

# Evaluating Diverse LLMs: How GDP Influences Responses to Gender and Diversity Questions

Alex Kaplunovich  
 Department of Computer Science  
 University of Maryland  
 Baltimore, MD 21250, USA  
 akaplun1@umbc.edu

**Abstract**—Generative large language models, like ChatGPT, are playing an increasingly important role in various aspects of our lives. They're democratizing machine learning and impacting an expanding segment of the population. Motivated by this, we embarked on a study, interviewing numerous open-source and OpenAI models. Our focus was on topics like diversity, inclusion, human rights, gender, LGBT leadership, and history to assess LLMs proficiency on these crucial subjects. Additionally, we investigated if there's a relationship between a country's GDP per capita and quality of the model's responses. For our experiments, we used the Huggingface model leaderboard to cherry-pick a diverse range of models and deploy them into AWS SageMaker GPU environment. We presented the same set of questions about nearly 200 countries and subsequently verified the responses for their accuracy and correlation to Gross Domestic Product (GDP). The diversity, quality, and quantity of the answers from these pretrained open-source LLMs were astonishing, especially given the complexity and sensitivity of the questions. Our exploration yielded valuable insights about model selection, streamlining the inference pipeline, GPU prerequisites and setups, and the depth of knowledge ingrained in models due to their training data. Finally, we discovered very peculiar correlation between generated answers and a country's GDP.

**Keywords**—LLM, Generative AI, GDP per capita, AWS Cloud, Huggingface, Diversity and Inclusion, ChatGPT

## I. INTRODUCTION

Generative AI has emerged as a center of innovation, drawing top-tier talent from academia and the industry. As a result, new pre-trained large language models (LLMs) are unveiled almost weekly. While some, like OpenAI's ChatGPT and Anthropic's Claude 2, operate on subscription models, incurring fees for invocation or fine-tuning, a plethora of open-source alternatives are available for free. Notably, these open-source LLMs often harness the standard Hugging Face API, enhancing their accessibility for researchers and AI enthusiasts. Some, like Falcon, even boast an Apache 2.0 license [1], permitting commercial use.

Although many researchers are utilizing LLMs, very few of them evaluate handling diversity and inclusion topics. We decided to investigate how well generative AI can handle these important topics.

Without platforms like the LLM Leaderboard [2] provided by Hugging Face, navigating the plethora of available models

would be difficult. This leaderboard presents a select list of models, ranked based on their performance across benchmarks: the AI2 Reasoning Challenge (25-shot), HellaSwag (10-shot), MMLU (5-shot), and TruthfulQA (0-shot). Evaluated on the Language Model Evaluation Harness [3], these benchmarks produce average scores spanning from 25 to 74.

Our study aimed to meticulously assess a range of open-source models, gauging them on performance and accuracy metrics. For comparison, we also integrated OpenAI's gpt-3.5-turbo, a leading model with 175 billion parameters. Publicized as the "most capable and cost-effective model in the GPT-3.5 family," [4] this commercial model has a pricing metric of \$0.0015 for 1K input tokens and \$0.002 for 1K output tokens.

GPT-3.5, released in November 2022 [5], revolutionized the AI domain. The avalanche of interest it garnered from both the scientific realm and the wider audience was unprecedented. Its ability in language comprehension and generation propelled it to iconic stature, making it a global focal point for research and imitation. It stands out for its ability to tackle intricate questions and untangle convoluted logical puzzles.

In the ever-evolving AI field, a rich variety of free LLMs presents itself as formidable contenders. They empower the AI community to generate outputs mirroring those of their commercial counterparts. In choosing our model pool, we aimed for diversity, selecting models with benchmark scores between 37 and 73, offering insights into the world of generative AI and the subtle distinctions between various LLMs.

## II. LARGE LANGUAGE MODELS

By 2023, the realm of generative models has transformed into a fiercely competitive landscape. Researchers and businesses must meticulously choose models that resonate with their unique requirements, considering elements like budget, technical specifications, and domain knowledge.

### A. Open source vs. Commercial models

The distinction between open-source and commercial language models is most apparent in the control and adaptability they afford. Open-source models provide developers with unlimited access to model weights, autonomy in deployment decisions, and the latitude to refine the inference framework. Additionally, tools such as Huggingface's Trainer and Pipeline

interfaces ([7] and [8]) provide methods for working with LLMs.

It is imperative to acknowledge the cost and rate constraints tied to the commercial LLMs. For example, gpt-3.5-turbo premium subscribers face a ceiling of twenty requests per minute in the initial two days [12]. Conversely, when you deploy models on proprietary infrastructure, you gain unmitigated control over the system's architecture, cost, scalability, data flow, and user interactions.

### B. Huggingface, memory and finetuning

The field of large language models has seen groundbreaking advancements, and at the forefront of this movement is Huggingface [7] and [8], renowned for its consistent and dependable interface. This platform has garnered substantial accolades for its resilience and reliability within the Data Science fraternity.

Popular models, such as Llama2\_70B [14], are enormous in size, often exceeding 100GB on disk storage. These dimensions introduce formidable challenges, particularly concerning generation times and memory constraints, both on RAM and GPU, during the inference phase. Nevertheless, the relentless pursuit of the research community has given birth to ingenious finetuning methodologies. A slew of memory-efficient strategies have emerged, facilitating the execution of these gargantuan tasks even on a singular, modestly-sized GPU.

TABLE I. LLM MODELS SELECTED FOR OUR EXPERIMENTS

<i>Model</i>	<i>Parameters (billions)</i>	<i>Disk Size (GB)</i>	<i>Leader Board average score</i>	<i>Hugging Face month downloads</i>
gpt-3.5-turbo	176	n/a	n/a	n/a
Llama-2-70b-instruct-v2	70	140	72.95	99,505
airoboros-c34b	34	70	58.09	3,321
llama-30b-instruct-2048	30	65	67.02	4,965
VicUnlocked-alpaca-30b	30	66	63.58	124
starchat-beta	16	32	55.78	76,423
gpt4-alpaca-lora-13B-HF	13	27	59.52	3,370
open llama 13b	13	26	52.06	65,323
dolly-v2-7b	7	14	43.56	12,806
falcon-7b-instruct	7	15	47.01	5,714,177
AlpacaGPT4-7B-elina	7	27	53.21	100
eachadea/vicuna-7b-1.1	7	14	56.43	6,445
Mistral-7B-Instruct-v0.1	7	15	60.45	144,533
mpt-7b-chat	7	13	49.95	17,392
gpt2-xl	1.5	6.4	36.66	304,828

### C. Selecting models

After careful consideration, we have picked the following representative subset of open source models (Table I). The models in the tables are sorted by number of parameters. Besides using a commercial gpt-3.5-turbo, we selected various models from 1.5 to 70 billion parameters, whose average leaderboard

score [2] ranges from 36.66 to 72.95. These models include Gpt2, Falcon, Dolly, Mistral, Alpaca and Llama2.

## III. GDP DATA, GROUPS AND QUESTIONS

### A. Grouping the nations

For a comprehensive comparison of countries, we began by obtaining a dataset on GDP per capita from the World Bank [6]. This dataset furnishes annual figures for approximately two hundred countries. Though the dataset predominantly showcases GDP per capita in numerical terms, there's a standardized method [13] for segmenting countries into four specific groups based on this metric. The categorization criteria for 2021 are detailed in Table II.

TABLE II. GDP GROUPING RANGES

<i>Group</i>	<i>GDP per capita range for 2021</i>
Low income	\$0 - \$1,045
Lower-middle income	\$1,046 - \$4,095
Upper-middle income	\$4,096 - \$12,695
High income	> \$12,695

### B. Interview Questions

Before any important interview, it's good to prepare a list of questions to properly test the candidates. We did the same when making questions for our LLMs. We wanted questions that checked the models' knowledge about different topics on diversity and inclusion for countries around the world. We were curious to see what these models knew about each country in the world and if their answers changed based on how wealthy a country was.

TABLE III. QUESTIONS AND THEIR SUBJECTS

<i>Subject</i>	<i>Question</i>
Civil Activists	Give me top 10 civil rights activists in {name}.
GLBT Activists	Give me top 10 gay and lesbian activists in {name}.
LGBT History	Tell me the LGBT history of {name}.
Race History	Tell me race and minority history of {name}.
Gender Distribution	What is the gender distribution in tech industry in {name}.

We also wanted to see if there was a link between a country's wealth (GDP) and how detailed the models' answers were. Table III lists the five questions we used, with {name} being a spot where we put the country's name.

## IV. AUTOMATION AND ANALYSIS FRAMEWORK

For consistent and systematic research execution, we designed multiple Jupyter notebooks. These notebooks perform inference for all countries and store the outcomes. We relied on AWS Sagemaker for a uniform experimental process. Additionally, whenever a new data file is added to the S3 bucket, we employed serverless Lambda event-driven programming to automate our workflow, following the approach recommended in [19].

### A. Environment configuration

To bolster the reliability of our framework and guarantee consistent replication of our experiments, we've incorporated

AWS Lifecycle management [16] into our notebook setup. This setup aligns seamlessly with AWS SageMaker notebooks and involves several key steps, outlined below:

- Install python libraries, configure instance to terminate if idle for certain interval, and mount large cloud volumes to the instance filesystem

Given that the lifecycle is a Python script, it allows us the flexibility to further customize our notebooks. Ensuring ample disk space on our notebook instance is vital, especially when accommodating multiple models that can each span up to 100GB for our experiments.

## V. GENERATED TEXT ANALYSIS AND CORRELATIONS

Let us clarify the data we will be analyzing. We have split our generated datasets into **5 subjects** – Civil Activists, GLBT Activists, LGBT History, Race History and Gender Distribution—each corresponding to one question from Table III. Each question was asked **196 times for each country**, substituting its name in the question instead of literal “{name}”. The experiment was conducted for each of the selected **15 models** from Table I. In total, we have generated  $5 \times 196 \times 15 = 14,700$  texts. Several of them (Civil Activists and GLBT Activists) supposedly should give an itemized list of ten answers (for example, answer to a question “Give me top 10 civil rights activists in France.”).

Throughout this research, we created an automated pipeline designed for juxtaposing LLM predictions using advanced text processing and cutting-edge metrics, details of which we'll discuss subsequently.

### A. Data preprocessing

Every model is a vast standalone system. For instance, the largest model we selected, Llama-2-70b-instruct-v2, occupies 140GB on disk and houses seventy billion parameters. While we're aware that LLMs can answer queries, their inner workings resemble a black box. These models undergo training on billions of texts and might be further finetuned using vast datasets. However, at the end of the day, we cannot pinpoint the exact books, websites, or human-generated data that trained each model, nor can we definitively determine the knowledge obtained from this training data.

Furthermore, many of the models prepend the answer with the question. To ensure a fair comparison, it is essential to preprocess the model responses by filtering out extraneous lines and retaining only the core answers. We created regular expressions to segregate to accomplish that.

### B. Text distance calculations

#### 1) Cosine similarity

Cosine similarity is a metric used to determine how similar two vectors are. Its return values can be interpreted as follows:

0: the vectors are orthogonal (not similar).

1: the vectors are identical.

Between 0 and 1: The value indicates the degree of similarity between the vectors.

To apply cosine similarity, we need to embed (vectorize) texts using BERT sentence transformers library [18] that converts an arbitrary answer into a vector of numbers.

#### 2) Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score

ROUGE [17] is a set of evaluation metrics designed to assess the quality of summaries by comparing them to reference summaries. There is a *rouge\_score* library that can easily calculate the score. Its return values can be interpreted as follows:

0: no match

1: perfect match

Between 0 and 1: A ROUGE-N score of 0.50 or above is generally considered good similarity

In our distance metrics, a higher numerical value indicates a closer match. This intuitive system facilitates straightforward comparisons, where the greatest value denotes the optimal result.

### C. Comparative Analysis with ChatGPT

ChatGPT stands out as a benchmark in the generative AI arena. We assess the efficacy of our chosen open-source models against gpt-3.5-turbo, employing the metrics detailed in Section V-B. Subsequently, we calculate average scores for all responses and also for each GDP group of countries. This layered approach aids in making a detailed comparison and discerning any potential correlations. Through this evaluation, we gauge how the responses from open-source LLMs measure up against those produced by OpenAI's flagship model. Essentially, it is like an interview where we are selecting the best performing candidates (models).

For single answer questions (subject “LGBT History” in Table III) we apply the metrics from section V-B (cosine similarity and ROUGE score) to cleaned texts.

For the questions containing multiple answers (“Give me top 10 ...”), we split cleaned answer into parts using regular expressions, and apply the above-mentioned metrics to single responses picking the highest scored pairs and adding them.

In each scenario, we quantify the comparison of answers, resulting in a numerical value that gauges the similarity between two responses. By calculating the average score for all responses from a particular LLM, we derive a unified performance metric for that model. With our metrics' design, a higher value signifies a closer resemblance between responses.

### D. Benchmarking Model Responses: A Confusion Matrix Approach

We saw the importance of comparing models against each other. Our goal is to ascertain the alignment in responses between each model pair. As a result, we crafted a confusion matrix, incorporating the previously mentioned metrics. This matrix offers a lens into the similarity of predictions across models.

The matrix's significance is crucial given the nature of large language models, which frequently function as 'black boxes.' Their inner workings often remain obscured, and the comprehensive training and finetuning datasets are rarely made public. This opacity complicates tracing back the roots of specific predictions or responses.

## VI. RESULTS

We looked into our data using several methods, like manual checks and automated tools. The amount of information we request from the models is enormous – information about human rights, LGBT activists and history for every country in the world. Going through the data was both interesting and insightful.

### A. Manual evaluation

The initial step entails a manual review of the answers, serving as a sanity check to gauge the general quality of the model's output. During this process, we encountered a few unique errors and noteworthy responses worth highlighting.

#### 1) Errors and Warnings

Every model has its limitations, what is important that the errors should be handled gracefully, hopefully without hallucinations to deceive a user. We found the following interesting use cases

- dolly-v2-7b answers very well for some countries providing full 10 items list or detailed history, while can give meaningless texts for some other countries
- falcon-7b-instruct can repeat the same name in a list multiple times.
- gpt-3.5-turbo gives even more comments like “As an AI language model, I can provide information on prominent civil rights activists in ... Please note that this list is by no means exhaustive, and the impact of these activists may vary depending on the context and the specific issues they focus on.”.
- gpt2-xl gives a variety of answers from “Read Full Article Here”, to valid lists of 10 people, or sometimes even details about a leader like “Martin "Mousa" Moussali-Titou was shot and killed as he stood up to take control of his community in the Central African Republic.”.

#### 2) Reviewing results

We were able to see very impressive and detailed answers to difficult questions. Let us look at the question:

“Give me top 10 civil rights activists in Madagascar”.

vicuna-7b answers:

1. Andry Rajoelina
2. Manandafy Rakotonirina
3. Marcellin Tsynda
4. Jean-Michel Ravelomanana
5. Fanomezana Rakotomalala
6. Flavien Rakotomalala
7. Elie **Rajaonarivelo**
8. Fanjolahy **Roland Ratsiraka**
9. **Monja Roindefo**
10. Jean-Joseph Rabearivelo

and llama-30b-instruct-2048 replies:

1. Jean-Luc Raharisoa
2. Lova Razafimandimby
3. **Monja Roindefo**

4. Pasteur Bazalais
5. Albert Zafy
6. Didier Ratsiraka
7. **Roland Ratsiraka**
8. Pierrot **Rajaonarivelo**
9. Herizo Razafimahaleo
10. Zavé Aznar

Reviewing thousands of answers manually would be both tedious and time-intensive. By automating the process and converting the results into a digital and visual format, we have streamlined our workflow. This allows us to assess our models effectively using just numerical metrics.

### B. Automatic evaluation

We have developed a reproducible automatic framework to compare model performance and results digitizing and visualizing various answers for selected models.

#### 1) Basic LLM Correlations

We have analyzed how generation time, length of answer and GDP are correlated in Fig. 1. It looks like for all the models time linearly correlates with length of text. It makes sense – the longer generated text, the longer time it takes to generate.

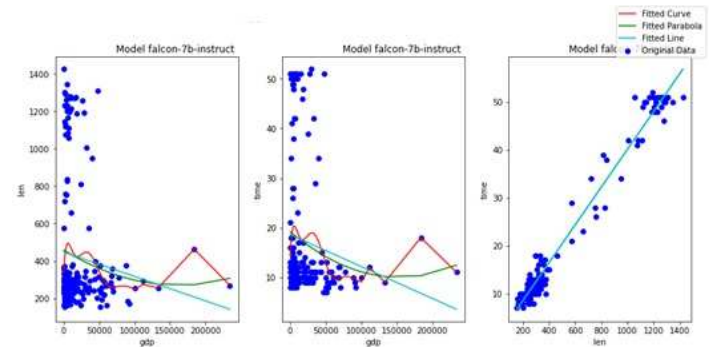


Fig. 1. GDP, time and length correlations for falcon-7b-instruct.

#### 2) Answer Comparison against ChatGPT

How do we evaluate thousands of answers? What constitutes a benchmark? To address these queries, we chose to measure model responses against the renowned OpenAI gpt-3.5-turbo model. The metric values will shed light on the extent to which the models' outputs are in sync with ChatGPT's.

Tables IV and V display average metric results for each question topic (refer to Table III for the questions). The top three scores are highlighted in red for easy identification of the front-runners.

It's clear that a model's effectiveness can fluctuate based on the subject matter. For instance, while VicUnlocked-alpaca-30b and Llama-2-70b-instruct-v2 consistently shine in various categories, others like mpt-7b-chat or starchat-beta stand out in only specific domains.

Additionally, we noticed pronounced differences in similarity scores depending on the subject. Take the "Civil Activists" category: the average cosine similarity score stands at 2.4, reaching a high of 4.41. In contrast, the "GLBT Activists" category averages at just 2.03, peaking at 3.56. Such variations underscore that these models have been trained on distinct datasets, each emphasizing different areas.

TABLE IV. MODEL ANSWERS COMPARED TO GPT-3.5-TURBO (CHATGPT) – CIVIL ACTIVISTS AND GLBT ACTIVISTS

Model	Civil Activists		GLBT Activists	
	Cosine Similarity	ROUGE Score	Cosine Similarity	ROUGE Score
airoboros-c34b-2.1	3.10	0.28	2.01	0.35
AlpacaGPT4-7B-elina	1.28	0.71	1.50	0.78
dolly-v2-7b	1.25	0.19	1.49	0.32
falcon-7b-instruct	3.42	0.22	2.76	0.13
gpt2-xl	0.38	0.15	0.51	0.23
gpt4-alpaca-lora-13B-HF	0.73	0.12	0.43	0.22
Llama-2-70b-instruct-v2	4.41	0.95	2.45	1.16
llama-30b-instruct-2048	3.49	0.69	2.23	1.23
Mistral-7B-Instruct-v0.1-GGUF	1.78	0.58	2.38	0.98
mpt-7b-chat	3.58	0.96	3.56	1.50
open_llama_13b	1.85	0.60	0.76	0.35
starchat-beta	1.91	0.99	2.22	1.39
vicuna-7b	2.87	0.69	3.35	1.01
VicUnlocked-alpaca-30b	3.53	0.29	2.83	0.31

TABLE V. MODEL ANSWERS COMPARED TO GPT-3.5-TURBO (CHATGPT) – LGBT HISTORY AND GENDER DISTRIBUTION

Model	LGBT History		Gender Distribution	
	Cosine Similarity	ROUGE Score	Cosine Similarity	ROUGE Score
airoboros-c34b-2.1	0.79	0.53	0.68	0.51
AlpacaGPT4-7B-elina	0.57	0.40	0.61	0.30
dolly-v2-7b	0.69	0.37	0.73	0.29
falcon-7b-instruct	0.76	0.59	0.59	0.46
gpt2-xl	0.53	0.38	0.57	0.27
gpt4-alpaca-lora-13B-HF	0.52	0.36	0.67	0.33
Llama-2-70b-instruct-v2	0.56	0.44	0.79	0.47
llama-30b-instruct-2048	0.74	0.58	0.82	0.53
Mistral-7B-Instruct-v0.1	0.77	0.54	0.71	0.50
mpt-7b-chat	0.75	0.50	0.67	0.34
open_llama_13b	0.59	0.30	0.58	0.25
starchat-beta	0.70	0.42	0.38	0.10
vicuna-7b	0.60	0.41	0.68	0.34
VicUnlocked-alpaca-30b	0.84	0.63	0.82	0.57

Table VI shows the average scores across all subjects, offering a clear view of the model boasting the highest overall average. We can see that mpt-7b-chat and Llama-2-70b-

Instruct-V2 emerge as the models most in sync with ChatGPT's outputs.

TABLE VI. MODEL ANSWERS COMPARED TO GPT-3.5-TURBO (CHATGPT) – AVERAGE FOR ALL THE SUBJECTS

Model	Average for all subjects	
	Cosine Similarity	ROUGE Score
airoboros-c34b-2.1	1.46	0.44
AlpacaGPT4-7B-elina	0.93	0.53
dolly-v2-7b	0.98	0.33
falcon-7b-instruct	1.67	0.42
gpt2-xl	0.50	0.28
gpt4-alpaca-lora-13B-HF	0.61	0.30
Llama-2-70b-instruct-v2	1.81	0.73
llama-30b-instruct-2048	1.63	0.74
Mistral-7B-Instruct-v0.1	1.28	0.63
mpt-7b-chat	1.87	0.77
open_llama_13b	0.86	0.36
starchat-beta	1.20	0.68
vicuna-7b	1.65	0.60
VicUnlocked-alpaca-30b	1.77	0.49

### 3) Comparing model answers and Visualisation

In our previous section, we looked at how different models stack up against gpt-3.5-turbo. We went one step further and developed an automated comparison pipeline to evaluate the predictions across all possible pairs of our selected models.

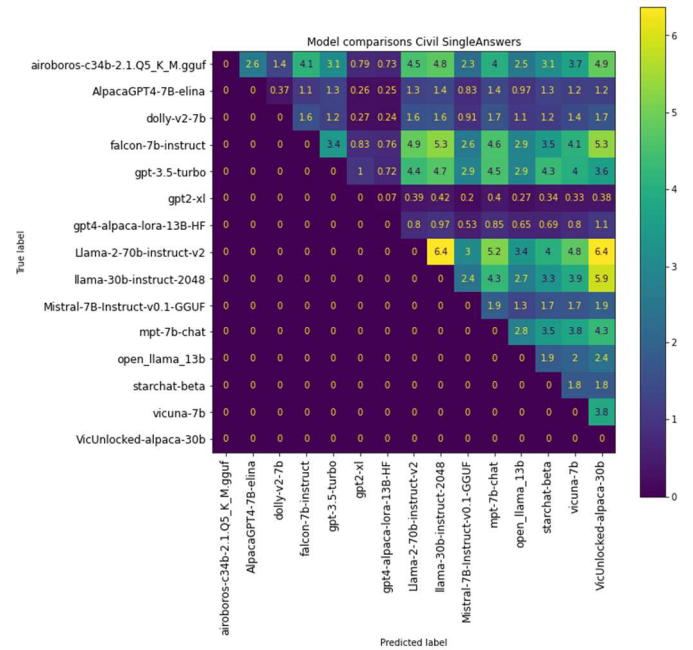


Fig. 2. Heatmap of pairwise model predictions comparison for Civil Activists using the average cosine similarity metric.

Using a confusion matrix approach, we made matrices that show how similar each model's predictions are to the others. A high value means two models predict alike. We turned these matrices into heatmaps: very similar predictions are yellow, while different ones are purple (Fig. 2).

We have taken our analysis a step further by introducing subject scores for each individual model. This scoring method assigns a "wellness" score to each large language model, providing an interpretative measure of how closely model predictions align with those of its counterparts. Our findings are illustrated in Table VII, where the top three open-source models are highlighted in red and ChatGPT scores are marked in blue.

Our findings highlight falcon-7b-instruct, llama-30b-instruct-2048, Llama-2-70b-instruct-v2, and VicUnlocked-alpaca-30b as the top-performing open-source models. Impressively, these models either rival or surpass the performance of the leading ChatGPT model, gpt-3.5-turbo. It's commendable that falcon-7b-instruct, with its seven billion parameters, can produce such stellar results while being efficient enough to operate on a single GPU. This efficiency reduces the dependency on high-end hardware.

Table VIII reveals that the best-performing model shifts based on the question's subject. No single LLM dominates in every area. Yet, there is a pattern: the same 3-4 models often top the charts, while another distinct group of 3-4 models usually ranks lower. There are exceptions; for example, Llama-2-70b-instruct-v2 is top-tier for the "Civil Activists" category but drops to 8th for "GLBT Activists". Such fluctuations highlight the need to assess each model based on specific topics. It is a reminder not to solely lean on metrics like parameter count, model size, or even a model's brand as the sole predictors of its effectiveness.

TABLE VII. TOTAL OF SCORES AMONG ALL MODELS FOR ALL SUBJECTS

Model	Overall score	
	Cosine Similarity	ROUGE Score
airoboros-c34b-2.1	19.068	6.116
AlpacaGPT4-7B-elina	11.586	6.13
dolly-v2-7b	12.66	4.384
falcon-7b-instruct	21.502	6.92
gpt-3.5-turbo	21.946	10.384
gpt2-xl	6.788	2.886
gpt4-alpaca-lora-13B-HF	8.034	3.528
Llama-2-70b-instruct-v2	21.272	9.588
llama-30b-instruct-2048	21.51	10.076
Mistral-7B-Instruct-v0.1	15.014	5.878
mpt-7b-chat	20.97	9.092
open_llama_13b	11.832	4.88
starchat-beta	17.252	6.644
vicuna-7b	19.712	8.542
VicUnlocked-alpaca-30b	21.458	10.148

TABLE VIII. MODEL SCORE RANK FOR SUBJECTS

Model Rank	Subject				
	Civil Activists	GLBT Activists	LGBT History	Race History	Gender Distribution
1	Llama-2-70b-instruct-v2	gpt-3.5-turbo	VicUnlocked-alpaca-30b	llama-30b-instruct-2048	VicUnlocked-alpaca-30b
2	llama-30b-instruct-2048	falcon-7b-instruct	gpt-3.5-turbo	VicUnlocked-alpaca-30b	gpt-3.5-turbo
3	falcon-7b-instruct	vicuna-7b	Mistral-7B-Instruct-v0.1	gpt-3.5-turbo	llama-30b-instruct-2048
4	VicUnlocked-alpaca-30b	mpt-7b-chat	mpt-7b-chat	Llama-2-70b-instruct-v2	Llama-2-70b-instruct-v2
5	mpt-7b-chat	VicUnlocked-alpaca-30b	falcon-7b-instruct	falcon-7b-instruct	dolly-v2-7b
6	airoboros-c34b-2.1	starchat-beta	llama-30b-instruct-2048	mpt-7b-chat	Mistral-7B-Instruct-v0.1
7	gpt-3.5-turbo	llama-30b-instruct-2048	airoboros-c34b-2.1	vicuna-7b	vicuna-7b
8	vicuna-7b	Llama-2-70b-instruct-v2	dolly-v2-7b	starchat-beta	mpt-7b-chat
9	starchat-beta	airoboros-c34b-2.1	starchat-beta	dolly-v2-7b	gpt4-alpaca-lora-13B-HF
10	open_llama_13b	Mistral-7B-Instruct-v0.1	vicuna-7b	Mistral-7B-Instruct-v0.1	AlpacaGPT4-7B-elina
11	Mistral-7B-Instruct-v0.1	dolly-v2-7b	open_llama_13b	airoboros-c34b-2.1	airoboros-c34b-2.1
12	dolly-v2-7b	AlpacaGPT4-7B-elina	AlpacaGPT4-7B-elina	gpt4-alpaca-lora-13B-HF	open_llama_13b
13	AlpacaGPT4-7B-elina	open_llama_13b	gpt2-xl	AlpacaGPT4-7B-elina	gpt2-xl
14	gpt4-alpaca-lora-13B-HF	gpt2-xl	Llama-2-70b-instruct-v2	open_llama_13b	falcon-7b-instruct
15	gpt2-xl	gpt4-alpaca-lora-13B-HF	gpt4-alpaca-lora-13B-HF	gpt2-xl	starchat-beta

#### 4) GDP group Correlations

After rigorously testing our pipeline to ensure its capability to compare model predictions, we turned our attention to exploring the potential correlation between the quality of model responses and a nation's GDP. Refer back to Table II in Section III-A, which categorizes countries into four groups based on



their GDP per capita. To clarify, Group 1 consists of low-income nations, while Group 4 comprises high-income countries.

We conducted an analysis of model scores, segmenting them by these country groups. For each LLM, we generated bar charts that display both the score values and the percentage change between groups. Fig. 3 presents such a bar chart for one model. The figure highlights the percentage changes in values across the different groups compared to the first group.

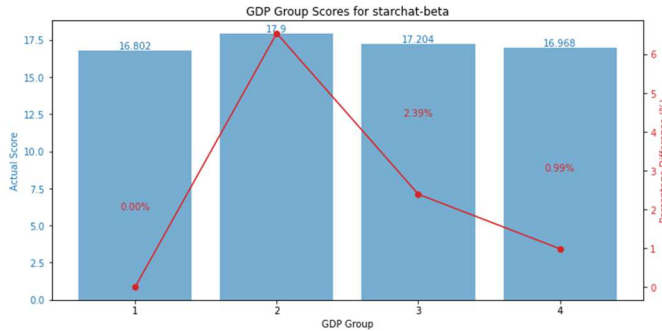


Fig. 3. starchat-beta GDP group bar chart for cosine similarity

Table IX consolidates data from all the models and names the best and worst groups. It contains two key metrics: cosine similarity and ROUGE score. The table reveals a strong correlation with GDP groups. We consistently get the best answers for Group 2 (Lower-middle income) and the worst for Groups 3 and 4 (Upper-middle income and High income). These results are very interesting and show that certain information was used (or omitted) to train models, and that different GDP group countries perform differently.

## VII. CONCLUSION

Interviewing LLMs is interesting and challenging. Evaluating the predictions of generative models can be a complex task. To simplify this process, we have created an automated pipeline that enables a universal assessment of response quality for a wide range of models and associated input prompts. Beyond its efficiency, our pipeline boasts full reproducibility, thereby optimizing both time and resource allocation. This innovative method enhances the research capabilities of professionals in the field of generative AI.

We have verified that open-source models can indeed rival the performance of state-of-the-art commercial counterparts. While the efficacy of a model may vary depending on the specific nature of your prompt subjects and application requirements, it's noteworthy that models with as few as seven billion parameters can produce high-quality responses comparable to ChatGPT. These models offer the added benefits of speed, efficiency, and cost-effectiveness, while being hosted on moderately-sized inexpensive single GPU instances. To ensure optimal performance, it is crucial to test various large language models on your specific datasets and topics to select the most suitable option.

TABLE IX. MODEL BAR CHARTS FOR GDP GROUP SCORES

Model	Cosine Similarity		ROUGE Score	
	GDP Group Bar chart	Best/Worst grps	GDP Group Bar chart	Best/Worst grps
airoboros-c34b-2.1		2/4		2/4
AlpacaGPT4-7B-elina		2/3		2/3
dolly-v2-7b		2/3		2/3
falcon-7b-instruct		2/4		2/3
gpt-3.5-turbo		3/1		2/1
gpt2-xl		2/1		2/1
gpt4-alpaca-lora-13B-HF		1/4		1/4
Llama-2-70b-instruct-v2		2/1		2/1
llama-30b-instruct-2048		2/4		2/3
Mistral-7B-Instruct-v0.1		2/3		2/3
mpt-7b-chat		2/4		2/3
open_llama_13b		1/4		1/3
starchat-beta		2/1		1/2
vicuna-7b		2/3		2/3
VicUnlocked-alpaca-30b		2/4		2/3

We have also discovered a correlation between quality of answer and nation's GDP. Most of the models we evaluated exhibited greater accuracy for Lower-middle income countries, and lower accuracy for Upper-middle income and High income states. While the specifics of the training data for these LLMs remain unknown, there appears to be a significant gap in information concerning diversity, inclusion and gender. Data scientists may need to give special attention to utilizing such data during the training process, equal for all countries independently of its GDP. This consideration is particularly crucial if we aim to promote diversity and inclusion, as well as to address the gender gap in the tech industry.

## REFERENCES

- [1] Apache 2.0 license, <https://www.apache.org/licenses/LICENSE-2.0>
- [2] Open LLM Leaderboard, web site from Huggingface [https://huggingface.co/spaces/HuggingFaceH4/open\\_llm\\_leaderboard](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard)
- [3] Language Model Evaluation Harness documentation, <https://github.com/EleutherAI/lm-evaluation-harness>
- [4] gpt-3-5-turbo documentation, <https://platform.openai.com/docs/models/gpt-3-5>
- [5] Ye, Junjie, et al. "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models." arXiv preprint arXiv:2303.10420 (2023), <http://export.arxiv.org/abs/2303.10420>.
- [6] GDP per capita (current US\$), web site, <https://data.worldbank.org/indicator/NY.GDP.PCAP>.
- [7] Huggingface Trainer documentation, [https://huggingface.co/docs/transformers/main\\_classes/trainer](https://huggingface.co/docs/transformers/main_classes/trainer)
- [8] Huggingface pipeline documentation, [https://huggingface.co/docs/transformers/main\\_classes/pipelines](https://huggingface.co/docs/transformers/main_classes/pipelines)
- [9] Hu, Zhiqiang, et al. "LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models." arXiv preprint arXiv:2304.01933 (2023).
- [10] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).
- [11] Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." arXiv preprint arXiv:2305.14314 (2023).
- [12] OpenAI rate limits documentation, <https://platform.openai.com/docs/guides/rate-limits/overview>
- [13] New World Bank country classifications by income level: 2021-2022, <https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2021-2022>
- [14] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).
- [15] Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." arXiv preprint arXiv:2306.01116 (2023).
- [16] Customize a Notebook Instance Using a Lifecycle Configuration Script, AWS documentation, <https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-lifecycle-config.html>
- [17] Lin, C. "ROUGE: A Package for Automatic Evaluation of Summaries. Proceedings of Workshop on Text Summarization Branches Out", Post-Conference Workshop of ACL 2004. Barcelona, Spain.
- [18] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.", Proceedings of the 2019 EMNLP-IJCNLP, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- [19] A. Kaplunovich and S. Kaplunovich, "Consolidating user data from social networks using Machine Learning and Serverless Cloud, 2023 ICCNS, Valencia, Spain, 2023, pp. 230-236, doi: 10.1109/ICCNS58795.2023.10193182.
- [20] Houshy, Neil, et al. "Parameter-efficient transfer learning for NLP." International Conference on Machine Learning. PMLR, 2019.