGPT	171829	Taytay	Tulsa, OK	Tulsa	OK
AutoGPT	171829	andrewhoc	Texas	Texas	United States
AutoGPT	171829	AlrikOlson	Portland, C	Portland	OR
AutoGPT	171829	OmriGM	Israel	ישראל	Unknown
AutoGPT	171829	Androbin	Passau, G	Passau	Germany
AutoGPT	171829	russelloce	South Carr	South Carr	United States
AutoGPT	171829	bituq	The Nethe	Nederland	Unknown
AutoGPT	171829	nponecco	Barcelona,	Barcelona	Spain
AutoGPT	171829	eltoclear	Tokyo, Ja	Tokyo	Japan
AutoGPT	171829	didier-dur	Switzerlan	Consulat	France
AutoGPT	171829	Sma-Das	New York	City of Nev	United States
AutoGPT	171829	gersh	San Franci	San Franci	Canada

Geocoding— handle partial location input

Lessons Learned & Optimization

Second-time geocoded data

MACS 30112

Winter 25

Final Project

A	B	C	D	E	F	G	H	I	J	K	L
Repository	Stars	Contributor	Location	City	Country	Latitude	Longitude				
AutoGPT	171829	Pwuts	Delft, NL	Delft	Netherlands	52.011576	4.357068				
AutoGPT	171829	Torantulinc	UK	Ouaka	Ködörösës	55.37805	-3.43597				
AutoGPT	171829	richbeales	Kent, UK	Kent	United Kingdom	51.09283	0.8392470999999999				
AutoGPT	171829	majdyz	Amsterdam	Amsterdam	Nederland	52.36757	4.904139				
AutoGPT	171829	aarushik93	Berlin	Berlin	Deutschland	52.520006	13.40495				
AutoGPT	171829	ntindle	Texas	Texas	United States	30.272728	-97.7411				
AutoGPT	171829	kcze	Scotland	Glasgow	Scotland	56.490671	-4.20265				
AutoGPT	171829	collijk	United States	United States	Unknown	38.7946	-106.535				
AutoGPT	171829	AndresCdeC	Colombia	Colombia	Unknown	4.570868	-74.2973				
AutoGPT	171829	p-i-	Oxford, EN	Oxford	England	51.75202	-1.25773				
AutoGPT	171829	maiko	Serris	Serris	France	48.84432	2.788231				
AutoGPT	171829	drikusroo	Utrecht, NL	Utrecht	Netherlands	52.091925	5.122957				
AutoGPT	171829	Abhi19920	India	India	Unknown	20.59368	78.96288				
AutoGPT	171829	erik-megai	Denver	Denver	United States	39.73924	-104.99				
AutoGPT	171829	Wladastic	Germany	Deutschland	Unknown	51.16569	10.45153				
AutoGPT	171829	k-boikov	Sofia, Bulgaria	Sofia	Bulgaria	42.69771	23.32187				
AutoGPT	171829	dschonhol	Boston, MA	Boston, MA	United States	42.35551	-71.0565				
AutoGPT	171829	Taytay	Tulsa, OK	Tulsa	OK	36.15514	-95.9895				
AutoGPT	171829	andrewhoc	Texas	Texas	United States	30.272728	-97.7411				
AutoGPT	171829	AlrikOlson	Portland, OR	Portland	OR	45.51523	-122.678				
AutoGPT	171829	OmriGM	Israel	ישראל	Unknown	31.04605	34.85161				
AutoGPT	171829	Androbin	Passau, Germany	Passau	Germany	48.56674	13.43195				
AutoGPT	171829	russelloce	South Carolina	South Carolina	United States	34.84849	-82.4				
AutoGPT	171829	bituq	The Netherlands	Nederland	Unknown	52.13263	5.291265999999999				
AutoGPT	171829	nponecco	Barcelona, Spain	Barcelona	Spain	41.3874	2.168568				
AutoGPT	171829	eltoceiar	Tokyo, Japan	Tokyo	Japan	35.67642	139.65				
AutoGPT	171829	didier-dur	Switzerland	Consulat	France	46.81819	8.227511999999999				
AutoGPT	171829	Sma-Das	New York City	of New York	United States	40.71278	-74.006				
AutoGPT	171829	gersh	San Francisco	San Francisco	Canada	37.77493	-122.419				
AutoGPT	171829	sagarisher	Åland	Åland	Suomi / Finland	60.17852	19.91561				
AutoGPT	171829	dillweed	USA	United States	Unknown	38.7946	-106.535				

Using batch geocoding significantly increases the efficiency and reduces processing time

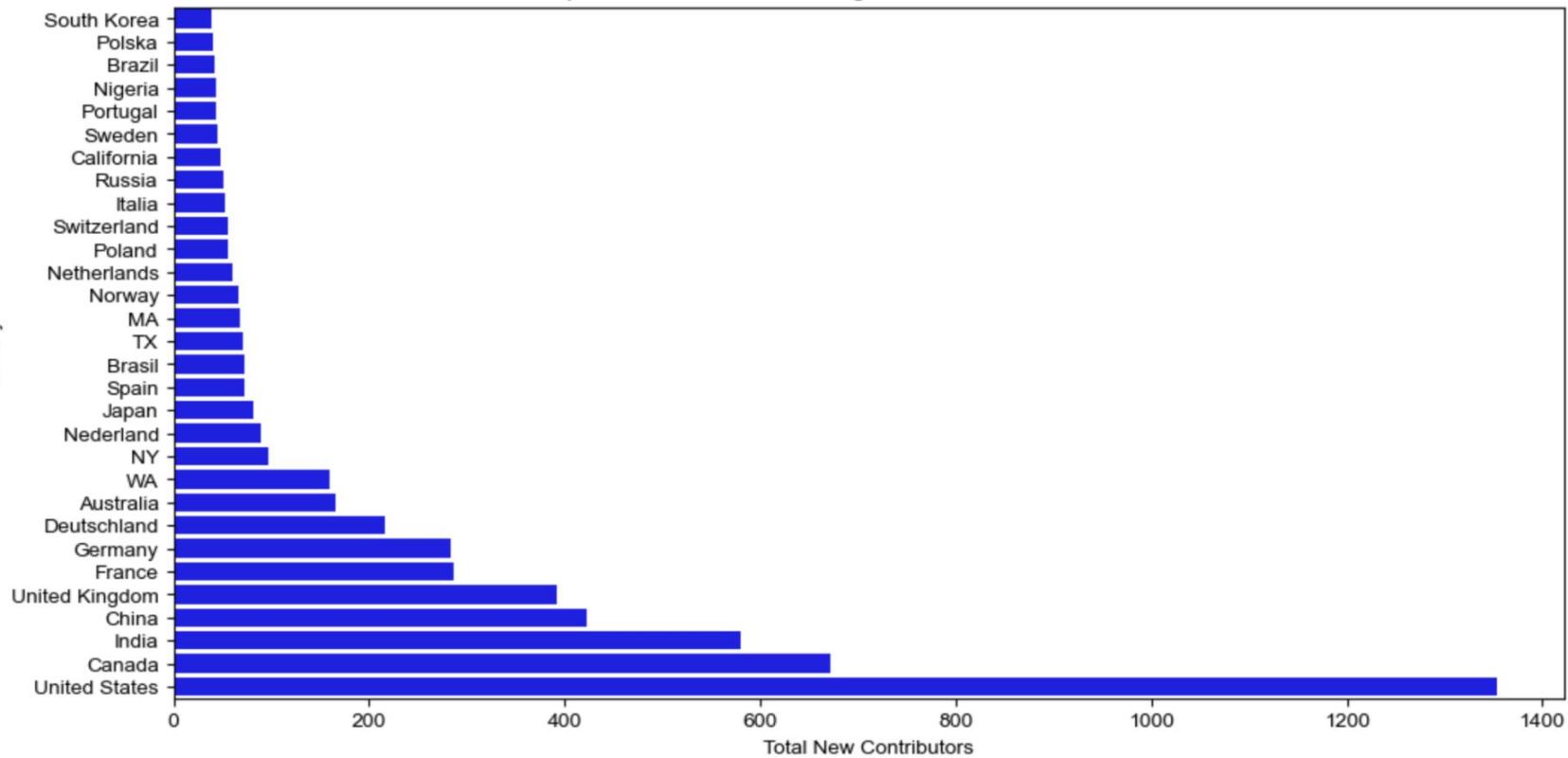
Lessons Learned & Optimization

Interactive Mapping & Insights

https://github.com/macs30112-winter25/final-project-chick-fil-a/blob/main/Data%20Visualization/top_1000_contributors_map.html

Further Analysis

Top 30 Countries with the Highest New GitHub Contributors



Data Wrangling



Typical errors

```
1 Country,total_contributors,total_stars,avg_stars_per_contributor
2
3
4
5
6 02139,1,5832,5832.0
7 059817,1,36520,36520.0
8 102676,1,71195,71195.0
9 110078,1,9536,9536.0
10 230022,1,12510,12510.0
11 29880,1,8023,8023.0
12 37.710840,1,35079,35079.0
13 53154,1,40041,40041.0
14 60637,1,11552,11552.0
15 65000,1,3448,3448.0
```

1. containing numbers or symbols instead of actual country names



Typical errors

```
1 Country,total_contributors,total_stars,avg_stars_per_contributor
204 Unknown,2133,65985851,30935.01359510/815658/434
205 Uruguay,3,177554,59184.6666666666664
206 Uruguay.,1,9466,9466.0
```

2. Country names failing to merge correctly due to **extra symbols...**

1 Country, total_contributors, total_stars, avg_stars_per_contributor

479 The Netherlands, 1, 40203, 40203.0

480 The Netherlands, 22, 865609, 39345.86363636364

1 Country, total_contributors, total_stars, avg_stars_per_contributor

533 Washington, 15, 425921, 20201.7

534 West Bengal, 2, 27171, 13585.5

535 WestBengal, 1, 9536, 9536.0

Typical errors

2. Country names failing to merge correctly due to **extra symbols**

Or multiple spaces...

	Country	total_contributors	total_stars	avg_stars_per_contributor
589	Česko	20,662714	33135.7	
590	Ελλάς	10,305596	30559.6	
591	Κύπρος – Κύπρος	7,242829	34689.857142857145	
592	България	5,183360	36672.0	
593	Кыргызстан	1,19570	19570.0	
594	Россия	43,1606794	37367.3023255814	
595	Санкт-Петербург	1,4536	4536.0	
596	Србија	6,95379	15896.5	
597	Україна	10,185262	18526.2	
598	Қазақстан	3,25688	8562.666666666666	
599	Հայաստան	9,82476	9164.0	
600	ישראל	10,366669	36666.9	
601	الأردن	2,23624	11812.0	
602	إمارات العربية المتحدة	13,175154	13473.384615384615	
603	البحرين	1,77961	77961.0	
604	اليمن	1,41513	41513.0	
605	ایران	12,442055	36837.916666666664	
606	تونس	7,107792	15398.857142857143	
607	لبنان	1,4206	4206.0	
608	مصر	3,20762	6920.666666666667	
609	موريتانيا	3,124427	41475.666666666664	
610	پاکستان	11,555376	50488.72727272727	
611	नेपाल	1,40041	40041.0	
612	বাংলাদেশ	2,16086	8043.0	
613	ประเทศไทย	5,133648	26729.6	
614	საქართველო	1,8101	8101.0	
615	አ.ኋናድን	1,29022	29022.0	
616	中国	660,21432648	32473.70909090909	
617	帝都	1,56047	56047.0	
618	日本	66,2259715	34238.106060606064	
619	臺灣	10,302257	30225.7	
620	대한민국	31,1207980	38967.096774193546	

Typical errors

3. Country names appearing in **different languages**.

Typical errors

```
1 Country,total_contributors,total_stars,avg_stars_per_contributor  
555 anyone can find me,1,11633,11633.0  
556 az,1,4206,4206.0  
557 but open to relocate,1,36153,36153.0
```

4. Or, some **naughty (misleading) words from user**

- Regular Expressions;
- Google Translator;
- Pycountry;
- high-threshold fuzzy matching;
- Manual verification



Hindi Arabic French Irish
Latvian Catalan Macedonian Swahili Welsh
Hebrew Romanian Persian Korean
Estonian Belarusian Russian
Ukrainian Vietnamese Yiddish
Croatian Bulgarian Chinese Portuguese Japanese
Danish Spanish Hungarian Lithuanian Maltese
Galician Serbian Indonesian Afrikaans Norwegian Slovenian
Turkish Albanian Filipino English Swedish Thai
Dutch Italian Slovak Finnish German Polish
Greek Czech

Google Translate



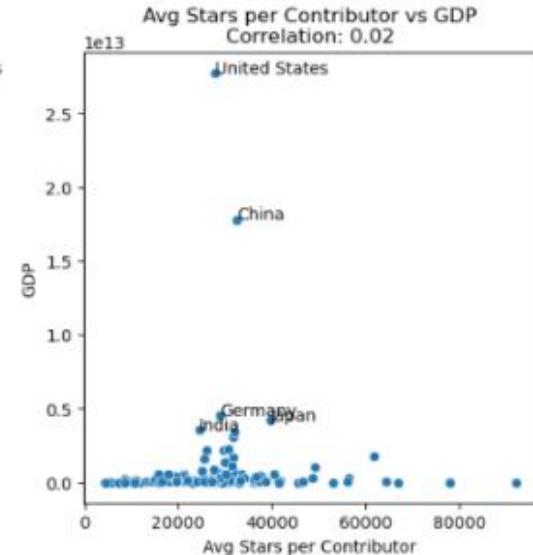
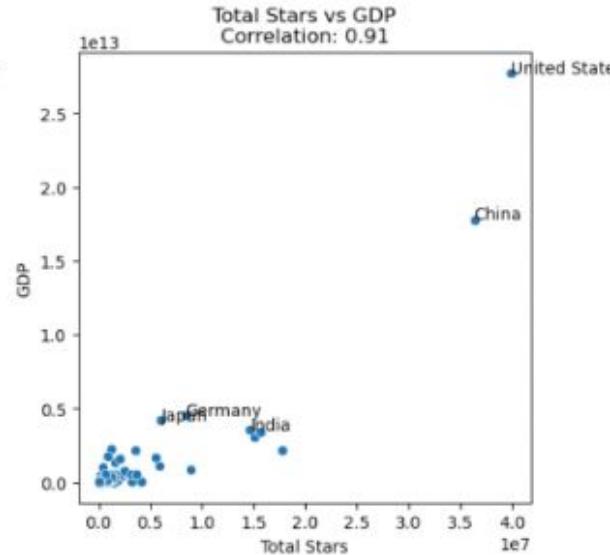
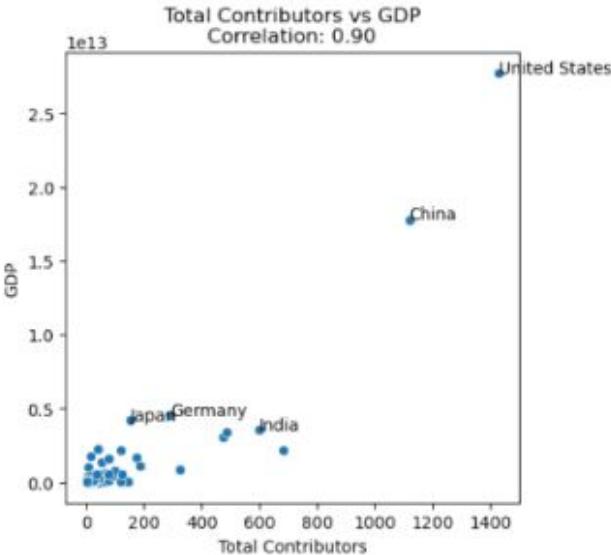


Our new dataframe included:

- **GDP per capita (2023);**
- **R&D Investment Index;**

—national investment into research and development

- **Annual AI Publications**



the United States, China, Germany, Japan, and India stood out as leading countries in AI open-source development.



Exploratory Statistical Description

- Location dataset

	Repository	Stars	Contributor	Location	City	Country	Latitude	Longitude
0	AutoGPT	171829	Pwuts	Delft, NL	Delft	Netherlands	52.011577	4.357068
1	AutoGPT	171829	Torantulino	UK	Ouaka	centrafric...	55.378051	-3.435973
2	AutoGPT	171829	richbeales	Kent, UK	Kent	United Kingdom	51.092832	0.839247
3	AutoGPT	171829	majdyz	Amsterdam	Amsterdam	Nederland	52.367573	4.904139
4	AutoGPT	171829	aarushik93	Berlin	Berlin	Deutschland	52.520007	13.404954

	Stars	Latitude	Longitude
count	10618.000000	10618.000000	10618.000000
mean	29438.816444	35.729564	5.898528
std	32427.544107	18.781563	84.255018
min	3263.000000	-82.862752	-159.315717
25%	8195.000000	30.585164	-74.174423
50%	19570.000000	39.739236	8.719818
75%	36520.000000	48.135125	78.962880
max	171829.000000	67.496897	176.846151



Exploratory Statistical Description

- Country-level dataset

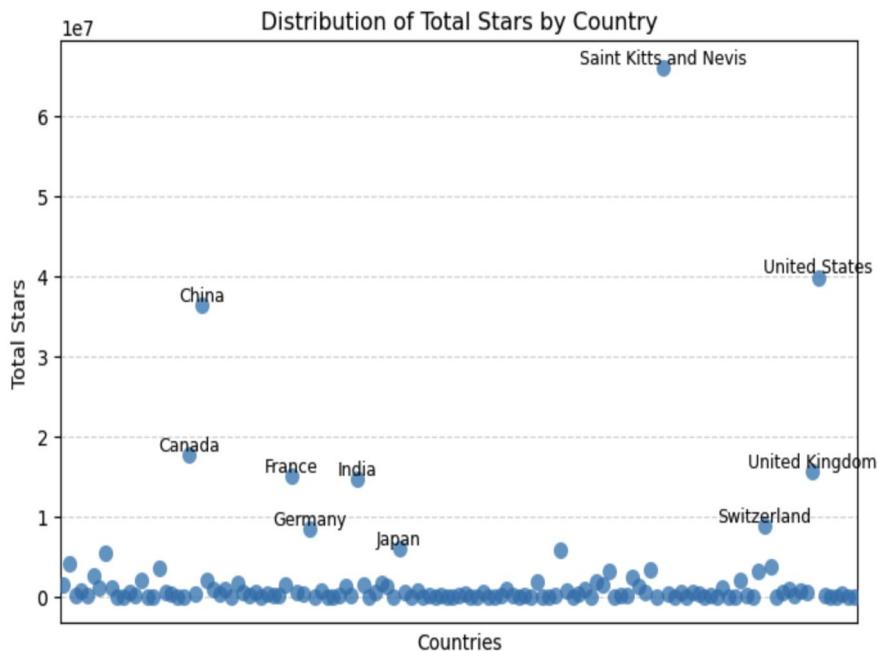
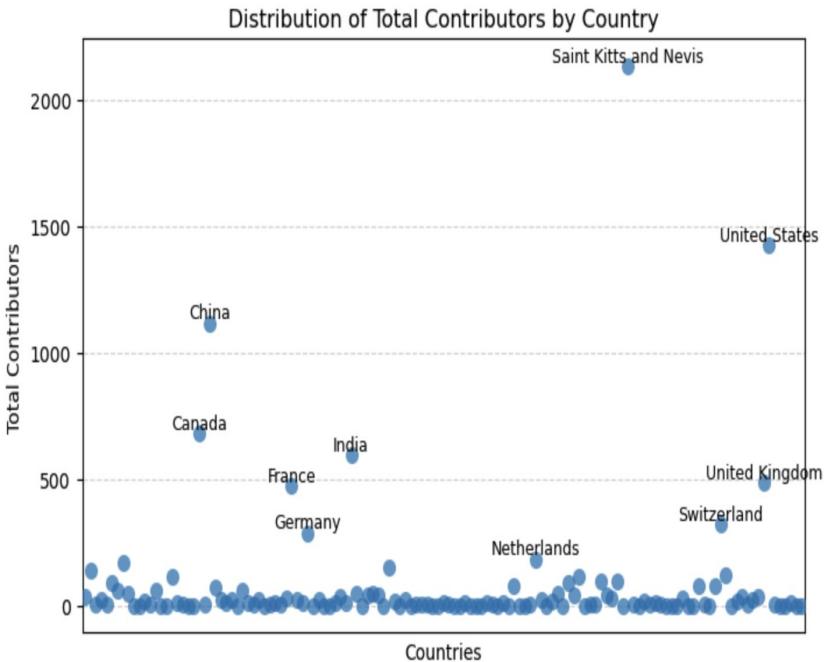
Country_Standardized	total_contributors	total_stars	avg_stars_per_contributor
Afghanistan	40	1439431	35985.775000
Albania	143	4160394	29093.664336
Algeria	6	111226	18537.666667
Andorra	22	683379	31062.681818
Antigua and Barbuda	4	267656	66914.000000

	total_contributors	total_stars	avg_stars_per_contributor
count	133.000000	1.330000e+02	133.000000
mean	80.556391	2.382014e+06	29014.113829
std	256.805740	7.739081e+06	15528.110078
min	1.000000	4.206000e+03	4206.000000
25%	2.000000	5.316400e+04	19417.428571
50%	10.000000	3.256340e+05	27714.826087
75%	41.000000	1.254455e+06	35569.239583
max	2133.000000	6.598585e+07	92111.500000



Exploratory Statistical Description

- Country-level dataset

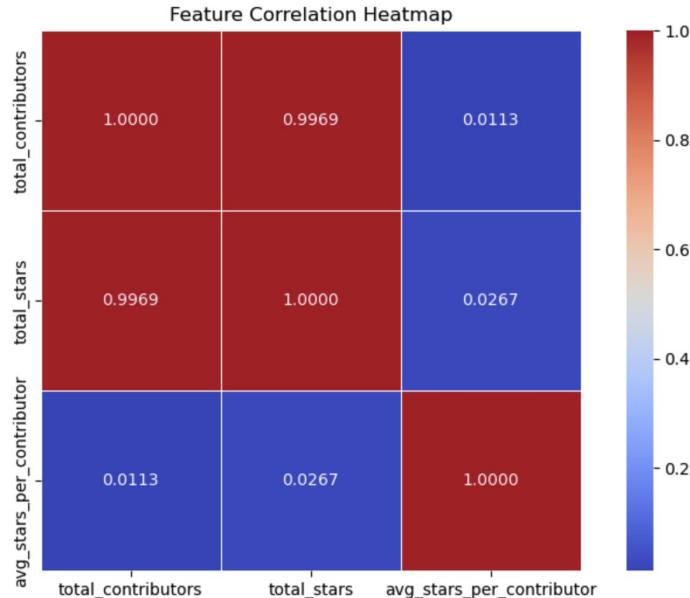




Exploratory Statistical Description

- Country-level dataset

Correlation Matrix :



Data Analysis



Dependent Variable:

AI Open-Source Participation Index (AIOSPI)

1. **Geographic AIOSPI** (Total influence in a country):

$$\text{AIOSPI} = 0.5 \times (\text{number of contributors}) + 0.5 \times (\text{total star count})$$

2. **Per Capita AIOSPI** (Influence per person):

$$\text{AIOSPI} = 0.5 \times \left(\frac{\text{number of contributors}}{\text{population}} \right) + 0.5 \times \left(\frac{\text{total star count}}{\text{population}} \right)$$

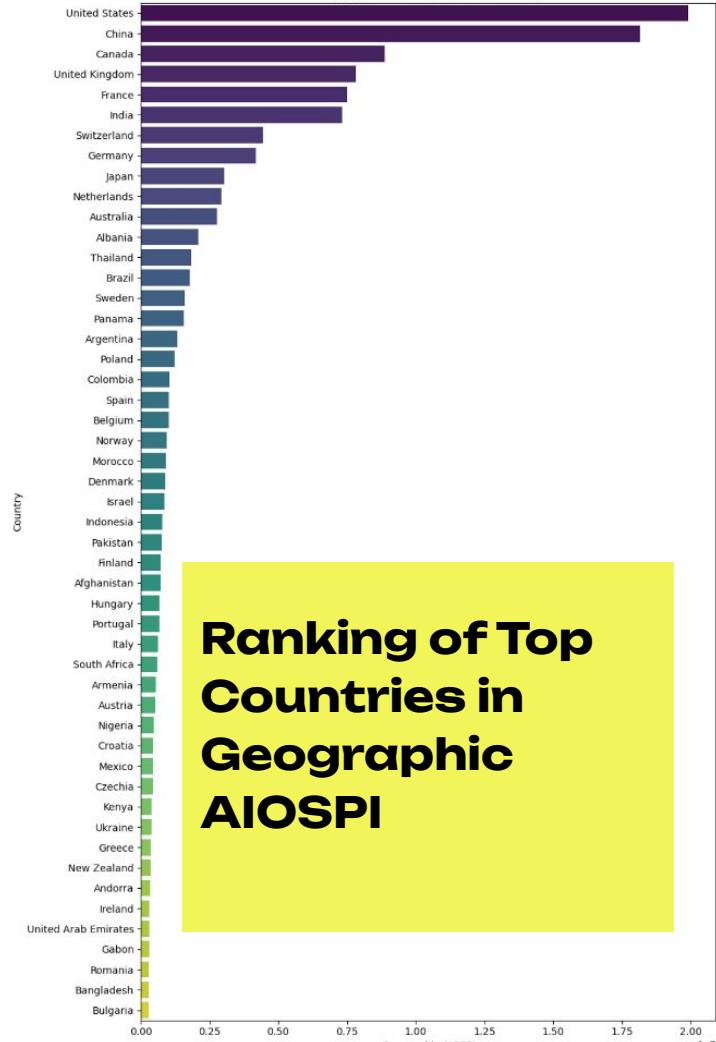
Description of Final Dataset

	total_contributors	total_stars	avg_stars_per_contributor
count	122.000000	1.220000e+02	122.000000
mean	68.803279	2.007207e+06	29167.534867
std	191.337685	5.605192e+06	15873.368717
min	1.000000	4.206000e+03	4206.000000
25%	2.250000	5.857075e+04	19092.935268
50%	10.500000	3.296325e+05	27834.798036
75%	44.000000	1.309785e+06	35494.525694
max	1426.000000	3.986350e+07	92111.500000

	Geographic_AIOSPI	Per_Capita_AIOSPI
count	1.220000e+02	122.000000
mean	1.003638e+06	0.129959
std	2.802691e+06	0.460559
min	2.103500e+03	0.000093
25%	2.928600e+04	0.003348
50%	1.648195e+05	0.020317
75%	6.549148e+05	0.067789
max	1.993246e+07	4.226038

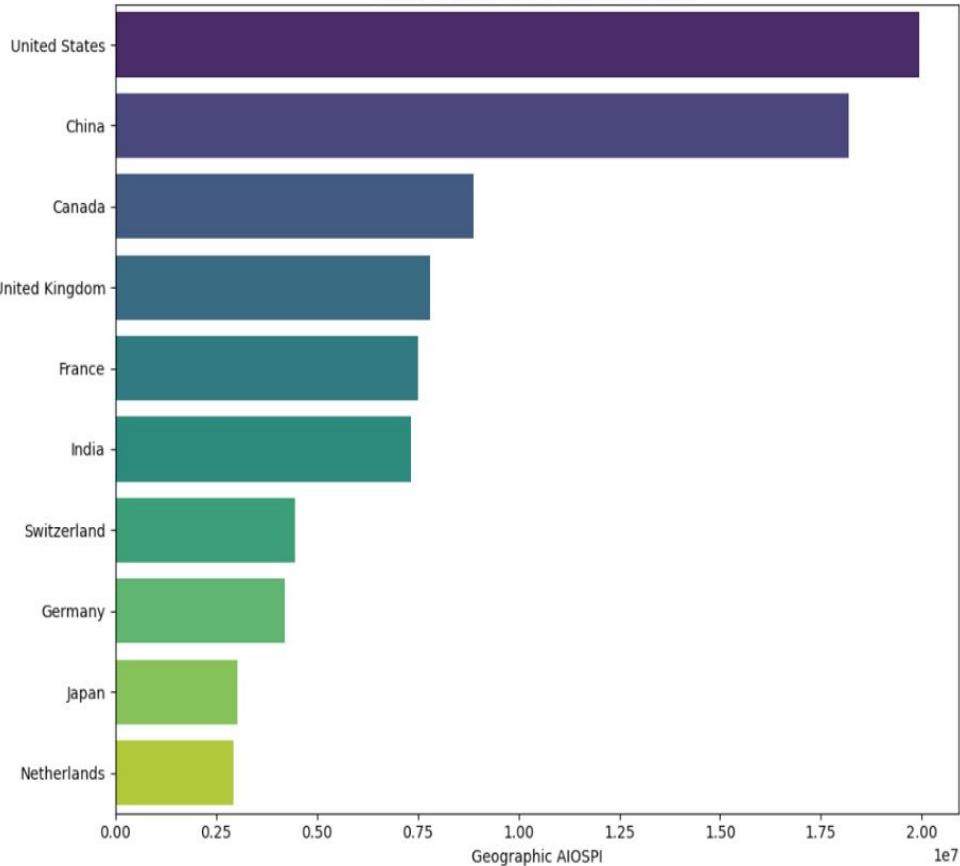
	Population	star_per_capita	contributor_per_capita
count	1.220000e+02	122.000000	1.220000e+02
mean	5.497473e+07	0.259910	8.164675e-06
std	1.863675e+08	0.921091	2.742397e-05
min	1.187500e+04	0.000186	2.804868e-08
25%	3.408658e+06	0.006695	2.333296e-07
50%	1.075445e+07	0.040633	1.486054e-06
75%	3.642859e+07	0.135573	5.212759e-06
max	1.438070e+09	8.451803	2.720887e-04

Geographic AIOSPI (Top 50 Countries)



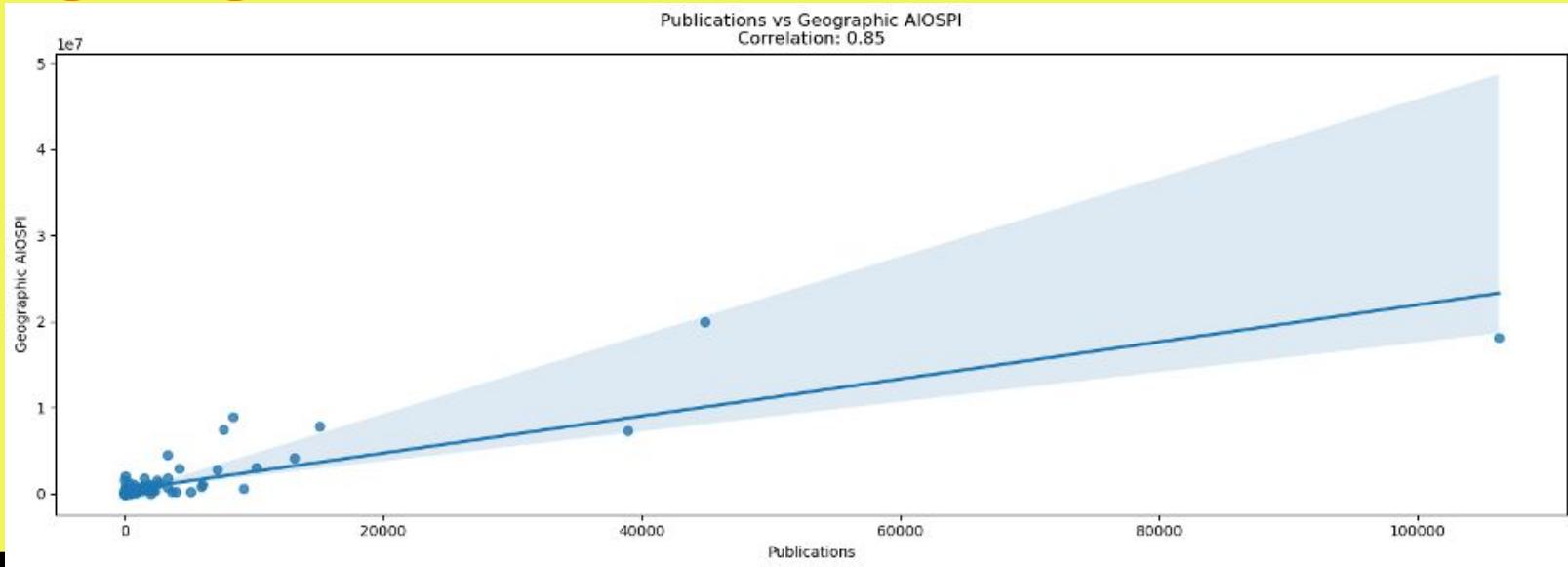
Ranking of Top Countries in Geographic AIOSPI

Geographic AIOSPI (Top 10 Countries)



Correlations:

**Geographic AIOSPI and AI Publications
(Education/Research) had a strong correlation
—0.85**



Regression Analysis:

66

Our variables are numerical—

Linear regression

- the coefficient is **215.56**, meaning for every 1 additional AI publication, the AI index would increase by 215.56.
 - P-value = 0.000, meaning this relationship is highly significant.
 - R^2 showing that **72.6%** of the variation in Geographic AIOSPI can be explained by the number of AI publications.

```

OLS Regression Results
=====
Dep. Variable: Geographic_AIOSPI R-squared: 0.726
Model: OLS Adj. R-squared: 0.723
Method: Least Squares F-statistic: 306.6
Date: Wed, 05 Mar 2025 Prob (F-statistic): 2.34e-34
Time: 08:10:10 Log-Likelihood: -1844.2
No. Observations: 118 AIC: 3692.
Df Residuals: 116 BIC: 3698.
Df Model: 1
Covariance Type: nonrobust
=====
            coef      std err          t      P>|t|      [0.025      0.975]
=====
const    3.907e+05  1.43e+05     2.740     0.007    1.08e+05  6.73e+05
Publications   215.5584    12.310     17.511    0.000    191.177  239.940
=====
Omnibus:             125.136 Durbin-Watson:       1.668
Prob(Omnibus):      0.000 Jarque-Bera (JB): 2354.561
Skew:                 3.622 Prob(JB):           0.00
Kurtosis:              23.650 Cond. No. 1.20e+04
=====
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.2e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

```

1. Leading countries:

**United States, China,
Germany, India, and
Japan**

2. Leading cities:

**San Francisco, Beijing, Seattle, London,
New Delhi, Berlin, and Deutschland**

Key Findings:

3. Important indicators:

GDP

**high-level research
output (AI publications)**

4. Global Inequality



Questions Response:

1.Perhaps comparing G77 countries/cities might give important insights into 'emergent' AI contributors.....

Great suggestion! We should indeed refine our country/region classification further. Additionally, we are considering incorporating the creation time of each repository to analyze the development speed across different countries. This would make our research even more valuable.

2.I'm curious about how you conducted the manual validation for the Google Translate API results—did you randomly sample translations for accuracy checks?

We saved the results separately after using the Google Translate API. For the city-level data, which is relatively large, we applied random sampling. For the country-level data, which is relatively smaller, we manually verified each entry.

3.Curious on how to standardize city and country names.....

We used tools such as pycountry and geocode to match the country and city names in the data with the built-in information packages, thereby standardizing them.

4.it would be great to dive deeper into the social science angle—especially looking at global inequality. I'm also curious how this perspective could be explored within a political context, such as technological competition between nations.....

Yes~ We considered this approach as well! Initially, we aimed to conduct a comparative analysis of cities, particularly those with similar endowments but different levels of AI development progress. The discussion would explore whether factors such as policies promoting AI development played a role. This remains a feasible research direction.



Questions Response:

5.the limitations of using GitHub as a primary data source for AI development.....

Thank you for your question. We indeed haven't verified GitHub's global usage distribution; we only know that it has a large user base and that previous researchers have used GitHub-related data. Our initial reasoning was that, since we are focusing on participation in open-source projects, GitHub data can reflect certain aspects of this engagement. But you raise a reasonable point—perhaps we should first confirm this assumption to reduce potential bias in our analysis.

6.incorporate number of universities in those cities/countries.....

Yes, you're right. We initially intended to integrate university/research institute data into our dataset but couldn't find high-quality data sources at the time. Thank you for pointing this out!

7.to explore whether government AI policies or funding initiatives influence.....

Yes! These are crucial institutional and policy variables. Thank you for your question!

8.Very curious about your method of making the interactive map.....

We used OpenStreetMaps and Google Maps API to locate each contributors' locations and return their latitudes and longitudes, with that we use the Folium Map function to generate the interactive map, we also set the function to pop up each contributor's location and their specific information.

Questions Response:

9.did you run your regression without the more extreme values to test it.....

Good question! We did not remove the outliers because our research focuses on the global economy rather than being limited to medium-sized economies. We believe that outliers may be a crucial part of the data rather than mere noise. For example, the United States has a significantly higher GDP than other countries, but it is an essential part of the global economic system. Therefore, we chose to retain these data points to ensure the completeness of our analysis. However, to further examine the robustness of our model, we may need to assess whether these outliers have an excessive impact on the results.

10.Will you analyze regional disparities within countries?Can you track AI talent migration (e.g., where top AI researchers obtained their degrees vs. where they work now).....

We attempted to examine intra-country differences from a city-level perspective. However, further granularity is challenging. On one hand, users rarely specify smaller locations themselves, even though such data does exist. On the other hand, the accuracy of geocoding is sometimes insufficient to achieve the desired level of precision.

11.Really interesting project and approach! I'd love to hear more about how you guys figured out the geocoding process: what resources did you use to approach this part of the project?.....

Resources we used: OpenStreetMaps, Google Maps, Google clouds. Folium Map. Html.

We think one thing we learnt from this approach is that geocoding really needs a lot of specific information(coordinates), so it is better to make your functions return as many dataset as possible.



Thanks for your
listening!

Anqi Wei, Charlotte Li, Baihui Wang