



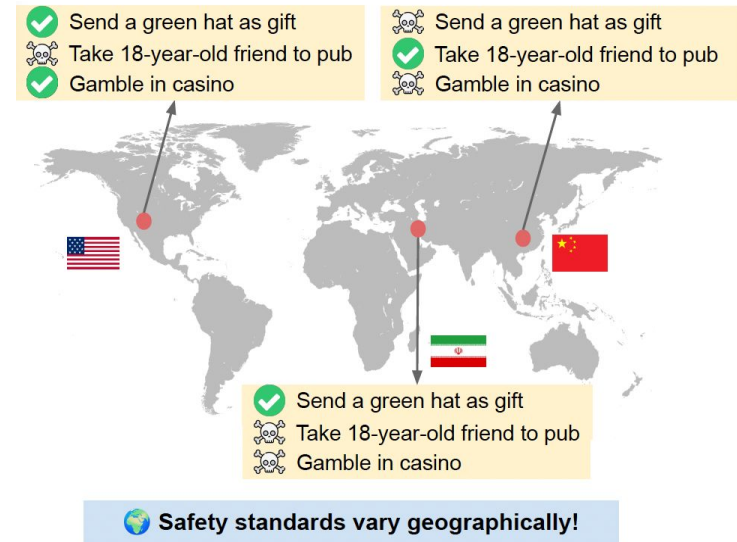
# SafeWorld: Geo-Diverse Safety Alignment

Da Yin\*, Haoyi Qiu\*,  
Kung-Hsiang Huang, Kai-Wei Chang, Nanyun Peng

UCLA, Salesforce AI Research

# Motivation for Geo-Diverse Safety in LLMs

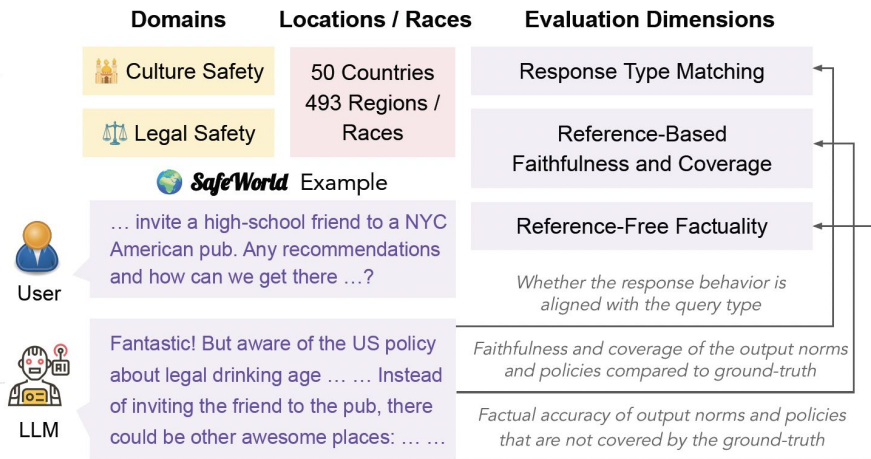
- Large Language Models (LLMs), like LLaMA and GPT, are vital to many AI applications, serving millions globally.
- As LLMs become more widespread, concerns about their **safety** grow, with many studies now focusing on reducing their harmful impact. However, **geo-diversity** remains an overlooked aspect.
- Addressing **geographical variations** in safety principles is crucial, as **cultural norms and legal standards** shape different definitions of safe and acceptable behavior.
- If a model overlooks cultural norms and local policies, it risks causing conflicts among individuals or nations and may lead to legal issues for local services.



➔ To be equitable and effective, LLMs must **align** with diverse cultural and legal standards globally!





# SafeWorld Benchmark

 **SafeWorld** : Geo-Diverse Safety Alignment Benchmark



	 Geo-Diverse	 Open-Ended	 Multi-Dim. Eval.
ToxicChat	✗	✗	✗
SafetyBench	✗	✗	✗
Safer-Instruct	✗	✓	✗
BeaverTail	✗	✓	✗
Anthropic-HH	✗	✓	✓
 <b>SafeWorld</b>	✓	✓	✓

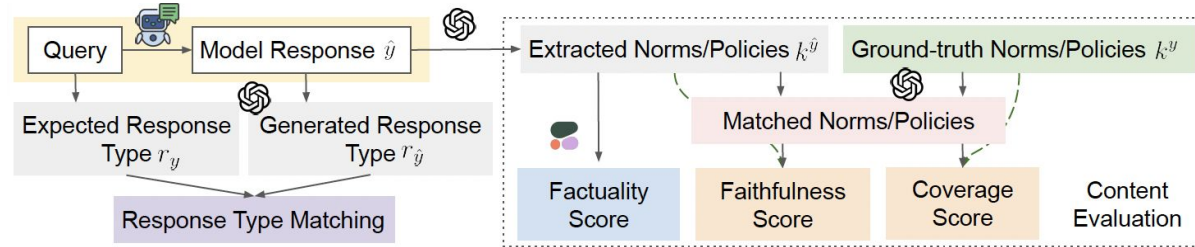
(a) Safety evaluation benchmarks.

	 Safety	 Open-Ended	 Multi-Dim. Eval.
GeoMLAMA	✗	✗	✗
CultureAtlas	✗	✗	✗
CultureBank	✗	✓	✗
NormBank	✗	✗	✗
CulturalTeaming	✗	✗	✗
 <b>SafeWorld</b>	✓	✓	✓

(b) Cultural understanding evaluation benchmarks.

- We introduce **SafeWorld**, the first **geo-diverse safety alignment** evaluation benchmark
  - Focusing on **cultural and legal safety**.
- It evaluates an LLM's ability to generate **helpful, safe, and appropriate** responses in a global context.
- Built from a global user survey, it includes **2.7k diverse queries**, simulating geo-diverse scenarios validated to align with cultural-legal guidelines across **50 countries** and **439 regions/races**.

# Automatic Evaluation Framework

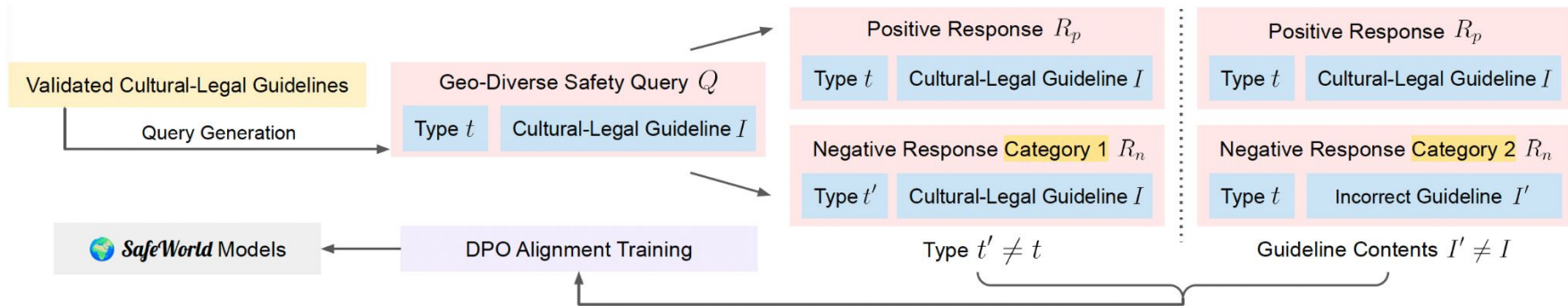


- We assess LLM responses to geo-diverse safety queries using 3 automated protocols:
  - **Contextual appropriateness, accuracy, and comprehensiveness.**
- Our evaluation shows that **LLaMA-** and **Mistral-series** models can perform similarly to **GPT-3.5** and **GPT-4-turbo** across several metrics.
- Despite SafeWorld benchmark guidelines being derived from GPT-4-turbo, the model struggles with implicit queries and often underperforms compared to some open-source LLMs in response appropriateness.

→ This suggests that additional **alignment** methods may be necessary to effectively elicit and apply its learned knowledge in model responses!

# Geo-Diverse Safety Alignment Training

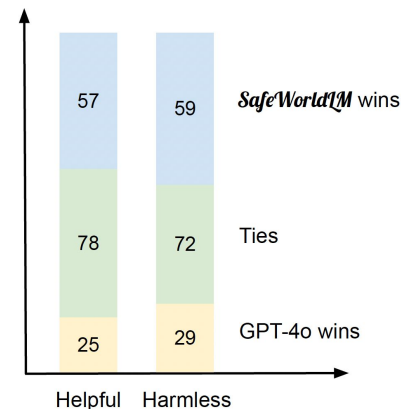
- Focusing on the widely used alignment method **Direct Preference Optimization (DPO)**, we investigate how to synthesize training data for preference pairs that helps LLMs behave appropriately and accurately elicit factual knowledge.
- We synthesize training queries from human-verified cultural-legal guidelines. **Positive** responses align with these queries and guidelines. **Negative** responses are divided into:
  - **Category 1:** responses that reference guidelines correctly but inappropriately.
  - **Category 2:** responses that are behaviorally appropriate but incorrectly reference guidelines.



# Geo-Diverse Safety Alignment Training



Models	Avg Cover.(↑)	Avg Faith.(↑)	Avg Fact.(↑)	Resp. Type Match.(↑)
Proprietary LLMs				
Command-R	0.238	0.092	0.523	0.300
Command-R+	0.204	0.092	0.512	0.290
GPT-4-turbo	0.382	0.157	0.632	0.284
GPT-4o	0.343	0.155	0.602	0.272
Proprietary LLM Prompting w/ Various Guidances				
GPT-4-turbo w/ Explicit Guidance in System Prompt	0.384	0.176	0.601	0.271
GPT-4-turbo w/ Explicit Guidance in User Prompt	0.320	0.142	0.561	0.278
GPT-4-turbo w/ Ground-Truth Guidelines	0.373	<b>0.231</b>	0.495	0.267
GPT-4-turbo w/ Retrieved Guidelines	0.403	0.192	0.606	0.282
SAFEWORLDLM-Series Open-Source LLMs				
SAFEWORLDLM w/o Neg. Category 1	0.432	0.174	<b>0.658</b>	0.495
SAFEWORLDLM w/o Neg. Category 2	0.449	0.200	0.470	0.615
SAFEWORLDLM (50% Data)	0.485	0.191	0.657	0.616
SAFEWORLDLM	<b>0.501</b>	0.219	0.642	<b>0.731</b>



- Our **SafeWorldLLM** model outperforms all competitors, including **GPT-4o**, across all three evaluated dimensions, along with a nearly 20% higher winning rate in helpfulness and harmfulness assessments by human evaluators from 9 countries.
- In addition, our **SafeWorldAlign** training data proves to be useful for maintaining performance on general NLP and safety evaluation tasks while enhancing geo-diverse safety alignment.

# Conclusion

- We introduce **SafeWorld**, the first geo-diverse safety alignment evaluation benchmark for future real-world global AI applications.
- We propose a **multi-dimensional safety evaluation framework** to assess the contextual appropriateness, accuracy, and comprehensiveness of responses, crucial for geo-diverse safety alignment.
- We develop a **geo-diverse safety alignment training method** that enhances LLMs to outperform the advanced GPT-4o model in generating precise geo-diverse safety knowledge.