Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Where the World went Missing: Geographic Inequality in Wikipedia and AI

Juan M. Lavista Ferres

jlavista@microsoft.com

Microsoft (United States)

Caleb Robinson

Microsoft (United States)

Anthony Ortiz

Microsoft (United States)

Marian Blazes

Microsoft (United States)

Inbal Becker-Reshef

Microsoft (United States)

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11413641/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. All authors are employees of Microsoft Corporation. Microsoft provided salary support and computational resources for this research. Microsoft has a commercial relationship with OpenAI, the developer of GPT, which was evaluated in this study.

Where the World went Missing: Geographic Inequality in Wikipedia and AI

Authors. Juan M. Lavista Ferres[1]*, Caleb Robinson[1]; Anthony Ortiz [1]; Marian Blazes [1], Inbal Becker-Reshef [1]

[1] Microsoft AI for Good Research Lab, Redmond, WA, USA

*Corresponding author: jlavista@microsoft.com

Abstract

Wikipedia is among the world's most important public knowledge datasets and a foundational source for search engines, maps, knowledge graphs, and AI systems, including LLMs. Geographic gaps in coverage may therefore influence what downstream systems can retrieve, infer, and represent.

We analyzed 21,270,978 geotagged records across 355 Wikipedia language editions by combining coordinates extracted from article wikitext and Wikidata. We release the dataset and interactive visualization as open resources for research on geographic representation and knowledge inequality. The Global North contained 275.6 records per 1,000 km², compared with 79.0 in the Global South, a density ratio of 3.49. Although this ratio declined from 5.60 in 2010 to 3.49 in 2026, it remained substantial and relatively stable over the past decade.

To examine whether this was reflected in AI performance, we asked GPT-5.4 to estimate the populations of 20,000 villages sampled from OpenStreetMap. Plausible estimates were 3.9 times more likely in the Global North than the Global South (37.9% vs. 9.8%). Villages with Wikipedia evidence were also substantially more likely to receive plausible estimates.

These findings suggest that geographic inequality in foundational knowledge sources is systematically associated with geographic inequality in AI capability, further marginalizing places that are already underrepresented.

Introduction

Wikipedia has become a major component of the modern digital knowledge ecosystem. The content is created to inform people but is also used by search engines, maps, linked-data systems, language technologies, and large-scale AI systems. Wikidata (launched in 2012) extends this infrastructure by linking Wikipedia pages to structured entities and properties, including geographic coordinates [1-4]. Because Wikipedia and Wikidata are open, widely used, and available in many languages, their coverage patterns can propagate into other systems.

Geographic coverage is especially important because coordinates make knowledge spatially addressable. A geotagged article can appear on a map, be found in a nearby search, be joined to other spatial data, or be used as a point in a geographic analysis. The set of geotagged Wikipedia articles can therefore be viewed as a spatial index of encyclopedic attention. It includes settlements, landmarks, administrative units, natural features, historical sites, institutions, and many other location-bearing topics. If this index is uneven, then systems that reuse it, inherit an uneven view of the world.

Imbalances in geographic representation in Wikipedia can occur due to multiple factors. The distribution of geotagged articles in Wikipedia is the result of a mixed process of human editorial attention, automated import practice, and source-data availability. Less populated or less digitally connected regions may have fewer human contributors generating content, including manually geotagging articles. Automated agents (bots) can import structured data (including geographic data) from external databases into Wikidata and generate short Wikipedia articles based on those data. Bots are often deployed by editors or volunteers attempting to address gaps in Wikipedia coverage, but these additions may be limited by the quality of the underlying datasets, which may have incomplete geographic information. The paper treats geotagged article count as a measure of spatial coverage, not as a measure of article quality. A short bot-generated stub and a long, carefully referenced article are not equivalent. Both, however, can place a topic in geographic space. That makes article-level geotags a useful measure of which places are visible to systems that index and reuse Wikipedia.

Prior work has documented geographic imbalance in Wikipedia. Graham et al. reported that about 84% of geotagged articles in the mid-2010s were located in Europe and North America and identified per-country informational disparities [5]. Hecht and Gergle showed that different language editions cover different concepts [6]. Miquel-Ribe and Laniado quantified cultural-context content across 40 language editions, finding that each edition contains unique cultural content that is often missing in other language editions. [7]. These studies establish that the imbalance exists. At least three questions remain open. First, most prior counts rely on wikitext coordinates (coordinates manually embedded in Wikipedia article markup) from a single or small number of language editions; the arrival of Wikidata as a coordinate source has changed the corpus composition materially, and the current multilingual imbalance has not been re-measured at scale. Second, the temporal trajectory over the full Wikidata era has not been

characterized; whether the gap is closing, stable, or widening is empirically undetermined. Third, whether the Wikipedia imbalance propagates into contemporary AI systems has been discussed conceptually but not audited directly with a paired corpus-and-model design [8-14].

This study focuses on the geotagged layer of multilingual Wikipedia. Building on earlier information-geography literature [5, 15-19] and on parallel work on volunteered geographic information and web-scale AI training corpora [8-14], the analysis specifically asks four empirical questions. Q1: How uneven is the spatial density of geotagged Wikipedia articles between the Global North and the Global South, Q2: Has this imbalance changed over time, Q3: How much do language-edition structure and bot-heavy article creation affect the observed distribution, and Q4: Is the same geographic pattern visible in LLM answers about small places? We address these questions using a 21.27M-record corpus that combines wikitext coordinates and Wikidata claims (structured statements specifying geographic location) across 355 language editions, a 16-year cumulative time series, and a paired language-model audit on 20,000 OpenStreetMap villages.

Results

Data: A cleaned multilingual corpus of 21.27 million geotagged records

The working merged corpus contained 21,335,644 raw geotagged records across 355 Wikipedia language editions. Records combined two sources: coordinates parsed from article wikitext and Wikidata coordinate claims for site-linked articles. A cleaning pass removed 64,666 records, or 0.30% of the raw corpus, leaving 21,270,978 records for analysis.

The removed records came from two classes of artifact. The first class contained 21,578 astronomy-related records whose celestial coordinates were erroneously included in the geographic corpus. The second contained 43,088 integer-degree open-ocean lattice records.

The cleaned dataset is a present-day geospatial article index. It records whether an article in a given language edition has a usable coordinate in the assembled corpus. It does not measure article length, citation quality, page views, local authorship, or reader demand.

Global North-South Classification: Global North countries have 3.49 times the geotag density of Global South countries

After country assignment using the classification described in Methods, Global North countries contained 14,693,485 geotagged records across 53,317,374 km^2 of land. This corresponds to 275.6 records per 1,000 km2. Global South countries contained 6,550,352 records across 82,942,339 km2, or 79.0 records per 1,000 km2. (Table 1) Antarctica was excluded from both groups and reported as N/A. The aggregate density ratio was 3.49. (Figure 1)

The imbalance was also present in English language Wikipedia alone. English Wikipedia contained 1,055,239 Global North records and 404,867 Global South records. Using the same

land areas, English Wikipedia had 19.8 records per 1,000 km2 in the Global North and 4.9 in the Global South, an aggregate ratio of 4.05.

Table 1. Wikipedia Geotag records per 1000 km2, Global South vs Global North.

Dataset	Group	Records	Land area, km2	Records per 1,000 km2
All languages	Global North	14,693,485	53,317,374	275.6
All languages	Global South	6,550,352	82,942,339	79.0
English only	Global North	1,055,239	53,317,374	19.8
English only	Global South	404,867	82,942,339	4.9

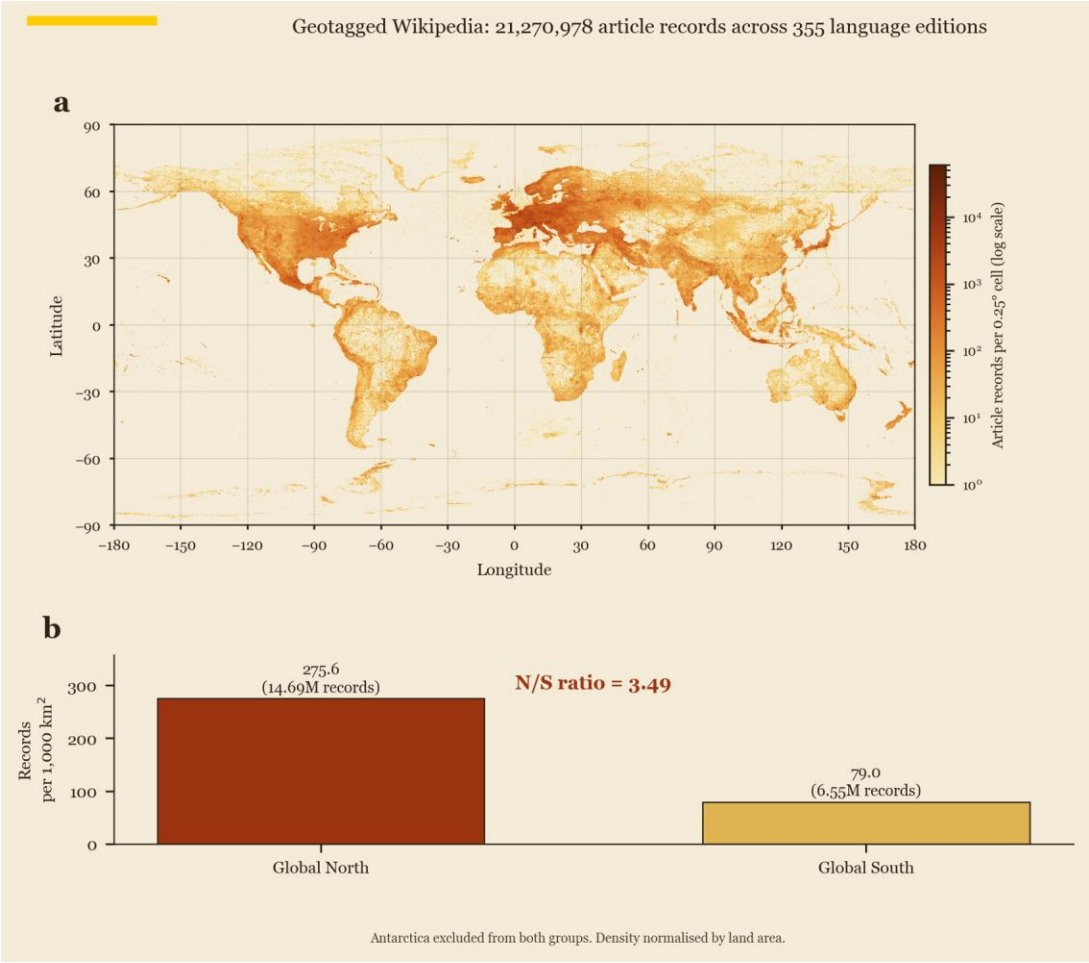


Figure 1. Global density of 21,270,978 geotagged Wikipedia records across 355 language editions. A world map showing the cleaned all-language geotag density (a), with a companion panel comparing aggregate Global North and Global South densities (b). Antarctica was excluded from both groups. Density normalized by land area.

These are aggregate area-normalized densities. They do not imply that every Global North country has a higher geotagged article density than every Global South country. Small territories and densely documented city-regions can have high per-area values in either group. The aggregate comparison asks a different question: given all land area in each group, how dense is the geotagged Wikipedia layer? On that measure, the Global North remains much more densely represented.

Because Global North/South classification is an analytical choice, we also recomputed the aggregate density ratio under three alternative groupings. The all-language ratio remained between 2.75 and 3.49, and the English-only ratio remained between 3.48 and 4.05. (Supplementary Table S2)

Absolute coverage and density rankings identify different places

The largest absolute record counts were concentrated in large, heavily documented, or heavily imported countries. The top countries by total geotagged records were the United States (2.59 million), France (1.75 million), Russia (1.13 million), Mexico (1.01 million), Germany (0.92 million), Italy (0.87 million), Spain (0.67 million), Poland (0.65 million), Norway (0.54 million), and the United Kingdom (0.53 million).

Density rankings gave a different view. Among countries with at least 1,000 km2 of land, the highest densities were Hong Kong, Luxembourg, Aland Islands, Slovenia, Faroe Islands, Palestine, Switzerland, Czechia, the Netherlands, France, Belgium, and South Georgia. These rankings are useful for identifying local outliers, but they are sensitive to small land areas and the treatment of territories. For that reason, the paper uses aggregate North/South density as its main statistic and treats country-level ranks as descriptive.

Temporal Analysis: The gap has narrowed since 2010 but remains large

The cumulative North/South density ratio decreased over time. In January 2010, records now present in the cleaned geotagged corpus had creation dates corresponding to 3,102,360 cumulative Global North records and 861,469 Global South records. Normalized by land area, these corresponded to 58.2 and 10.4 records per 1,000 km2, respectively, a ratio of 5.60.

By May 2026, the cumulative series reached 14,678,253 Global North records and 6,544,636 Global South records, corresponding to 275.3 and 78.9 records per 1,000 km2. The ratio had narrowed to 3.49. The change indicates relative improvement in Global South coverage, but the final ratio still shows a large imbalance. It is important to note in Figure 2 that most of the improvement occurred before the year 2016, after which the ratio remained relatively stable.

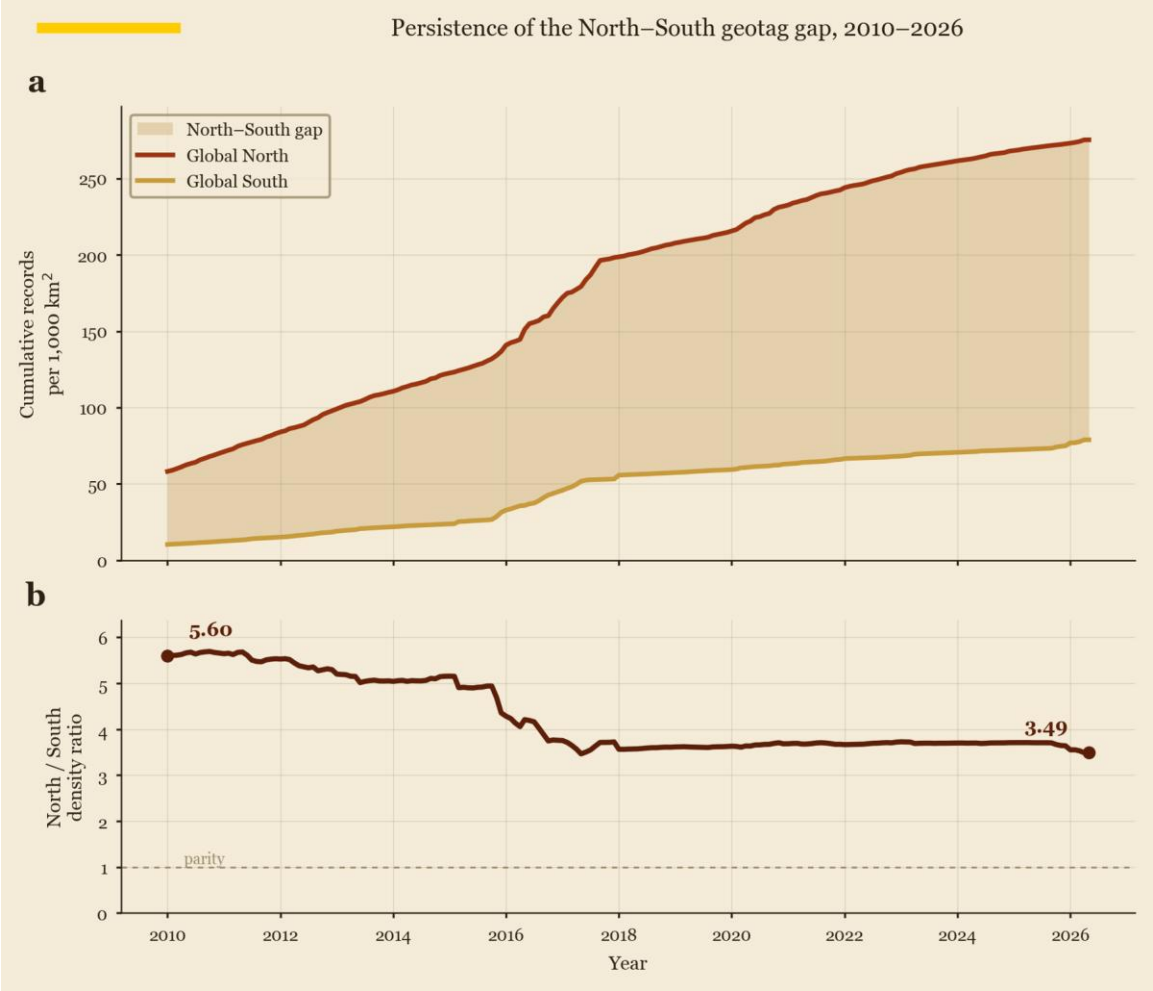


Figure 2. Persistence of the North/South gap. Monthly cumulative geotagged records per 1,000 km2 for the Global North and Global South based on article creation dates (a), plus the North/South ratio from 2010 to 2026 (b).

This temporal analysis uses article creation dates. It does not identify the date when each coordinate was added to the article. The time series therefore describes the creation dates of articles that are geotagged in the current corpus.

Country-level economic analysis Country-level geotagged coverage was positively associated with GDP per capita. For 207 countries with population, and GDP data, and at least 50 geotagged records, a log-log regression of geotagged articles per million people on GDP per capita gave slope 0.617, Pearson r = 0.621, R2 = 0.385, and p = 1.96e-23. (Figure 3)

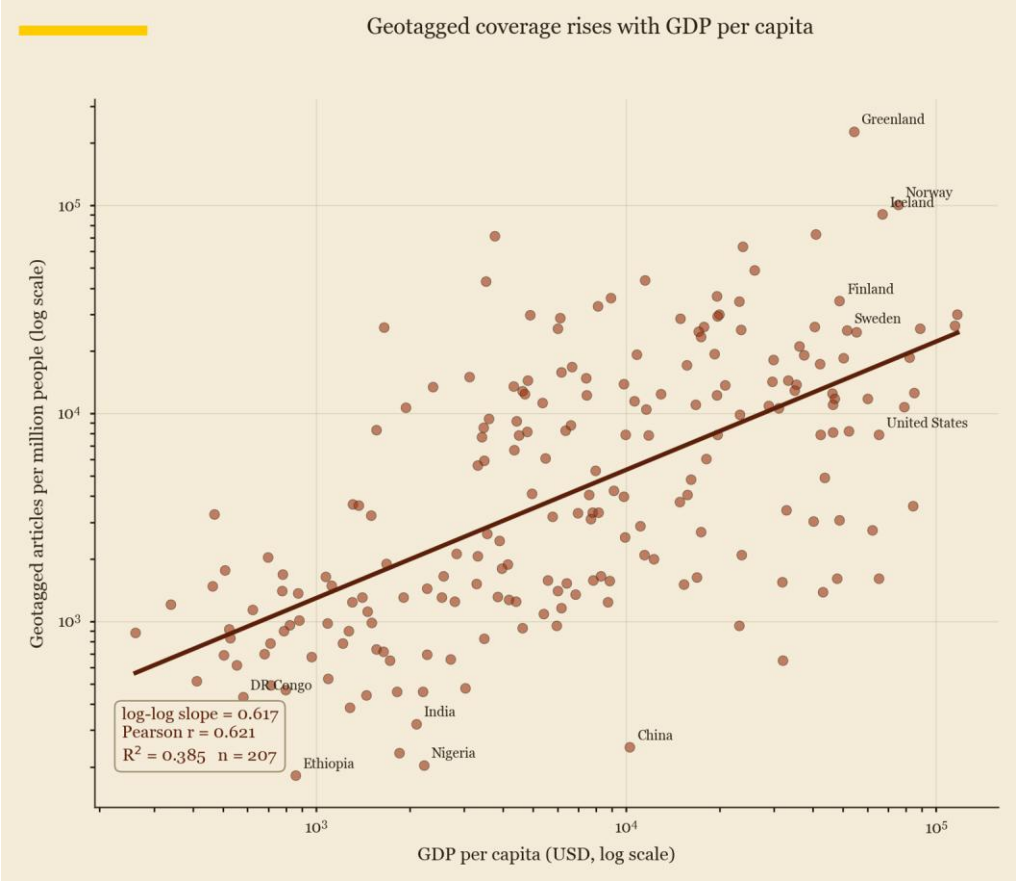


Figure 3. Economic association. Log-log scatter of geotagged articles per million people versus GDP per capita, with the fitted regression line and labelled outliers.

Sensitivity Analyses

Sensitivity of the aggregate density ratio to country composition

We assessed whether the aggregate ratio was robust to country composition and to the skewed distribution of country densities. The distribution of country-level densities was significantly higher in the Global North (Mann-Whitney $p = 6.26 \times 10^{-9}$; rank-biserial correlation = 0.468). Bootstrap resampling and sensitivity analyses yielded similar conclusions.

The aggregate result was not driven by very small territories. Excluding countries and territories below 1,000 km2 gave a ratio of 3.52; excluding units below 10,000 km2 gave 3.56; excluding units below 50,000 km2 gave 3.44; and excluding units below 100,000 km2 gave 3.29. The result was also not driven by a single high-count country. Excluding the United States gave a ratio of 3.50. Excluding the top five countries by geotagged record count (United States, France, Russia, Mexico, and Germany) gave 4.70, and excluding the top ten gave 3.07.

Deduplication to unique geographic locations preserves the imbalance

Because a single geographic entity can contribute one record per language edition, the record-level count of 21,270,978 includes multilingual duplication. We assessed how much of the

North/South gap survives when each distinct coordinate is counted exactly once. Grouping records by exact `(lat, lon)` collapsed the corpus to 8,670,787 unique locations, a records-per-location ratio of 2.45. Rounding coordinates to four decimal places (approximately 11 m at the equator) gave 8,317,851 unique locations, indicating minimal sensitivity to small per-edition coordinate typos.

Using unique locations as the unit of analysis, Global North countries contained 5,630,997 locations across the same 53,317,374 km2 (105.6 unique locations per 1,000 km2), and Global South countries contained 3,030,127 locations across 82,942,339 km2 (36.5 unique locations per 1,000 km2). The aggregate ratio was 2.89, and the rounded-4-decimal ratio was 2.85. Restricting to unique locations that have at least one English-Wikipedia record gave 1,009,376 in the Global North and 382,211 in the Global South, a ratio of 4.11, which is essentially unchanged from the record-level English-only ratio of 4.05. The unique-location time series moved from 6.03 in January 2010 to 2.89 in May 2026. The unique-location GDP association was similar in direction and magnitude to the record-level version (log-log slope 0.561, Pearson r = 0.596, R2 = 0.355, n = 207, p = 2.64e-21). (Supplementary Table S3)

Language editions and bot-heavy article creation strongly shape the corpus

The multilingual corpus was not dominated by English. Cebuano, driven by bot automation, was the largest language edition by geotagged record count, with 4,187,812 records, or 19.69% of the cleaned corpus. English had 1,463,869 records, or 6.88%. Cebuano therefore contained 2.86 times as many geotagged records as English. The largest language editions by geotagged record count were Cebuano, English, Chechen, German, French, Swedish, Tatar, Polish, Russian, Spanish, Ukrainian, Chinese, Dutch, Serbian, Italian, Min Nan Chinese, Persian, Egyptian Arabic, Serbian (Latin), and Portuguese. Several of these editions are known to have grown primarily through automated article creation or mass import rather than through natural editor activity. We developed a human-curated exploratory list of nine such language editions for use in the sensitivity analyses below: Cebuano, Swedish, Waray, Chechen, Tatar, Min Nan Chinese, Egyptian Arabic, Ladin, and Volapuk. (Supplementary Table S1) Because the sensitivity analysis reports the North/South ratio both with and without these editions, the paper's headline result does not depend on the list.

Sensitivity to coordinate source and bot-heavy language editions

Because the merged corpus combines two coordinate sources (article wikitext and Wikidata P625) (Wikidata Property P625 represents the coordinate location of an item) and because a small number of language editions are dominated by automated article creation, the headline density ratio could in principle be driven by one of these components. We recomputed every headline statistic under six filters: all sources, wikitext only, Wikidata only, all sources excluding bot-heavy editions, wikitext excluding bot-heavy editions, and Wikidata excluding bot-heavy editions. (Table 2) The bot-heavy exclusion list is the human-curated exploratory set documented in Supplementary Table S1 . All rows in this table use the same country classification and land-area denominators: 53,317,374 km2 for the Global North and 82,942,339 km2 for the Global South; Antarctica is excluded from both denominators.

Table 2. Sensitivity analysis of North/South density ratios.

Subset	Records	Ratio all langs	Ratio English	Ratio 2010-01	Ratio 2026-05	GDP slope	GDP r
All sources	21,270,978	3.49	4.05	5.60	3.49	0.617	0.621
Wikitext only	8,449,186	2.63	4.27	6.20	2.63	0.539	0.573
Wikidata only	12,821,792	4.28	2.55	5.34	4.28	0.690	0.637
All sources no bots	14,205,076	4.25	4.05	5.56	4.25	0.693	0.648
Wikitext no bots	3,778,199	3.34	4.27	6.29	3.34	0.725	0.669
Wikidata no bots	10,426,877	4.67	2.55	5.24	4.67	0.677	0.632

The Global North/South imbalance is present under every subset. The temporal trajectory, however, is not uniform. Bot-heavy wikitext editions, notably Cebuano and Waray from the Lsjbot project, added large numbers of geotagged articles for locations that fall in Global South countries and account for much of the narrowing observed in the all-source and wikitext-only series. Wikidata coordinates, by contrast, have been added disproportionately for Global North articles and have widened the gap relative to wikitext alone.

Population query experiment: Language-model answers about village populations show a similar geographic gap

The paper's central measurement concerns Wikipedia, but the practical concern is broader: uneven geographic documentation may affect downstream systems. To examine this possibility, we sampled 20,000 OpenStreetMap (OSM) villages with numeric population tags, split evenly between the Global North and Global South. We asked two language-model deployments for the population of each village.

GPT-4o-mini returned an answer for 19,998 of 20,000 villages. A simple numeric extractor found numeric-looking population guesses for 2,255 of 10,000 Global North villages and 420 of 10,000 Global South villages. The corresponding rates were 22.6% and 4.2%, a 5.4-fold difference. GPT-5.4 returned 19,998 answers and was evaluated with a separate answer classifier. The classifier produced concrete population estimates for 37.9% of Global North villages and 9.8% of Global South villages. (Table 3)

Table 3. Language model results for the Global North vs. Global South.

Model and measure	Global North	Global South	Ratio
GPT-4o-mini, numeric-looking answer	22.6%	4.2%	5.4

GPT-5.4, real-looking population answer	37.9%	9.8%	3.9

The GPT-5.4 results were strongly associated with Wikipedia evidence. Among villages with any Wikipedia evidence, 4,543 of 14,265 received real-looking population answers (31.9%). Among villages with no Wikipedia evidence in any language, 227 of 5,735 did so (4.0%). English Wikipedia evidence was associated with a still larger difference: 48.3% for villages with English Wikipedia evidence versus 8.3% for villages without it. (Supplementary Table S4)

We then compared parsed population claims with the OSM population tag used to construct the village sample. OSM population tags are not perfect ground truth, but they provide a consistent reference value. The North/South gap remained visible after this stricter comparison. (Table 4, Figure 4)

Table 4. Population data comparison – parsed claims vs. OSM tags.

Model	Group	Parsed claims	Within 25% of OSM	Within 50% of OSM	Within 100% of OSM	Median relative error among parsed claims
GPT-5.4	Global North	5,473	25.6%	31.2%	37.9%	0.315
GPT-5.4	Global South	1,661	4.2%	5.6%	7.9%	1.216
GPT-4o-mini	Global North	2,255	6.2%	12.1%	17.0%	0.468
GPT-4o-mini	Global South	420	0.3%	0.7%	2.6%	0.989

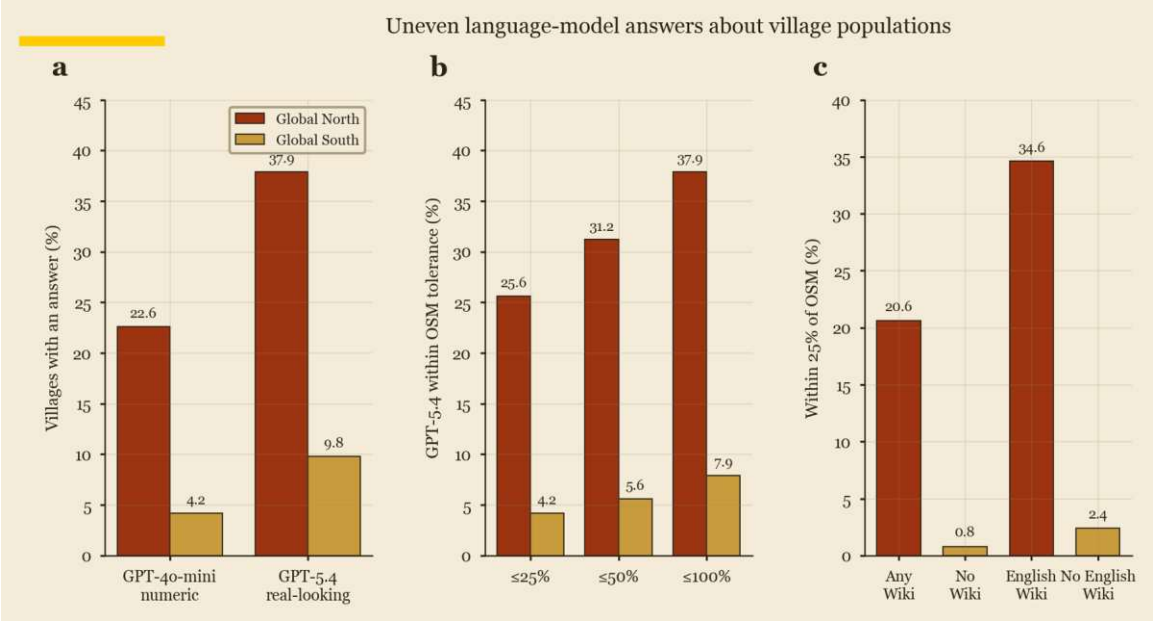


Figure 4. Village-population model audit. Global North (red)/South (yellow) comparison of GPT-4o-mini numeric-looking answers (a), GPT-5.4 real-looking answers (a), and parsed model claims within 25%, 50%, and 100% of the OSM population tag (b), plus GPT-5.4 accuracy rates by Wikipedia evidence condition (Any Wiki, red: No Wiki, yellow) (c).

Wikipedia evidence was also associated with answer accuracy relative to OSM. (Figure 4) For GPT-5.4, answers were within 25% of OSM for 20.6% of villages with any Wikipedia evidence and 0.8% of villages with no Wikipedia evidence; within 100% for 31.2% versus 2.3%. English Wikipedia evidence showed an even larger separation: 34.6% within 25% and 48.6% within 100% for villages with English Wikipedia evidence, compared with 2.4% and 6.6% for villages without English Wikipedia evidence.

Discussion

This study shows that the geotagged layer of multilingual Wikipedia remains spatially unequal at global scale. The main result is simple to measure and large in magnitude: after normalizing by land area under a single country classification and area source, and excluding Antarctica from both groups, Global North countries have 3.49 times the geotagged article-record density of Global South countries. The same imbalance appears in English Wikipedia alone. Collapsing the corpus to unique `(lat, lon)` locations gives a ratio of 2.89, indicating that both more unique places are represented in the Global North and more language editions describe each represented place. Both effects increase the representation of Global North locations in downstream systems that consume article-level records. The ratio has narrowed since 2010, but the 2026 value still indicates a large difference in spatial coverage. These results extend prior work on information geographies and Wikipedia representation in three ways: by using article-

level coordinates rather than only page counts or editor locations, combining coordinates from article wikitext and Wikidata across 355 language editions, and adding a time-resolved view using article creation dates.

The temporal results require careful interpretation because the article-creation timestamp used here does not identify when each coordinate was added, and because the merged corpus combines wikitext coordinates and Wikidata coordinate claims created under different infrastructures. The sensitivity analysis revealed that the apparent narrowing is not uniform across subsets. Under the most conservative subset (Wikidata coordinates excluding bot-heavy editions), the ratio moved from 5.24 to 4.67 between 2010 and 2026. Under wikitext coordinates excluding bot-heavy editions, the ratio moved from 6.29 to 3.34, a substantive reduction. The apparent narrowing in the all-source series (5.60 to 3.49) is therefore driven more by bot-heavy wikitext creation in the Global South than by Wikidata coordinate additions. The main empirical conclusion is that the North/South gap in geotagged Wikipedia remains large under every filter, and that observed improvement over time reflects specific production mechanisms rather than a general redistribution of encyclopedic attention.

Country-level geotagged coverage was significantly associated with GDP per capita. The observed relationship is consistent with prior evidence linking digital content production to economic factors including internet access, institutional capacity, tourism, and data availability[5]. However, it is important to note that this analysis is correlational and does not imply causation.

The village-population experiment provides an empirical illustration of how geographic information disparities may be reflected in downstream language-model behavior. GPT-4o-mini and GPT-5.4 were both much more likely to produce concrete population answers for Global North villages than for Global South villages. The gap persisted when parsed model claims were compared with OSM population tags: GPT-5.4 was within 25% of OSM for 25.6% of Global North villages and 4.2% of Global South villages. Wikipedia evidence was also strongly associated with both answer availability and OSM-relative accuracy. This result should be interpreted with caution because it does not establish causality, and OSM population tags are a reference value rather than definitive official ground truth. Language models are trained on many sources, and OSM itself may overlap with other web data. It is still an important audit: the result shows that model answer availability and OSM-relative answer accuracy are uneven in a way that mirrors the Wikipedia geotag imbalance. It also shows that Wikipedia evidence is a strong marker of whether a model can produce a concrete answer about a small place.

Several limitations matter. First, geotag count is a coverage measure, not a quality measure. It does not capture article length, citations, page views, maintenance, local authorship, or cultural significance. Also, geotagged Wikipedia data contains systematic errors, mass imports, and valid but analytically special batches (Supplemental Methods). These artifacts do not drive the headline North/South density result, but they show why large-scale open knowledge analyses require spatial quality control. Second, the time series uses article creation dates for articles

currently known to be geotagged. Third, country assignment depends on boundary data, treatment of territories, and area definitions. Fourth, the bot-heavy language list is a human-curated exploratory analytical choice rather than a criterion derived from article-level provenance metadata; the paper's headline result does not depend on this list because sensitivity analyses report ratios both including and excluding these editions, but a future data release should replace the curated list with an objective bot-flag or mass-creation-window measure. Fifth, the GPT village experiment measures answer availability, plausibility, and OSM-relative accuracy, not ground-truth accuracy against official census records.

The findings have practical implications for researchers and system builders who use Wikipedia-derived geodata. A geotagged Wikipedia article is often treated as a convenient point in space, but the point set is not a neutral sample of the world. It is shaped by economic conditions, language communities, automated imports, editorial practices, data-source availability, and quality-control failures. Analyses using this layer should document these biases and avoid interpreting article presence as a direct proxy for importance or need.

The findings also suggest constructive paths for improvement. The gap has narrowed over time, which means coverage can change. Better local-language documentation, transparent bot policies, Wikidata quality control, and targeted efforts in under-documented regions could make the geotagged layer more balanced. The same interventions should record provenance and article quality so that increased coverage does not simply add more opaque stubs.

Methods

Wikipedia data collection and open dataset

We assembled a harmonized dataset of geotagged Wikipedia content using public Wikimedia dumps and Wikidata coordinate claims [1–4]. Geographic coordinates were extracted from article wikitext, including language-specific coordinate-template variants. For articles without a usable wikitext coordinate, we used the Wikidata coordinate-location claim (P625) associated with the article's corresponding Wikidata item.

The initial merged corpus contained 21,335,644 records. Each record included the article title, URL, latitude and longitude, language edition, coordinate source, and article creation timestamp when available. After coordinate validation and artifact removal, the final corpus contained 21,270,978 records from 355 Wikipedia language editions, comprising 8,449,186 records with coordinates derived from article wikitext and 12,821,792 with coordinates derived from Wikidata. The primary unit of analysis is a language-edition geotagged article record. A single geographic entity may contribute multiple records, one per language edition in which it appears with a coordinate.

We release the cleaned and harmonized geotagged Wikipedia dataset, together with the country classifications and associated analysis code, as an open resource for research on geographic representation and knowledge inequality [Data: https://huggingface.co/datasets/jlavista/wikipedia-geo-atlas, code repository:

https://github.com/jlavista/wikipedia-geo-atlas, and live map: https://jlavista.github.io/wikipedia-geo-atlas/]. The complete dataset schema, data provenance, processing pipeline, and release files are described in the Supplementary Methods.

Coordinate validation and geographic assignment

We retained coordinates within the valid geographic ranges of −90° to 90° latitude and −180° to 180° longitude. Malformed records and coordinates outside these ranges were excluded. We also removed two systematic classes of non-geographic artifacts: coordinates associated with astronomical objects and regularly spaced integer-degree coordinates located over the open ocean. These procedures removed 21,578 astronomy-related records and 43,088 ocean-grid records. Detailed detection rules and additional artifact investigations are provided in the Supplementary Methods.

Coordinates were assigned to countries or administrative level-zero units using Natural Earth geographic boundaries and country attributes [20,21]. Records falling outside the country polygons used in the analysis were not included in country-level or Global North–South comparisons.

Global North–South classification

Our primary classification followed the developed-economy grouping used in the United Nations World Economic Situation and Prospects framework, informed by the broader Global North–South and Brandt-line literature [22–24]. The Global North included Europe, including the Russian Federation; the United States and Canada; Japan and the Republic of Korea; Israel; Australia and New Zealand; and Singapore, Taiwan, and Hong Kong. All remaining countries were assigned to the Global South. Antarctica was excluded from both groups.

The classification of territories and the complete country-level assignment are provided in the released classification table. (Data and source code will be made available at a public repository; URL/DOI pending) Because no single definition of the Global North and Global South is universally accepted, we repeated the analysis using alternative classifications based on the UN M49 developed-regions convention, the World Bank high-income grouping, and the UNCTAD developed-economies grouping [28–30].

Geographic coverage measures are described in the Supplementary Methods. The density measure reflects the multilingual article-record layer that downstream systems index and query, not the density of unique geographic entities.

Temporal analysis

For records with an available article creation timestamp, we extracted the year and month of creation and calculated cumulative record counts and densities for the Global North and Global South. Records without a creation timestamp were excluded from the temporal analysis.

The timestamp represents the creation date of the Wikipedia article, not necessarily the date on which its geographic coordinate was added. The resulting time series should therefore be

interpreted as the creation history of articles that were geotagged in the corpus at the time of data collection, rather than as a direct history of coordinate insertion.

Country-level economic association

As a secondary analysis, we examined the association between geographic Wikipedia coverage and national income. Country record counts were joined to population and gross domestic product estimates from the Natural Earth country attributes. Countries with fewer than 50 geotagged records or populations below 50,000 were excluded.

The dependent variable was the number of geotagged records per million residents, and the independent variable was gross domestic product per capita. Both variables were log10-transformed, and their association was estimated using ordinary least squares:

$$[\log_{10}(\text{records per million}) = \alpha + \beta \log_{10}(\text{GDP per capita}).]$$

We report the regression slope, Pearson correlation, coefficient of determination, standard error, and (p) value. This analysis measures association and is not interpreted causally.

Sensitivity and robustness analyses

We evaluated whether the main geographic disparity depended on the coordinate source, multilingual duplication, automated article creation, country composition, or the definition of the Global North and Global South.

First, we repeated the analysis separately for coordinates extracted from article wikitext and coordinates obtained from Wikidata. We also recalculated the estimates after excluding language editions, identified through prior Wikipedia community documentation and our artifact investigations, as containing large numbers of automated or mass-imported articles. Because this designation was not based on article-level bot provenance, results both including and excluding these editions are reported.

Second, because the same geographic entity may appear in multiple Wikipedia language editions, we performed a unique-location analysis in which each distinct latitude–longitude pair was counted once. We additionally repeated this analysis after rounding coordinates to four decimal places to account for small coordinate differences across language editions.

Third, we used 10,000 country-level bootstrap replicates to examine the sensitivity of the aggregate density ratio to country composition. Countries were resampled with replacement separately within the Global North and Global South, preserving the observed number of countries in each group. Percentile-based 95% intervals from this procedure describe robustness to the composition of the country groups; they do not represent sampling uncertainty in the Wikipedia corpus itself.

We further recalculated the density ratio after excluding small territories and the countries contributing the largest numbers of records. Country-level density distributions were

summarized using medians and interquartile ranges and compared using a one-sided Mann–Whitney U test, with rank-biserial correlation reported as an effect-size measure. Complete specifications and results for these analyses are provided in the Supplementary Methods and Supplementary Results

OpenStreetMap language model audit

Village sampling

To examine whether geographic differences in foundational knowledge coverage were reflected in an artificial intelligence system, we constructed a sample of villages from OpenStreetMap planet data [25]. We selected point features tagged place=village that had a village name and a clean numeric population value. Each village was assigned to a country using the same Natural Earth boundaries and then classified as belonging to the Global North or Global South using the primary classification described above.

Reservoir sampling was used to select 10,000 eligible villages from each group, producing a balanced sample of 20,000 villages. OpenStreetMap population tags were used for evaluation but were not included in the prompts sent to the models.

Population-query experiment

For each sampled village, the model received the village name and country and was asked to provide its population. The system instruction requested a concise response and instructed the model to state that the population was unknown when it did not have a reliable value, rather than guessing. The exact system and user prompts are provided in the Supplementary Methods.

The primary experiment used Azure OpenAI deployment for GPT-5.4. We additionally repeated the full 20,000-village experiment using GPT-4o-mini, to assess whether the principal geographic patterns depended on the GPT-5.4 deployment. Both models were queried with deterministic generation settings where supported. Complete model parameters, API versions, retry procedures, and execution metadata are reported in the Supplementary Methods.

Classification of model answers

GPT-5.4 responses were classified using GPT-4o-mini into four mutually exclusive categories: a concrete population claim for the requested village; an explicit statement that the population was unknown; an ambiguous answer or a population referring to a different geographic scope; and no usable answer or an API error.

The classifier evaluated the content and scope of the answer but did not determine whether the population value was factually correct. We therefore distinguish between answer availability, defined by whether the model supplied a concrete population claim for the requested village, and OSM-relative accuracy, defined by comparison with the OpenStreetMap population value. The classifier prompt, category definitions, output schema, and quality-control procedures are provided in the Supplementary Methods.

Wikipedia evidence for sampled villages

We measured whether each sampled village had evidence of representation in Wikipedia using three complementary procedures. First, we checked for a Wikipedia tag on the OpenStreetMap feature. Second, where an OpenStreetMap Wikidata identifier was available, we checked the associated Wikidata item for Wikipedia sitelinks. Third, we searched the cleaned multilingual geotagged Wikipedia corpus for normalized title matches located within 25 km of the village coordinate.

We constructed separate indicators for evidence in English Wikipedia, evidence in a non-English Wikipedia edition, and evidence in any Wikipedia language edition. Because title and proximity matching may produce false matches or miss differently named places, these indicators are treated as evidence of likely Wikipedia representation rather than definitive entity resolution.

Validation against OpenStreetMap population values
Numeric population claims were extracted from model responses using a separate deterministic parser. The parser distinguished candidate population values from years and other numerical content by considering their range and surrounding textual context.

For each extracted population estimate $\hat{P}$ we calculated relative error against the OpenStreetMap population value (P):

$$[\text{relative error} = \frac{|\hat{P} - P|}{\max(P, 1)}.]$$

We also calculated absolute logarithmic error:

$$[\left|\log_{10}(\hat{P} + 1) - \log_{10}(P + 1)\right| .]$$

We report the proportion of all sampled villages, and the proportion of answers containing a parsed population claim, that fell within 25%, 50%, and 100% of the OpenStreetMap value. Responses containing evidence that the model may have answered for a municipality, district, commune, city, or other administrative unit rather than the requested village were flagged as potentially referring to a different geographic scope.

OpenStreetMap population tags are community-contributed and may be outdated, approximate, or derived from heterogeneous sources. Accordingly, these comparisons are interpreted as OSM-relative validation rather than validation against definitive census ground truth.

Statistical interpretation
The analyses are descriptive and observational. The Global North–South comparisons measure differences in aggregate geographic coverage, and the village experiment evaluates geographic differences in model answer availability and OSM-relative accuracy. Associations between Wikipedia evidence and model performance do not establish that Wikipedia coverage caused

the observed model behavior. They indicate that inequalities in a foundational knowledge resource and inequalities in model performance are systematically related.

Data Availability

The study is based on publicly available Wikimedia and Wikidata data. The processed dataset generated for this study, including the article-level geospatial dataset and the aggregated data underlying all reported analyses and figures, has been deposited in the study repository at https://huggingface.co/datasets/jlavista/wikipedia-geo-atlas. The code repository is available at https://github.com/jlavista/wikipedia-geo-atlas. A live interactive map is available at https://jlavista.github.io/wikipedia-geo-atlas/. The source Wikimedia and Wikidata records used to generate these datasets are available from public Wikimedia data dumps and Wikidata coordinate-location claims. All data needed to reproduce the findings are available from the repository. More information about the data and code used in this study is available in the Supplementary Information.

References

1. Wikimedia Foundation. Data dumps. https://meta.wikimedia.org/wiki/Data_dumps (accessed 20 May 2026).
2. Wikimedia Foundation. Wikimedia database backup dumps. https://dumps.wikimedia.org/backup-index.html (accessed 20 May 2026).
3. Vrandecic, D. & Krotzsch, M. Wikidata: a free collaborative knowledgebase. *Communications of the ACM.* **57**, 78-85 (2014). https://doi.org/10.1145/2629489.
4. Wikidata contributors. Property:P625, coordinate location. https://www.wikidata.org/wiki/Property:P625 (accessed 20 May 2026).
5. Graham, M., Hogan, B., Straumann, R. K. & Medhat, A. Uneven geographies of user-generated information: patterns of increasing informational poverty. *Annals of the Association of American Geographers.* **104**, 746-764 (2014). https://doi.org/10.1080/00045608.2014.910087.
6. Hecht, B. & Gergle, D. The tower of Babel meets Web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 291-300 (ACM, 2010). https://doi.org/10.1145/1753326.1753370.
7. Miquel-Ribe, M. & Laniado, D. Wikipedia Culture Gap: quantifying content imbalances across 40 language editions. *Frontiers in Physics.* **6**, 54 (2018). https://doi.org/10.3389/fphy.2018.00054.
8. Brown, T. B. et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems.* **33**, 1877-1901 (2020). https://arxiv.org/abs/2005.14165.

9. Dodge, J. et al. Documenting large webtext corpora: a case study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286-1305 (2021). https://arxiv.org/abs/2104.08758.
10. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at *arXiv:2108.07258* (2021). https://arxiv.org/abs/2108.07258.
11. Bender, E. M., Gebru, T., McMillan-Major, A. & Mitchell, M. On the dangers of stochastic parrots: can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 610-623 (ACM, 2021). https://doi.org/10.1145/3442188.3445922.
12. Bender, E. M. & Friedman, B. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics.* **6**, 587-604 (2018). https://aclanthology.org/Q18-1041/.
13. Johnson, I. L., McMahon, C., Schoning, J. & Hecht, B. The effect of population and structural biases on social media-based algorithms: a case study in geolocation inference across the urban-rural spectrum. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* 1167-1178 (ACM, 2017). https://doi.org/10.1145/3025453.3026015.
14. Thebault-Spieker, J., Hecht, B. & Terveen, L. Geographic biases are born, not made: exploring contributors' spatiotemporal behavior in OpenStreetMap. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork.* 71-82 (ACM, 2018). https://doi.org/10.1145/3148330.3148350.
15. Graham, M., De Sabbata, S. & Zook, M. A. Towards a study of information geographies: (im)mutable augmentations and a mapping of the geographies of information. *Geo: Geography and Environment.* **2**, 88-105 (2015). https://doi.org/10.1002/geo2.8.
16. Graham, M., Hale, S. A. & Stephens, M. Geographies of the world's knowledge. *Oxford Internet Institute* (2011). https://ora.ox.ac.uk/objects/uuid:a737b5c3-069f-4044-ade8-2ce6c95a0d4c.
17. Halfaker, A., Geiger, R. S., Morgan, J. T. & Riedl, J. The rise and decline of an open collaboration system: how Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist.* **57**, 664-688 (2013). https://doi.org/10.1177/0002764212469365.
18. Reeves, N. & Simperl, E. Machines in the margins: a systematic review of automated content generation for Wikipedia. Preprint at *arXiv:2509.22443* (2025). https://arxiv.org/abs/2509.22443.
19. Wagner, C., Graells-Garrido, E., Garcia, D. & Menczer, F. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science.* **5**, 5 (2016). https://doi.org/10.1140/epjds/s13688-016-0066-4.
20. Kelso, N. V. & Patterson, T. Natural Earth Vector. *Cartographic Perspectives.* **64**, 45-50 (2009). https://doi.org/10.14714/CP64.148.
21. Kelso, N. V. & Patterson, T. *Introducing Natural Earth Data*. Natural Earth project paper. http://shadedrelief.com/borsa/natural_earth_paper.pdf.
22. United Nations Department of Economic and Social Affairs. *World Economic Situation and Prospects 2024*. United Nations, New York (2024).

https://desapublications.un.org/publications/world-economic-situation-and-prospects-2024. Country classification "developed economies" defined in Statistical Annex, Table A.
23. Dados, N. & Connell, R. The Global South. *Contexts*. **11**, 12-13 (2012). https://doi.org/10.1177/1536504212436479.
24. Brandt, W. (Chair), Independent Commission on International Development Issues. *North-South: A Programme for Survival*. MIT Press, Cambridge, MA (1980).
25. OpenStreetMap contributors. Planet dump. https://planet.openstreetmap.org/ (accessed 2026).
26. Wikimedia Foundation. Terms of Use. https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use (accessed 20 May 2026).
27. Wikidata contributors. Wikidata:Copyrights. https://www.wikidata.org/wiki/Wikidata:Copyrights (accessed 20 May 2026).
28. United Nations Statistics Division. Standard country or area codes for statistical use (M49). https://unstats.un.org/unsd/methodology/m49/ (accessed 2 July 2026).
29. World Bank. World Bank Country and Lending Groups. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups (accessed 2 July 2026).
30. United Nations Conference on Trade and Development. UNCTADstat classifications. https://unctadstat.unctad.org/EN/Classifications.html (accessed 2 July 2026).

Funding

The authors received no specific funding for this work.

Author contributions

Conceptualization: J.M.L.F., C.R., A.O., I.B.-R. ; Methodology: J.M.L.F., C.R., A.O., I.B.-R. Validation: J.M.L.F. Investigation: J.M.L.F. Resources: J.M.L.F. Writing-original draft: J.M.L.F.; Writing – review and editing: J.M.L.F., C.R., A.O., M.B., I.B.-R.; Supervision: J.M.L.F.

Additional Information

Competing interests statement: All authors are employees of Microsoft Corporation. Microsoft provided salary support and computational resources for this research. Microsoft has a commercial relationship with OpenAI, the developer of GPT, which was evaluated in this study.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx