WORLDBENCH: QUANTIFYING GEOGRAPHIC DISPARITIES IN LLM FACTUAL RECALL

Anonymous authors
Paper under double-blind review

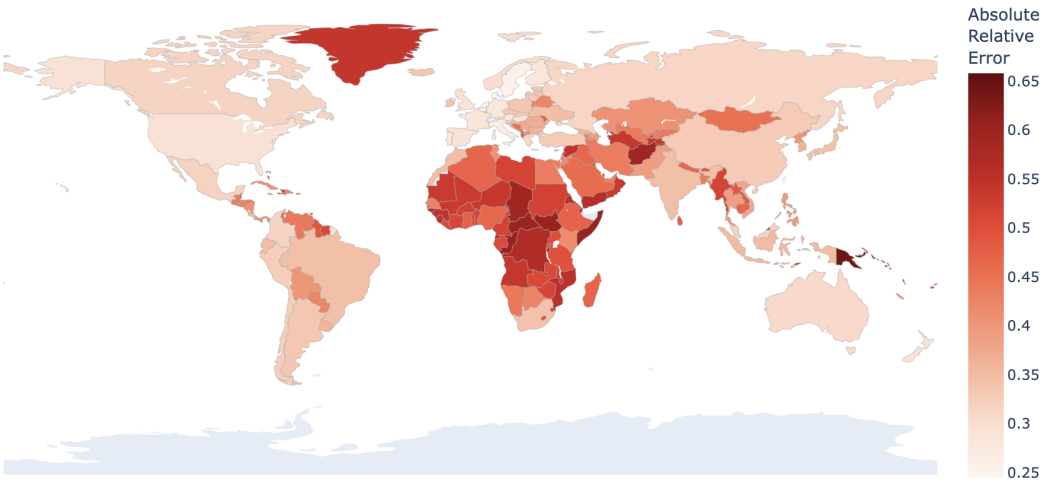

Figure 1: WORLDBENCH leverages World Bank data to assess the ability of LLMs to recall factual information about specific countries. Above, we plot the absolute relative error per country, averaged over 11 global development statistics queried to 20 state of the art open source and private LLMs. WORLDBENCH 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). We hope our benchmark will draw attention to geographic disparities in existing LLMs and facilitate the remedying of these biases.

1 INTRODUCTION

Large language models (LLMs) are the public face of AI and foundation models. They are easy to access and have demonstrated impressive performance on a wide array of tasks, including real-world benchmarks like legal and medical licensing exams (Katz et al., 2023; Nori et al., 2023; Kasai et al., 2023). However, LLMs are known to hallucinate, posing risk for risks for factual recall tasks. Beyond correctness, AI has also had well-documented challenges with performance disparities, at times manifesting in fairness issues (Buolamwini & Gebru, 2018b; DeVries et al., 2019; Gustafson

et al., 2023; Ojo et al., 2023). That is, models that seem strong according to summary metrics fail disproportionately for certain groups, potentially amplifying harmful societal biases. The responsible development of AI hinges on ensuring that models work *for all*, and a key first step in this pursuit is creating benchmarks that 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 Bank (2024b). We build and validate (via human inspection) an automated, indicator-agnostic prompting and parsing pipeline to interface with the World Bank data. 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, to private ones accessible via API, like GPT-4 OpenAI (2023). As visualized in Figure 1, 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 $1.5\times$ 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 individually. That is, *all* 20 *LLMs exhibit geographic disparities in factual recall*. We hope our benchmark can facilitate further research on the fairness of LLMs, towards building models that work reliably *for all*.

2 WORLDBENCH: A FLEXIBLE AND DYNAMIC BENCHMARK

We utilize data from the World Bank, which provides per-country statistics for numerous global development indicators (e.g. population, education expenditure, CO_2 emissions). We design a standard prompting and parsing protocol that can be applied for any indicator and country. In short, we provide an example question and response before asking the question of interest, so that the model's output can be automatically parsed. Through three manual inspection studies of $\sim 2k$ responses, we verify the completeness and correctness of our pipeline: we extract a numeric value in 98.2% of cases where an answer can be parsed, and in 98.7% of these cases, the parsed value was correct. Appendix A has more details on indicators studied, prompting, parsing, groundtruth selection, etc.

Our primary metric is *absolute relative error*, defined as $\frac{|a-b|}{\max(a,b)}$ for two scalars a, b. We also compute *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 member of a given category. To categorize countries, we again appeal to the World Bank, which assigns each country to one of 7 regions and 4 income groups (see appendix A). To contextualize disparity scores, we compute a baseline of the disparity for a random categorization (approximated over ten trials) of countries into k groups; we set $k = 7$ for Regions and $k = 4$ for Income groups.

Compared to other evaluations, WORLDBENCH has a few unique advantages. First and foremost, all countries are *equally represented*. Secondly, data quality and objectiveness is assured, as they are collected from a reputable third party (the World Bank) with no stake (or part) in the handling and outcome of the evaluations. Lastly, WORLDBENCH is *flexible*, since one can choose indicators as they please, and *dynamic*, since data is maintained yearly, enabling the longevity of the benchmark as well as analyses along a temporal dimension (e.g. to inspect if an LLM's knowledge is out of date). We refer readers to Appendix H for a review of other relevant benchmarks.

3 RESULTS

Figure 2 visualizes our central finding. We observe substantially disparate average performance based on the Region and Income group of the country of interest. Namely, mean error for countries from Sub-Saharan Africa is roughly $1.5\times$ higher than for countries form from North America and

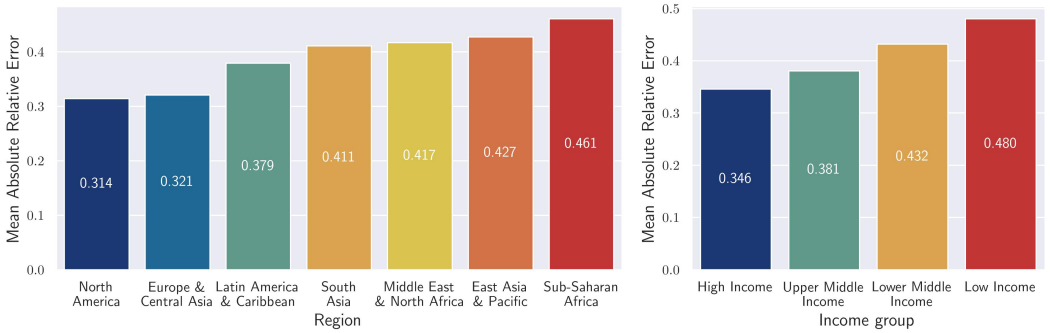

Figure 2: Language models exhibit disparate performance for countries from different regions and income groups. Error rates are lower for western and high income countries. Mean absolute relative error rate per region and income group reported over all 11 queries and 20 language models studied.

Europe & Central Asia. Error rises steadily as the income level drops, with the the highest error being 0.480 for low income countries vs. 0.346 for high income countries. Disparities are more extreme across countries, with error nearly triples between countries with the least and most error. Notably, of the 15 countries that incur lowest error, 13 of them are European, and *all* of them fall in the high income category (see figure 13 in Appendix C, which contains a complete breakdown of results). When inspecting each LLM individually (visualized in Figure 3), we find that all models (i) struggle with WORLDBENCH, and (ii) exhibit disparate performance. The left panel shows that the lowest mean absolute relative error achieved is 0.19, and the value for most models is near 0.4. The middle and right panels show disparities across regions and income groups. In nearly all cases, the observed disparity far exceeds the expected disparity for a random categorization (blue dashed lines).

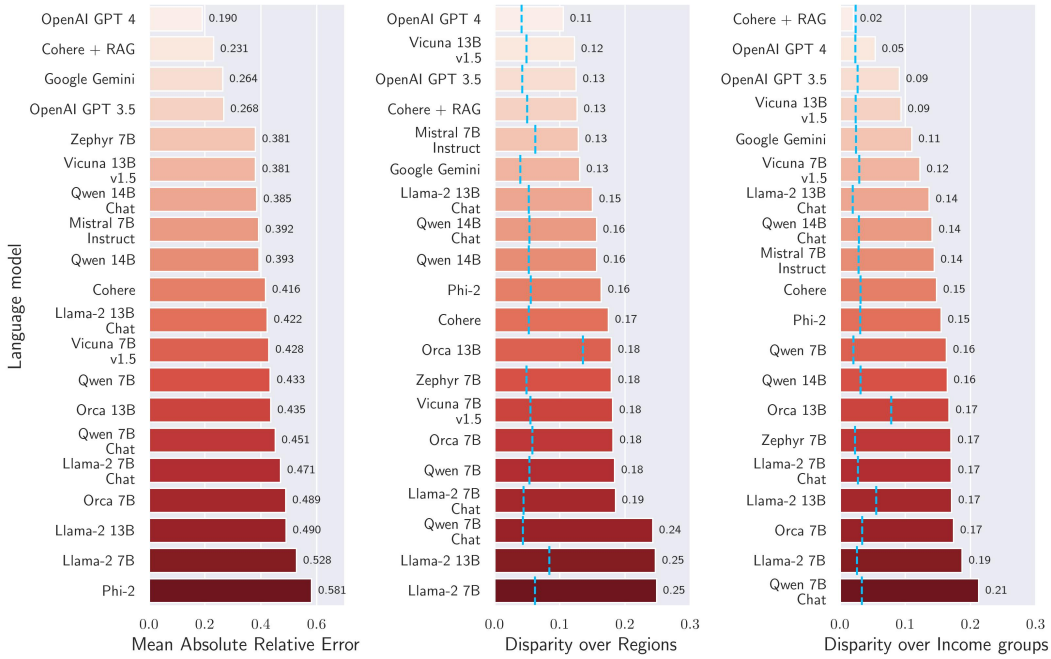

Figure 3: Performance of 20 LLMs averaged over 11 indicators from WORLDBENCH. We present the absolute relative error (left), as well as disparities across regions (middle) and income groups (right). For disparities, the blue dashed lines correspond to the disparity incurred using a random categorization of countries, which is exceeded by observed disparity across nearly all LLMs.

In comparing LLMs, expected trends emerge: base models are outperformed by their chat-tuned versions; smaller models are outperformed by their larger versions. One trend we highlight is the

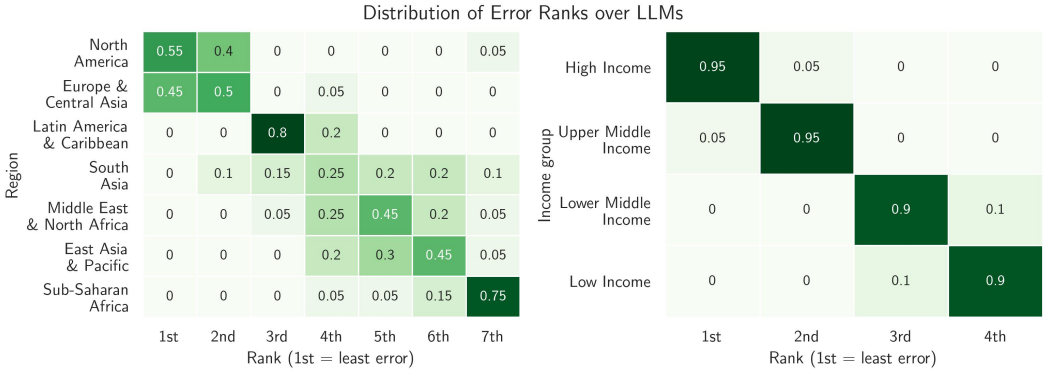

Figure 4: The order of regions and income groups by absolute relative error is largely consistent per LLM. The regions with the lowest errors are most frequently North America and Europe & Central Asia, while the regions with the highest error are most frequently Sub-Saharan Africa and East Asia & Pacific. For Income groups, error nearly always increases as income decreases.

impact of retrieval augmented generation (RAG), utilized to augment the Cohere LLM. Incorporating RAG reduces mean absolute error by nearly a factor of two, from 0.416 to 0.231. RAG also 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*.

Not only do disparities emerge for all LLMs, but the regions and income groups with the highest/lowest error are also the same for each LLM. In figure 4, 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 errors, 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, with the high income group having lowest error for 95% of LLMs and the low income group having highest error for 90% of LLMs. In appendix B, we inspect performance per-indicator, again finding consistent and pervasive disparities.

4 Noteworthy Observations from Additinoal Analyses

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". Overall, *responses with "citations" were no more accurate than those without "citations"*, still incurring substantial mean absolute relative errors. Second, because we have data per-country *per-year*, we could compute error rates while selecting groundtruths from specific years. For 13 out of 20 LLMs, lowest error occurs when comparing to groundtruths from 2021, suggesting that some LLMs may already be slightly out of date. See appendix G for details.

5 Discussion

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. Even if LLMs may never ace this task (due to the volatility of some indicators and the challenge of memorizing numbers), we can believe our benchmark can offer valuable signal in measuring geographic disparities. Moreover, we hope WorldBench can aide in reducing those disparities in future generations of LLMs, towards models that work well *for all*.

REFERENCES

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. Do language models know when they're hallucinating references?, 2023.

Cohere AI. Cohere api, 2023. URL https://github.com/cohere-ai/cohere-python.

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. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

World Bank. World bank country and lending groups, 2024a. URL https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups#:~:text=For%20the%20current%202024%20fiscal,those%20with%20a%20GNI%20per.

World Bank. World bank open data, 2024b. URL https://data.worldbank.org/.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson (eds.), *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 2018a. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 23–24 Feb 2018b. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone?, 2019.

Laura Gustafson, Megan Richards, Melissa Hall, Caner Hazirbas, Diane Bouchacourt, and Mark Ibrahim. Pinpointing why object recognition performance degrades across income levels and geographies, 2023.

Melissa Hall, Candace Ross, Adina Williams, Nicolas Carion, Michal Drozdzal, and Adriana Romero-Soriano. DIG in: Evaluating disparities in image generations with indicators for geographic diversity. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=FDt2UGM1Nz. Featured Certification.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023.

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. Mistral 7b, 2023.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge, 2023.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations, 2023.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 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.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories, 2023.

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. Orca 2: Teaching small language models how to reason, 2023.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *ArXiv*, abs/2303.13375, 2023. URL https://api.semanticscholar.org/CorpusID:257687695.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David I. Adelani. How good are large language models on african languages?, 2023.

OpenAI. Openai api, 2023. URL https://openai.com/product.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. Head-to-tail: How knowledgeable are large language models (llm)? a.k.a. will llms replace knowledge graphs?, 2023.

Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F. Rousseau, and Greg Durrett. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors, 2023.

Gemini Team. Gemini: A family of highly capable multimodal models, 2023.

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. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. URL https://api.semanticscholar.org/CorpusID:259950998.

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. Zephyr: Direct distillation of lm alignment, 2023.

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. Huggingface's transformers: State-of-the-art natural language processing, 2020.

Yifan Zhang, Cheng Wei, Shangyou Wu, Zhengting He, and Wenhao Yu. Geogpt: Understanding and processing geospatial tasks through an autonomous gpt. *ArXiv*, abs/2307.07930, 2023. URL https://api.semanticscholar.org/CorpusID:259937048.

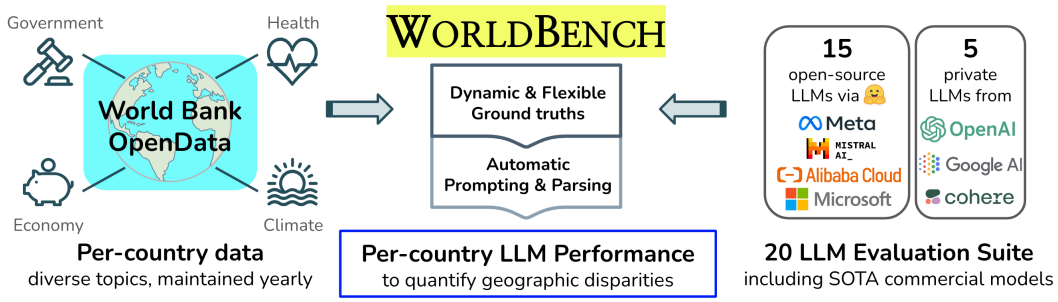

Figure 5: Overview of WORLDBENCH. Our benchmark provides a manner to quantify the performance of large language models (LLMs) on a per-country basis. We disentangle data collection from evaluation by utilizing the World Bank's data bank, which contains indicators pertaining to numerous diverse aspects 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.

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: Indicators in WORLDBENCH. Each query corresponds to a World Bank defined and maintained global development indicator.

A BENCHMARK DETAILS

We now present full details on our benchmark. Figure 5 provides an overview of our study as a whole.

A.1 DATA

WORLDBENCH 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 Bank (2024b). 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. 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 grountruth data per indicator studied.

Country categorization. The World Bank also provides various categorizations of countries, based on geographic or economic reasons Bank (2024a). We focus on two high level categorizations, visualized in Figure 6, 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

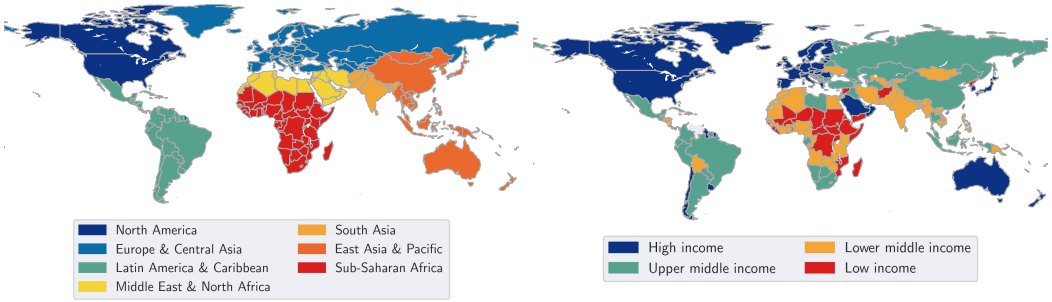

Figure 6: World Bank categorizations of countries into 7 regions (top) and 4 income groups (bottom).

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.

A.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 7 displays the base instruction and example. We always use Switzerland as the example country and exclude it from our analysis, given its history of neutrality[1]. 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 include other text, such as special tokens, or also undesirable behavior, like 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

[1]This is a joke - the choice is arbitrary, and we chose Switzerland, as it had data for all indicators we studied. We confirm that results are similar when using alternate example countries in Appendix F

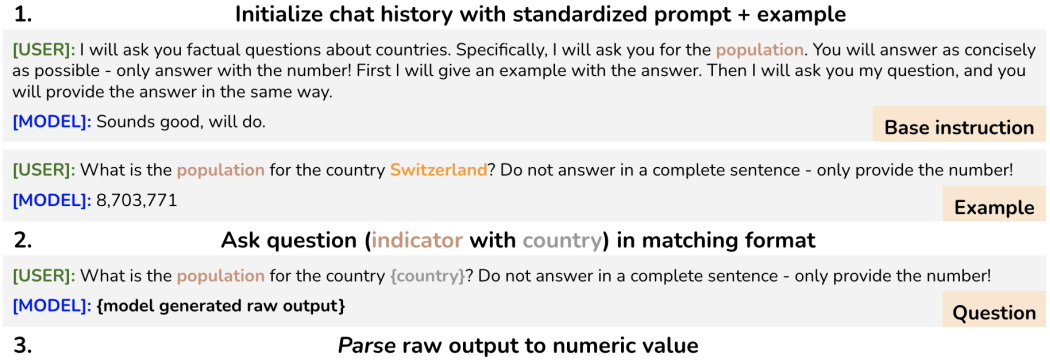

Figure 7: Standard pipeline for extracting numeric answers from LLMs. Each question is defined by a query (i.e. Before asking a language model a question, we prompt it with a base instruction and example. Then, we automatically parse the raw output to obtain a numeric value which can be compared to the groundtruth data.

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 (a total of $44.5k$ questions), we retrieve a numeric answer for 88.9% of the questions we ask. We provide further validation of 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, our parsing is mostly complete, as **we obtain a numeric 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**. 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 D, 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 E, 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 G.2, to gain insight on if LLM responses are dated (i.e. more accurate for a prior year than the most recent year).

A.3 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 Wolf et al. (2020) to obtain and operate 15 models (and respective tokenizers). Namely, from Meta's LLama-2 Touvron et al.

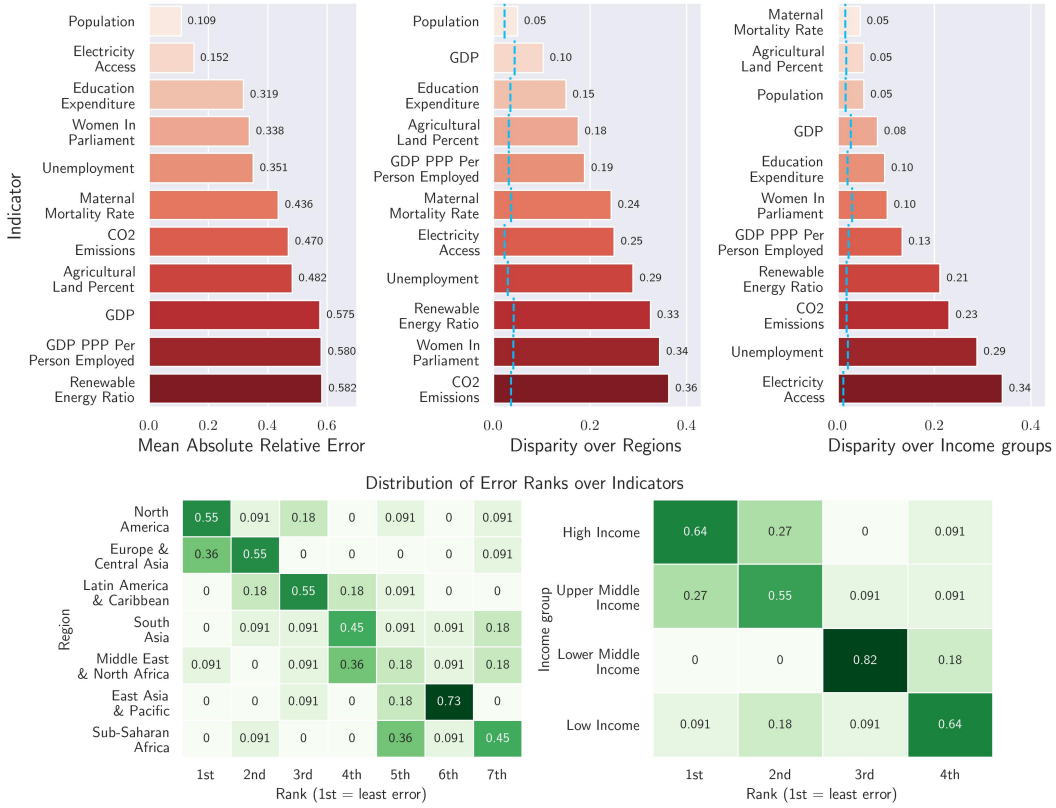

Figure 8: (**top**) Error rates and disparities per indicator, averaged over LLMs. For disparities, the blue dashed lines correspond to the disparity incurred using a random categorization of countries (into 7 groups for Regions and 4 for Income groups), averaged over 10 trials. (**bottom**) Order of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

(2023), 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 Mitra et al. (2023), 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 Jiang et al. (2023). We also study Zephyr-$7B$ β Tunstall et al. (2023), tuned from a Mistral-AI model. Lastly, we include $7B$ and $14B$ Qwen models from Alibaba Cloud, both with and without chat-tuning Bai et al. (2023). For **closed source models**, we include the following LLMs. From OpenAI, we evaluate gpt-3.5-turbo and gpt-4 OpenAI (2023). From Google, we evalute Gemini Team (2023). From Cohere, we evaluate the 'command' model, as well as the same model equipped with retrieval augmented generation (RAG) AI (2023). 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.

B Results per Indicator

Figure 8 shows errors and disparities per indicator, as well as the distribution of ranks per region/income group over indicators. Mean absolute relative error exceeds 0.3 for all but two of the indicators. Similar to the per-LLM invesitgation, disparities are present for most cases, though they are more pronounced across Regions 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 C – the section below this one.

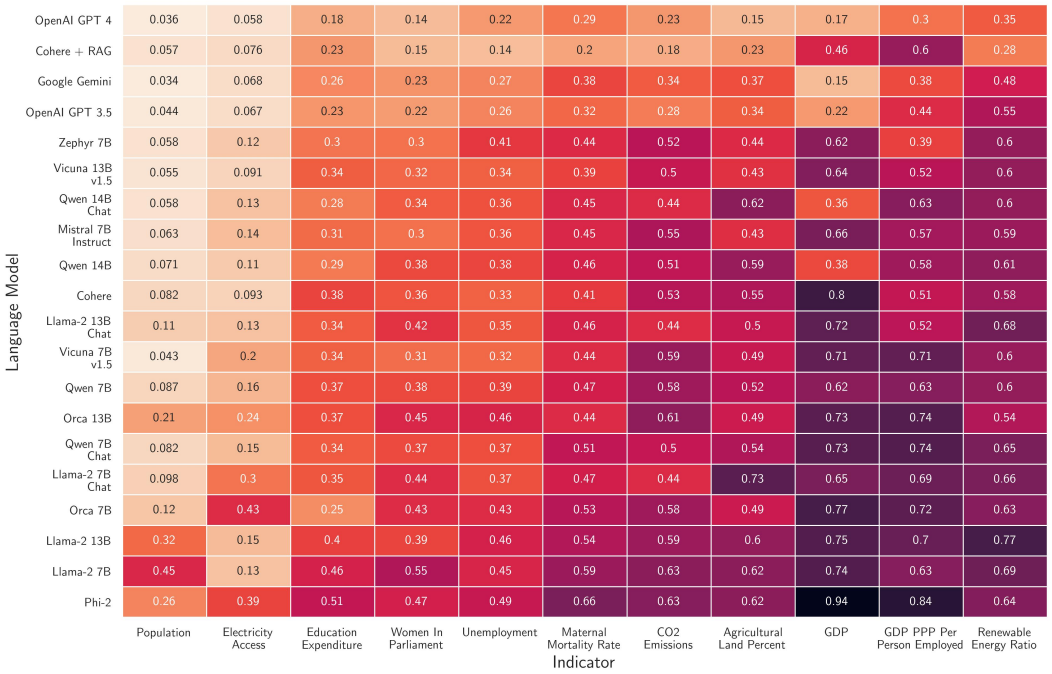

Figure 9: Absolute relative error averaged over countries per LLM and Indicator. Language models and indicators are each sorted by overall average error respectively.

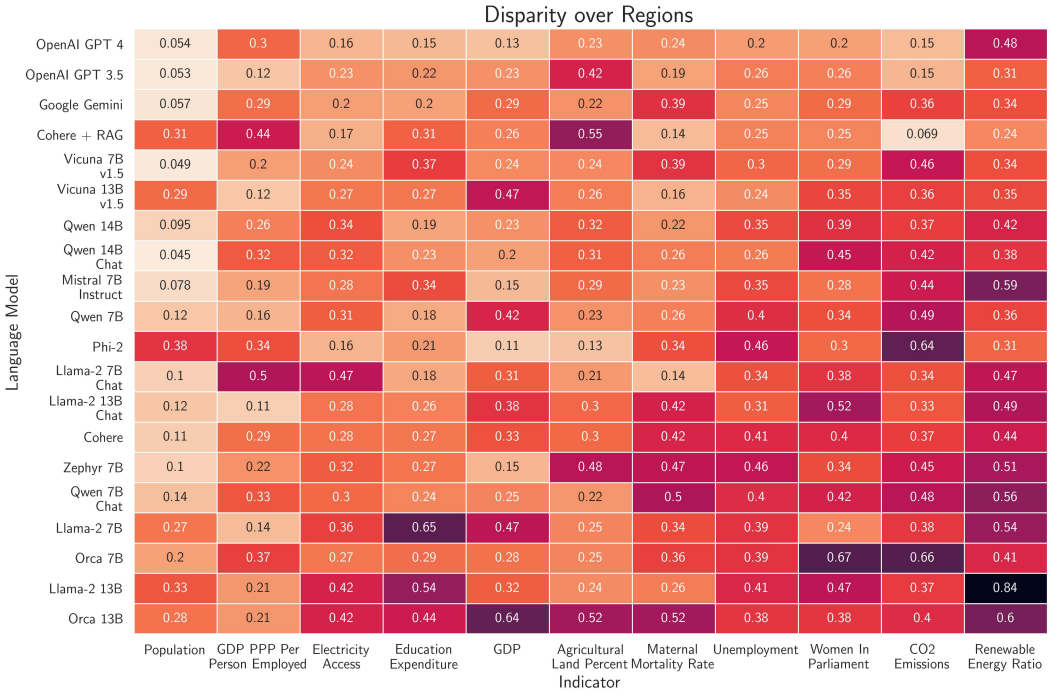

Figure 10: Disparities over Regions per LLM and Indicator. Language models and indicators are each sorted by overall average error respectively.

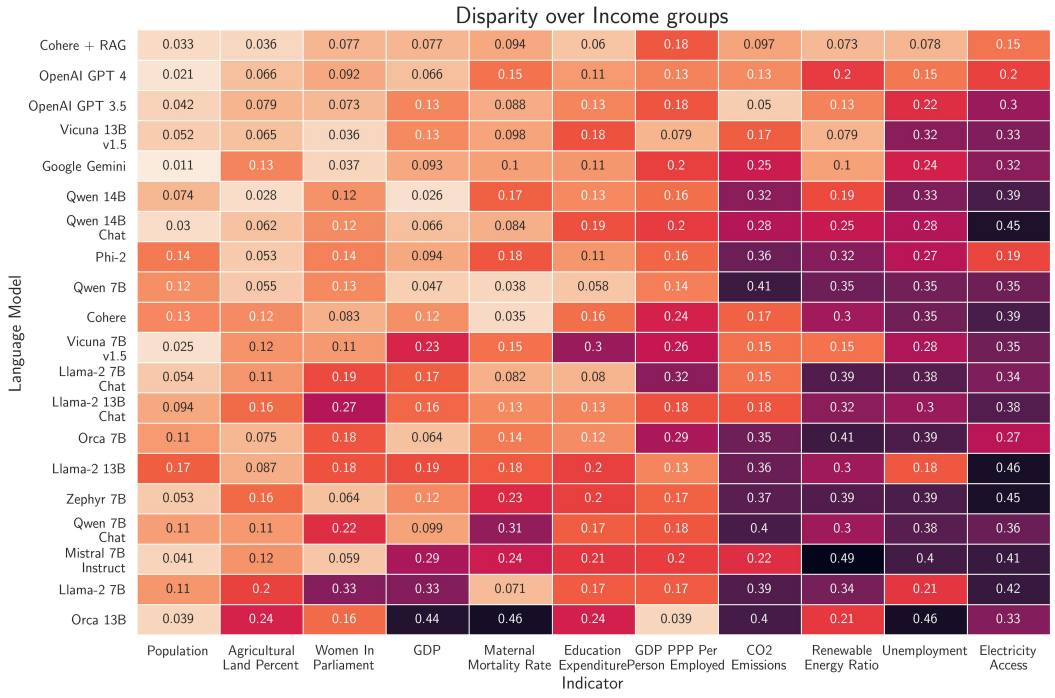

Figure 11: Disparities over Regions per LLM and Indicator. Language models and indicators are each sorted by overall average error respectively.

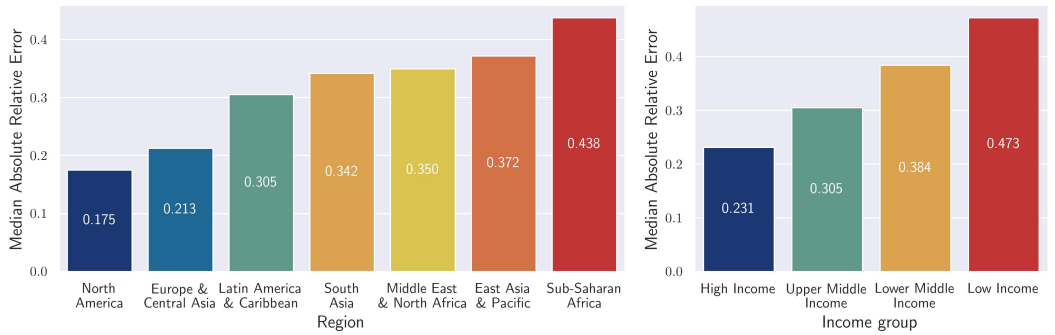

Figure 12: Median absolute relative error per region and income group. See figure 2 for mean errors.

C Complete Results Breakdown

We now present the results as completely as possible. In Figure 9, we present mean absolute relative error per LLM per indicator. In Figure 10, we present disparities over regions per LLM per indicator, and in Figure 11 we show the same for disparities over income groups. In general, the indicators that are most challenging are challenging for all LLMs.

D Larger Disparities when using Median instead of Mean Error

We now present results when aggregating with median instead of mean. Figure 12 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.

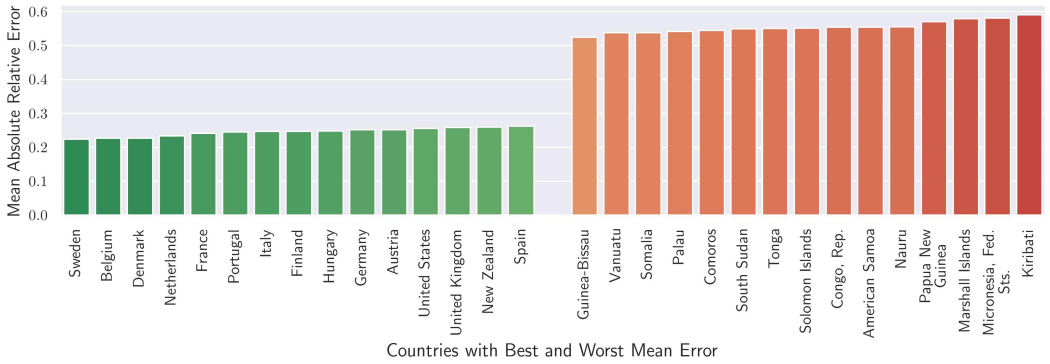

Figure 13: Error rates can vary significantly across countries, with some countries experiencing nearly $3\times$ higher absolute relative error than others. Strikingly, all of the 15 countries with the lowest error rates fall in the high income category, while all of the 15 countries with the highest error rates fall in the low income category.

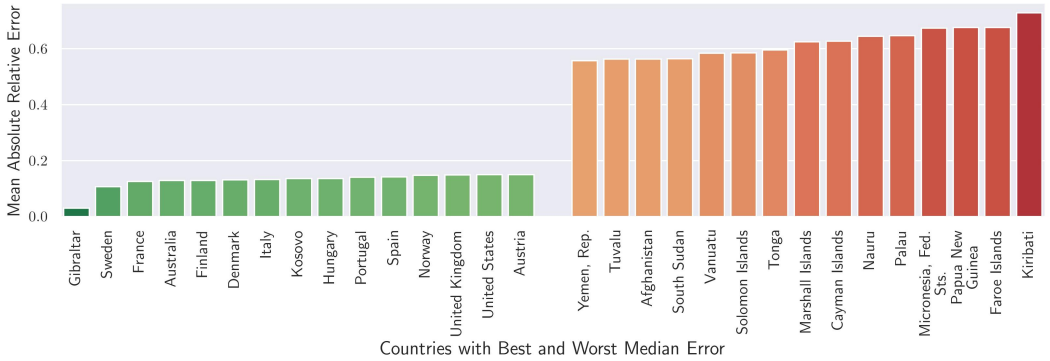

Figure 14: Median absolute relative error per country for countries with most and least median error. Again, the countries with least error belong to Western regions and the high income category.

E Alternate Groundtruth Selection Strategies

E.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 2 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 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 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.

E.2 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 3 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.

Indicator	Specified Year for Alternate Groundtruth					
	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 2: 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.

F 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 4 shows the results. Again, main trends are consistent, with Western and High income countries 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.

G Additional Analyses

G.1 Citation Hallucination

Despite our prompting to only return a numeric value, often additional text was still produced by the LLMs studied. Interestingly, sometimes, generated text would resemble a citation[2], 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 15 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 few examples, the

[2]Such behavior has been observed in Agrawal et al. (2023)

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 3: 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.

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 4: 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.

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

LLM answer is way off, despite the arguably convincing "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 (yielding a relative error of about 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 value encoded 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.

G.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 16 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. 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. 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.

H Related Work

Evaluating Factual Recall. Recent works have documented the performance of LLMs in factual recall: Mallen et al. (2023), Kandpal et al. (2023), Sun et al. (2023), Tang et al. (2023). 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 Huang et al. (2023). Zhang et al. (2023) specifically investigated the recall of geographic information, though their study is limited to GPT-4 and does not inspect disparities. Some works (e.g., Mallen et al. (2023), Sun et al. (2023)) 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.

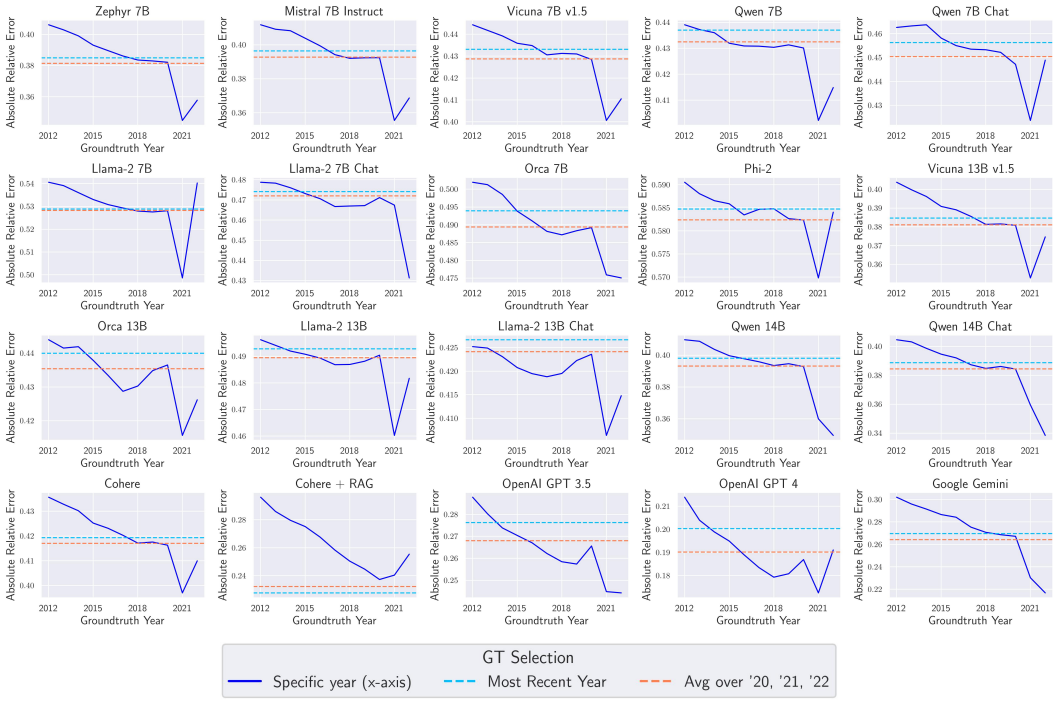

Figure 16: Error rate of LLM outputs compared to year from which groundtruth is extracted. Many models show the lowest error rate when their outputs are compared to groundtruths from 2021, indicating that models may already be slightly out of date.

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 Buolamwini & Gebru (2018a); Gustafson et al. (2023); DeVries et al. (2019); Hall et al. (2024). In the realm of language processing, Ojo et al. (2023) 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 recalls, which in turn enables inspection of disparities across regions and income groups.

Benchmark. 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) the categories inputs are grouped by are salient, and (iii) the groundtruth answers are accurate and up-to-date.