WorldBench: Quantifying Geographic Disparities in LLM Factual Recall

Mazda Moayeri
University of Maryland
USA

Elham Tabassi
Michigan State University
USA

Soheil Feizi
University of Maryland
USA

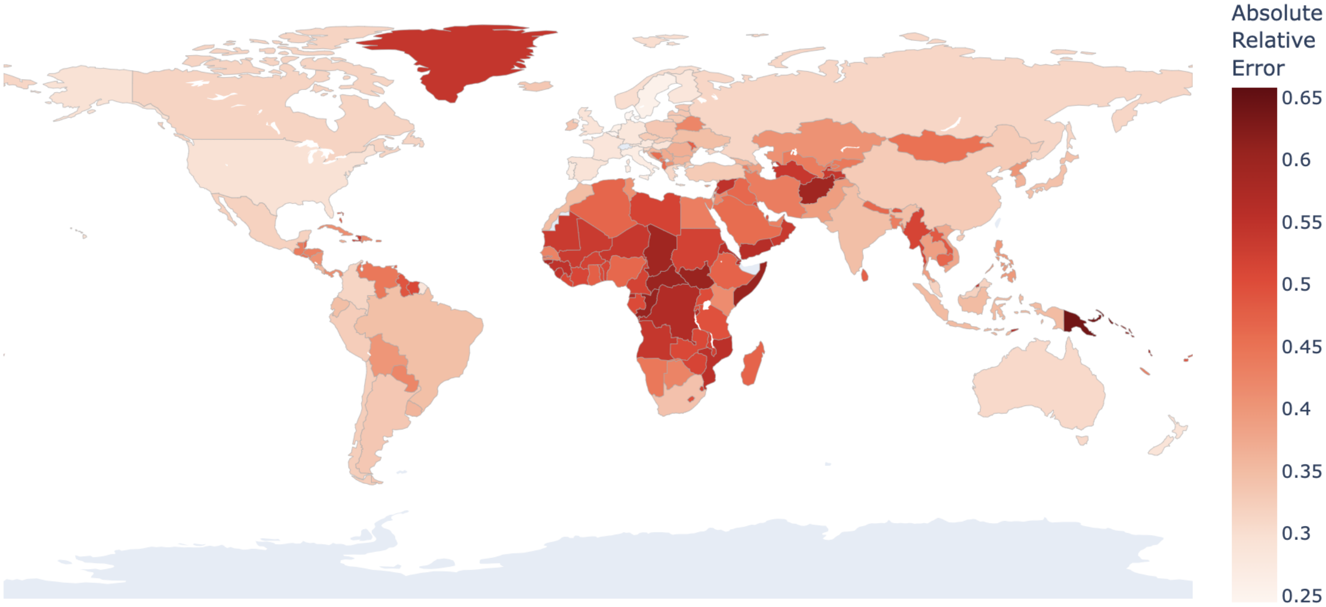


Figure 1: WorldBench leverages World Bank data to assess the ability of large language models (LLMs) to recall factual information about specific countries. Above, we plot the absolute relative error per country, averaged over 11 global development statistics queried to 20 state of the art open source and private LLMs. WorldBench reveals significant geographic disparities in LLM factual recall.

ABSTRACT

As large language models (LLMs) continue to improve and gain popularity, some may use the models to recall facts, despite well documented limitations with LLM factuality. Towards ensuring that models work reliably *for all*, we seek to uncover if geographic disparities emerge when asking an LLM the same question about different countries. To this end, we present WorldBench, a dynamic and flexible benchmark composed of per-country data from the World Bank. In extensive experiments on state of the art open and closed source models, including GPT-4, Gemini, Llama-2, and Vicuna, to name a few, we find significant biases based on region and income level. For example, error rates are 1.5 times higher for countries from Sub-Saharan Africa compared to North American countries. We observe these disparities to be consistent over 20 LLMs and 11 individual World Bank indicators (i.e. specific statistics, such as population or CO_2 emissions). WorldBench also enables automatic detection of citation hallucination, where models cite the World Bank itself while providing false statistics, and a manner to assess when an LLM's stored facts begin to go out of date. We hope our benchmark will draw attention to geographic disparities in existing LLMs and facilitate the remedying of these biases.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**; *Information extraction.*

KEYWORDS

Geographic Disparity, Bias, Fairness, Large Language Models, Factuality

ACM Reference Format:
Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. WorldBench: Quantifying Geographic Disparities in LLM Factual Recall. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June*

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658967

03–06, 2024, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3630106.3658967

1 INTRODUCTION

Large language models (LLM) exhibit remarkable performance on a wide array of tasks, from summarizing the news to writing code to answering trivia questions [22, 24, 30]. Impressively, LLMs have also been effective on real-world benchmarks. For example, GPT-4 [21] has been shown to pass the licensing exams for both legal [16] and medical professions [15, 19]. However, LLMs are also known to hallucinate, where they generate inaccurate text in a plausible manner [12]. This can pose particular risks for factual recall tasks. Given the black box nature of LLMs, continued development and application of diverse benchmarks is instrumental in understanding when LLMs can be trusted to answer reliably.

In addition to issues with correctness, AI in general has well documented challenges with performance disparities, in which seemingly strong models fail more frequently for some subset of inputs than others. Performance disparities can manifest as fairness issues when the subset of inputs where the model underperforms is characterized by sharing a sensitive attribute. For example, Buolamwini and Gebru [7] identified widespread performance disparities along race and gender lines across commercial facial recognition systems, while others have shown that object recognition models suffer performance drops when images originate in lower income countries [9, 10]. Similarly, Ojo et al. [20] show LLMs are less performant when tasks are posed using African languages instead of English. A key first step to building models that work *for all* is creating benchmarks to quantify not only performance, but also performance disparities.

To this end, in this work, we introduce a novel benchmark called WorldBench to uncover if *geographic* disparities emerge in LLM factual recall. In other words, we ask, *are LLMs more accurate in answering questions about some parts of the world than others?* To systematically tackle this question, we compute LLM performance on a country-wise level, by way of utilizing per-country indicators (i.e. statistics) from the World Bank [6]. We build and validate (via human inspection) an automated, indicator-agnostic prompting and parsing pipeline to interface with the World Bank data, summarized in Figure 2. This way, any set of indicators can be used in future variations of WorldBench, without having to change our code, which we will make public. In our study, we incorporate 11 diverse indicators, each having data for about 200 countries, resulting in a total of 2, 225 questions per LLM.

We evaluate 20 state of the art LLMs released in 2023, ranging from open-source models like Llama-2 and Vicuna [27, 32], to private commercial ones accessible via API, including GPT-4 and Gemini [21, 26]. As visualized in Figure 1, when averaging over all LLMs and indicators, we observe substantial differences in per-country error, with African countries seemingly incurring the largest errors. Using country categorizations defined by the World Bank, we quantify disparities across 7 regions and 4 income groups, finding that LLMs are most accurate for countries from Western regions and the high income category. Problematically, these error rates rise by a factor of about 1.5× when moving to the region (Sub-Saharan Africa) and income group (low income) for which models are least accurate. Moreover, we find these disparities and their order (i.e. which groups have most/least error) to be consistent when inspecting LLMs or indicators individually. That is, *all* 20 *LLMs exhibit geographic disparities in factual recall.*

In addition to our main result, we utilize the temporal aspect of the World Bank data to conduct extra analyses, such as automatically cross-checking LLM generated "citations" which turn out to be hallucinated, and inspecting error as a function of the groundtruth year, finding that some LLMs in our suite may already be slightly out of date.

In summary, we present WorldBench, a flexible benchmark for understanding LLM factual recall abilities on a per-country basis. With WorldBench, we conduct a large scale evaluation of 20 LLMs, and find pervasive geographic disparities across regions and income levels. We hope our benchmark can facilitate further research on the fairness of LLMs, towards building models that work well *for all.*

2 RELATED WORK

Evaluating Factual Recall. Recent works have documented the performance of LLMs in factual recall: [17], [14], [23], [25]. The general conclusion to these works is that while existing LLMs appear capable in answering certain factual question, their factual recall is less than perfect, as models can hallucinate completely fabricated information [12]. Zhang et al. [31] specifically investigated the recall of geographic information, though their study is limited to GPT-4 and does not inspect disparities. Some works (e.g., [17], [23]) linked factual recall to 'popularity', showing that error rate increases for less popular entities. While those studies categorize facts by popularity, each question in our benchmark has an associated country, as well as Region and Income group. These additional annotations enable going beyond overall error, so to assess geographic performance disparities in factuality.

Bias. The issues of bias and fairness in AI are of immense societal impact. Several studies have observed computer vision models to exhibit disparate performance when grouping inputs by race, gender, and across income levels and geographies, for tasks like facial recognition, object classification, and diverse image generation [8–11]. In the realm of language processing, Ojo et al. [20] observed a performance gap when tasks are presented in African languages. To the best of our knowledge, our study is the first to propose an automated and systematic examination of country-wise disparities in LLM factual recall, which in turn enables inspection of disparities across regions and income groups.

Benchmarks. Other works have noted and sought to improve challenges associated with evaluating factuality, primarily for tasks like summarization, where constructing a similarity metric between generated and reference texts is nontrivial. In our case, we design our benchmark to obtain numeric answers from LLM repsonses, with which we can compare to groundtruth values with the simple metric of absolute relative error. Further, we utilize a reputable third party (the World Bank), so that (i) the questions asked are relevant, (ii) inputs are grouped into salient cateogries, and (iii) groundtruth answers are accurate and up-to-date.

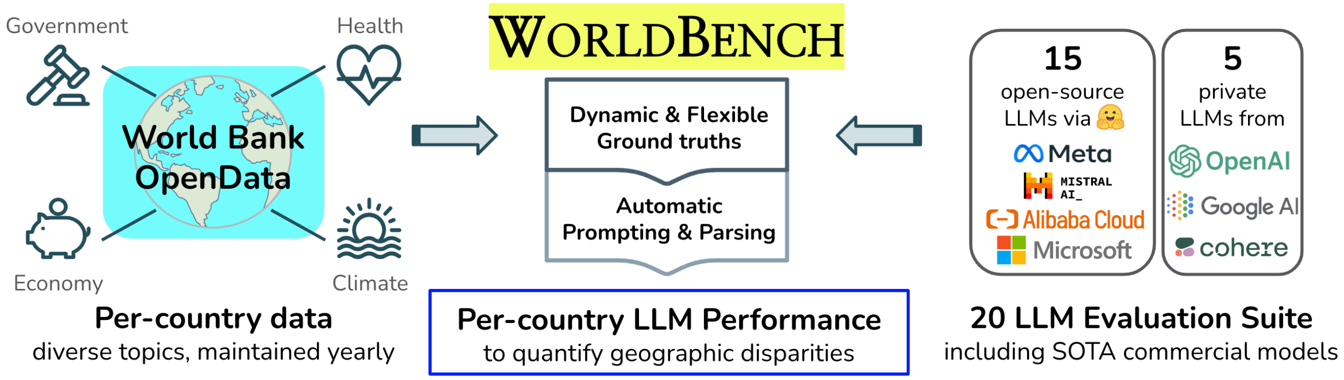

Government
Health
World Bank OpenData
Economy
Climate
Per-country data
diverse topics, maintained yearly
WorldBench
Dynamic & Flexible Ground truths
Automatic Prompting & Parsing
Per-country LLM Performance
to quantify geographic disparities
15 open-source LLMs via 🤗
Meta
MISTRAL AI_
Alibaba Cloud
Microsoft
5 private LLMs from
OpenAI
Google AI
cohere
20 LLM Evaluation Suite
including SOTA commercial models

Figure 2: Overview of WorldBench. Our benchmark provides a manner to quantify the performance of large language models (LLMs) on a per-country basis. We disentangle data collection from evaluation by utilizing the World Bank's data bank, which contains statistics (called indicators) pertaining to numerous diverse aspects of global development. Crucially, the data is available for nearly all countries and is updated year to year. With WorldBench, one can flexibly select specific statistics of interest, and dynamically re-evaluate models as time passes to see if they remain up to date. In this work, we uncover substantial geographic disparities in LLM performance for a wide range of models released by industry leaders, revealing the inequities pervasive across state of the art LLMs.

Indicator	Metric
Population	Total Population
Unemployment	Unemployment As A Percent Of The Total Labor Force
Maternal Mortality Rate	Maternal Mortality Ratio As Number Of Deaths Per 100,000 Live Births
Women In Parliament	Proportion Of Seats Held By Women In National Parliaments (As A Percent)
Education Expenditure	Government Expenditure On Education As A Total Percent Of Gdp
Electricity Access	Percent Of The Total Population That Has Access To Electricity
Agricultural Land Percent	Percent Of Total Land Area That Is Agricultural
CO_2 Emissions	Amount Of Carbon Dioxide Emissions In Metric Tonnes Per Capita
GDP	Gdp Measured In Us Dollars
GDP PPP Per Person Employed	Gdp At Purchasing Power Parity (Ppp) Per Person Employed
Renewable Energy Ratio	Renewable Energy Consumption As A Percent Of Total Final Energy Consumption

Table 1: Global development indicators in WorldBench, each defined and maintained by the World Bank.

3 METHODS: WORLDBENCH

3.1 Data

Our benchmark is constructed directly from statistics collected and maintained by the World Bank. The World Bank is a global organization with nearly 200 member countries, whose mission is to reduce extreme poverty via sustainable solutions to promote shared prosperity, particularly in developing countries [6]. The World Bank tracks numerous global development **indicators**, from 20 wide ranging categories, such as Climate, Health, and Poverty, to name a few. These statistics are freely available to the public and updated yearly. Importantly, the data are collected *per country*, meaning that regardless of the size, wealth, or location of a country, it is represented in the World Bank's data. We leverage this publicly available open data to build WorldBench, a benchmark to quantify the degree to which language models can recall facts about *all* countries in the world.

Our benchmark offers a few unique advantages to most existing benchmarks. First, and most importantly, WorldBench equitably represents all countries. Thus, we can query a language model for the same exact statistic for completely different countries, enabling direct comparisons across countries to uncover disparities in performance. Next, data quality and licensing is assured, as it comes from a globally reputable source which explicitly allows for its use by the public. Third, our benchmark is dynamic and flexible. The dynamic nature comes from the fact that the statistics are updated on a yearly basis, enabling the longevity of our benchmark, as well as analysis of LLM factual recall along a temporal dimension (see §6.2). The flexibility is borne out of the vast number of indicators one could choose from. In other words, if one sought to better examine the ability of language model to recall facts about the environment, they can elect to choose indicators from the Climate category. In contrast, if a language model is being developed for financial purposes, one could focus on indicators from the Economy and Growth categories.

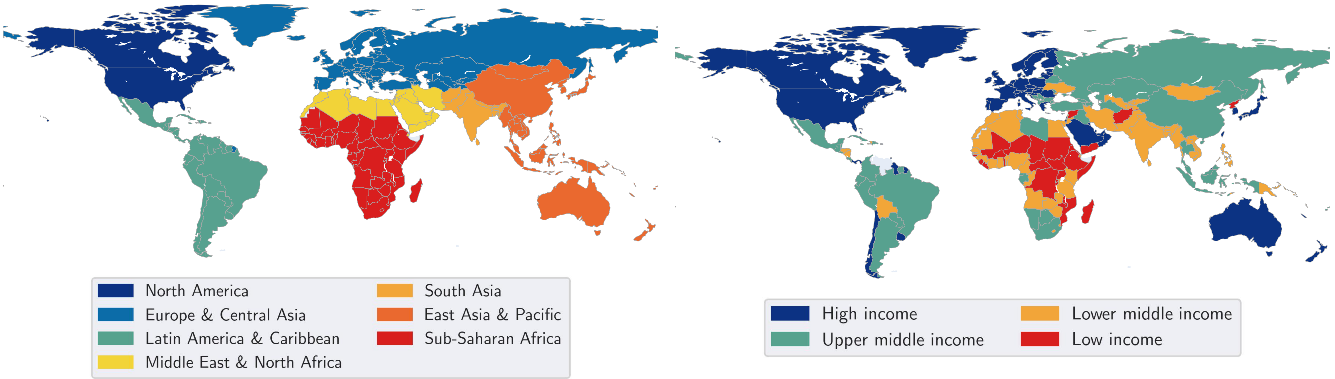


Figure 3: World Bank categorizations of countries into 7 regions (left) and 4 income groups (right).

In this study, we select 11 indicators, as shown in Table 1. The indicators are chosen to represent multiple different categories, and qualitatively are amongst the indicators that are easier to understand for lay people (i.e. non-experts in global development, like AI researchers). In total, there are 2, 225 questions, reflecting an average of 202 countries with groundtruth data per indicator studied.

Country categorization. The World Bank also provides various categorizations of countries, based on geographic or economic reasons [5]. We focus on two high level categorizations, visualized in Figure 3, which divide the world into 7 *Regions* and 4 *Income groups*. We note that, like the collection and maintenance of the groundtruth data for our benchmark, country categorization is carried out by an external body (i.e. the World Bank) to the model producers and evaluators. We hope that the disentanglement of these three parties enables a more objective comparative analysis, informed by experts on global development.

3.2 Language Model Evaluation

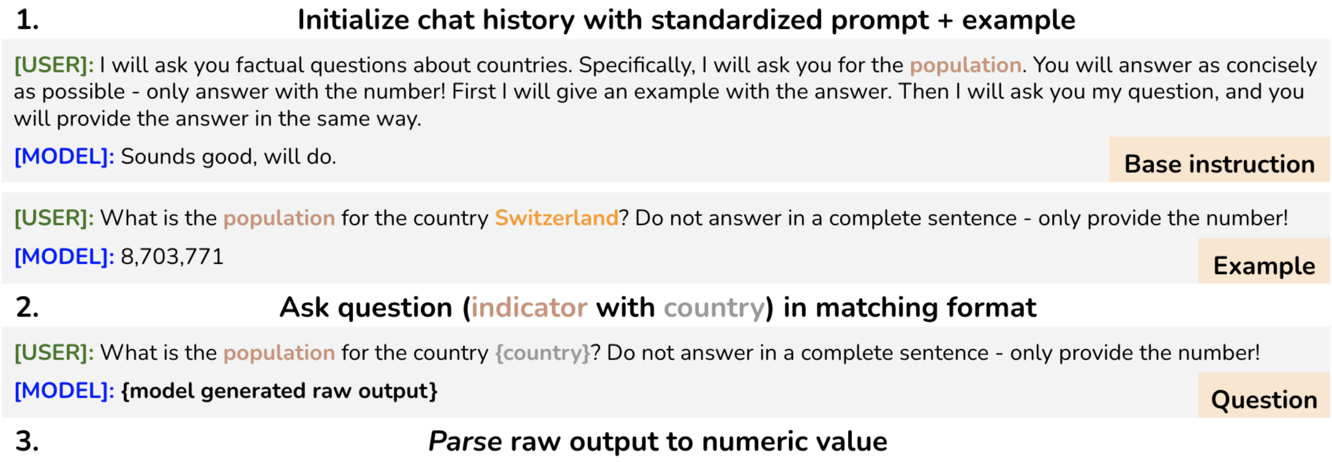
While the World Bank's open data is crucial to our analysis, additional steps are needed to interface with the available data scalably. To enable large scale evaluation of LLMs, we design a procedure to obtain a numeric answer given an arbitrary indicator, country, and LLM of interest. Namely, we utilize a template prompt to guide models to provide answers in a mostly uniform fashion, and then apply an automated parsing method to extract the numeric value from the raw LLM output. We detail these steps below, as well as results from human studies to validate the correctness of our pipeline. We also explain how we compute errors, given numeric answers from LLMs and the World Bank's groundtruth data.

Prompting. Our standard prompt consists of a base instruction, an example, and a template question filled in with values for the indicator and country of interest. Figure 4 displays the base instruction and example. For consistency, we fix our choice of example country, electing Switzerland, as it has groundtruth data for all indicators in our study; we confirm results are similar when using alternate example countries in Appendix E. Importantly, we prompt the model to only provide the number in its response. Without this instruction, models generate longer free-from responses, increasing the difficulty of automatically extracting numeric values and the the computational cost of our benchmark. For every question (i.e. combination of an indicator and country), we first initialize the chat history of the LLM of interest with the base instruction and example, and then ask the question. Notably, all three components are modular with respect to the country and indicator of interest, allowing for them to work for any World Bank indicator.

Parsing. Despite the instruction to 'only provide the number', LLMs at times exhibit undesirable, like including other text (e.g. special tokens) or repeating the question with new countries and responding to itself again and again. We design an automated parsing method to scalably extract a numeric value from the raw LLM outputs. The parsing method removes special characters, and in most cases, extracts the first numeric value provided. We also account for special cases like, for example, where a suffix (e.g. 'million' or 'billion') is used. In a small number of cases, the LLM either provides no output, an invalid output (e.g. a number with two decimal points), or abstains from answering. For these outputs and any others where the parsed number cannot be converted to a float, we exclude them from further analysis.

Error metric. To compare numeric values, we utilize absolute relative error, computed as follows: given two scalars a, b, we define *Absolute Relative Error* as $\frac{|a-b|}{\max(a,b)}$. Essentially, this metric conveys by what percent two measures are different from one another. For example, an absolute relative error of 0.1 means that one value was 10% larger or smaller than the other. Notice that absolute relative error always falls between 0 (because all values we encounter are non-negative) and 1 (because the denominator is the maximum of the two positive values). We elect to use relative error over absolute error because the ranges of values varies dramatically across indicators, with the population indicator having some groundtruth values in the millions and billions, while others (e.g. unemployment) take on values under 10.

Validation. Over 20 LLMs and 11 indicators (44.5k total questions), automated parsing obtains a numeric answer 88.9% of the time. We further validate the correctness and completeness of our pipeline via three manual inspection studies. First, we check 450 random cases where a numeric answer could not be extracted. In 85.2% of cases, the LLM did not provide a parseable answer. Thus,



1. Initialize chat history with standardized prompt + example

[USER]: I will ask you factual questions about countries. Specifically, I will ask you for the population. You will answer as concisely as possible - only answer with the number! First I will give an example with the answer. Then I will ask you my question, and you will provide the answer in the same way.

[MODEL]: Sounds good, will do.

[USER]: What is the population for the country Switzerland? Do not answer in a complete sentence - only provide the number!

[MODEL]: 8,703,771

2. Ask question (indicator with country) in matching format

[USER]: What is the population for the country {country}? Do not answer in a complete sentence - only provide the number!

[MODEL]: **{model generated raw output}**

3. *Parse* raw output to numeric value

Figure 4: Standard pipeline for extracting numeric answers from LLMs. Each question is defined by a query (i.e. Before asking a language model a question, we prompt it with a base instruction and example. Then, we automatically parse the raw output to obtain a numeric value which can be compared to the groundtruth data.

our parsing is mostly complete, as **we obtain a numeric value in** 98.2% **of cases where an answer can be parsed**. To verify the correctness of the parsing, we first check 945 randomly selected raw LLM outputs where a numeric value was parsed. **In** 98.7% **of these cases, the parsed value was correct** (details in appendix C). Then, we take a closer look at parsed responses that incurred high (over 0.85) absolute relative error compared to the groundtruth value. For 825 randomly selected high error cases, the parsing was manually verified to be correct 93.7% of the time. Motivated by this slightly lower correctness rate, we also analyze median errors over groups in Appendix B, where observed trends are consistent (and disparities over Regions and Income groups are even larger). We conclude that our prompting and parsing pipeline is largely complete and correct. Nonetheless, when evaluating a new LLM, we recommend verifying the parsing behavior using the four validations we outline above, as individual LLMs can have unique idiosyncracies (e.g. special tokens or output patterns) that potentially could affect parsing. Along with all code, we will also publicly release methods to facilitate automatic and manual verification of parsing.

Groundtruth selection. For each indicator and country, data is available over a span of many years, though certain values are missing. To define a single groundtruth value for per country per indicator, we average the statistic over the past three years. The primary motivation for this strategy is to maximize the number of countries included in our study. Alternatively, one could select a specific year to draw all groundtruths from, though the number of countries considered would be lower than the averaging strategy. In Appendix D, we compare groundtruth values obtained via different selection methods, and observe groundtruths to only vary by a small amount. We also explore specifying a year when querying LLMs, and observe consistent results with respect to performance disparities to those observed without year specification in the query. Lastly, we more closely inspect overall error rates between LLM responses and groundtruths selected by specifying a year in section 6.2, to gain insight on if LLM responses are dated (i.e. more accurate for a prior year than the most recent year).

4 EVALUATION SUITE

We seek to evaluate a wide array of language models, including both open source and private. For the **open source models**, we utilize Huggingface's transformers library [29] to obtain and operate 15 models (and respective tokenizers). Namely, from Meta's LLama-2 [27], we include both base and chat-tuned versions of the $7B$ and $13B$ models, where $7B$ indicates 7 billion parameters. We also include two Vicuna models ($7B$ and $13B$), which are fine-tuned from Llama-2. From Microsoft, we have $7B$ and $13B$ Orca-2 models [18], as well as Phi-2, the smallest model in our suite with just $2.7B$ parameters. From Mistral-AI, we include the $7B$ instruction-tuned model [13]. We also study Zephyr-$7B$ β [28], tuned from a Mistral-AI model. Lastly, we include $7B$ and $14B$ Qwen models from Alibaba Cloud, both with and without chat-tuning [4]. For **closed source models**, we include the following LLMs. From OpenAI, we evaluate gpt-3.5-turbo and gpt-4 [21]. From Google, we evalute Gemini [26]. From Cohere, we evaluate the 'command' model, as well as the same model equipped with retrieval augmented generation (RAG) [2]. RAG is a procedure where a langauge model can retrieve relevant documents (in this case, from the internet) and look over them before generating a response.

5 RESULTS: PERVASIVE AND CONSISTENT GEOGRAPHIC DISPARITIES

5.1 Large disparities across Regions and Income groups

Figure 5 visualizes our central finding. Over 20 LLMs and 11 World Bank indicators, we observe substantially disparate average performance based on the Region and Income group of the country of interest. Namely, the mean absolute relative error is 0.316 and 0.321 for countries from North America and Europe & Central Asia

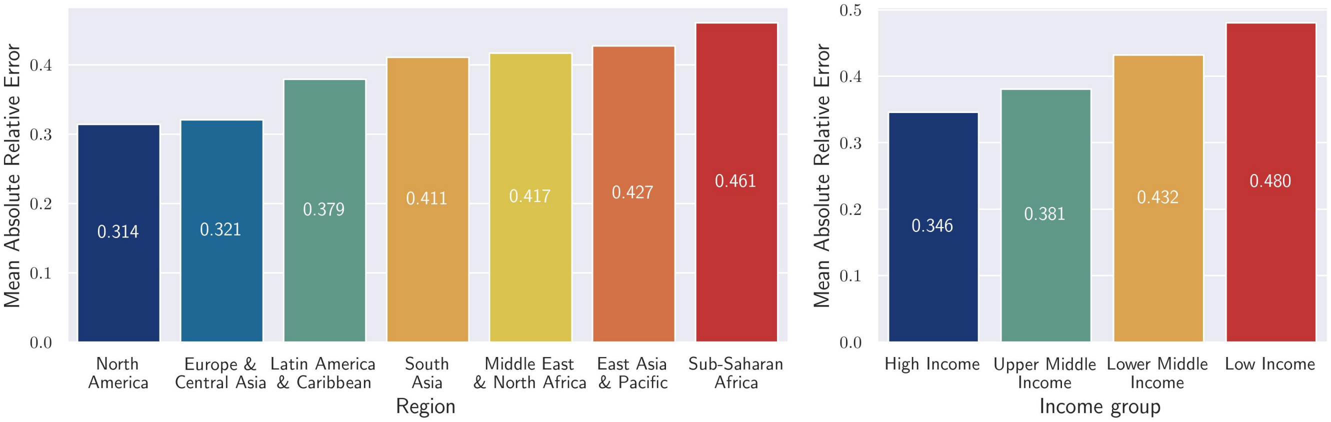


Figure 5: Language models exhibit disparate performance for countries from different regions and income groups. Error rates are lower for western and high income countries. Mean absolute relative error rate per region and income group reported over all 11 queries and 20 language models studied. When computing median instead of mean, similar trends hold, with even larger disparities (see Figure 15). We note that the best performing LLMs have much lower error rates than the averages presented above (see figure 7).

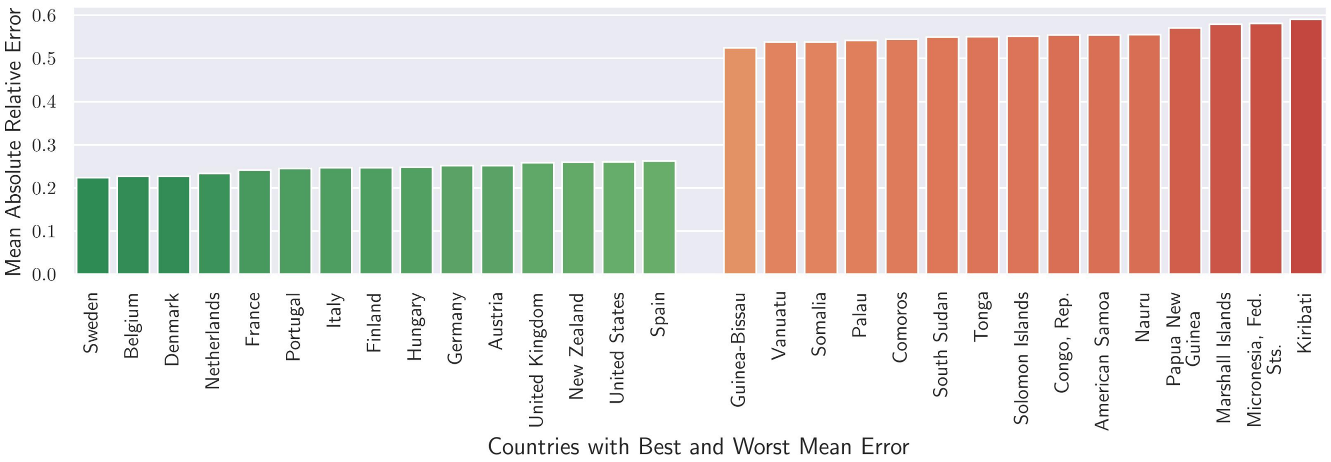


Figure 6: Error rates can vary significantly across countries, with some countries experiencing nearly 3× higher absolute relative error than others. Strikingly, all of the 15 countries with the lowest error rates fall in the high income category, while all of the 15 countries with the highest error rates fall in the low income category.

respectively. In contrast, the mean absolute relative error rises to 0.461 for countries from Sub-Saharan Africa, which is about 1.5× higher than the error for North America. For Income groups, mean absolute error rises steadily as the income level drops, with the lowest error being 0.346 for high income countries, and the highest error being 0.480 for low income countries.

5.2 Error nearly triples between some countries

On a per country basis, disparities can become even more pronounced. Figure 6 visualizes mean absolute error rate per country for the countries that, when asked about, language models (on average) have the most and least amount of error. We observe that 13 of the 15 countries that incur the least amount of error are European, while all 15 of these countries fall are categorized as high income. On the other hand, countries that incur the most error are all categorized as low income. Strikingly, error rises by a factor of nearly 3 across the two groups.

5.3 Consistent disparities across LLMs and indicators

Previously, we presented results averaged over all LLMs and indicators, grouped either by country or category (i.e. Region or Income group). We now inspect performance along the axes of LLMs and indicators separately, starting with LLMs. In addition

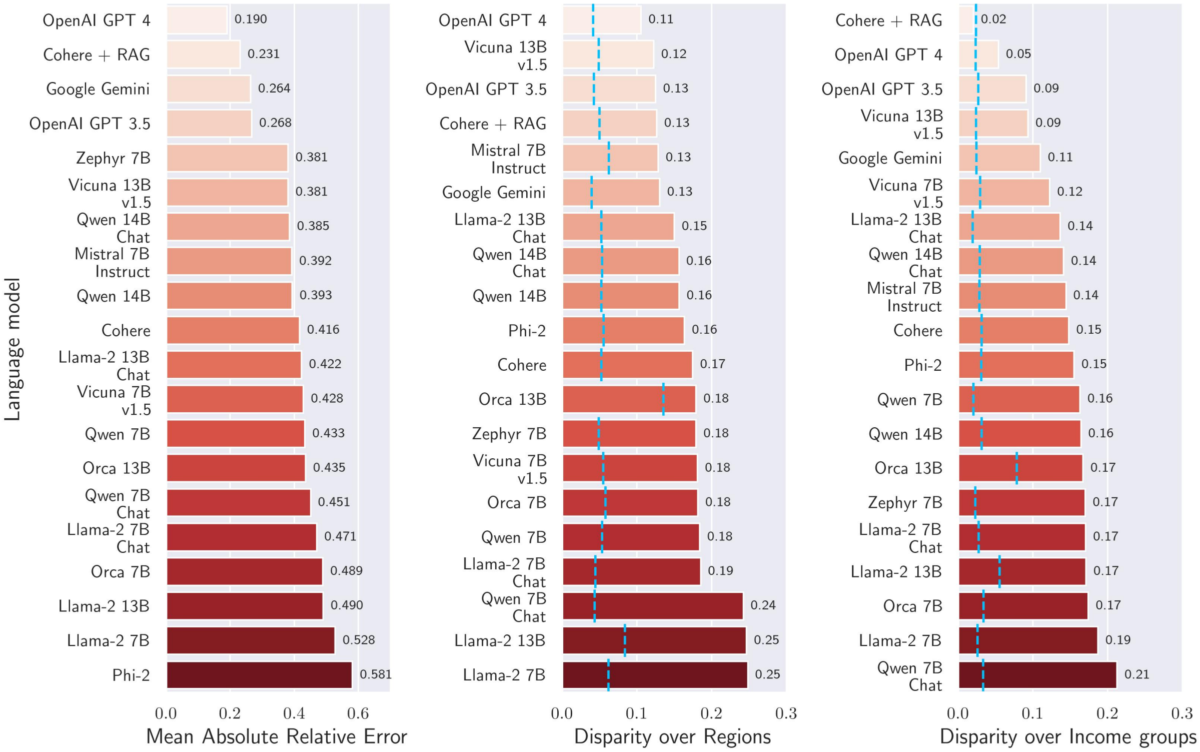


Figure 7: Performance of 20 LLMs averaged over 11 indicators from WorldBench. We present the absolute relative error (left), as well as disparities across regions (middle) and income groups (right). For disparities, the blue dashed lines correspond to the disparity incurred using a random categorization of countries (into 7 groups for Regions and 4 for Income groups), averaged over ten trials. Observed disparities far exceed the amount expected for a random categorization of countries across nearly all LLMs.

to absolute relative error, we also employ a second metric to summarize differences in performance across certain categories. Namely, we define *Disparity* as $\max_{e_i,e_j \in E} e_i - e_j$, where E is the set of mean absolute relative errors for each category of a given categorization. In other words, for example, *Disparity* over *Regions* is the gap between the mean absolute relative errors for the region with the greatest error and the region with the least error. Disparity also always falls between 0 and 1. To contextualize disparity scores, we compute a baseline corresponding to the disparity achieved using a random categorization of countries into k groups; we set $k = 7$ for Regions and $k = 4$ for Income groups. We approximate the baseline disparity given a set (i.e. for one LLM of interest) of per-country errors by applying ten random country categorizations and averaging the observed disparity over all trials.

Figure 7 visualizes average error and disparities per LLM. From the left most panel, we see that the lowest mean absolute relative error achieved is 0.19, and the value for most models is near 0.4, indicating that there is substantial room for improvement for this task. Shifting from error to disparity (middle and right panels), we observe that **all models exhibit disparate performance over regions**, with gaps of at least 0.1 between the regions with the most and least error per LLM. Across income groups, disparities are also consistently present, though to a lesser degree, with only 4 of the 20 models studied achieving a disparity below 0.1. Nonetheless, both over Regions and Income groups, observed disparity almost always far exceeds the expected disparity for a random categorization (blue dashed lines).

A few expected trends emerge: base models are outperformed by their chat-tuned versions; smaller models are outperformed by their larger versions. One such trend we highlight is the impact of retrieval augmented generation (RAG), which is utilized for to augment the Cohere LLM. Incorporating RAG reduces mean absolute error by nearly a factor of two, reducing it from 0.416 to 0.231. Impressively, RAG causes disparity across Income groups to nearly vanish, going from 0.15 to 0.02, the lowest such disparity observed across our model suite, and *on par with a random categorization of countries.* However, it is worth noting that RAG comes at the cost of latency, as internet searches are required and the LLM must review retrieved documents in addition to the provided prompt. Nonetheless, RAG appears to be a promising direction for reducing errors and also potentially disparities.

Turning our attention now to indicators, Figure 8 shows errors and disparities per indicator. Mean absolute relative error exceeds 0.3 for all but two of the indicators. Again, disparities are present for most cases, though they are more pronounced across Regions

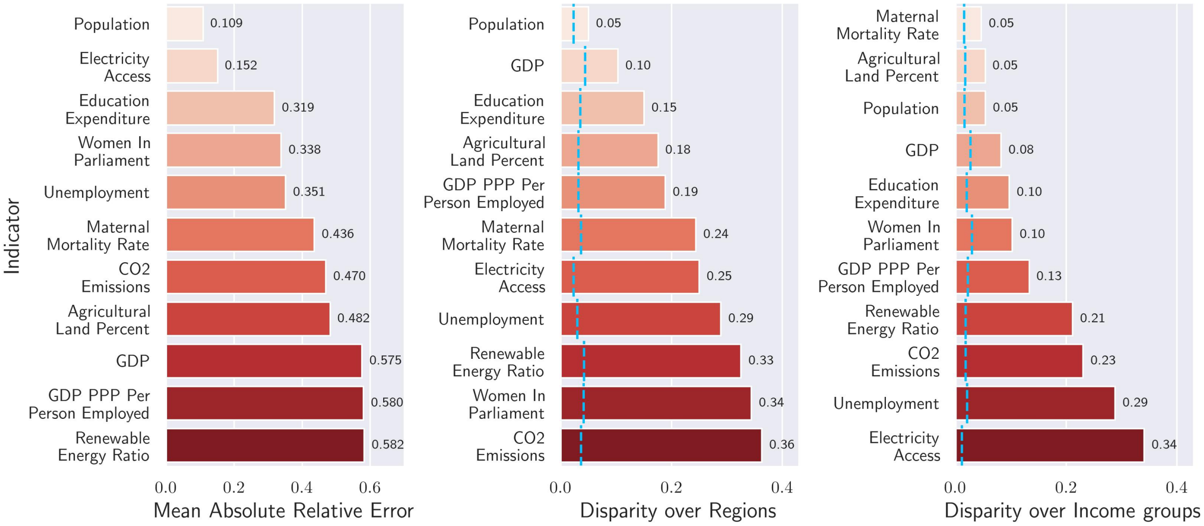


Figure 8: Error rates and disparities per indicator, averaged over LLMs. For disparities, the blue dashed lines correspond to the disparity incurred using a random categorization of countries (into 7 groups for Regions and 4 for Income groups), averaged over 10 trials.

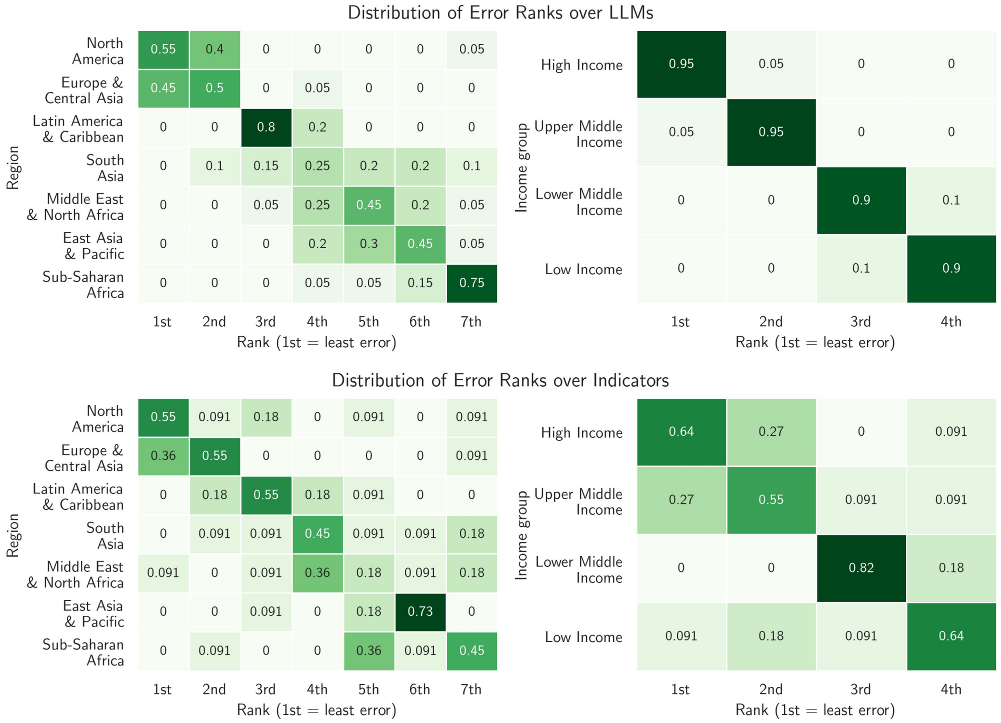
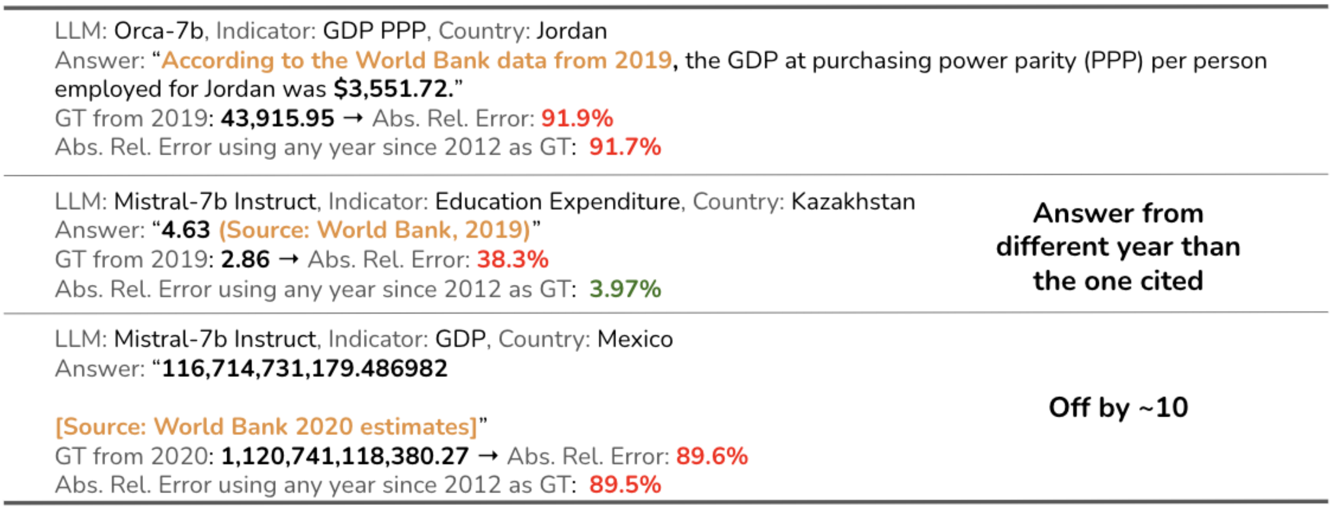


Figure 9: The order of regions and income groups by absolute relative error is largely consistent per LLM (top) and per indicator (bottom). For both LLMs and indicators, the regions with the lowest errors are most frequently North America and Europe & Central Asia, while the regions with the highest error are most frequently Sub-Saharan Africa and East Asia & Pacific. For Income groups, error nearly always increases as income decreases.



LLM: Orca-7b, Indicator: GDP PPP, Country: Jordan
Answer: "According to the World Bank data from 2019, the GDP at purchasing power parity (PPP) per person employed for Jordan was $3,551.72."
GT from 2019: 43,915.95 → Abs. Rel. Error: 91.9%
Abs. Rel. Error using any year since 2012 as GT: 91.7%

LLM: Mistral-7b Instruct, Indicator: Education Expenditure, Country: Kazakhstan
Answer: "4.63 (Source: World Bank, 2019)"
GT from 2019: 2.86 → Abs. Rel. Error: 38.3%
Abs. Rel. Error using any year since 2012 as GT: 3.97%

Answer from different year than the one cited

LLM: Mistral-7b Instruct, Indicator: GDP, Country: Mexico
Answer: "116,714,731,179.486982

[Source: World Bank 2020 estimates]"
GT from 2020: 1,120,741,118,380.27 → Abs. Rel. Error: 89.6%
Abs. Rel. Error using any year since 2012 as GT: 89.5%

Off by ~10

Figure 10: In addition to hallucinating false answers, we also observe LLMs to occassionally hallucinate *citations*. Above, a few examples of hallucination citation are shown.

than across Income groups. Moreover, over both Regions and Income groups, *observed per-Indicator disparity far exceeds the random baseline in almost all cases.* Indicators that seem to be driving the observed disparities include CO_2 Emissions, Renewable Energy Ratio, and Unemployment. For a complete breakdown of performance and disparities for each (LLM, indicator) pair, we refer to Appendix A.

5.4 Ordering of Regions and Income groups by error are consistent per-LLM and per-Indicator: Lowest error is with Western and high income groups

Having demonstrated that significant disparities are present for each LLM and each indicator separately, we now show that the order of disparity is consistent as well. that is, the regions and income groups with highest and lowest error (respectively) are the same within each subset. Namely, LLMs achieve the lowest error when answering questions about Western or high income countries, and they suffer the greatest error when answering questions about countries from the low income category. In figure 9, we show the distribution of *error ranks.* That is, e.g. in the top right heatmap, for each LLM, we rank the regions by their mean absolute relative erros, and then report the fraction of LLMs for which a region obtains a specific rank. Thus, we see that for 75% of the LLMs, the highest error occurs for Sub-Saharan African countries. Strikingly, the pattern across income groups is strongly pronounced. Error ranks are almost perfectly inversely related to amount of income, wiht the high income group having lowest error for 95% of LLMs and the low income group having highest error for 90% of LLMs. Again, the same trends emerge when inspecting error rates per indicator.

Thus, the original trends we observe when averaging over all LLMs and indicators, visualized in Figure 5, appear to hold when we inspect each LLM individually and each indicator individually. These results suggest that geographic and income-based disparities in LLM factul recall are pervasive throughout existing LLMs.

6 NOTEWORTHY OBSERVATIONS

In analyzing per-country performance and geographic disparities in LLM factual recall, we additionally came across a number of noteworthy observations made possible by our benchmark. First, we found that LLMs occasionally offer what resembles citations in their responses, including instances where the WorldBank itself was mentioned. Since we have that exact data, we were able to cross-check the LLM "citations". Second, because we have data per-country *per-year*, we could compute error rates while selecting groundtruths from specific years, so to see how up-to-date LLM responses are. We explore these observations in more detail below.

6.1 Citation Hallucination

Despite being prompted to only return a numeric value, the LLMs we studied still would often produce additional text. Interestingly, sometimes generated text would resemble a citation[1], claiming the provided answer was sourced from institutes like the World Health Organization, the International Monetary Fund, and even, the World Bank. In the last case, we cross-checked the provided responses to see if the numeric response matched the groundtruth World Bank data, contained in WorldBench. Overall, *responses with "citations" were no more accurate than those without "citations"*, still incurring substantial mean absolute relative errors. Specifically, in 650 instances where the string "World Bank" (case insensitive) was mentioned, mean absolute error rate was 0.465. This suggests that the LLM-produced "citations" are hallucinated, as the provided responses do not actually come from the sources listed. Figure 10 displays a few examples of LLM produced "citations". For each example, we highlight the "citation", and provide the absolute relative error of the parsed answer compared to (1) the groundtruth value from the specific year cited, and (2) the lowest absolute relative error to groundtruths for any of the past ten years. In the first example, the LLM answer is way off, despite the arguably convincing

[1] Such behavior has been observed in [1]

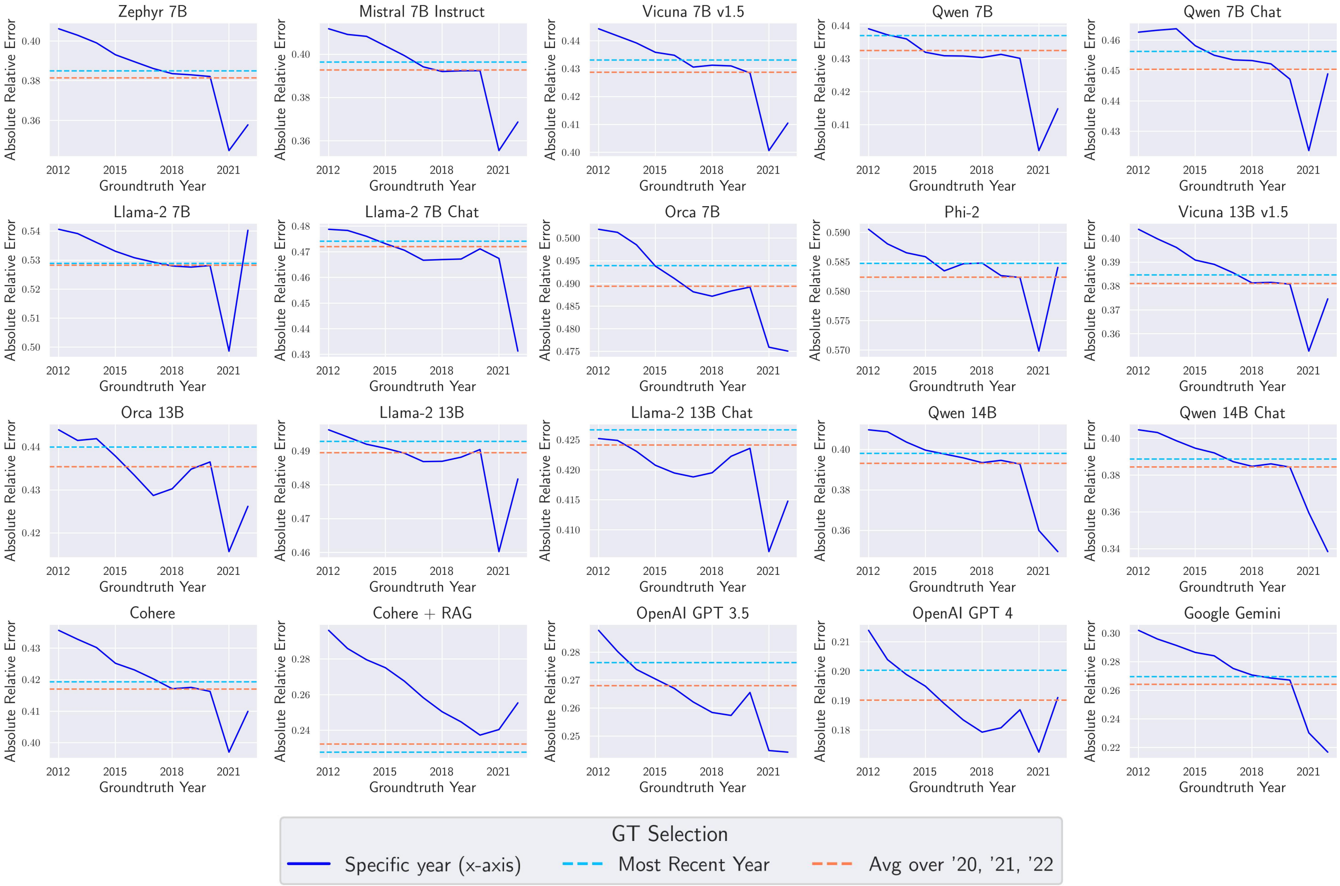


Figure 11: Error rate of LLM outputs compared to year from which groundtruth is extracted. Many models show the lowest error rate when their outputs are compared to groundtruths from 2021, indicating that models may already be slightly out of date.

"citation". Interestingly, we also observe an instance where the provided answer does not match the groundtruth from the cited year, yielding an error of 0.383, but it does match the groundruth from the following year, with error dropping to 3.97%. Finally, we see an example where the provided answer is off by almost exactly a factor of 10 (relative error of ~ 0.9). This highlights a pitfall in using LLMs to return numeric information, as the difference in tokens between two numbers can be very small, while the resultant encoded value can be very large.

In summary, hallucinated citations pose a serious challenge in LLM reliability. On one hand, producing false citations obfuscates model errors, and generally denigrates the overall trust the end user has in the system. On the other, that the LLMs appear to know what sources would contain the answer seem to be an encouraging sign to the potential benefits of retrieval-augmented systems.

6.2 Are some LLMs already out of date?

Now, we compare LLM responses to groundtruths from specific years for all LLM responses, not just the rare few where "citations" are present. Figure 11 shows the mean absolute relative error over indicators and all countries per LLM, computed using groundtruths selected in a variety of ways. The orange dashed line corresponds to the default groundtruth selection (averaging over any available data from the past three years), while the light blue one corresponds to using data from the most recent year (per country; details in D.2). The solid blue lines correspond to using the groundtruth value from the year on the x-axis. A trend that emerges in 13 of the 20 LLMs is that the lowest error occurs when comparing to data from 2021. In one extreme, error increases from 0.5 to 0.54 when changing the groundtruth year from 2021 to 2022. These results suggest that the facts internally stored in some LLMs may already be out of date, reporting statistics closer to previous years, especially if their training data was curated in years past. Of course, an LLM cannot recall a fact that did not exist at the time of its training. Nonetheless, as the use of LLMs continues to grow, the ability to stay up to date will be paramount. We hope WorldBench can aide in this pursuit.

Correlation with Indicator			Self-consistency		
Indicator	Pearson's r	p-value	Indicator	Pearson's r	p-value
Maternal Mortality Rate	0.383	0.029	Population	0.749	0.000
Renewable Energy Ratio	0.135	0.180	Electricity Access	0.653	0.000
Unemployment	0.078	0.344	Women In Parliament	0.386	0.036
Education Expenditure	-0.047	0.453	CO2 Emissions	0.338	0.012
Agricultural Land Percent	-0.075	0.372	Renewable Energy Ratio	0.337	0.004
Population	-0.138	0.112	Education Expenditure	0.325	0.016
CO2 Emissions	-0.189	0.106	Agricultural Land Percent	0.324	0.057
GDP	-0.194	0.074	Unemployment	0.319	0.004
Women In Parliament	-0.344	0.046	GDP	0.246	0.136
Electricity Access	-0.348	0.005	Maternal Mortality Rate	0.155	0.245
GDP PPP Per Person Employed	-0.396	0.015	GDP PPP Per Person Employed	-0.016	0.206

Table 2: (Left) Correlation between per-country mean absolute relative error and individual indicator values. (Right) Per-indicator, correlation between per-country mean absolute relative error and normalized standard deviation of responses obtained over five trials.

6.3 What kinds of countries experience high error rates?

We now present a purely correlational study to better understand what countries experience the highest error rates. Using the per-country data for each indicator studied, we compute the correlation between these values and per-country error. We also compare the normalized (by mean) standard deviation of responses per country per indicator, with responses taken over five trials. The hypothesis here is that LLMs will have greater variance in answering questions about countries they are less accurate for, similar to [1]; we call this self-consistency. We compute correlation to country-wise errors for each (LLM, indicator) pair separately, as the values can take on substantially different ranges as either LLM or indicator changes, and then average over all such pairs. Results are reported in table 2. We find that most indicators are not correlated with per-country error. The strongest correlation is -0.396 for GDP PPP per person employed, suggesting that LLMs perform worse on countries with lower per-person wealth. Notably, neither population nor GDP are correlated well with error. As for self-consistency, in most cases, correlations are within $0.3 - 0.4$. In a couple instances, high correlations are observed, suggesting that sampling multiple outputs and inspecting variance can sometimes (but not always reliably) aide in estimating the uncertainty of the LLM.

In summary, our simple correlational analyses do not shed much insight in to why particular countries incur higher error rates for LLMs. We conjecture that the availability of training data plays a large role. However, the groundtruths are available *for all countries*, and World Bank data is likely in the training sets of many LLMs, as indicated by the hallucinated citations to them. We leave investigation to the cause of the geographic disparities we observe to future work.

7 LIMITATIONS

Is it reasonable to expect language models to perform this task? LLMs are not directly optimized for information retrieval, and developers often caution that LLMs many not always provide factual answers. Furthermore, retrieving specific numbers can be challenging, given the fact that many sequences of numbers are feasible/reasonably likely to appear in natural language, where as the distribution of words has far less entropy. Nonetheless, LLMs have been observed to produce factual responses to certain queries, achieving as high as 86% exact match on TriviaQA [3]. Indeed, in our experiments, we observe mean absolute error rates as low as 3.6% for the Population indicator and 5.8% for the Electricity Access indicator (see Appendix A), suggesting that LLM-based factual recall is feasible. We emphasize that the point of our benchmark is to enable comparison in LLM performance across countries, so to uncover systemic disparities. Moreover, despite warnings from developers, as LLMs become more ubiquitous, end users will likely still make factual queries, to which we'd hope language models respond accurately, and importantly, without substantial differences in performance due to factors like geography or wealth of the country of interest. Thus, we hope our benchmark aide in assuring that LLMs exhibit fair performance when deployed.

Can LLMs ever ace this task? Some of the indicators studied are volatile, in the sense that they change non-trivially from year to year. Also, some metrics can take on slightly different values based on which organization measured them (e.g. the World Bank's numbers may differ from the United Nation's numbers). Thus, we do not expect LLMs to achieve perfect performance on this metric. Nonetheless, we believe our benchmark can offer valuable signal in measuring geographic disparities. That is, even though error rates may never be exactly zero, we can hope that they will not vary substantially across countries.

8 CONCLUSION

We present WorldBench, a benchmark to quantify geographic disparities in LLM factual recall. We find pervasive and consistent biases across 20 evaluated LLMs, with Western and higher income countries experiencing lower error rates. By utilizing World Bank data, our benchmark is flexible and will remain up to date. Thus, we hope our benchmark can aide in reducing geographic disparities of future generations of LLMs, towards models that work well *for all*.

ACKNOWLEDGMENTS

This project was supported in part by a grant from an NSF CAREER AWARD 1942230, ONR YIP award N00014-22-1-2271, ARO's Early Career Program Award 310902-00001, HR00112090132 (DARPA/RED), HR001119S0026 (DARPA/GARD), Army Grant No. W911NF2120076, the NSF award CCF2212458, NSF Award No. 2229885 (NSF Institute for Trustworthy AI in Law and Society, TRAILS), an Amazon Research Award and an award from Capital One.

REFERENCES

[1] Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2023. Do Language Models Know When They're Hallucinating References? arXiv:2305.18248 [cs.CL]

[2] Cohere AI. 2023. Cohere API. https://github.com/cohere-ai/cohere-python

[3] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report. arXiv:2305.10403 [cs.CL]

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609* (2023).

[5] World Bank. 2024. World Bank Country and Lending Groups. Retrieved January 16, 2024 from https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups#:~:text=For%20the%20current%202024%20fiscal,those%20with%20a%20GNI%20per

[6] World Bank. 2024. World Bank Open Data. Retrieved January 16, 2024 from https://data.worldbank.org/

[7] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[9] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does Object Recognition Work for Everyone? arXiv:1906.02659 [cs.CV]

[10] Laura Gustafson, Megan Richards, Melissa Hall, Caner Hazirbas, Diane Bouchacourt, and Mark Ibrahim. 2023. Pinpointing Why Object Recognition Performance Degrades Across Income Levels and Geographies. arXiv:2304.05391 [cs.CV]

[11] Melissa Hall, Candace Ross, Adina Williams, Nicolas Carion, Michal Drozdzal, and Adriana Romero-Soriano. 2024. DIG In: Evaluating Disparities in Image Generations with Indicators for Geographic Diversity. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=FDt2UGM1Nz Featured Certification.

[12] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv:2311.05232 [cs.CL]

[13] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL]

[14] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large Language Models Struggle to Learn Long-Tail Knowledge. arXiv:2211.08411 [cs.CL]

[15] Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. arXiv:2303.18027 [cs.CL]

[16] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. GPT-4 Passes the Bar Exam. (March 15 2023). Available at SSRN: https://ssrn.com/abstract=4389233 or http://dx.doi.org/10.2139/ssrn.4389233.

[17] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. arXiv:2212.10511 [cs.CL]

[18] Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agrawal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching Small Language Models How to Reason. arXiv:2311.11045 [cs.AI]

[19] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. *ArXiv* abs/2303.13375 (2023). https://api.semanticscholar.org/CorpusID:257687695

[20] Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David I. Adelani. 2023. How good are Large Language Models on African Languages? arXiv:2311.07978 [cs.CL]

[21] OpenAI. 2023. OpenAI API. https://openai.com/product

[22] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code Llama: Open Foundation Models for Code. arXiv:2308.12950 [cs.CL]

[23] Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-Tail: How Knowledgeable are Large Language Models (LLM)? A.K.A. Will LLMs Replace Knowledge Graphs? arXiv:2308.10168 [cs.CL]

[24] Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-Augmented Language Models. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=-cqvvvb-NkI

[25] Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F. Rousseau, and Greg Durrett. 2023. Understanding Factual Errors in Summarization: Errors, Summarizers, Datasets, Error Detectors. arXiv:2205.12854 [cs.CL]

[26] Gemini Team. 2023. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL]

[27] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv* abs/2307.09288 (2023). https://api.semanticscholar.org/CorpusID:259950998

[28] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct Distillation of LM Alignment. arXiv:2310.16944 [cs.LG]

[29] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu,

Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs.CL]

[30] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. arXiv:2301.13848 [cs.CL]

[31] Yifan Zhang, Cheng Wei, Shangyou Wu, Zhengting He, and Wenhao Yu. 2023. GeoGPT: Understanding and Processing Geospatial Tasks through An Autonomous GPT. *ArXiv* abs/2307.07930 (2023). https://api.semanticscholar.org/CorpusID:259937048

[32] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL]

A COMPLETE RESULTS BREAKDOWN

We now present the results as completely as possible. In Figure 12, we present mean absolute relative error per LLM per indicator. In Figure 13, we present disparities over regions per LLM per indicator, and in Figure 14 we show the same for disparities over income groups. In general, the indicators that are most challenging are challenging for all LLMs.

B LARGER DISPARITIES WHEN USING MEDIAN INSTEAD OF MEAN ERROR

We now present results when aggregating with median instead of mean. Figure 15 shows that disparities grow larger when inspecting median absolute relative error instead of mean. We attribute this difference to some outlier countries, such as Bermuda for North America and Greenland for Europe & Central Asia.

C VALIDATION DETAILS FOR THE PROMPTING AND PARSING PIPELINE

We propose a general (i.e. for any LLM) pipeline for prompting LLMs responses to flexible (with respect to the country or indicator in question) queries. We seek to validate two aspects of this pipeline: completeness, where the parsing successfully extracts numeric answers in all instances where a numeric answer was provided, and correctness, where the parsed number should match the original numeric value embedded in the text. By simply running our parsing method, we can obtain our first statistic: parsing extracted a numeric answer for 88.9% of responses. For the 11.1% of responses where parsing failed, failures are either due to the LLM not providing a parseable answer (e.g. abstaining from answering or providing gibberish) or due to incompleteness of the parser. Through our manual verification, we find that 85.2% of failed parsing instances are due to the LLM provding unparseable responses. Thus, we obtain parseable responses in $88.9\% + (1 - 0.852) * 11.1\% = 90.54\%$ of total instances. Of this portion, $88.9/90.54 = \mathbf{98.2}\%$ of times, we obtain a numeric value when it is possible. Thus, our parsing is mostly complete. Correctness is easier to demonstrate, via manual verification of parsing outputs to original raw responses. Here, we find that 98.7% of parsed numbers are indeed correct, matching the number embedded in the raw response. This parsing allows us to employ the simple metric of relative absolute error, as it operates directly on numeric values, as opposed to more opaque metrics that leverage LLMs as a judge.

D ALTERNATE GROUNDTRUTH SELECTION STRATEGIES

D.1 Variance across groundtruth values selected from different years

We confirm that variance due to alternate groundtruth selection strategies is minimal. Groundtruths can be selected by specifying a particular year, or by averaging over the past three years, as we do in the main text. Table 3 shows the mean absolute relative error obtained by comparing the groundtruth value obtained by selecting a specific year and the groundtruth value obtained by averaging over the past three years. We find that, averaged over all indicators, the absolute relative error between two different groundtruth values

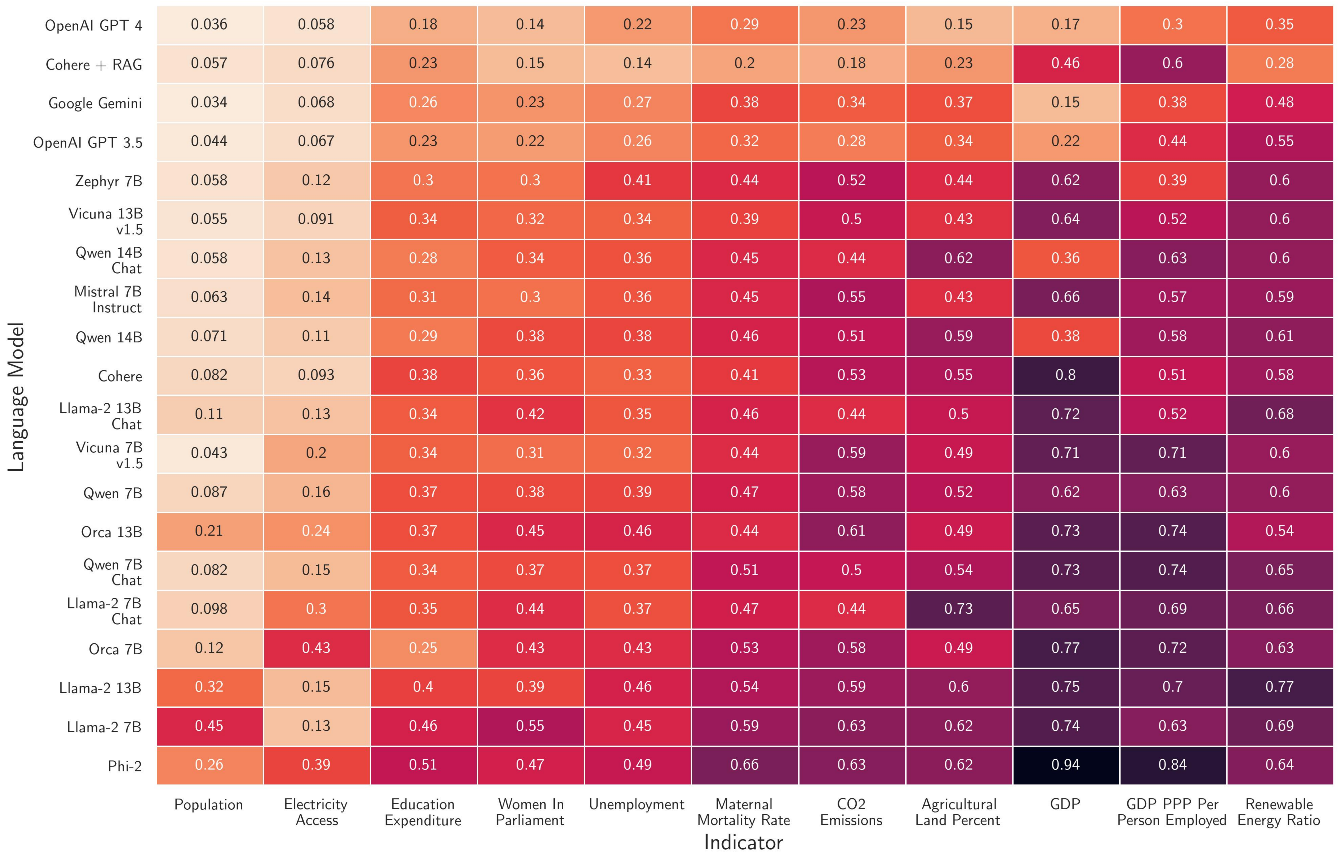


Figure 12: Absolute relative error averaged over countries per LLM and Indicator. Language models and indicators are each sorted by overall average error respectively.

	Specified Year for Alternate Groundtruth					
Indicator	2018	2019	2020	2021	2022	Average
Agricultural Land Percent	0.011	0.008	0.002	0.002	NaN	0.006
CO_2 Emissions	0.104	0.092	0.000	NaN	NaN	0.065
Education Expenditure	0.108	0.095	0.054	0.053	0.075	0.077
Electricity Access	0.026	0.017	0.006	0.006	NaN	0.014
GDP	0.106	0.093	0.109	0.038	0.086	0.087
GDP PPP Per Person Employed	0.059	0.047	0.033	0.014	0.028	0.036
Maternal Mortality Rate	0.087	0.074	0.000	NaN	NaN	0.054
Population	0.037	0.025	0.013	0.001	0.012	0.018
Renewable Energy Ratio	0.114	0.093	0.009	0.025	NaN	0.060
Unemployment	0.140	0.139	0.063	0.033	0.082	0.091
Women In Parliament	0.170	0.130	0.084	0.066	0.068	0.103
Average	0.087	0.074	0.034	0.027	0.059	0.056

Table 3: Comparing alternative groundtruth values to the value computed using our method (averaging over any available groundtruth numbers from 2020 to 2022). Using groundtruths from earlier years invokes higher error. On average, absolute relative error is only 5.6% **between different choices for groundtruth.**

is 5.6%, driven by the Unemployment and Women in Parliament indicators. We conclude that it may be unreasonable for any LLM to achieve zero error on this benchmark, as values can change from

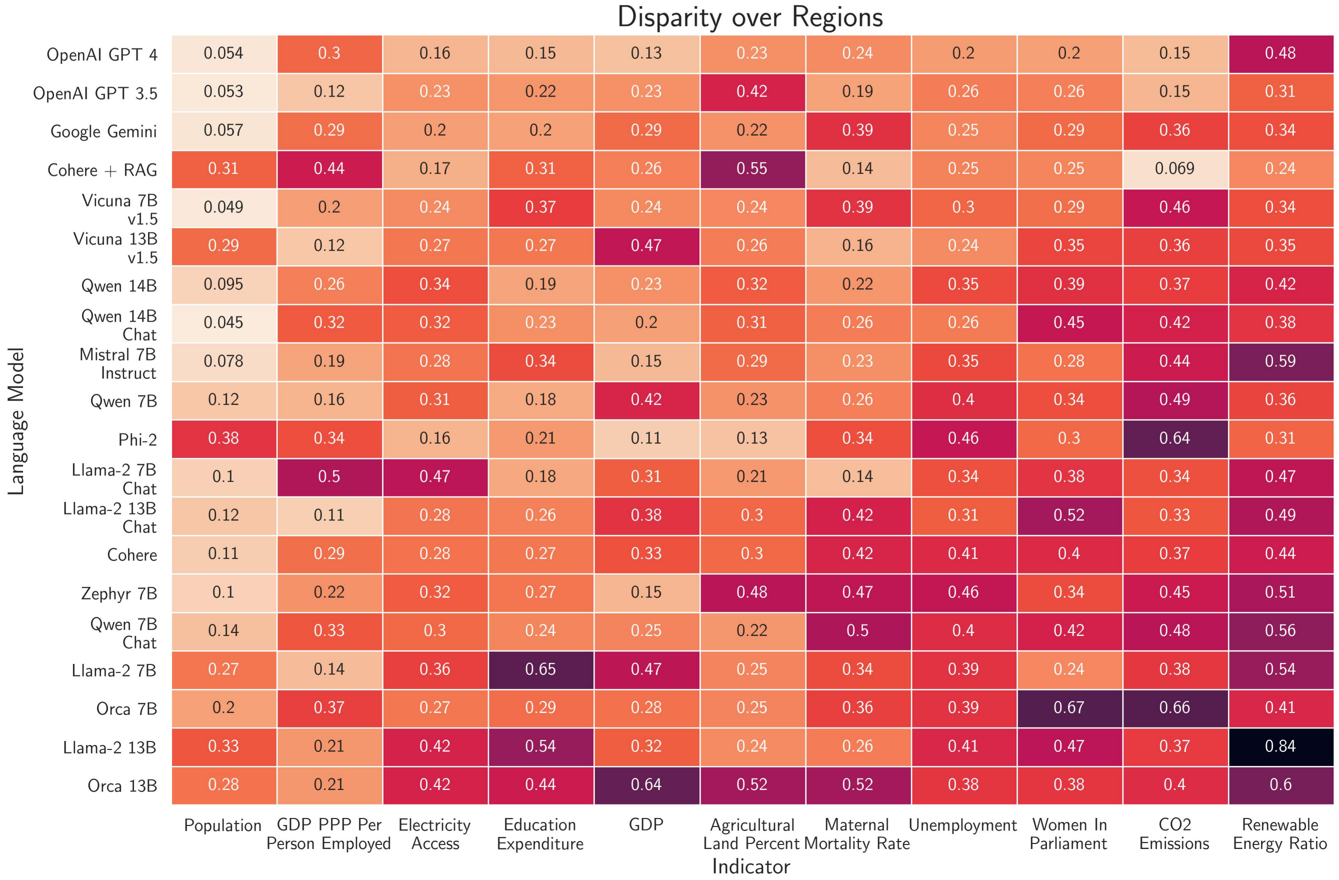


Figure 13: Disparities over Regions per LLM and Indicator. Language models and indicators each sorted by overall average error.

year to year, with some indicators being more volatile. Nonetheless, our benchmark can still offer valuable signal for measuring disparities (its intended purpose), as volatilities are present for all countries.

D.2 Clarification on using 'most recent year'

In Figure 11, we plot the error incurred when comparing to groundtruth values selected over different years, so to investigate if LLM reported statistics are closer to values from previous years (see section 6.2). One baseline selection strategy was termed 'most recent year'. We now clarify how this value is computed. We pick the most recent available statistic *per-country*, as some countries may have more recent statistics than others. We exclude any countries that have no statistics for each of the past five years. Note that at the time of this study (December '23), the most recent available statistics for any country was from 2022. Thus, for some countries, the 'most recent year' groundtruth was be drawn from as early as 2017, though in the vast majority of cases, it was drawn from 2022.

D.3 Specifying a year *in the question*

We also investigate if observed errors or disparities by LLMs could be caused by ambiguity in our prompt. Namely, in our prompt, we do not specify the *year* from which we desire the LLM to provide the requested metric for the given country. In the absence of a specification, we believe it is reasonable to assume that the most recent value is desired. Nonetheless, we conduct extra experiments where a specific year is mentioned in the prompt. We ask for values from 2021 and from 2016. Table 4 shows the results. Trends are very similar for both cases where a year is specified, and the case where no year is specified (matching the results we present in the main text). Note: GPT-4 was excluded in this ablation, purely for reasons of reducing cost.

E SIMILAR RESULTS WHEN USING DIFFERENT EXAMPLE COUNTRIES

We also verify that changing the choice of example country does not alter our main findings. Recall that we provide an example in our standard prompt. We originally chose Switzerland, as it had data for all indicators in the study. Now, we also inspect results when using Colombia and Mali as example countries. We choose these countries as they pertain to Regions that experience different levels of error (Colombia incurs around an average level of error, while Mali incurs high error). Table 5 shows the results. Again, main trends are consistent, with Western and High income countries

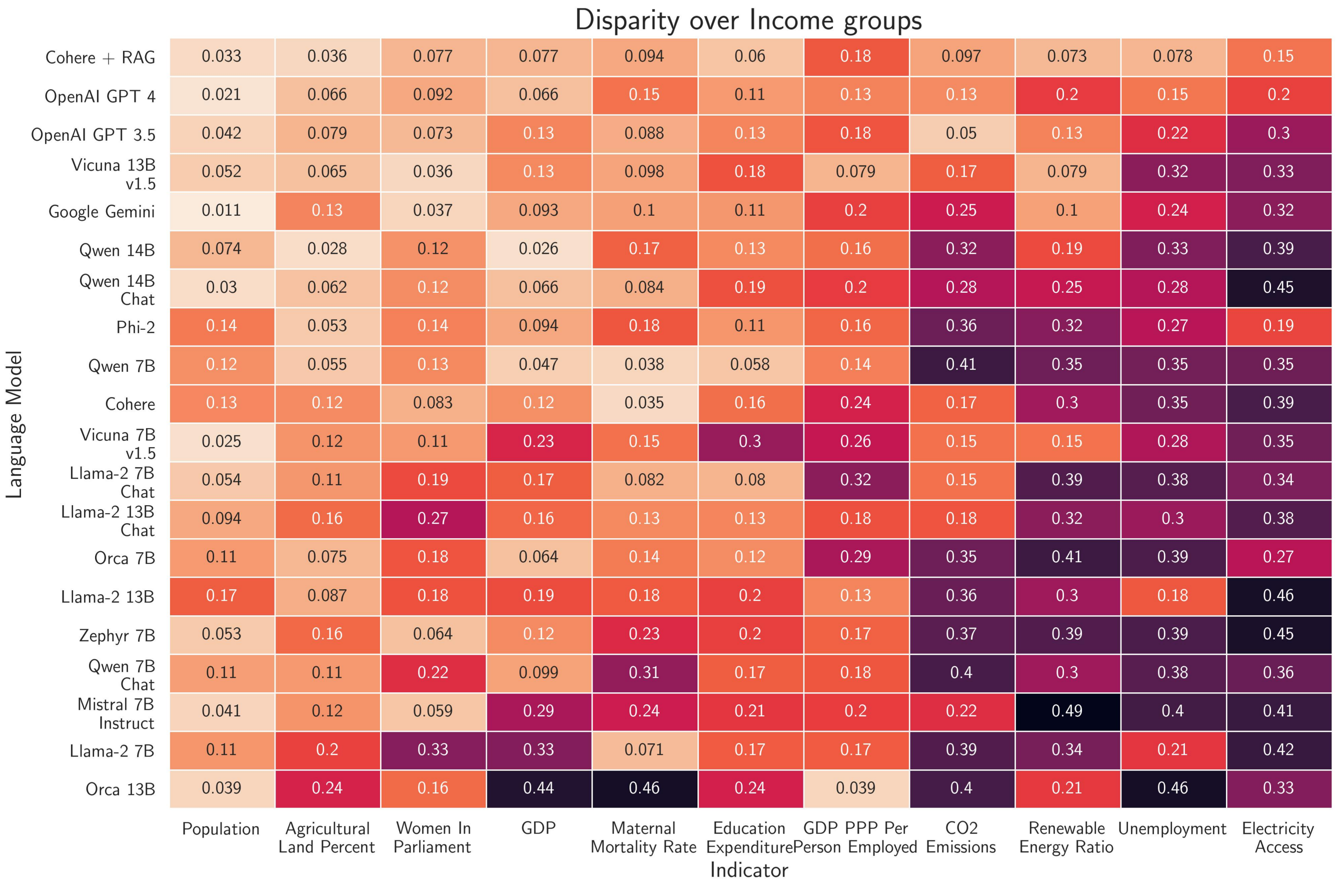


Figure 14: Disparities over income groups per LLM and Indicator. Language models and indicators each sorted by overall average error.

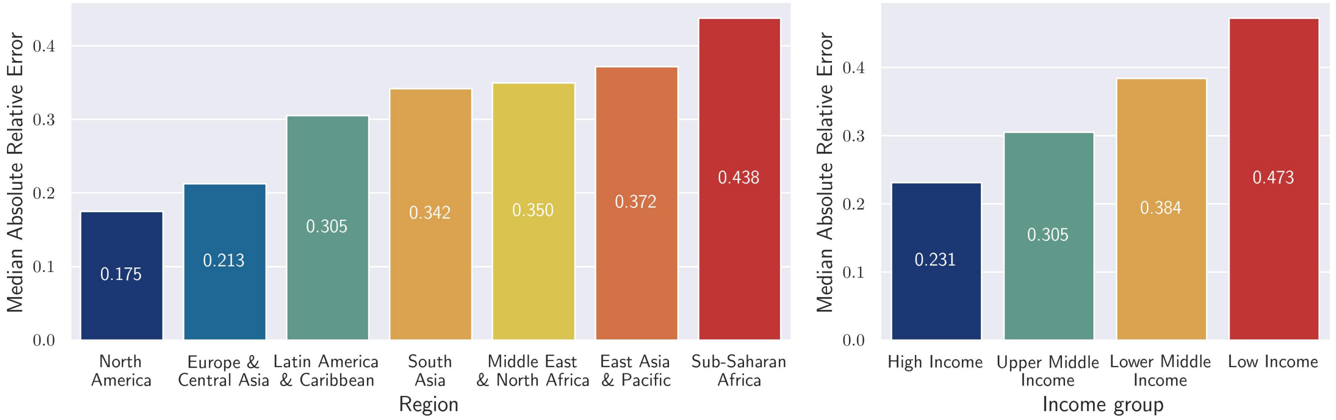


Figure 15: Median absolute relative error per region and income group. See figure 5 for mean errors.

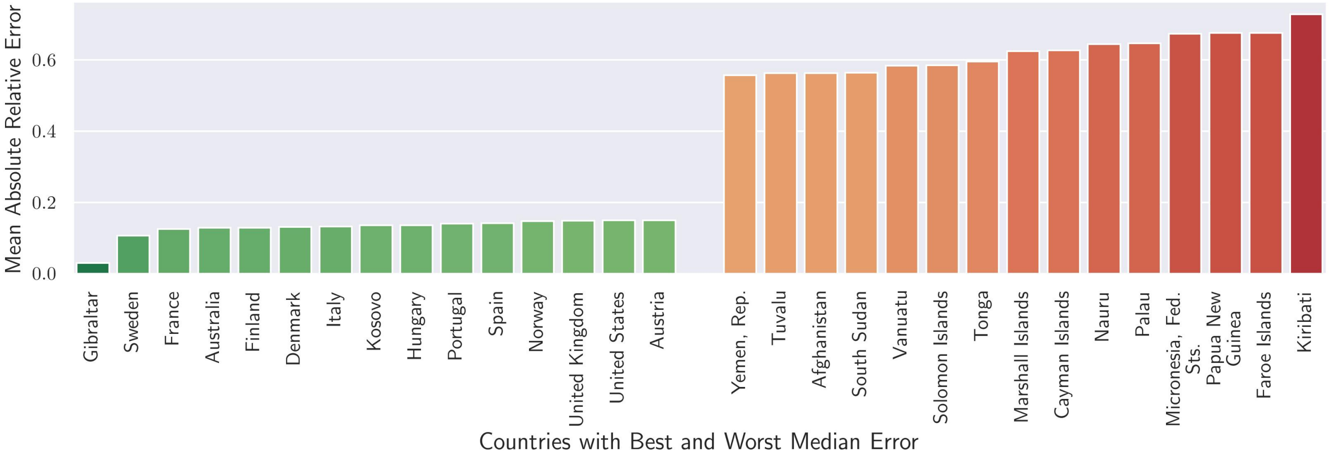


Figure 16: Median absolute relative error per country for countries with most and least median error. Again, the countries with least error belong to Western regions and the high income category. See figure 5 for mean errors.

Groundtruth Year →	2021	2016	Average	2021	2016	Average
Category ↓	Mean Abs. Rel. Error			Median Abs. Rel. Error		
North America	0.336	0.302	0.325	0.175	0.133	0.19
Europe & Central Asia	0.303	0.329	0.331	0.177	0.233	0.227
Latin America & Caribbean	0.367	0.389	0.39	0.274	0.324	0.326
South Asia	0.378	0.405	0.425	0.277	0.339	0.363
Middle East & North Africa	0.396	0.427	0.429	0.299	0.378	0.373
East Asia & Pacific	0.403	0.434	0.439	0.319	0.391	0.394
Sub-Saharan Africa	0.44	0.475	0.475	0.417	0.461	0.462
High income	0.333	0.351	0.356	0.193	0.249	0.25
Upper middle income	0.364	0.389	0.393	0.276	0.32	0.326
Lower middle income	0.405	0.442	0.444	0.339	0.41	0.405
Low income	0.466	0.498	0.497	0.462	0.503	0.498

Table 4: Mean and median absolute relative errors when using different groundtruth years. Importantly, in the columns for 2021 **and** 2016**, those specific years are included *in the question.* That is, we instruct the LLM to provide the statistic for a specific year, and compute error with respect to the groundtruth from that year. General trends are the same compared to when a year is not specified (denoted 'Average', our usual strategy), with Western and high income countries achieving lower error rates.**

incurring lowest error. The size of disparity is slightly reduced when using Mali as the example country, though this effect is not as strong when inspecting median errors, suggesting that outliers may be effecting the exact size of the disparity. Note: closed source LLMs were excluded in this ablation, purely for reasons of reducing costs.

E.g. Country →	Switzerland	Colombia	Mali	Switzerland	Colombia	Mali
Category ↓	Mean Abs. Rel. Error			Median Abs. Rel. Error		
North America	0.346	0.356	0.379	0.215	0.23	0.252
Europe & Central Asia	0.361	0.362	0.389	0.272	0.291	0.322
Latin America & Caribbean	0.42	0.395	0.429	0.374	0.348	0.389
South Asia	0.468	0.431	0.443	0.443	0.39	0.421
Middle East & North Africa	0.463	0.451	0.482	0.441	0.419	0.489
East Asia & Pacific	0.472	0.446	0.472	0.452	0.406	0.465
Sub-Saharan Africa	0.515	0.485	0.466	0.52	0.48	0.456
High income	0.382	0.388	0.412	0.289	0.32	0.348
Upper middle income	0.428	0.402	0.437	0.378	0.35	0.393
Lower middle income	0.483	0.442	0.458	0.47	0.41	0.448
Low income	0.537	0.517	0.479	0.554	0.525	0.474

Table 5: Mean and median absolute relative errors when using different example country (columns) in prompt. General trends are the same across choice of example country, with Western and high income countries achieving lower error rates.