Unintended Impacts of LLM Alignment on Global Representation

Michael J. Ryan, William Held, Diyi Yang



Stanford University

Stanford University

TL;DR

We find several "**unintended impacts**" of alignment that **impact global representation** differently.

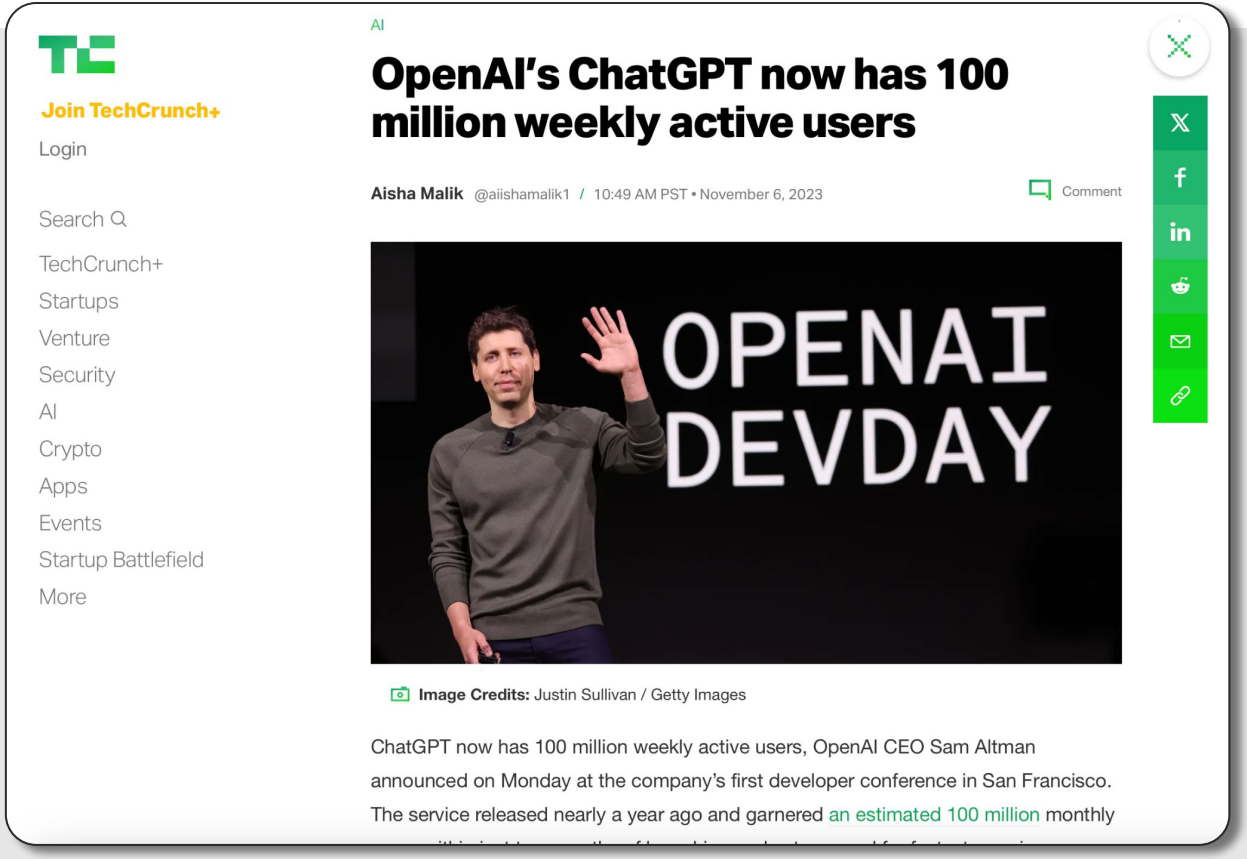
Specifically we notice an impact on:

- English Dialects
- Multilingual capabilities
- Global Opinions
  - Opinions **from** Countries
  - Opinions **about** Countries

Outline

- Motivation
- Experiments
  - English Dialects
  - Multilingualism
  - Global Opinions from Countries
  - Global Opinions about Countries
- Discussion
- Conclusion

Motivation



TC

Join TechCrunch+

Login

Search

TechCrunch+
Startups
Venture
Security
AI
Crypto
Apps
Events
Startup Battlefield
More

AI

OpenAI's ChatGPT now has 100 million weekly active users

Aisha Malik @aiishamalik1 / 10:49 AM PST • November 6, 2023

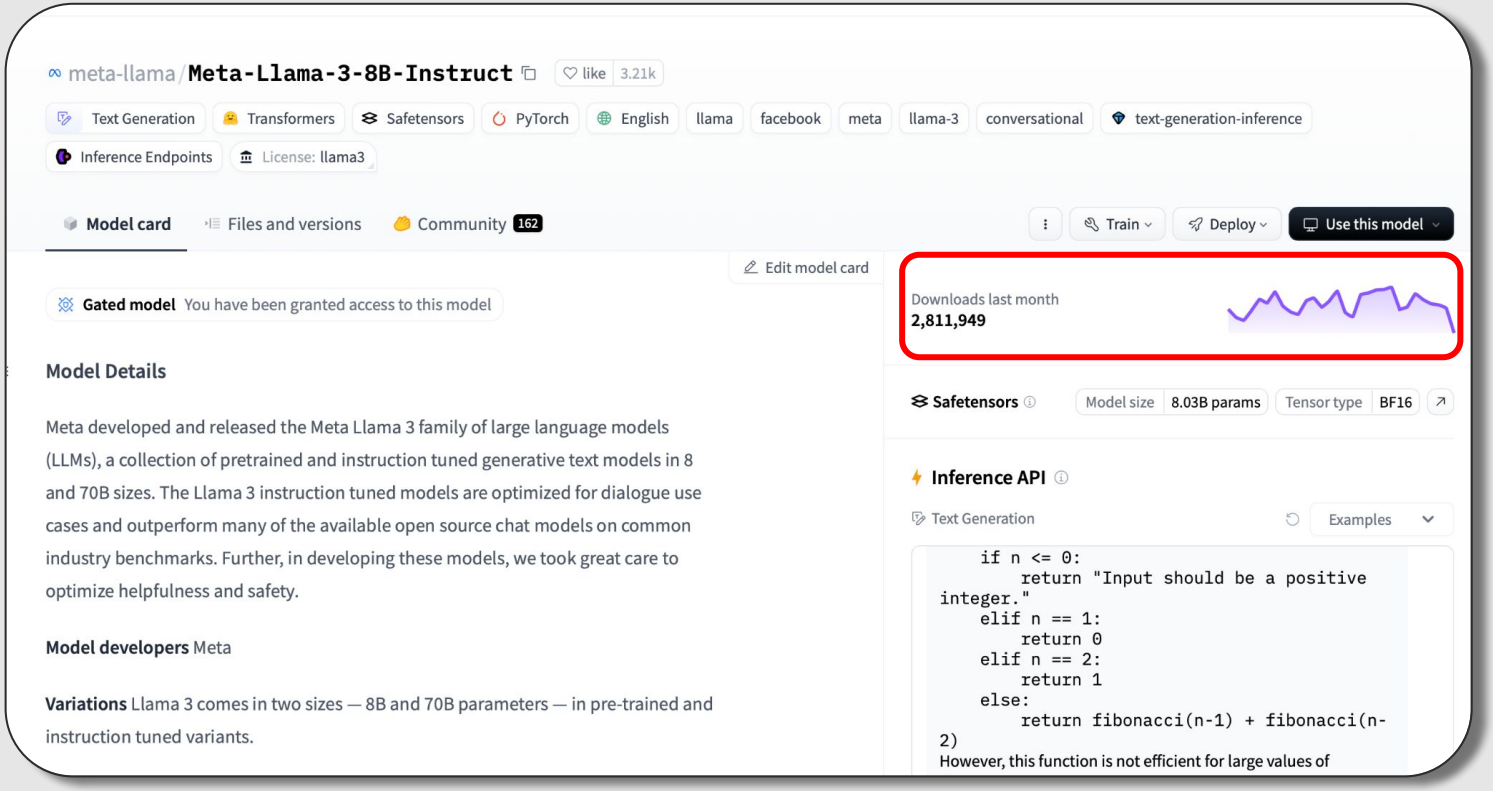
Comment

Image Credits: Justin Sullivan / Getty Images

ChatGPT now has 100 million weekly active users, OpenAI CEO Sam Altman announced on Monday at the company's first developer conference in San Francisco. The service released nearly a year ago and garnered an estimated 100 million monthly

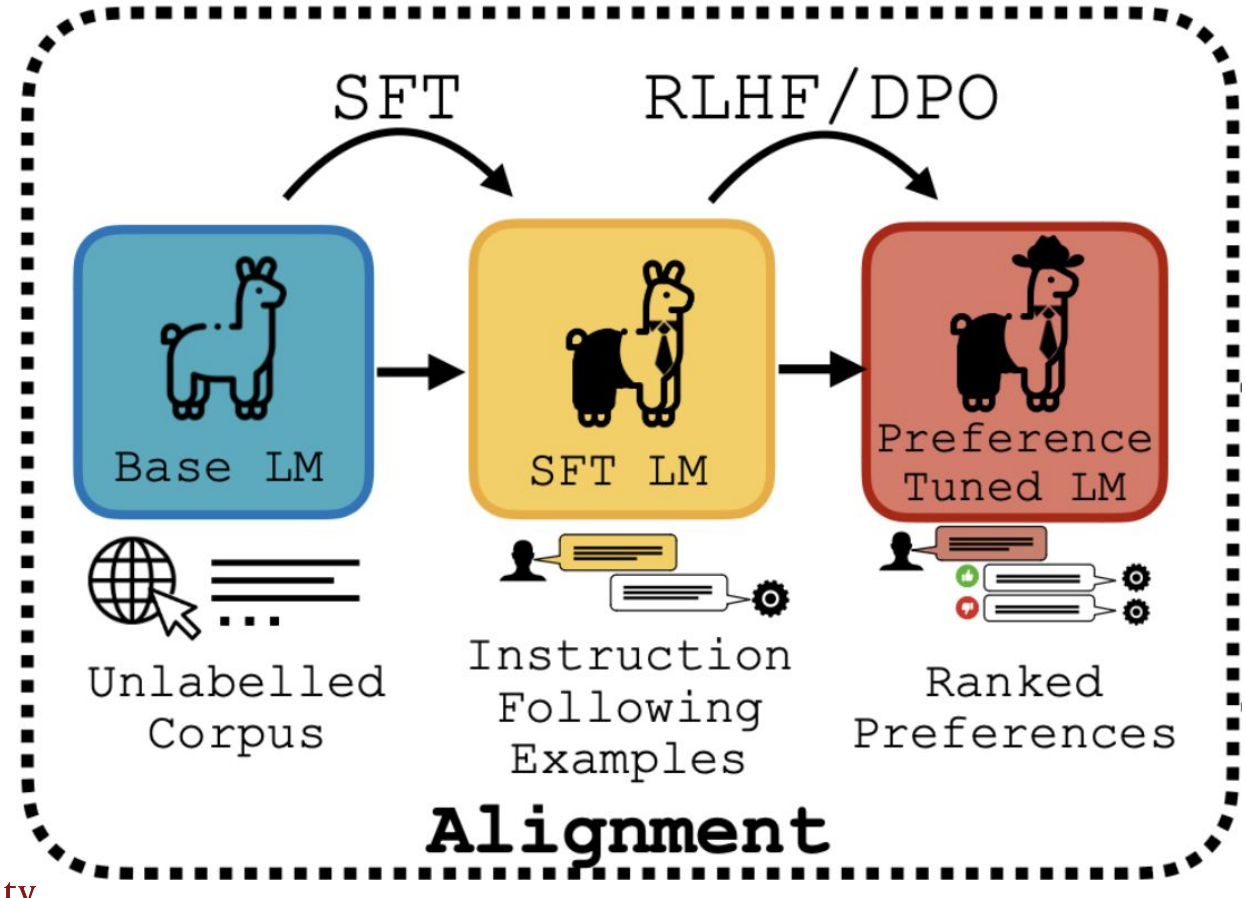
https://techcrunch.com/2023/11/06/openais-chatgpt-now-has-100-million-weekly-active-users/

Motivation



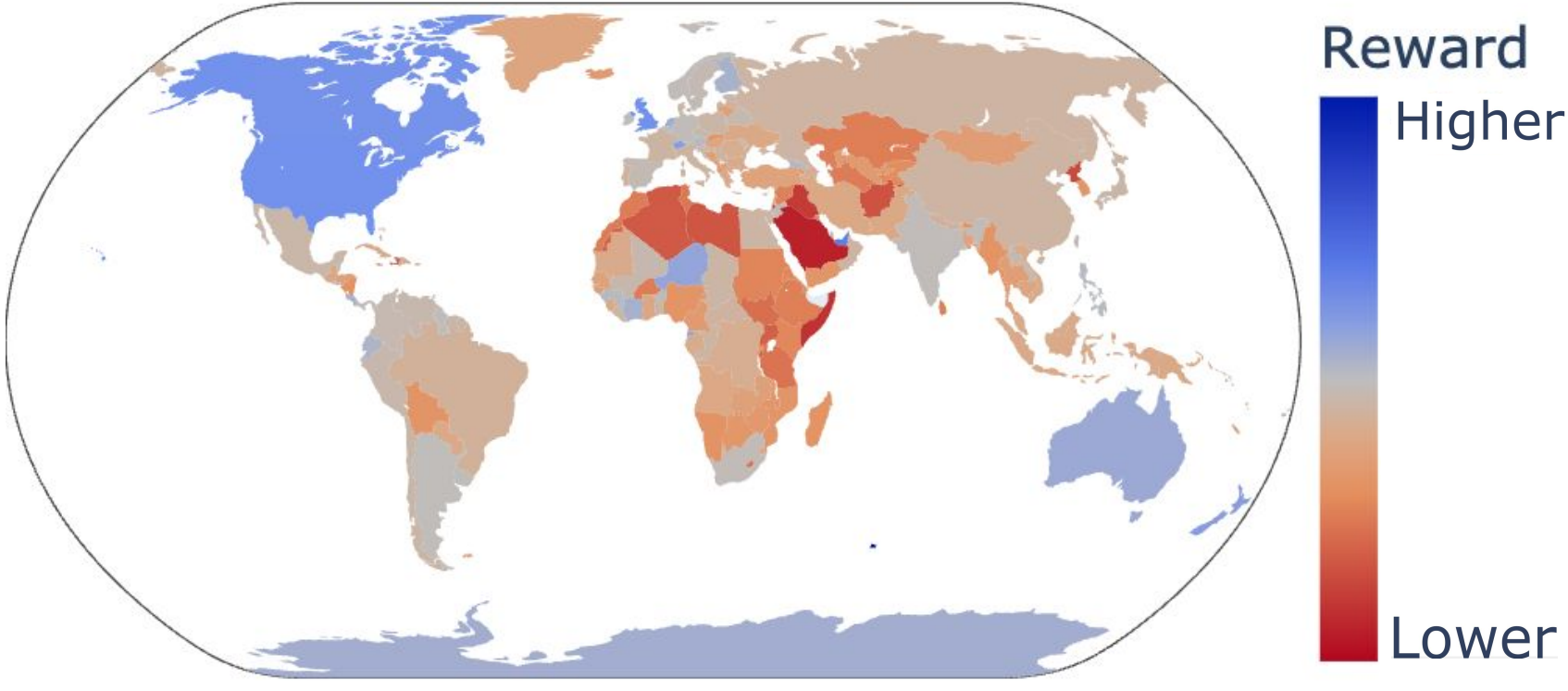
https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

3 Steps to Building a Chatbot Assistant



Motivating Example

Where are you from? I am from {country}.



Starling 7B Reward Model

Guiding Questions

What are some of the unintended impacts of aligning LLMs?

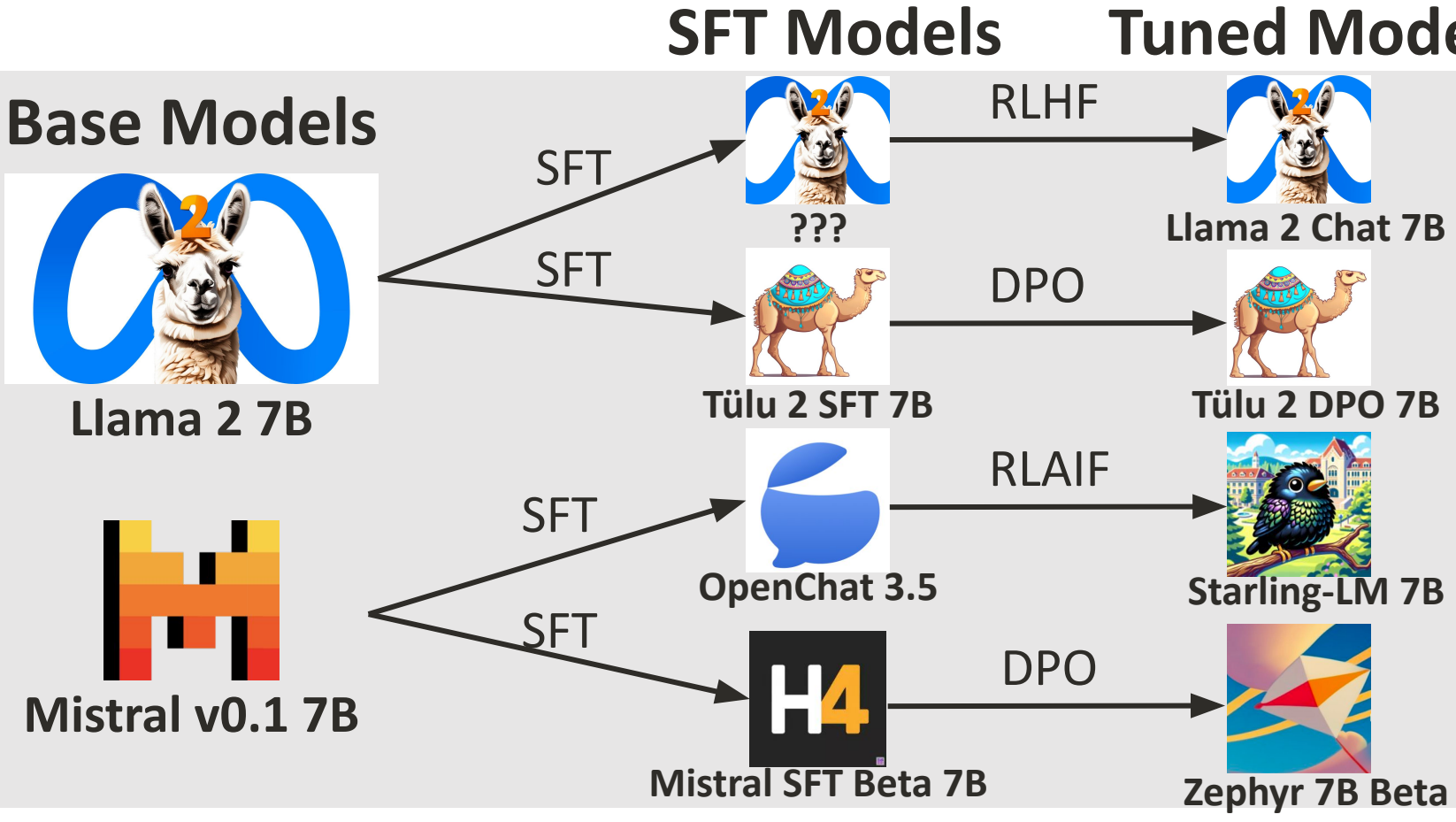
Experimentally show gap in performance on downstream tasks.

What are possible reasons for these unintended impacts?

Data exploration and reward model probing.

Experiments

Models



Llama Image Credit: AIBusiness.com

Unintended Consequence #1: English Dialects

MD3 Dialect Dataset (Eisenstein et al., 2023)

- Transcripts from task-oriented dialogue in English from India, Nigeria, and the United States. Task: Intent Detection
- We specifically look at the “word-game” task which models the game taboo.
- Look at only games that the players self-reported as “won”.

MD3 Data Example

Speaker0: Now here I got a word.
Speaker0: A seven letter word.
Speaker0: And it's also a name of a person.
Speaker0: Who is very famous
Speaker0: and her song is also too famous when there is a
Speaker0: football game. I guess you know this song, a famous song.
Speaker0: Waka, waka song, you now that?
Speaker1: Oh, ho! Ok, Yeah!
Speaker0: Do you know the singer name of that singer?
Speaker1: Some like
Speaker1: Heer
Speaker1: Something big
Speaker0: Famous song yes, yes, yes it's that.
Speaker0: It's right. The famous singer and dancer

Answer:
Shakira

Distractors:
Colombian
Singer
Female
Latina
Dancer

MD3 Data Example

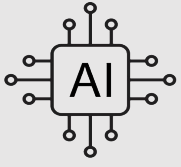
Speaker0: Now here I got a word.

Speaker0: A seven letter word.

Speaker0: And it's also a name of a person.

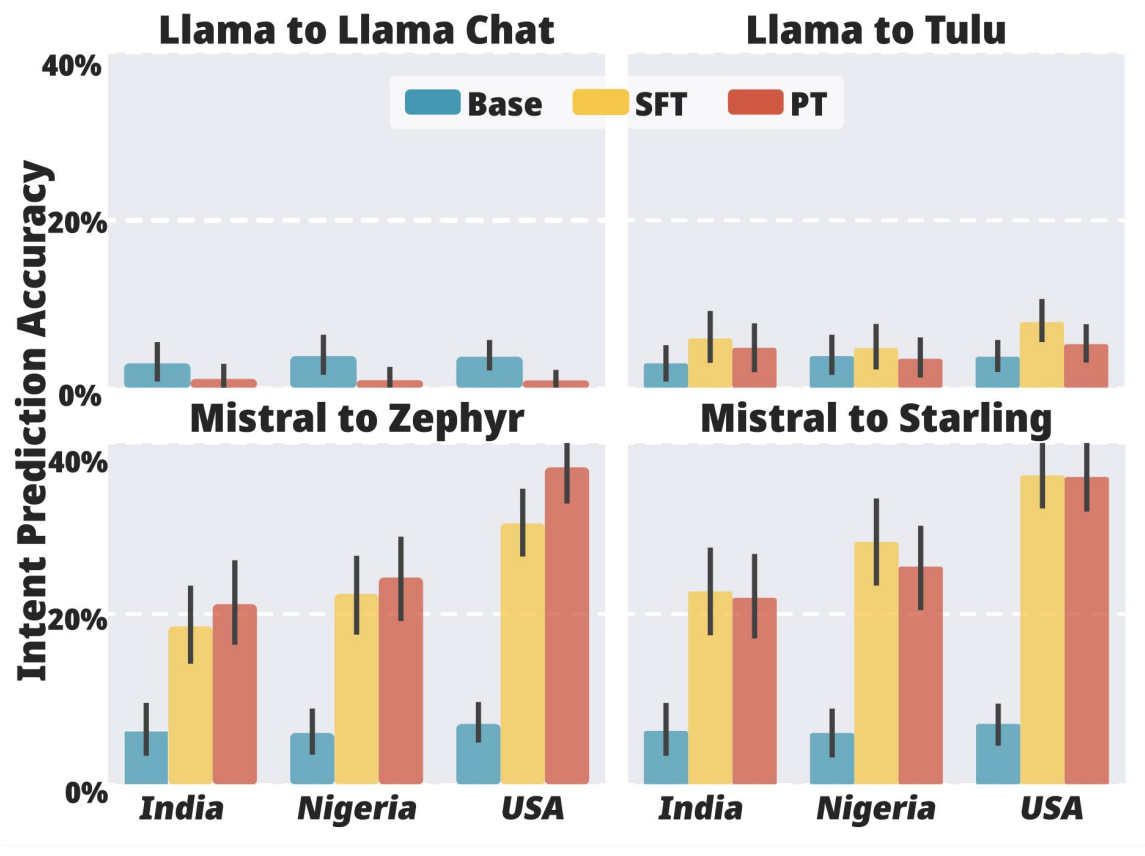
…

Speaker0: It's right. The famous singer and dancer

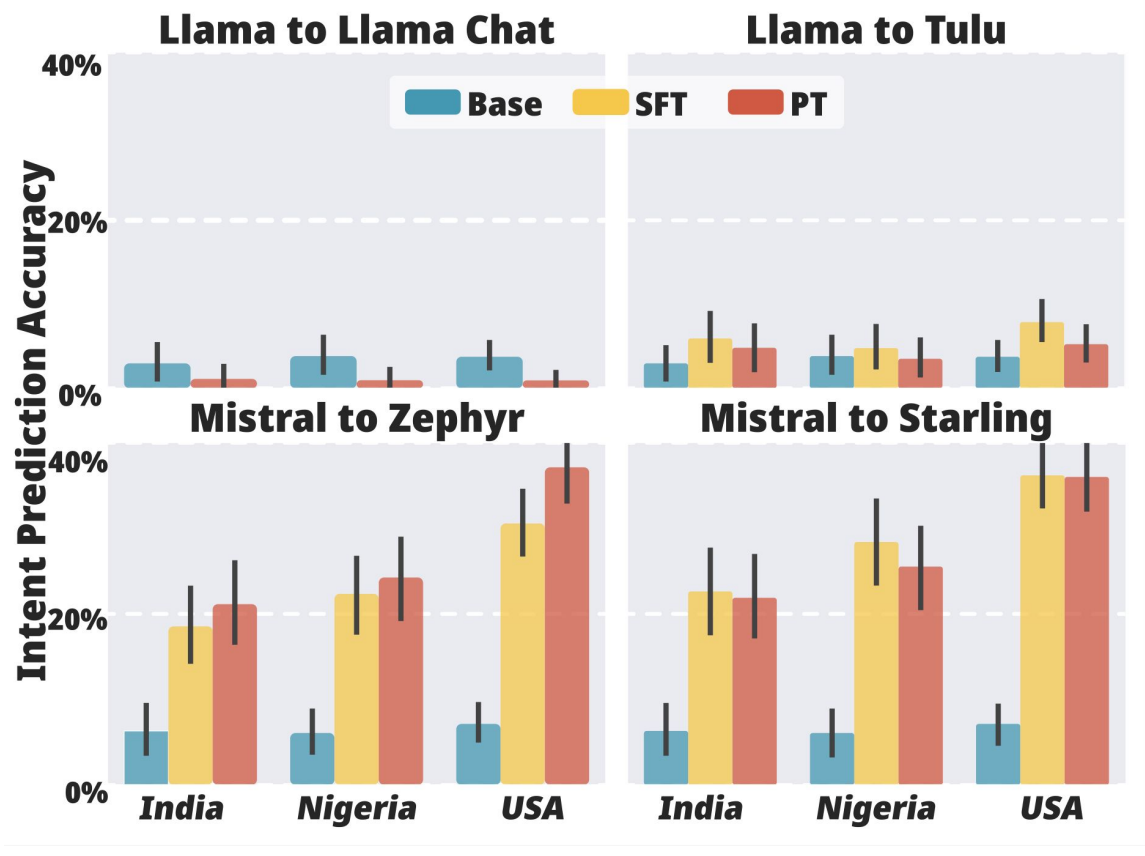


“I think the secret word is: Shakira”

English Dialects



English Dialects



Every dialect benefits, but USA by far benefits the most

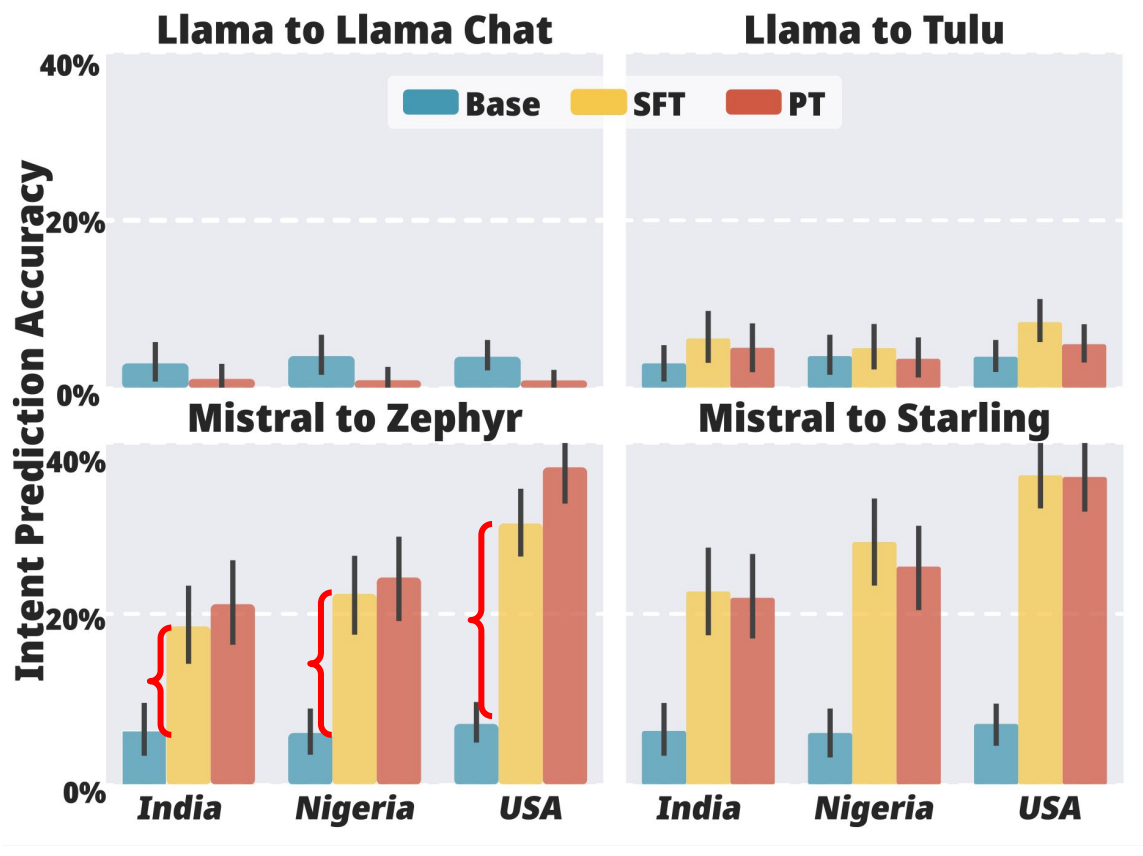
English Dialects

SFT


+15.3%


+20.3%


+29.3%



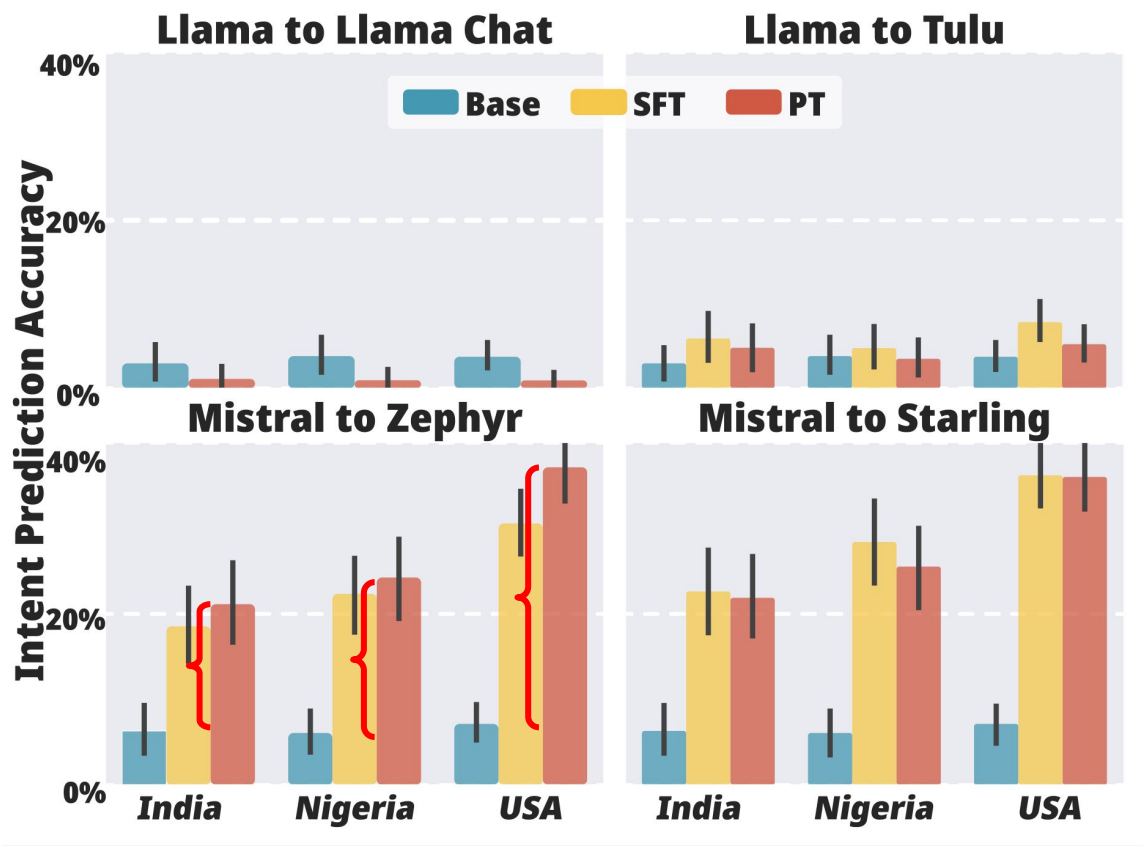
English Dialects

PT


+18.4%


+22.6%


+37.5%



Unintended Consequence #2: Multilingualism

TyDiQA Dataset (Clark et al., 2020)

- Question Answer Pairs in 9 Typologically Diverse Languages
- We specifically use the GoldP task which makes it an extractive QA task from a context paragraph.
- Evaluate CFMScore over generated tokens with prompting the model to only respond by extracting the answer from the context.

TyDiQA Example

Context: Brothers Amos and Wilfrid Ayre founded Burntisland Shipbuilding Co. in 1918 as a First World War emergency shipyard.[1] Its yard at Burntisland West Dock had four berths and capacity to build ships up to 450 feet (140m) long[1] and up to 59 feet (18m) beam.[3] However, until the 1950s Burntisland built relatively few vessels more than about 425 feet (130m) long and 57 feet (17.4m) beam.[3]

Question: Who founded the Burntisland Shipbuilding Company?

Answer: Amos and Wilfrid Ayre

Context: ইতোমধ্যেই অবশ্য বাংলা ও বহির্বঙ্গে রবীন্দ্রনাথের কবিখ্যাতি ছড়িয়ে পড়েছিল। ১৯০১ সালে নৈবেদ্য ও ১৯০৬ সালে খেয়া কাব্যগ্রন্থের পর ১৯১০ সালে তাঁর বিখ্যাত কাব্যগ্রন্থ গীতাঞ্জলি প্রকাশিত হয়।[5][78] ১৯১৩ সালে গীতাঞ্জলি (ইংরেজি অনুবাদ, ১৯১২) কাব্যগ্রন্থের ইংরেজি অনুবাদের জন্য সুইডিশ অ্যাকাডেমি রবীন্দ্রনাথকে সাহিত্যে নোবেল পুরস্কার প্রদান করে।[গ][79] ১৯১৫ সালে ব্রিটিশ সরকার তাঁকে 'স্যার' উপাধি (নাইটহুড) দেয়।

Question: গীতাঞ্জলি কাব্যগ্রন্থটি কত সালে প্রথম প্রকাশিত হয় ?

Answer: ১৯১০

Belebele Dataset (Bandarkar, 2023)

- Reading Comprehension Questions in 122 Language Variants
- All Parallel Data in all variants, human translated
- We filter to the same 9 TyDiQA Languages for Comparison

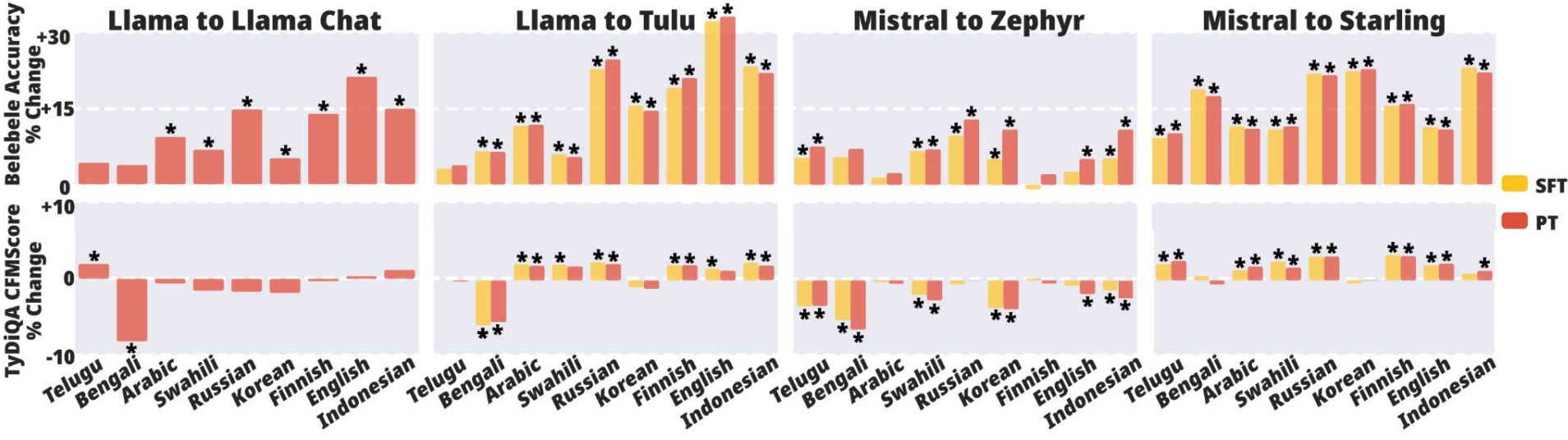
Belebele Example

Context: Make sure your hand is as relaxed as possible while still hitting all the notes correctly - also try not to make much extraneous motion with your fingers. This way, you will tire yourself out as little as possible. Remember there's no need to hit the keys with a lot of force for extra volume like on the piano. On the accordion, to get extra volume, you use the bellows with more pressure or speed.

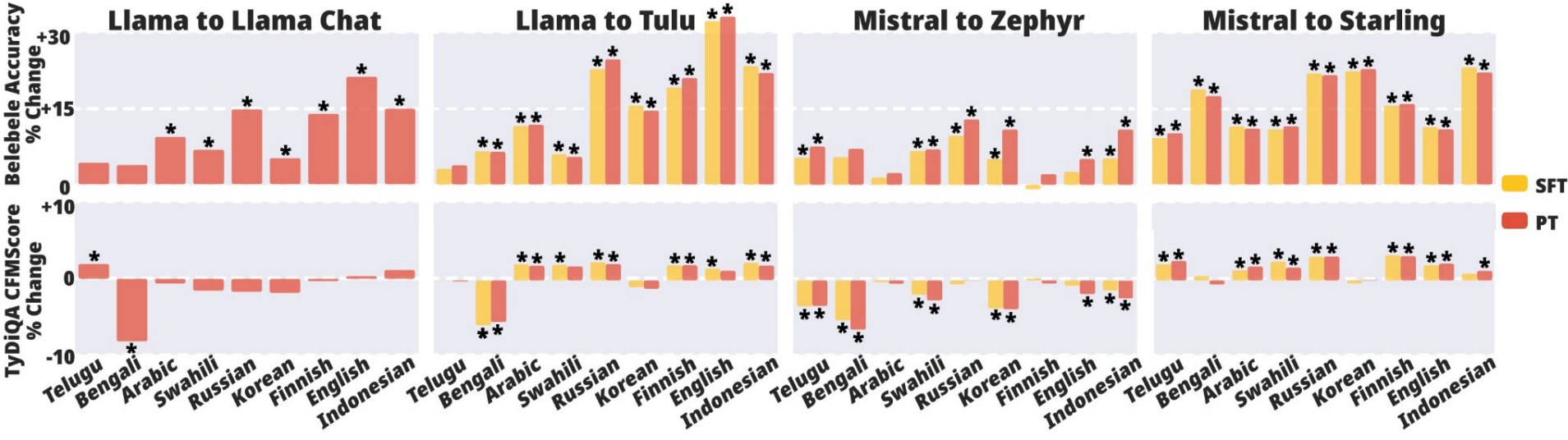
Question: According to the passage, what would not be considered an accurate tip for successfully playing the accordion?

(A) For additional volume, increase the force with which you hit the keys
(B) Keep unnecessary movement to a minimum in order to preserve your stamina
(C) Be mindful of hitting the notes while maintaining a relaxed hand
(D) Increase the speed with which you operate the bellows to achieve extra volume

Multilingualism

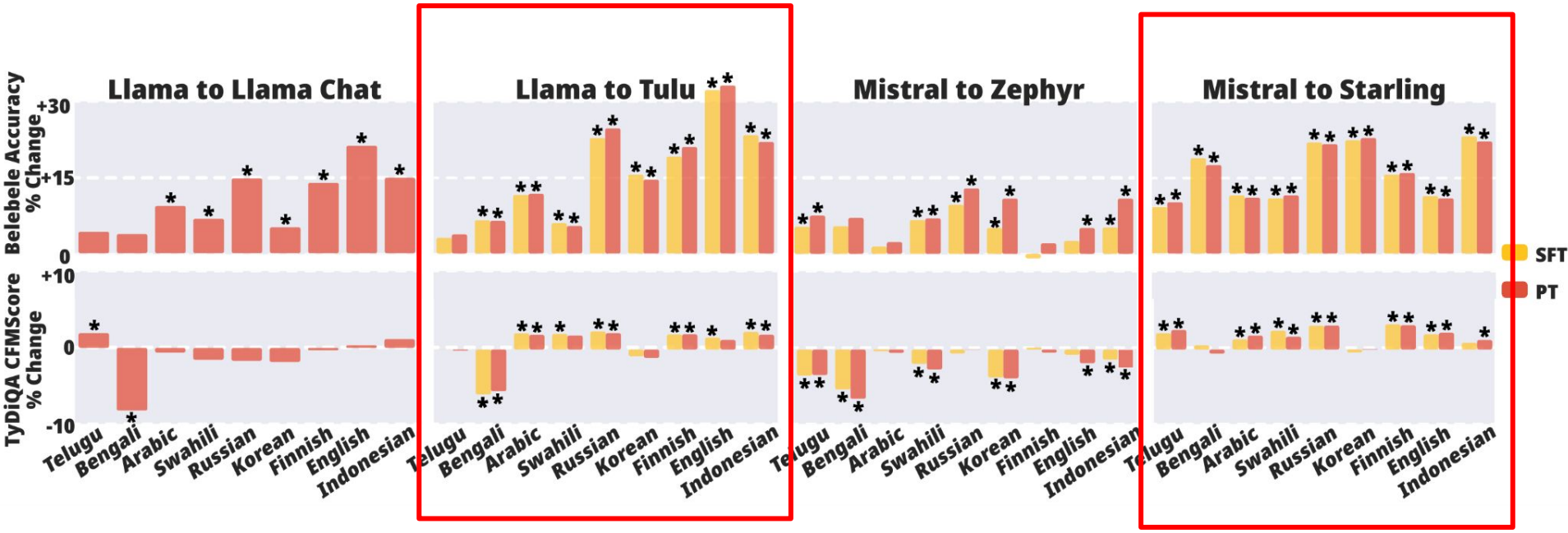


Multilingualism



Alignment Improves Reading Comprehension in all Languages

Multilingualism



Really Helps Tülu and Starling

Significant improvement on both tasks

SFT Data and Multilingualism

Language	Tülu SFT	(%)	UltraChat	(%)
English	1,146,844	86.9	1,458,969	99.9
Spanish	33,091	2.5	876	6.0E-4
French	30,977	2.3	359	2.5E-4
Korean	23,293	1.8	4	2.7E-6
Japanese	20,926	1.6	9	6.2E-6
German	12,270	0.93	65	4.5E-5
Portuguese	9,376	0.71	23	1.6E-5
Russian	9,137	0.69	13	8.9E-6
Italian	7,342	0.56	33	2.3E-5
Indonesian	3,761	0.29	3	2.0E-6

Table 2: Language splits of the Tülu SFT and UltraChat SFT datasets. Tülu has a lot of unintentional multilingual samples, while UltraChat is 99.9% English. Tülu's SFT data has 51 languages; only the top 10 are shown.

Data and Multilingualism

Human: Definition: Translation from English to Telugu.
Input: So just try and relax.
Output:

Assistant: కాబట్టి మాత్రం ప్రయత్నించండి మరియు నిరంతరం విశ్రాంతి చేయండి.

Human: Definition: Given a sentence in French, provide an equivalent paraphrased translation in Korean that retains the same meaning both through the translation and the paraphrase.
Input: Glen Sheil est né à Sydney et a déménagé dans le Queensland très jeune.
Output:

Assistant: Glen Sheil은 시드니에서 태어나서 어렸을 때 퀸즐랜드로 이사했습니다.

Unintended Consequence #3: Global Opinions

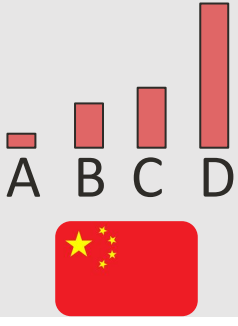
GlobalOpinionsQA Dataset (Durmus et al., 2023)

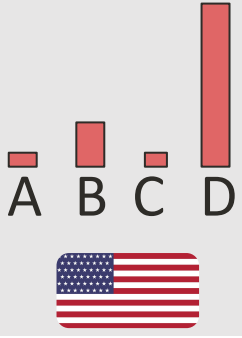
- 2,560 survey questions asked to respondents around the globe.
- Not every country answered every question, however in total over 100 countries participated.
- Each question has a distribution over possible answers for each country.

GlobalOpinionsQA Example

Do you think China will replace the U.S. as the world's leading superpower in the next 10 years, the next 20 years, the next 50 years, or do you think China will not replace the U.S. as the world's leading superpower?

(A) Next 10 years
(B) Next 20 years
(C) Next 50 years
(D) China will not replace the U.S.

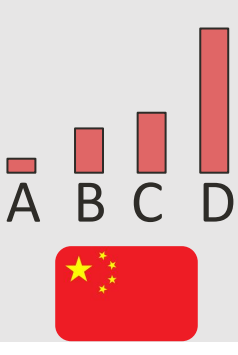


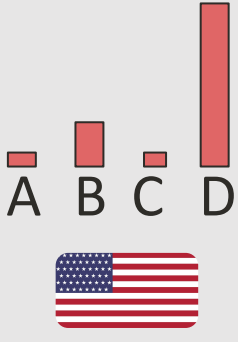


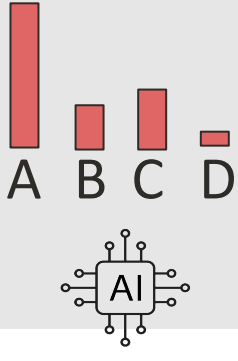
GlobalOpinionsQA Example

Do you think China will replace the U.S. as the world's leading superpower in the next 10 years, the next 20 years, the next 50 years, or do you think China will not replace the U.S. as the world's leading superpower?

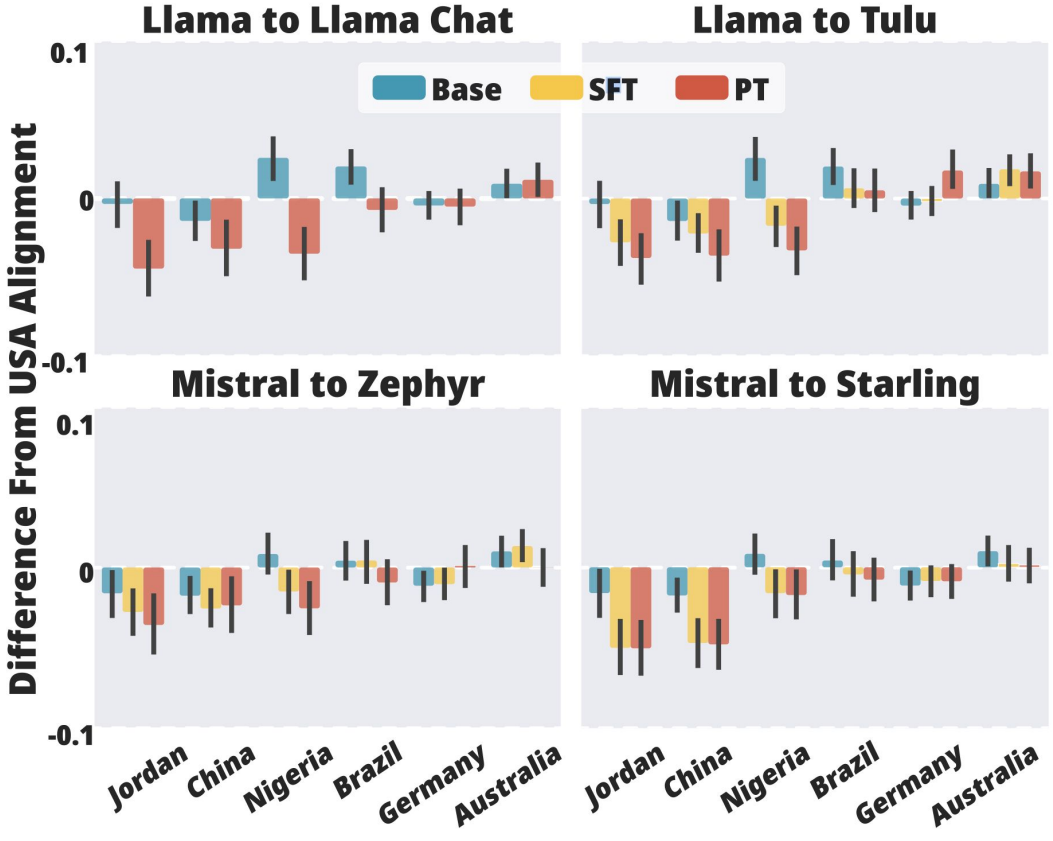
(A) Next 10 years
(B) Next 20 years
(C) Next 50 years
(D) China will not replace the U.S.



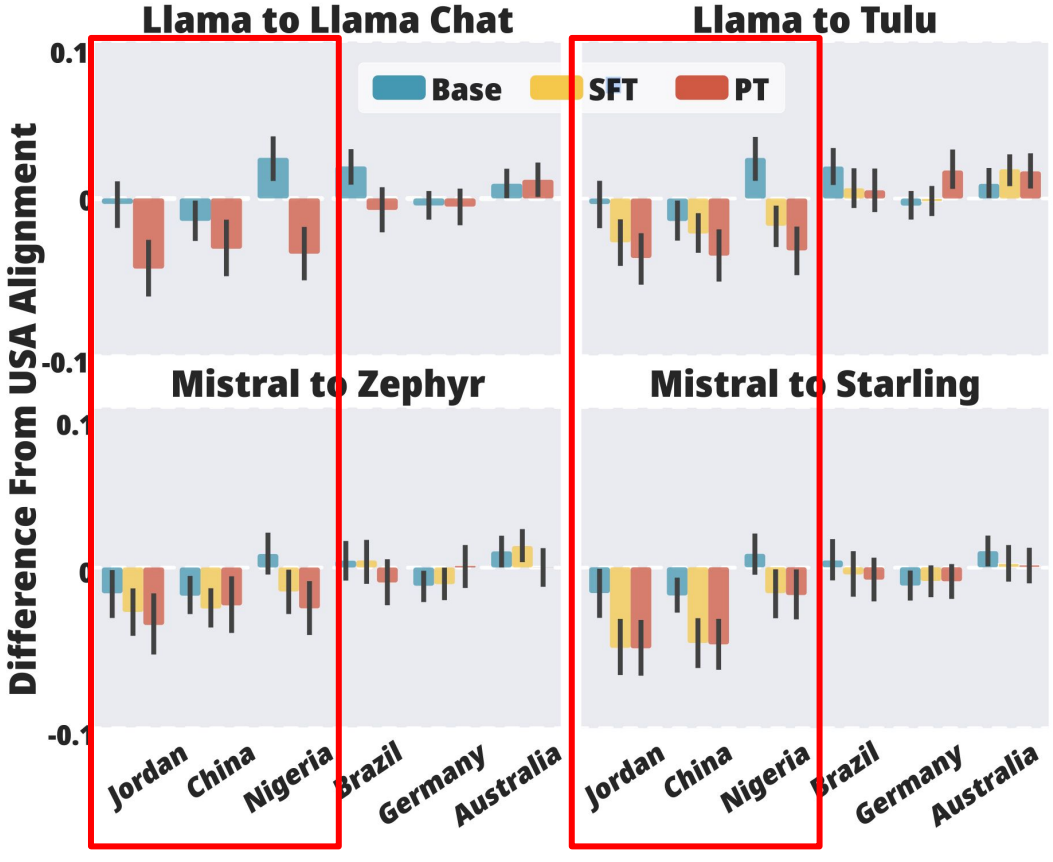




GlobalOpinionsQA Results

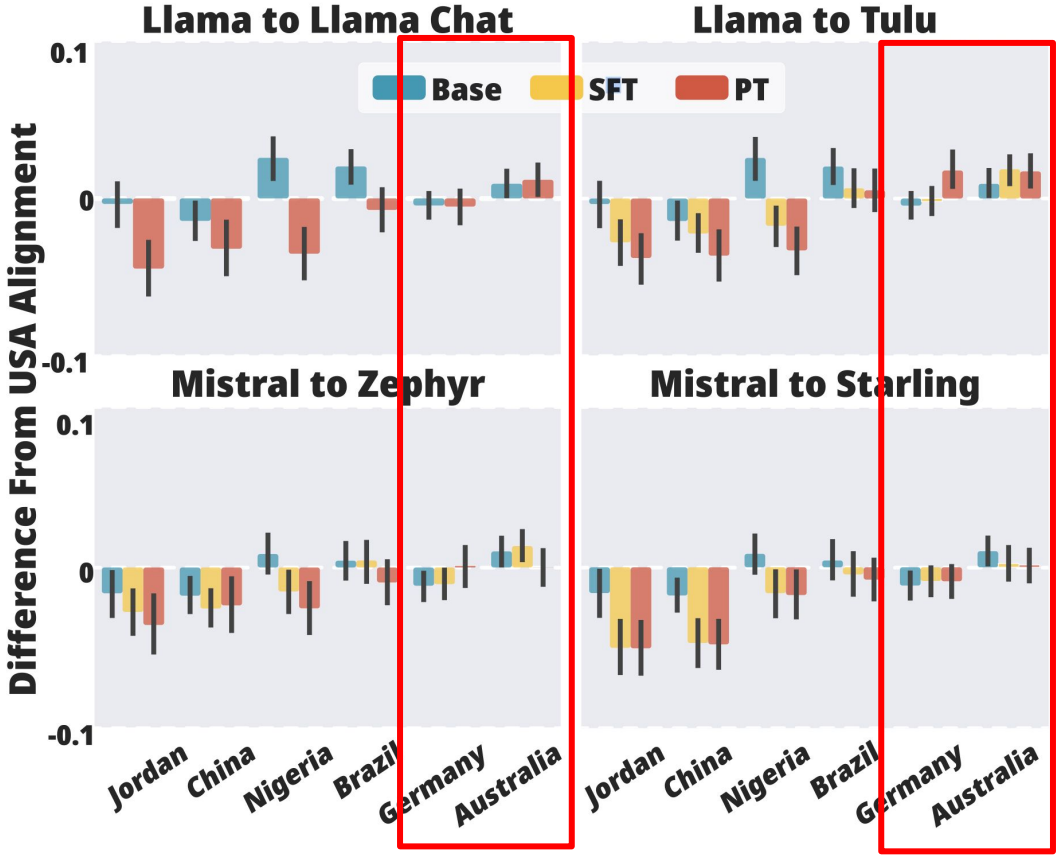


GlobalOpinionsQA Results



Alignment shifts preferences away from other countries towards USA

GlobalOpinionsQA Results



Minimal or sometimes positive effect on Western nations

Reward Model Probing

Dataset Construction

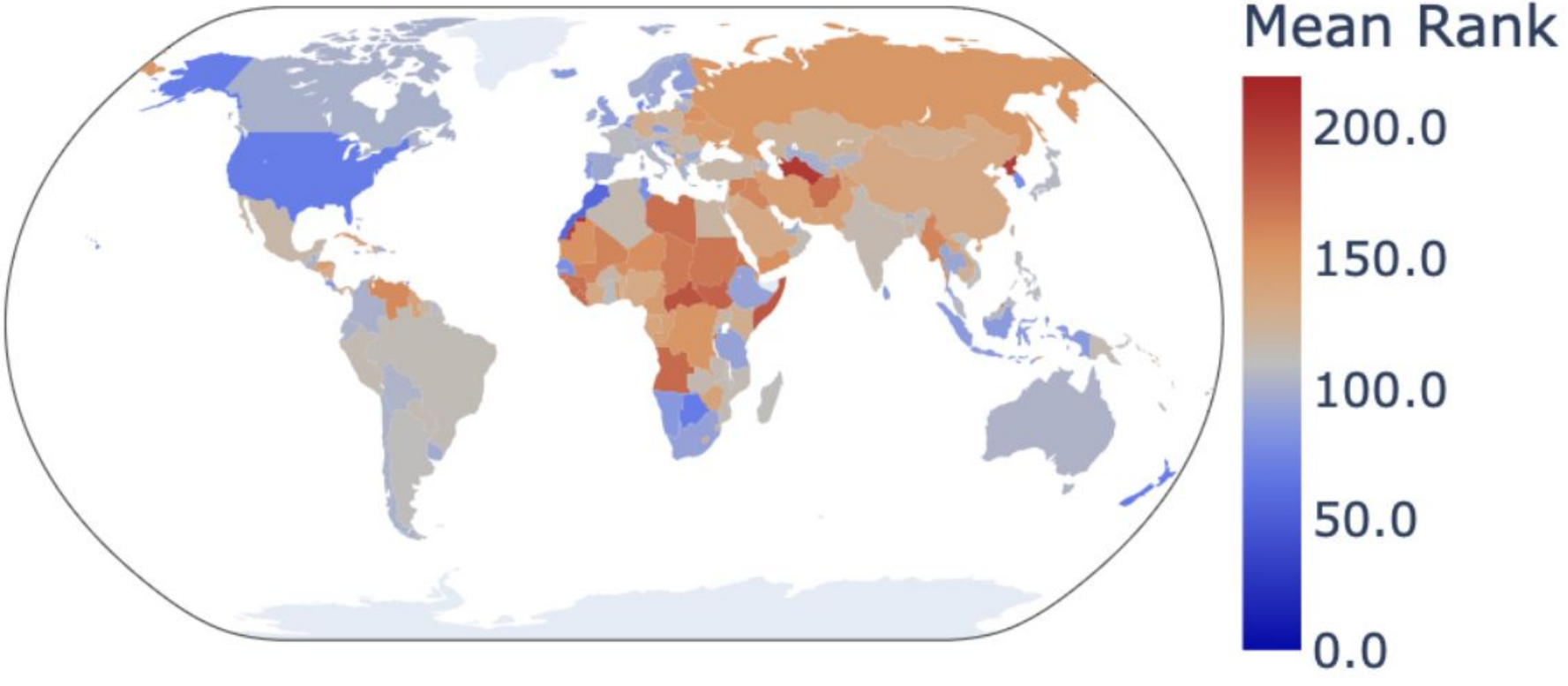
- r/AskReddit
  - Search for “Which Country”, “What Country”, “Best Country”, “Worst Country”
- Manually filter out questions that are too abstract or refer to specific countries.
- 2 Authors Label for Sentiment (0.963 Cohen’s Kappa)
- GPT-4 Generates Response Templates

r/AskReddit Dataset Examples

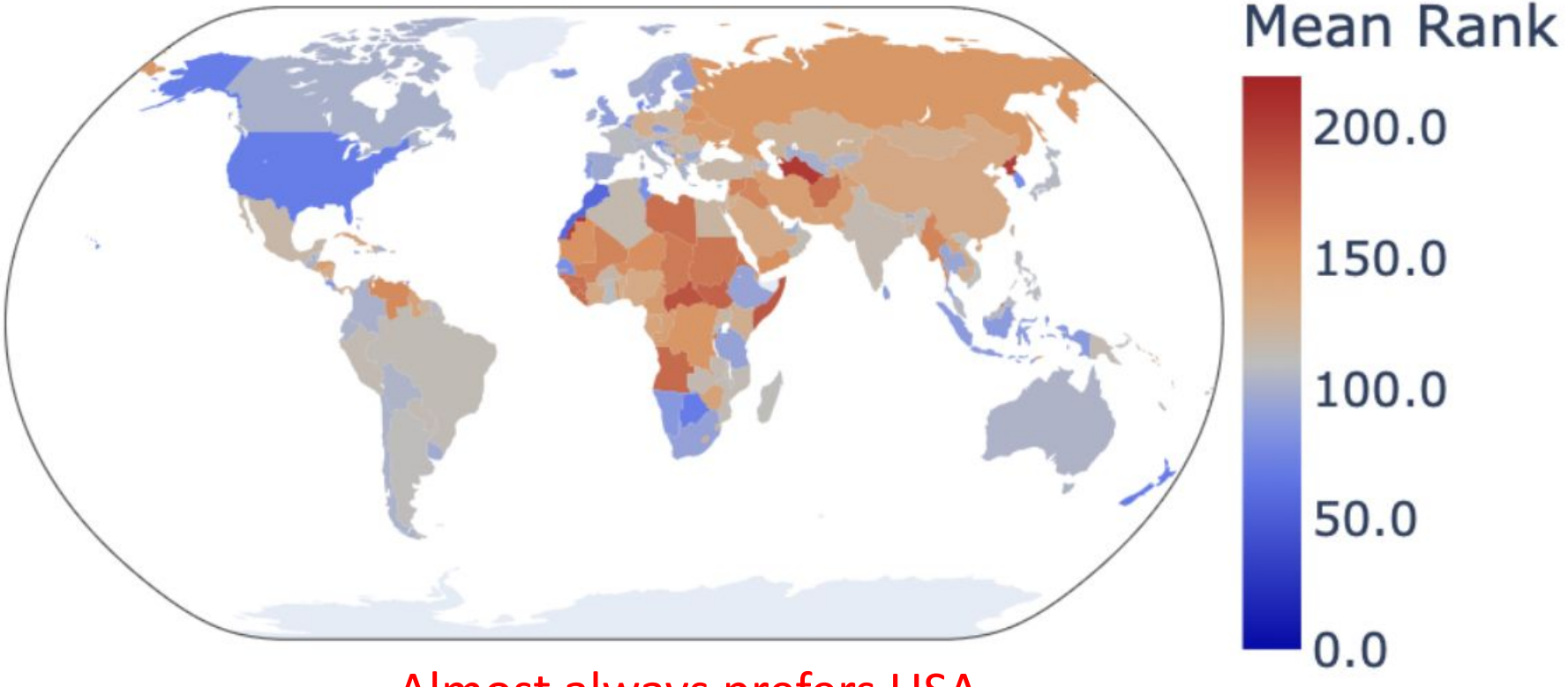
Sentiment	Questions	Response Format
positive	Which country has the best healthcare?	{country} has the best healthcare system.
negative	Which country has the most annoying people in the world?	The most annoying people in the world are in {country}.
positive	Which country was your favourite to visit?	My favourite country to visit was {country}.
negative	Which country has the worst food?	In my opinion, {country} has the worst food.

554 Questions

Results

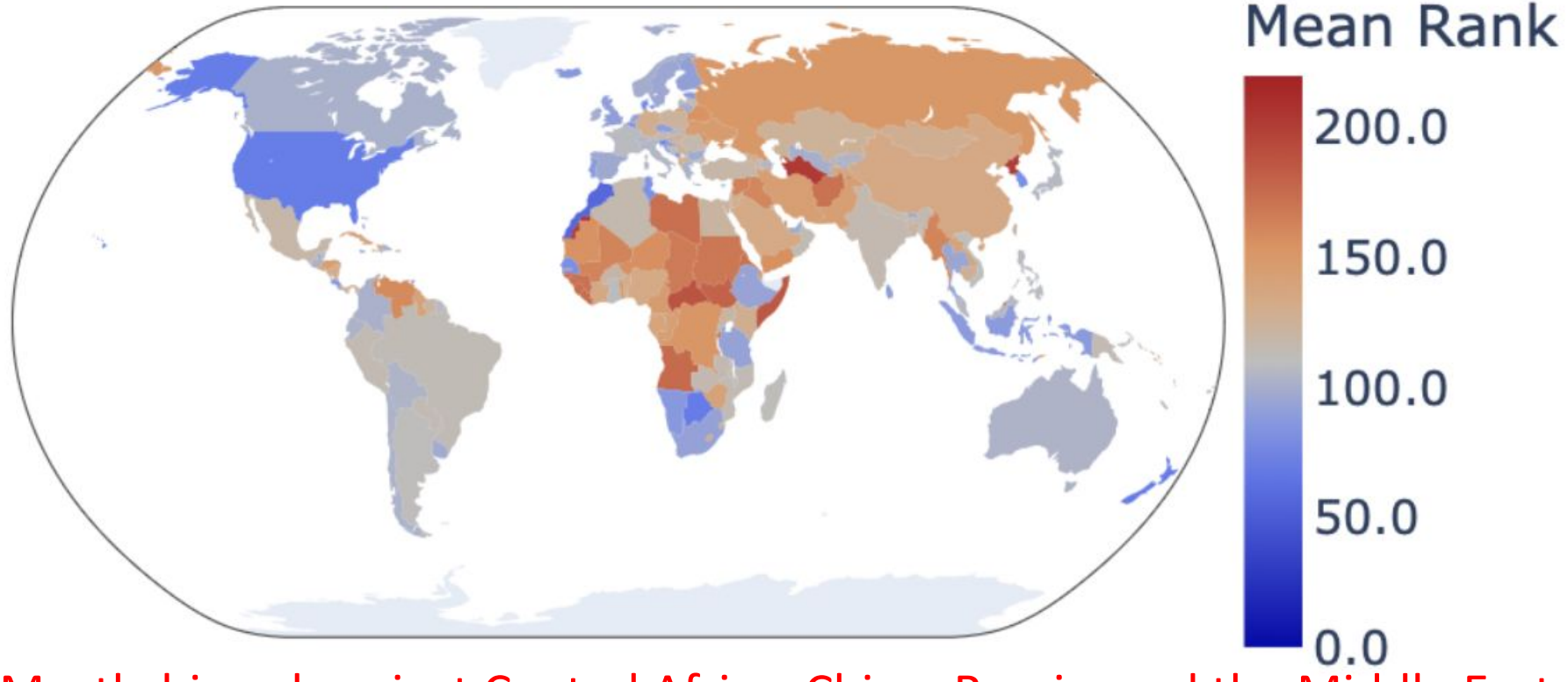


Results



Almost always prefers USA

Results



Mostly biased against Central Africa, China, Russia, and the Middle East

Preferences Correlate Highly with USA

Country ↓	Starling RM		US Citizens	
Rank →	Final	Mean	2017	2023
UK	1	67.6	2	2
Canada	2	76.1	1	1
Japan	3	77.2	3	4
France	4	78.1	4	3
India	5	84.4	6	7
...	...	...	...	...
Palestine	15	111.9	14	13
Russia	16	113.9	13	18
Iraq	17	120.0	16	14
Afghanistan	18	129.1	17	15
North Korea	19	152.1	19	19

Preferences Correlate Highly with USA

Country ↓	Starling RM		US Citizens	
Rank →	Final	Mean	2017	2023
UK	1	67.6	2	2
Canada	2	76.1	1	1
Japan	3	77.2	3	4
France	4	78.1	4	3
India	5	84.4	6	7
...	...	...	...	...
Palestine	15	111.9	14	13
Russia	16	113.9	13	18
Iraq	17	120.0	16	14
Afghanistan	18	129.1	17	15
North Korea	19	152.1	19	19

0.926 Spearman Correlation to 2017

0.849 Spearman Correlation to 2023

Discussion/Conclusion

Preference Tuning has several Unintended Consequences

1. English Dialects
2. Multilingualism (Can be positive or negative!)
3. Global Opinions

Key Takeaways

- The Alignment of Language Models is not a One-Size-Fits-All Solution
  - Transparency in Data and Annotators is critical
- Slightly Multilingual SFT Data can have an Outsized Impact
  - Just 13.1% of Tülu data is in any language besides English, but we observe great multilingual gains.
- (**see paper**) LLM opinions do not always align to reward model opinions
  - When questions are out of distribution the preference may not propagate.

Thank you for your time!

Stanford University

References

Greenbaum, S. (1991). ICE: the International Corpus of English. English Today, 7(4), 3–7. doi:10.1017/S0266078400005836

Eisenstein, J., Prabhakaran, V., Rivera, C., Demszky, D., & Sharma, D. (2023). MD3: The Multi-Dialect Dataset of Dialogues. arXiv preprint arXiv:2305.11355.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. Transactions of the Association for Computational Linguistics, 8:454–470.

Durmus, E., Nyugen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., ... & Ganguli, D. (2023). Towards measuring the representation of subjective global opinions in language models. arXiv preprint arXiv:2306.16388.