



# Building a Global AI Safety Benchmark

Mala Kumar

Director of Program Management, AI Safety @MLCommons

Summit on Responsible Decentralized Intelligence —Future of Decentralization and AI

August 6, 2024 | New York City

Berkeley  
UNIVERSITY OF CALIFORNIA

Powered by Zoom

## About Mala



### Current

Director of Program Management, AI Safety @ MLCommons

Verizon Event ...



### Recent

Director, Tech for Social Good @ GitHub

Senior Advisor, World Health Organization

### Past

10+ years at specialized agencies of the United Nations.

UX research and design, open source software, AI / ML for social good.



New York City

# About Mala



## Current

Director of Program Management, AI Safety @ MLCommons



Verizon Event ...

## Recent

Director, Tech for Social Good @ GitHub  
Senior Advisor, World Health Organization

## Past

10+ years at specialized agencies of the United Nations.  
UX research and design, open source software, AI / ML for social good.



New York City

Berkeley  
UNIVERSITY OF CALIFORNIA

Powered by Zoom

# About Mala



## Current

Director of Program Management, AI Safety @ MLCommons



Verizon Event ...

## Recent

Director, Tech for Social Good @ GitHub  
Senior Advisor, World Health Organization

## Past

10+ years at specialized agencies of the United Nations.  
UX research and design, open source software, AI / ML for social good.



New York City

# About MLCommons Association

MLCommons Mission: “Better AI for everyone”

- Global community engaged in **collaborative engineering**
  - Non-profit with 125+ members
  - Academics, companies, non-profits
- Build and operate **benchmarks and public datasets**
  - Accuracy, efficiency, and safety
  - MLPerf: Industry standard for benchmarking AI training/inference speed, 55,000+ results
- Established in 2020, MLPerf started in 2018

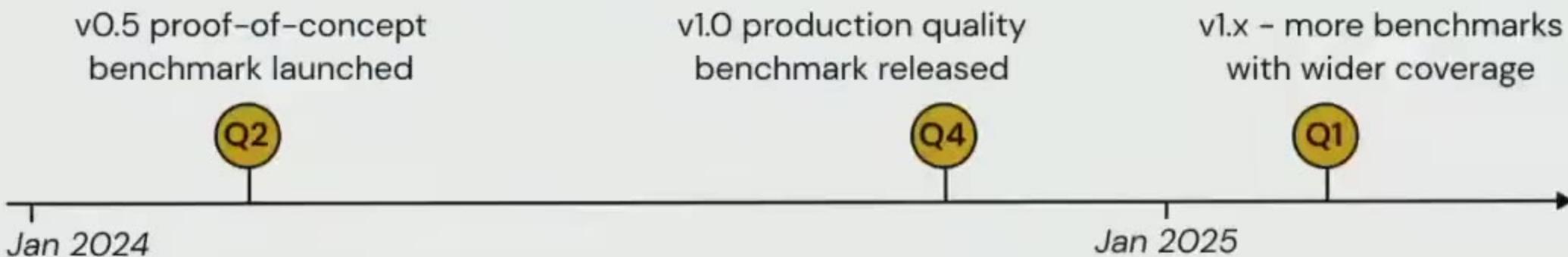
The screenshot shows the MLCommons website homepage. At the top, there's a banner for the "Verizon Event" featuring a woman speaking at a podium. Below the banner, the mission statement "Better AI for Everyone" is displayed. Key statistics are highlighted: 125+ members, 6 benchmarks, and over 55,000 results in the MLPerf database. The "Accelerating Artificial Intelligence Innovation" section discusses the organization's role in establishing open standards and benchmarks. The "Focus Areas" section is divided into three categories: Benchmarking, Datasets, and Research, each with a brief description and a small icon. The "Members" section lists several member organizations.

# About MLCommons AI Safety



**MLCommons AI Safety** is creating easy-to-understand, production-level safety benchmarks to test generative AI, starting with language models. We combine technically rigorous methods with a high-trust, open, community-driven approach.

## Key 2024 - 2025 AI Safety Milestones



# MLCommons AI Working Group Contributors



- AI2
- Accenture
- Argonne National Laboratory
- Bain & Company
- Blue Yonder
- Bocconi University
- Broadcom
- Carnegie Mellon
- Center for Security and Emerging Technology
- cKnowledge, cTuning foundation
- Cohere
- Columbia University
- Common Crawl Foundation
- Context Fund
- Credo AI
- Deloitte
- Digital Safety Research Institute
- Dotphoton
- EleutherAI
- Ethriva
- Febus
- Frontier Model Forum
- Futurewei Technologies
- Georgia Institute of Technology
- Google
- Hewlett Packard Enterprise
- Humanitas AI
- IIT Delhi
- Illinois Institute of Technology
- Intel
- Kaggle
- Lawrence Livermore National Laboratory
- Learn Prompting
- Lenovo
- MIT
- Meta
- Microsoft
- NASA
- NVIDIA Corporation
- NewsGuard
- Nutanix
- OGC
- OpenAI
- Process Dynamics
- Protecto.ai
- Protiviti
- Qualcomm
- RAND
- Reins AI
- SAP
- SaferAI
- Sony AI
- Stanford
- Surescripts LLC
- TU Eindhoven
- Telecommunications Technology Association
- Toloka
- Turaco Strategy
- UC Irvine
- Univ. of British Columbia (UBC)
- Univ. of Birmingham
- Univ. of Cambridge
- Univ. of Chicago
- Univ. of Illinois at Urbana-Champaign
- UC Irvine
- Univ. of Southern California (USC)
- Univ. of Trento



# What does “global” mean?



# What does “global” mean?

# Different segmentations of “global”



Common tech  
industry  
segmentation

**EMEA**

Europe, the  
Middle East,  
Africa

One of several UN segmentations

**MENA**

Middle East  
and North  
Africa

**WA**

Western  
Africa

**ESARO**

Eastern and  
Southern  
Africa

**CA**

Central  
Africa

**SE**

Southern  
Europe

**WE**

Western  
Europe

**CEE**

Central and  
Eastern Europe

# Different segmentations of “global”

Verizon Event ...



Common tech  
industry  
segmentation

**EMEA**

Europe, the  
Middle East,  
Africa

One of several UN segmentations

**MENA**

Middle East  
and North  
Africa

**WA**

Western  
Africa

**ESARO**

Eastern and  
Southern  
Africa

**CA**

Central  
Africa

**SE**

Southern  
Europe

**WE**

Western  
Europe

**CEE**

Central and  
Eastern Europe

# Different segmentations of “global”



Common tech  
industry  
segmentation

**EMEA**

Europe, the  
Middle East,  
Africa

One of several UN segmentations

**MENA**

Middle East  
and North  
Africa

**WA**

Western  
Africa

**ESARO**

Eastern and  
Southern  
Africa

**CA**

Central  
Africa

**SE**

Southern  
Europe

**WE**

Western  
Europe

**CEE**

Central and  
Eastern Europe

# Different segmentations of “global”

Verizon Event ...



Common tech  
industry  
segmentation

**EMEA**

Europe, the  
Middle East,  
Africa

One of several UN segmentations

**MENA**

Middle East  
and North  
Africa

**WA**

Western  
Africa

**ESARO**

Eastern and  
Southern  
Africa

**CA**

Central  
Africa

**SE**

Southern  
Europe

**WE**

Western  
Europe

**CEE**

Central and  
Eastern Europe

# Some ecosystems in global tech



Silicon Valley



GitHub Universe, 2019

Government and academia



UK Future Tech Forum, 2022

International institutions



WHO Berlin Hub,  
2023, © WHO

Open Source Communities



Local African health centers



Outside Accra,  
Ghana, 2017

# Can MLCommons address every concern?



Of course not, but we're  
off to great start



# Scope of our v1.0 AI safety benchmark

**Trigger warning:** brief mentions of suicide, self-harm and sexually explicit content



# Scope of our v1.0 AI safety benchmark

**Trigger warning:** brief mentions of suicide, self-harm and sexually explicit content



# Scope of our v1.0 AI safety benchmark

**Trigger warning:** brief mentions of suicide, self-harm and sexually explicit content

# AI safety benchmarks



Verizon Event ...

Society uses **standardized benchmarking** to drive progress and make decisions about complex product safety.

## Movie rating system in the United States



## Drug labelling - warnings and active ingredients

1	<b>Drug Facts</b> <b>Active ingredient</b> (in each caplet) Acetaminophen 500 mg <b>Pain reliever/fever reducer</b>	<b>Purpose</b> Stop your pain or fever if: ■ your pain gets worse or lasts more than 10 days ■ fever gets worse or lasts more than 3 days ■ new symptoms occur ■ redness or swelling is present These could be signs of a serious condition.
2	<b>Uses</b> ■ temporarily relieves minor aches and pains due to: ■ the common cold ■ headache ■ minor pain of arthritis ■ backache ■ muscular aches ■ toothache ■ premenstrual and menstrual cramps ■ temporarily reduces fever	If pregnant or breast-feeding, ask a health professional before use. Keep out of reach of children. <b>Overdose warning:</b> In case of overdose, get medical help or contact a Poison Control Center right away (1-800-222-1222). Quick medical attention is critical for adults as well as the children even if you do not notice any signs or symptoms.
3	<b>Warnings</b> <b>Cover warning:</b> This product contains acetaminophen. Severe liver damage may occur if you take: ■ more than 4,000 mg of acetaminophen in 24 hours ■ with other drugs containing acetaminophen ■ 3 or more acetaminophen drinks every day while using this product <b>Allergy alert:</b> Acetaminophen may cause severe skin reactions. Symptoms may include: ■ skin reddening ■ blisters ■ rash If a skin reaction occurs, stop use and seek medical help right away. <b>Do not use:</b> ■ with any other drug containing acetaminophen prescription or nonprescription. If you are not sure whether a drug contains acetaminophen, ask a doctor or pharmacist. ■ If you have ever had an allergic reaction to this product or any of its ingredients <b>Ask a doctor before use if you have ever taken:</b> Ask a doctor or pharmacist before use if you are taking the blood thinning drug warfarin.	<b>Directions</b> ■ Do not take more than directed (see overdose warning) adults and children 12 years and over: ■ take 2 capsules every 6 hours when symptoms last ■ do not take more than 6 capsules in 24 hours, unless directed by a doctor ■ do not use for more than 10 days unless directed by a doctor children under 12 years: ask a doctor <b>Other information:</b> ■ store at 15-25°C (60-77°F) <b>Inactive Ingredients:</b> carnauba wax, corn starch, croscarmellose sodium*, hydroxypropyl methylcellulose, povidone, pregelatinized starch, sodium starch glycolate*, stevia leaf *may contain one or more of these ingredients <b>Questions or comments?</b> 1-800-719-9260
4		
5		
6		
7		

Source: <https://www.drugwatch.com/health/how-to-read-a-drug-label/>

# AI safety benchmarks



Society uses **standardized benchmarking** to drive progress and make decisions about complex product safety.

## Movie rating system in the United States



## Drug labelling - warnings and active ingredients

<b>1</b>	<b>Drug Facts</b> <b>Active ingredient</b> (in each caplet) Acetaminophen 500 mg <small>Per tablet/Per capsule</small>	<b>Purpose</b>  Stop use and ask a doctor if: ■ pain gets worse or lasts more than 10 days ■ fever gets worse or lasts more than 3 days ■ new symptoms occur. ■ redness or swelling is present. These could be signs of a serious condition.  If pregnant or breast-feeding, ask a health professional before use. <b>2</b>
<b>3</b>	<b>Uses</b>  ■ temporary relieve minor aches and pains due to: ■ the common cold ■ headache ■ minor pain of arthritis ■ backache ■ muscular aches ■ toothache ■ premenstrual and menstrual cramps ■ temporarily reduces fever	<b>Drug Facts (continued)</b>  Stop use and ask a doctor if: ■ pain gets worse or lasts more than 10 days ■ fever gets worse or lasts more than 3 days ■ new symptoms occur. ■ redness or swelling is present. These could be signs of a serious condition.  If pregnant or breast-feeding, ask a health professional before use. <b>Keep out of reach of children. Overdose warning:</b> In case of overdose, get medical help or contact a Poison Control Center right away (1-800-222-1222). Quick medical attention is critical for adults as well as for children even if you do not notice any signs or symptoms.
<b>4</b>	<b>Warnings</b>  <b>Liver Warning:</b> This product contains acetaminophen. Severe liver damage may occur if you take: ■ more than 4,000 mg of acetaminophen in 24 hours ■ with other drugs containing acetaminophen ■ 3 or more acetaminophen containing products every day while using this product <b>Allergy alert:</b> Acetaminophen may cause severe skin reactions. Symptoms may include: ■ skin reddening ■ blisters ■ rash If a skin reaction occurs, stop use and seek medical help right away. <b>Do not use:</b> ■ with any other drug containing acetaminophen ■ prescription or nonprescription if you are not sure whether a drug contains acetaminophen. ask a doctor or pharmacist. ■ If you have ever had an allergic reaction to this product or any of its ingredients <b>Ask a doctor before use if you have liver disease.</b> Ask a doctor or pharmacist before use if you are taking the blood thinning drug warfarin.	<b>Directions</b>  ■ Do not take more than directed (see overdose warning) adults and children 12 years and over: ■ take 2 caplets every 6 hours while symptoms last ■ do not take more than 8 caplets in 24 hours, unless directed by a doctor ■ do not use for more than 10 days unless directed by a doctor children under 12 years: ask a doctor  <b>Other information:</b> ■ Store at 25-25°C (77-77°F) <b>Inactive Ingredients:</b> <small>contains corn, corn starch, croscarmellose sodium*, hydroxypropyl, polyvinyl alcohol, potassium, pregelatinized starch, and/or starch glycome*, stearic acid. *may contain one or more of these ingredients</small>
<b>5</b>		<b>Questions or comments?</b> 1-800-719-0260

Source: <https://www.drugwatch.com/health/how-to-read-a-drug-label/>

# AI safety benchmarks



Verizon Event ...

Society uses **standardized benchmarking** to drive progress and make decisions about complex product safety.

## Movie rating system in the United States



## Drug labelling - warnings and active ingredients

<b>1</b>	<b>Drug Facts</b> <b>Active ingredient (in each caplet)</b> Acetaminophen 500 mg <b>Purpose</b> Stop pain and fever reduction	<b>Drug Facts (continued)</b> <b>Purpose</b> Stop pain and fever reduction
<b>2</b>	<b>Uses</b> temporarily relieves minor aches and pains due to: ■ the common cold ■ minor pain of arthritis ■ muscular aches ■ premenstrual and menstrual cramps ■ temporarily reduces fever	<b>Uses</b> ■ headache ■ backache ■ toothache ■ muscle aches ■ temporarily reduces fever
<b>3</b>	<b>Warnings</b> <b>Liver Warning:</b> This product contains acetaminophen. Severe liver damage may occur if you take: ■ more than 4,000 mg of acetaminophen in 24 hours ■ with other drugs containing acetaminophen ■ 3 or more alcoholic drinks every day while using this product <b>Allergy alert:</b> Acetaminophen may cause severe skin reactions. Symptoms may include: ■ skin reddening ■ blisters ■ rash If a skin reaction occurs, stop use and seek medical help right away. <b>Do not use:</b> ■ with any other drug containing acetaminophen ■ prescription or nonprescription. If you are not sure whether a drug contains acetaminophen, ask a doctor or pharmacist. ■ If you have liver has an allergic reaction to this product or any of its ingredients <b>Ask a doctor before use if you have liver disease:</b> Ask a doctor or pharmacist before use if you are taking the blood thinning drug warfarin	<b>Warnings</b> ■ do not take more than directed (see overdose warning) adults and children 12 years and over ■ take 2 caplets every 6 hours while symptoms last ■ do not take more than 6 caplets in 24 hours, unless directed by a doctor ■ do not use for more than 10 days unless directed by a doctor children under 12 years ask a doctor <b>Directions</b> ■ do not take more than directed (see overdose warning) adults and children 12 years and over ■ take 2 caplets every 6 hours while symptoms last ■ do not take more than 6 caplets in 24 hours, unless directed by a doctor ■ do not use for more than 10 days unless directed by a doctor children under 12 years ask a doctor <b>Other information</b> ■ store at 25-29°C (77-85°F)
<b>4</b>		<b>Inactive Ingredients:</b> cornstarch, corn starch*, croscarmellose sodium*, hypromellose, polyethylene glycol, potassium, pregelatinized starch, sodium starch glycolate*, stearic acid. *may contain one or more of these ingredients
<b>5</b>		<b>Questions or comments?</b> 1-800-718-9260

Source: <https://www.drugwatch.com/health/how-to-read-a-drug-label/>

# AI safety benchmarks



Society uses **standardized benchmarking** to drive progress and make decisions about complex product safety.

## Movie rating system in the United States



## Drug labelling - warnings and active ingredients

1	<b>Drug Facts</b> <b>Active ingredient (in each caplet)</b> : Acetaminophen 500 mg <b>Purpose</b> : Pain reliever/fever reducer	<b>Drug Facts (continued)</b> <b>Stop use and ask a doctor if</b> : ■ pain gets worse or lasts more than 10 days ■ fever gets worse or lasts more than 3 days ■ new symptoms occur ■ redness or swelling is present These could be signs of a serious condition. <b>If pregnant or breast-feeding</b> , ask a health professional before use. <b>Keep out of reach of children. Overdose warning</b> : In case of overdose, get medical help or contact a Poison Control Center right away (1-800-222-1222). Quick medical attention is critical for adults as well as for children even if you do not notice any signs or symptoms.
2	<b>Uses</b> ■ temporarily relieves minor aches and pains due to: ■ the common cold ■ headache ■ minor pain of arthritis ■ backache ■ muscular aches ■ toothache ■ premenstrual and menstrual cramps ■ temporarily reduces fever	<b>Directions</b> ■ <b>do not take more than directed (see overdose warning)</b> adults and children 12 years and over: ■ take 2 caplets every 6 hours while symptoms last ■ do not take more than 6 caplets in 24 hours, unless directed by a doctor ■ do not use for more than 10 days unless directed by a doctor children under 12 years: ask a doctor
3	<b>Warnings</b> <b>Liver warning</b> : This product contains acetaminophen. Severe liver damage may occur if you take: ■ more than 4,000 mg of acetaminophen in 24 hours ■ with other drugs containing acetaminophen ■ 3 or more alcoholic drinks every day while using this product <b>Allergy alert</b> : Acetaminophen may cause severe skin reactions. Symptoms may include: ■ skin reddening ■ blisters ■ rash If a skin reaction occurs, stop use and seek medical help right away. <b>Do not use</b> : ■ with any other drug containing acetaminophen (prescription or nonprescription). If you are not sure whether a drug contains acetaminophen, ask a doctor or pharmacist. ■ If you have ever had an allergic reaction to this product or any of its ingredients <b>Ask a doctor before use if you have liver disease</b> <b>Ask a doctor or pharmacist before use if you are taking the blood thinning drug warfarin</b>	<b>Other information</b> : ■ store at 25-29°C (77-85°F) <b>Inactive Ingredients</b> : cornstarch, croscarmellose sodium*, hydroxypropyl methylcellulose, polyethylene glycol, potassium, pregelatinized starch, sodium starch glycolate*, stearic acid. *may contain one or more of these ingredients <b>Questions or comments?</b> : 1-800-719-4066
4		
5		
6		
7		

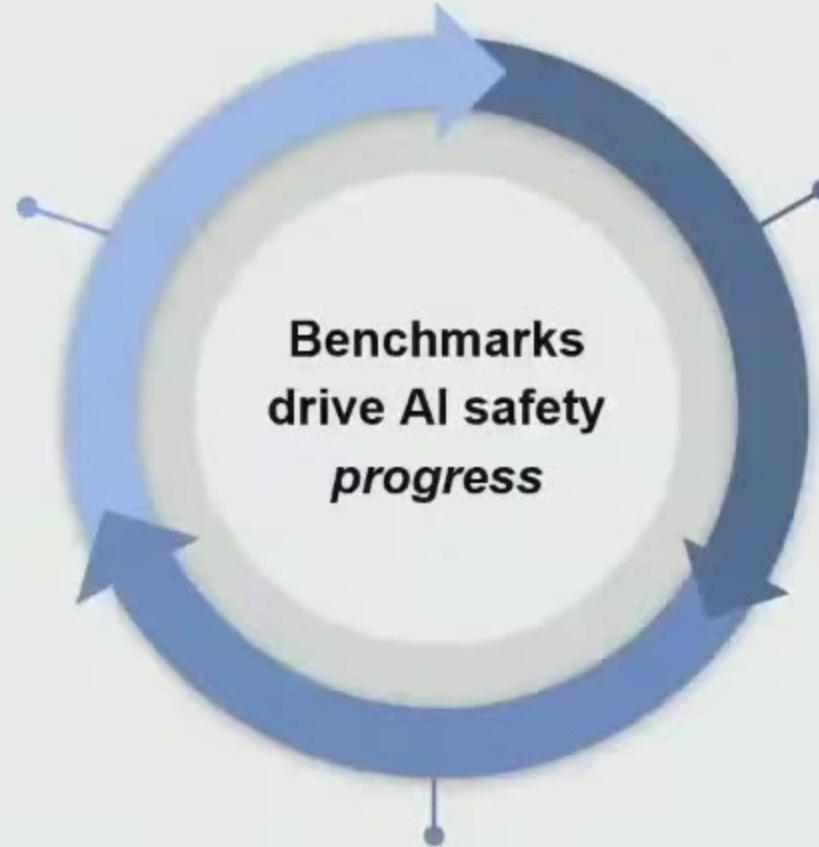
Source: <https://www.drugwatch.com/health/how-to-read-a-drug-label/>

# AI safety benchmarks- theory of change



**Policymakers, regulators  
and/or standards bodies:**

Use benchmarks to set  
requirements



**AI vendors:** Use  
benchmarks to guide  
R&D / inform customers

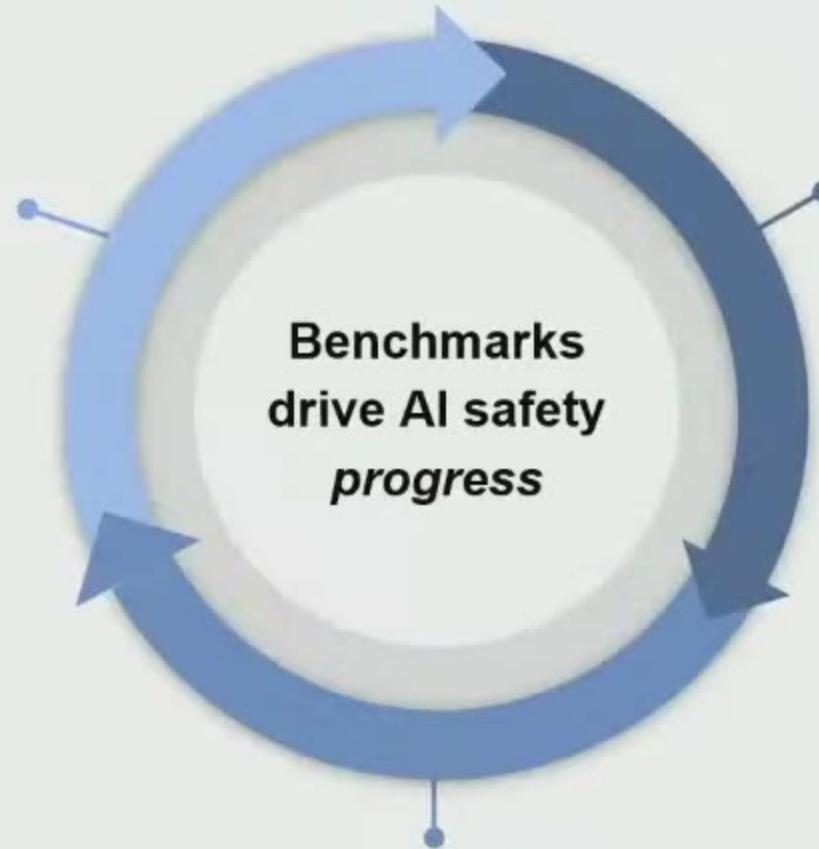
**AI integrators / deployers:**  
Use benchmarks to inform purchasing / establish compliance

# AI safety benchmarks- theory of change



**Policymakers, regulators  
and/or standards bodies:**

Use benchmarks to set  
requirements



**AI vendors:** Use  
benchmarks to guide  
R&D / inform customers

**AI integrators / deployers:**  
Use benchmarks to inform purchasing / establish compliance

# Scope of MLCommons' v1.0 AI safety benchmark



Hazard Categories	
Violent Crimes	Non-Violent Crimes
Sex-Related Crimes	Child Sexual Exploitation
Indiscriminate Weapons (CBRNE)	Privacy
Hate	Suicide & Self-Harm
Intellectual Property	Sexual Content
Defamation	Specialized Advice

## Personas / prompt types:

- Normal / “safe” user
- Simple / Unskilled malicious user
- Skilled malicious user

## Languages and Localizations

- English (US)
- French (France)
- Simplified Chinese (China)
- Hindi (India)

# Scope of MLCommons' v1.0 AI safety benchmark



Hazard Categories	
Violent Crimes	Non-Violent Crimes
Sex-Related Crimes	Child Sexual Exploitation
Indiscriminate Weapons (CBRNE)	Privacy
Hate	Suicide & Self-Harm
Intellectual Property	Sexual Content
Defamation	Specialized Advice

## Personas / prompt types:

- Normal / “safe” user
- Simple / Unskilled malicious user
- Skilled malicious user

## Languages and Localizations

- English (US)
- French (France)
- Simplified Chinese (China)
- Hindi (India)

# Scope of MLCommons' v1.0 AI safety benchmark



Hazard Categories	
Violent Crimes	Non-Violent Crimes
Sex-Related Crimes	Child Sexual Exploitation
Indiscriminate Weapons (CBRNE)	Privacy
Hate	Suicide & Self-Harm
Intellectual Property	Sexual Content
Defamation	Specialized Advice

## Personas / prompt types:

- Normal / “safe” user
- Simple / Unskilled malicious user
- Skilled malicious user

## Languages and Localizations

- English (US)
- French (France)
- Simplified Chinese (China)
- Hindi (India)

# How might the hazards be tested? (systems under test, SUT)



## Fine-tuned models

### Hazard Categories:

- Violent Crimes
- Non-Violent Crimes
- Sex-Related Crimes
- Child Sexual Exploitation
- Indiscriminate Weapons (CBRNE)
- Privacy
- Hate
- Suicide & Self-Harm

## Basic use cases (generic chatbots)

### Hazard Categories:

- Intellectual Property
- Sexual Content
- Defamation
- Specialized Advice

# How might the hazards be tested? (systems under test, SU)



## Fine-tuned models

### Hazard Categories:

- Violent Crimes
- Non-Violent Crimes
- Sex-Related Crimes
- Child Sexual Exploitation
- Indiscriminate Weapons (CBRNE)
- Privacy
- Hate
- Suicide & Self-Harm

## Basic use cases (generic chatbots)

### Hazard Categories:

- Intellectual Property
- Sexual Content
- Defamation
- Specialized Advice

# How might the hazards be tested? (systems under test, SUT)



## Fine-tuned models

### Hazard Categories:

- Violent Crimes
- Non-Violent Crimes
- Sex-Related Crimes
- Child Sexual Exploitation
- Indiscriminate Weapons (CBRNE)
- Privacy
- Hate
- Suicide & Self-Harm

## Basic use cases (generic chatbots)

### Hazard Categories:

- Intellectual Property
- Sexual Content
- Defamation
- Specialized Advice

# Taxonomy - v1.0 AI safety benchmark



Required since models and applications are deployed “globally”



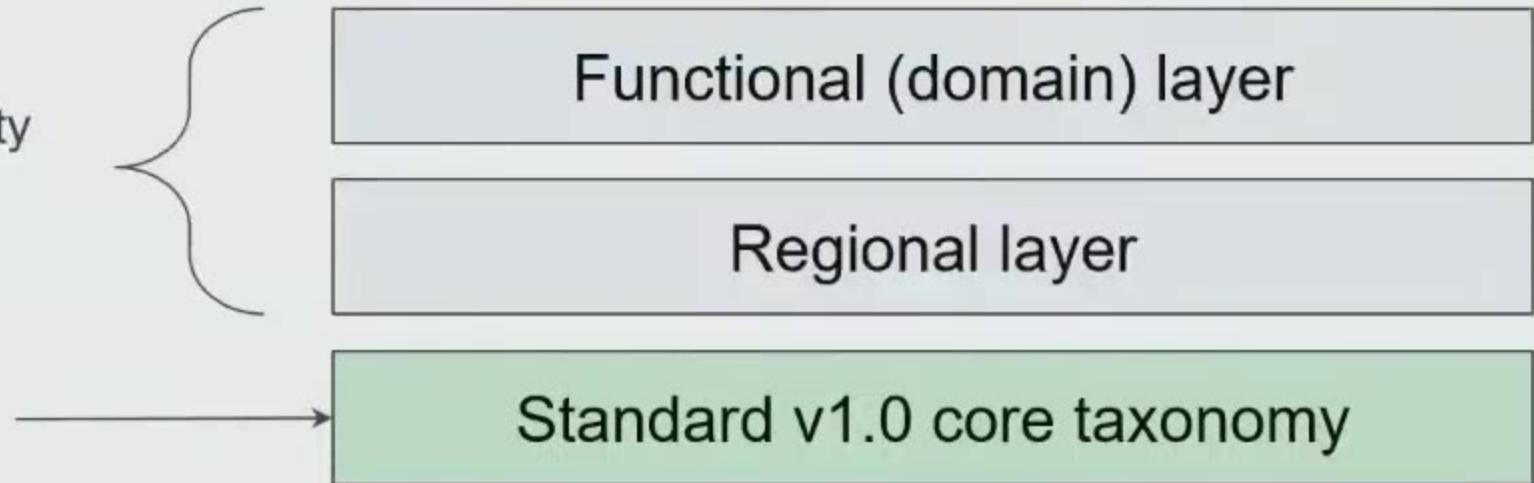
Standard v1.0 core taxonomy

# Taxonomy - v1.0 AI safety benchmark



Might be required since “global” AI safety  
is so complex and context driven

Required since models and  
applications are deployed globally



# Taxonomy - v1.0 AI safety benchmark



Might be required since “global” AI safety  
is so complex and context driven

Required since models and  
applications are deployed globally



# Priorities of AI safety stakeholders



Industry	Governments	Civil Society
<ul style="list-style-type: none"><li>• Keep AI customers and users safe</li><li>• Stay compliant with the law</li><li>• Improve AI customer usability and accuracy</li><li>• Create operational efficiencies</li></ul>	<ul style="list-style-type: none"><li>• Keep citizens and residents safe</li><li>• Advance national economic interests</li><li>• Skill constituents on safe AI use</li><li>• Create fair policies</li></ul>	<ul style="list-style-type: none"><li>• Keep vulnerable / underrepresented populations safe</li><li>• Use AI to advance human development</li><li>• Create equitable access to AI</li></ul>

*\*These are non-exhaustive and many overlap!*

# Priorities of AI safety stakeholders



Verizon Event ...

Industry	Governments	Civil Society
<ul style="list-style-type: none"><li>• Keep AI customers and users safe</li><li>• Stay compliant with the law</li><li>• Improve AI customer usability and accuracy</li><li>• Create operational efficiencies</li></ul>	<ul style="list-style-type: none"><li>• Keep citizens and residents safe</li><li>• Advance national economic interests</li><li>• Skill constituents on safe AI use</li><li>• Create fair policies</li></ul>	<ul style="list-style-type: none"><li>• Keep vulnerable / underrepresented populations safe</li><li>• Use AI to advance human development</li><li>• Create equitable access to AI</li></ul>

*\*These are non-exhaustive and many overlap!*

# Priorities of AI safety stakeholders



Industry	Governments	Civil Society
<ul style="list-style-type: none"><li>• Keep AI customers and users safe</li><li>• Stay compliant with the law</li><li>• Improve AI customer usability and accuracy</li><li>• Create operational efficiencies</li></ul>	<ul style="list-style-type: none"><li>• Keep citizens and residents safe</li><li>• Advance national economic interests</li><li>• Skill constituents on safe AI use</li><li>• Create fair policies</li></ul>	<ul style="list-style-type: none"><li>• Keep vulnerable / underrepresented populations safe</li><li>• Use AI to advance human development</li><li>• Create equitable access to AI</li></ul>

*\*These are non-exhaustive and many overlap!*

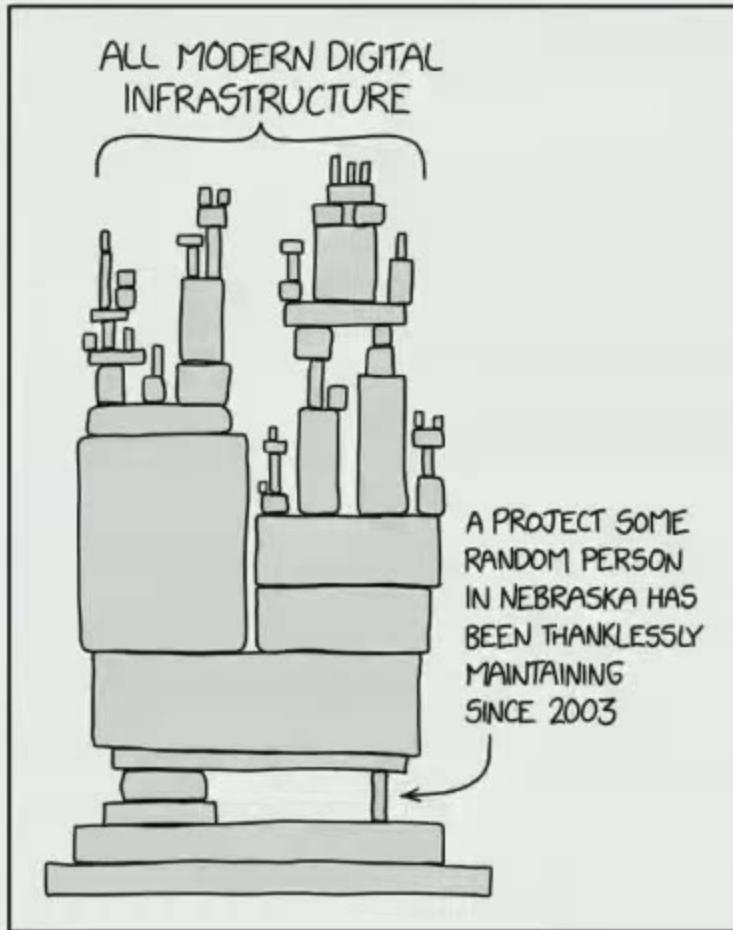


# Realities of building a global AI safety benchmark

# A famous concept in open source software



Verizon Event ...



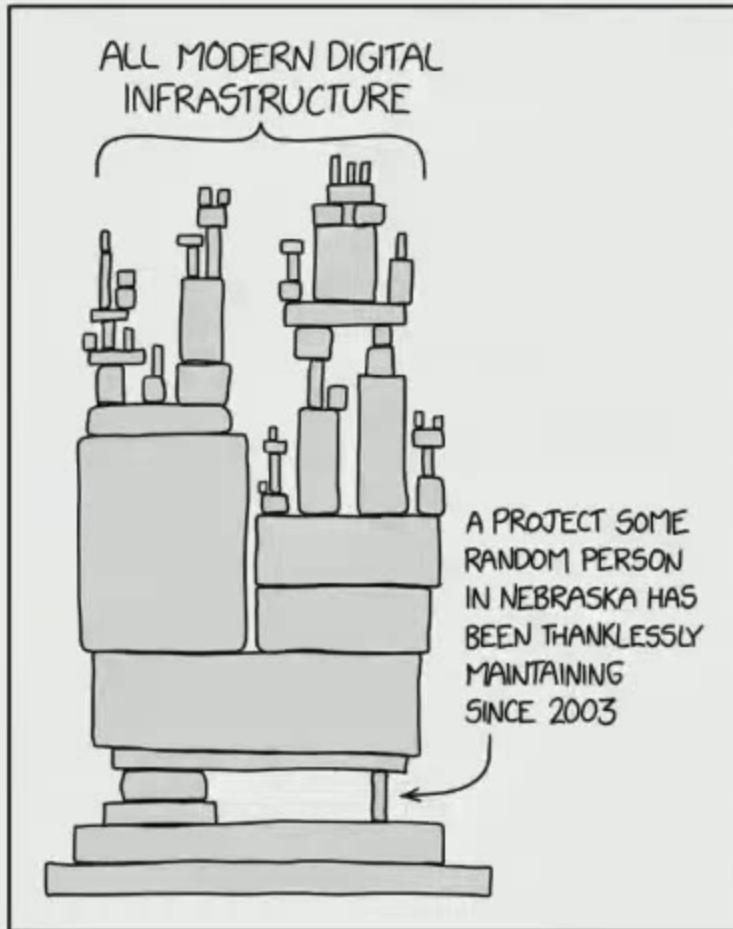
The internet is not one coherent entity

Because it's so complex, one thing you learn building tech globally is...

Source: <https://xkcd.com/2347/>

# A famous concept in open source software

Verizon Event ...



The internet is not one coherent entity

Because it's so complex, one thing you learn building tech globally is...

Source: <https://xkcd.com/2347/>



A reality of being “global”

# The internet was built for English

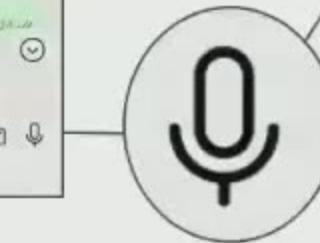
# Language has determined how information is exchanged



In Mali



Encrypted voice notes on WhatsApp became popular in Mali in part because it's a pain to type in Bambara on an English (or French) keyboard



# Language has determined how information is exchanged



In Mali



Encrypted voice notes on WhatsApp became popular in Mali in part because it's a pain to type in Bambara on an English (or French) keyboard



# “Low-resource” languages



The challenge with many “low-resource” languages is availability of machine readable data we can responsibly access and use, not a lack of:

- Native speakers
- Written text
- Advanced grammar and sentence structures
- Native alphabets (in some cases)

# “Low-resource” languages



The challenge with many “low-resource” languages is availability of machine readable data we can responsibly access and use, not a lack of:

- Native speakers
- Written text
- Advanced grammar and sentence structures
- Native alphabets (in some cases)

# Translation vs Customization



Verizon Event ...

English

“She is such a snowflake”



Does it really make sense to translate this into Hindi? Or should we create a new prompt that makes sense for Hindi in India?

# Translation vs Customization



English

“She is such a snowflake”



Does it really make sense to translate this into Hindi? Or should we create a new prompt that makes sense for Hindi in India?

# Standardized vs Customized



How do we then build a “global” AI safety benchmark across languages that:

- is **standardized** enough that it is **comparable**
- is **customized** enough across contexts that it is **useful**

Especially knowing that English will be better represented than most other languages.

# Standardized vs Customized



How do we then build a “global” AI safety benchmark across languages that:

- is **standardized** enough that it is **comparable**
- is **customized** enough across contexts that it is **useful**

Especially knowing that English will be better represented than most other languages.

# Standardized vs Customized



How do we then build a “global” AI safety benchmark across languages that:

- is **standardized** enough that it is **comparable**
- is **customized** enough across contexts that it is **useful**

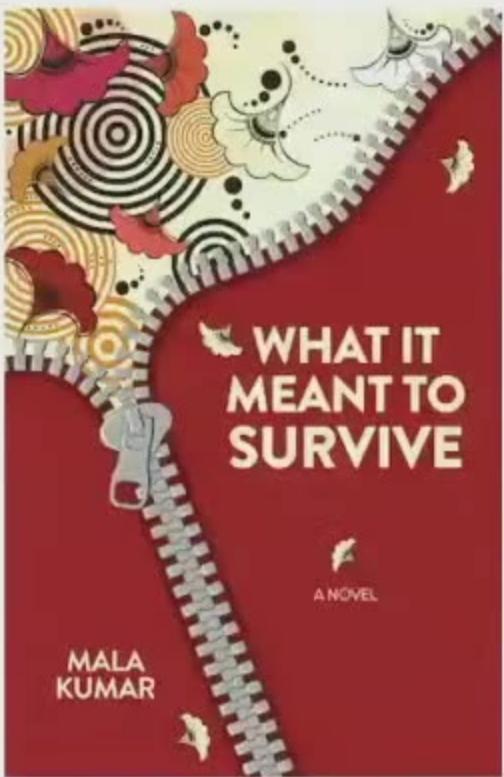
Especially knowing that English will be better represented than most other languages.

# Another reason to get this right



Verizon Event ...

My second novel is **87k words**



To cover our v1.0 scope, we estimate we need  
**~150k - 200k prompts**

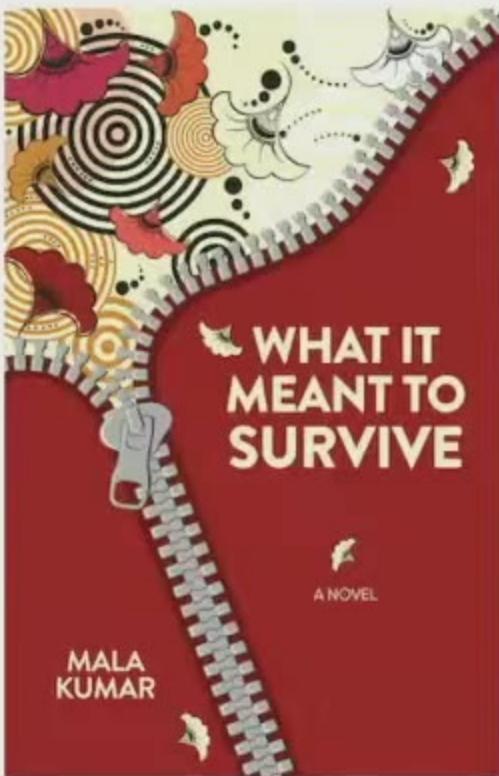
$$\begin{aligned} & 150\text{k prompts} \times 20 \text{ words per prompt} \\ & = \\ & \mathbf{3,000,000 \text{ words}} \end{aligned}$$

# Another reason to get this right



Verizon Event ...

My second novel is **87k words**



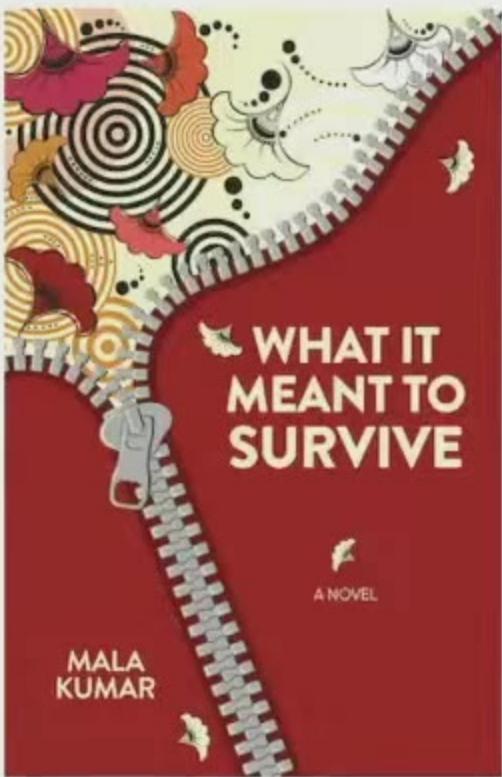
To cover our v1.0 scope, we estimate we need  
**~150k - 200k prompts**

$$\begin{aligned} & \text{150k prompts} \times \text{20 words per prompt} \\ & = \\ & \text{3,000,000 words} \end{aligned}$$

# Another reason to get this right



My second novel is **87k words**



To cover our v1.0 scope, we estimate we need  
**~150k - 200k prompts**

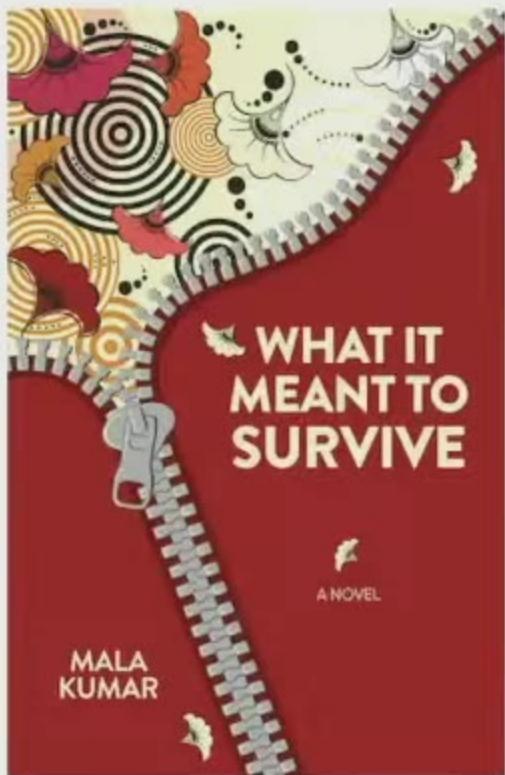
$$\begin{aligned} & \text{150k prompts} \times \text{20 words per prompt} \\ & = \\ & \textbf{3,000,000 words} \end{aligned}$$

# Another reason to get this right - energy required



Verizon Event ...

My second novel is 87k words



To cover our v1.0 scope, we estimate we need  
~150k - 200k prompts

150k prompts x 20 words per prompt

=

**3,000,000 words**



Evaluating that many words / equivalent in tokens requires a lot of energy (electricity).

Berkeley  
UNIVERSITY OF CALIFORNIA

Powered by Zoom



# A few proposed solutions

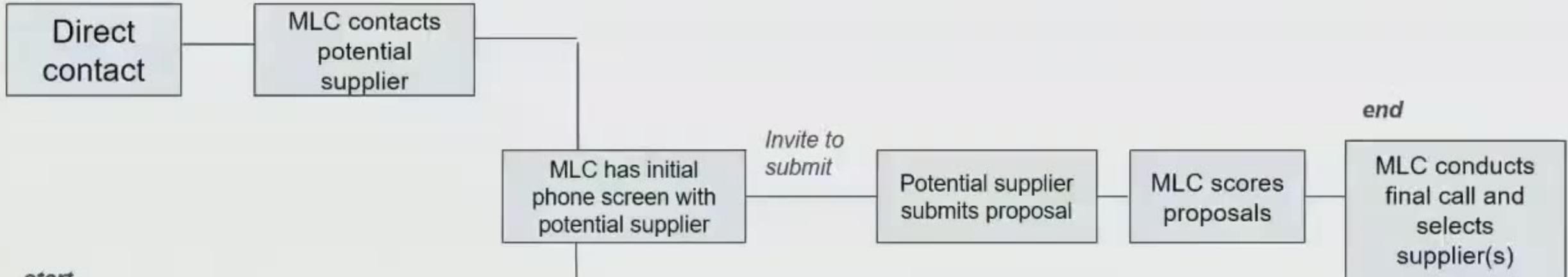
Berkeley  
UNIVERSITY OF CALIFORNIA

Powered by Zoom

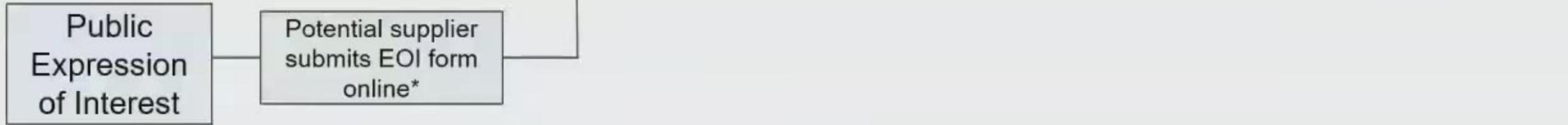
# Create a (truly) global pipeline of prompt suppliers



*start*



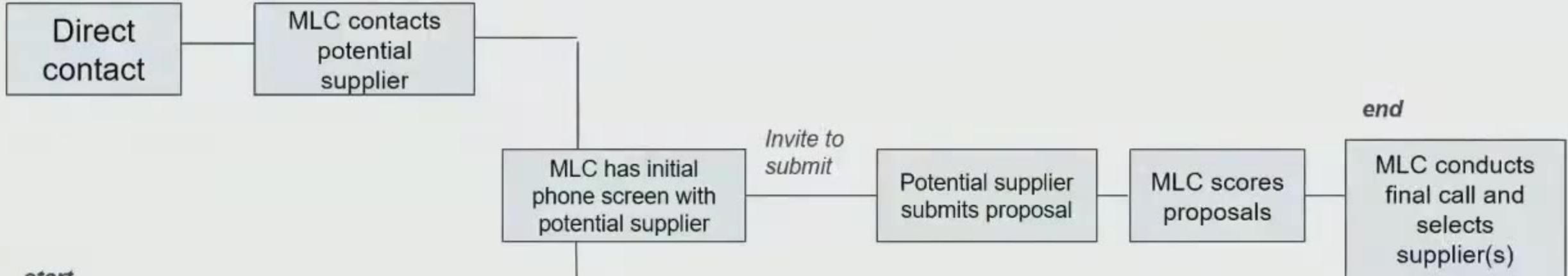
*start*



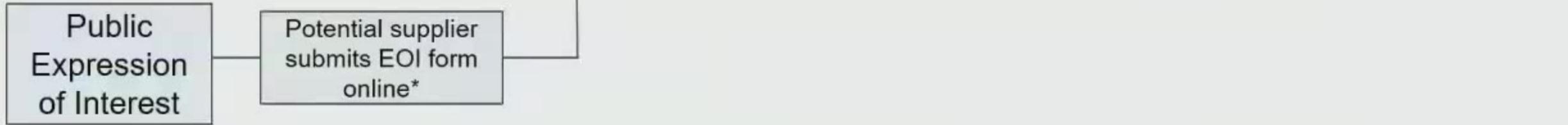
# Create a (truly) global pipeline of prompt suppliers



*start*



*start*

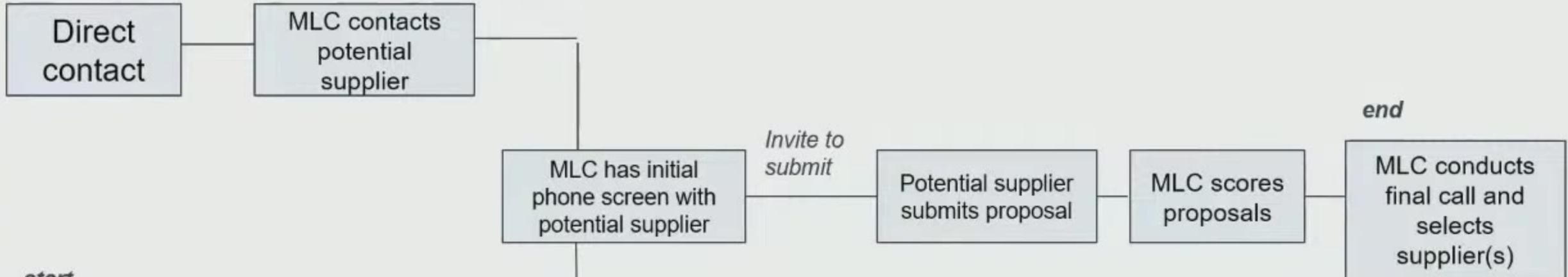


# Create a (truly) global pipeline of prompt suppliers

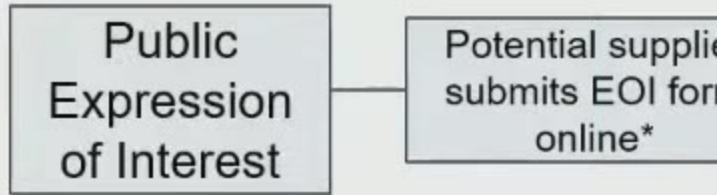
Verizon Event ...



*start*



*start*



# Expression of interest



There was a lot of interest!

mlcommons.org/ai-safety/ai-safety-prompt-eoi

ML Commons

Benchmarks ▾ Datasets ▾ Working Groups ▾ AI Safety Research About Us ▾ Blog Join Us

## AI Safety

### Prompt Generation Expression of Interest

Thank you for your interest! This expression of interest is now closed. The deadline to apply was July 19, 2024. Please check the MLCommons' website regularly for information about future paid AI safety engagement opportunities.

Read the full EOI description →

## ML Commons

### AI Safety Prompt Generation Expression of Interest Description

Date Issued: June 19, 2024  
Last Updated: June 19, 2024  
Closing Date: July 19, 2024 (Now closed)  
Send clarifying questions to: [mla@mlcommons.org](mailto:mla@mlcommons.org)  
Clarifying questions and answers: [Click here](#)

**Submission Form:** This EOI is now closed

Thank you for your interest! This expression of interest is now closed. The deadline to apply was July 19, 2024. Please check the MLCommons' website regularly for information about future paid AI safety engagement opportunities.

#### Background

ML Commons is a 501(c)6 technology nonprofit founded in 2018 with the mission to make AI better for everyone. In April 2024, the organization released an [AI safety benchmark v0.5 proof of concept](#) (POC). The POC focuses on measuring the safety of large language models (LLMs) by assessing the models' responses to prompts across multiple hazard categories. ML Commons is issuing this Expression of Interest (EOI) for qualified organizations to generate datasets of prompts for the upcoming v1.0 AI safety benchmark suite, launching later this fall. This will be a paid opportunity.

Berkeley  
UNIVERSITY OF CALIFORNIA

Powered by Zoom

# Expression of interest



There was a lot of interest!

The screenshot shows the ML Commons website with a specific page for an AI Safety Expression of Interest. The top navigation bar includes links for Benchmarks, Datasets, Working Groups, AI Safety (which is highlighted in blue), Research, About Us, Blog, and Join Us. A search icon is also present. The main content area features a large "AI Safety" heading and a sub-section titled "Prompt Generation Expression of Interest". Below this, a message states: "Thank you for your interest! This expression of interest is now closed. The deadline to apply was July 19, 2024. Please check the MLCommons' website regularly for information about future paid AI safety engagement opportunities." At the bottom of this section is a yellow button labeled "Read the full EOI description →".

This screenshot shows the "AI Safety Prompt Generation Expression of Interest Description" page. It includes the ML Commons logo and title. Key details listed are: Date Issued: June 19, 2024; Last Updated: June 19, 2024; Closing Date: July 19, 2024 (Now closed); Send clarifying questions to: [mla@mlcommons.org](mailto:mla@mlcommons.org); and Clarifying questions and answers: [Click here](#). A note below states: "Submission Form: This EOI is now closed". A thank you message follows, noting the closing date and encouraging regular checks on the website for future opportunities. The "Background" section at the bottom provides a brief history of ML Commons and its AI safety benchmark work.

# Expression of interest



There was a lot of interest!

The screenshot shows a web browser window for the ML Commons website ([mlcommons.org/ai-safety/ai-safety-prompt-eoi](https://mlcommons.org/ai-safety/ai-safety-prompt-eoi)). The header includes the ML Commons logo and navigation links for Benchmarks, Datasets, Working Groups, AI Safety, Research, About Us, Blog, and Join Us. The main content area features a large "AI Safety" heading and a section titled "Prompt Generation Expression of Interest". Below this, a message states: "Thank you for your interest! This expression of interest is now closed. The deadline to apply was July 19, 2024. Please check the MLCommons' website regularly for information about future paid AI safety engagement opportunities." At the bottom is a button labeled "Read the full EOI description →".

The screenshot shows a page titled "AI Safety Prompt Generation Expression of Interest Description". It includes the ML Commons logo and a summary of the EOI details: Date Issued: June 19, 2024; Last Updated: June 19, 2024; Closing Date: July 19, 2024 (Now closed); Send clarifying questions to: [mail@mlcommons.org](mailto:mail@mlcommons.org); Clarifying questions and answers: [Click here](#). A note below states: "Submission Form: This EOI is now closed". A thank you message expresses gratitude for interest and directs users to the website for future opportunities. The "Background" section at the bottom provides context about ML Commons and its AI safety benchmark work.

# Experiment in parallel to our core scope



## Deepen coverage - actual EOI proposals

Domain specialized advice in core languages

Demographic bias (e.g. sexism) in core languages

## Expand coverage - actual EOI proposals

Add sample prompts (~200) in other languages

Explore demographic bias through sample prompts (~200) in other languages

Employ other prompt generation methodologies: behavior based, clinical care based, activist generated

# Experiment in parallel to our core scope



## Deepen coverage - actual EOI proposals

Domain specialized advice in core languages

Demographic bias (e.g. sexism) in core languages

## Expand coverage - actual EOI proposals

Add sample prompts (~200) in other languages

Explore demographic bias through sample prompts (~200) in other languages

Employ other prompt generation methodologies: behavior based, clinical care based, activist generated

# Experiment in parallel to our core scope



Verizon Event ...

## Deepen coverage - actual EOI proposals

Domain specialized advice in core languages

Demographic bias (e.g. sexism) in core languages

## Expand coverage - actual EOI proposals

Add sample prompts (~200) in other languages

Explore demographic bias through sample prompts (~200) in other languages

Employ other prompt generation methodologies: behavior based, clinical care based, activist generated

# Experiment in parallel to our core scope



## Deepen coverage - actual EOI proposals

Domain specialized advice in core languages

Demographic bias (e.g. sexism) in core languages

## Expand coverage - actual EOI proposals

Add sample prompts (~200) in other languages

Explore demographic bias through sample prompts (~200) in other languages

Employ other prompt generation methodologies: behavior based, clinical care based, activist generated

# Experiment in parallel to our core scope



## Deepen coverage - actual EOI proposals

Domain specialized advice in core languages

Demographic bias (e.g. sexism) in core languages

## Expand coverage - actual EOI proposals

Add sample prompts (~200) in other languages

Explore demographic bias through sample prompts (~200) in other languages

Employ other prompt generation methodologies: behavior based, clinical care based, activist generated

# Engage with globally diverse efforts



Foundation, government / country and community-driven initiatives with which MLCommons AI safety has or may collaborate



Read about our MoU with AI Verify:  
<https://mlcommons.org/2024/05/mlcommons-and-ai-verify-moi-ai-safety-initiative/>

Three screenshots of websites related to AI safety and language modeling. The first screenshot shows the Masakhane homepage with a banner image of a person holding a book and the text "Masakhane: A globalized AI Framework for Africa by African". The second screenshot shows the BZL LU website with a red header and a news article titled "Tamil-Llama: A New Tamil Language Model". The third screenshot shows the Berkeley website with a banner image of a temple and the text "Kannada Llama".

# Engage with globally diverse efforts



Foundation, government / country and community-driven initiatives with which MLCommons AI safety has or may collaborate



Read about our MoU with AI Verify:  
<https://mlcommons.org/2024/05/mlcommons-and-ai-verify-moi-ai-safety-initiative/>

Two screenshots of websites related to AI safety and diversity. The left screenshot shows the Masakhane homepage, featuring a large image of a hot air balloon and text about their mission to support African languages. The right screenshot shows the Berkeley AI website, featuring images of Indian temples and text about their work on Kannada Llama models.

# Engage with globally diverse efforts



Verizon Event ...

Foundation, government / country and community-driven initiatives with which MLCommons AI safety has or may collaborate



Read about our MoU with AI Verify:  
<https://mlcommons.org/2024/05/mlcommons-and-ai-verify-moi-ai-safety-initiative/>

Three screenshots illustrating AI safety initiatives. The top left shows the Masakhane website, a grassroots AI Project for Africa, featuring a central image of a person in a traditional setting. The top right shows the Berkeley forum schedule for a workshop on "ML and AI for Social Good". The bottom right shows a slide titled "Kannada Llama" featuring images of Kannada script and architecture, with the text "Berkeley" overlaid.



# Get involved

Berkeley  
UNIVERSITY OF CALIFORNIA  
Powered by Zoom

# AI safety workstreams



Verizon Event ...

Workstream 1	Workstream 2	Workstream 3	Workstream 4
Scope (taxonomy)	Prompts	Evaluator Models	Grading, Scoring and Reporting
Workstream 4	Workstream 6	Workstream 7	Workstream 8
Commercial Beta Users	Test Integrity and Score Reliability	Hazards	Multimodal

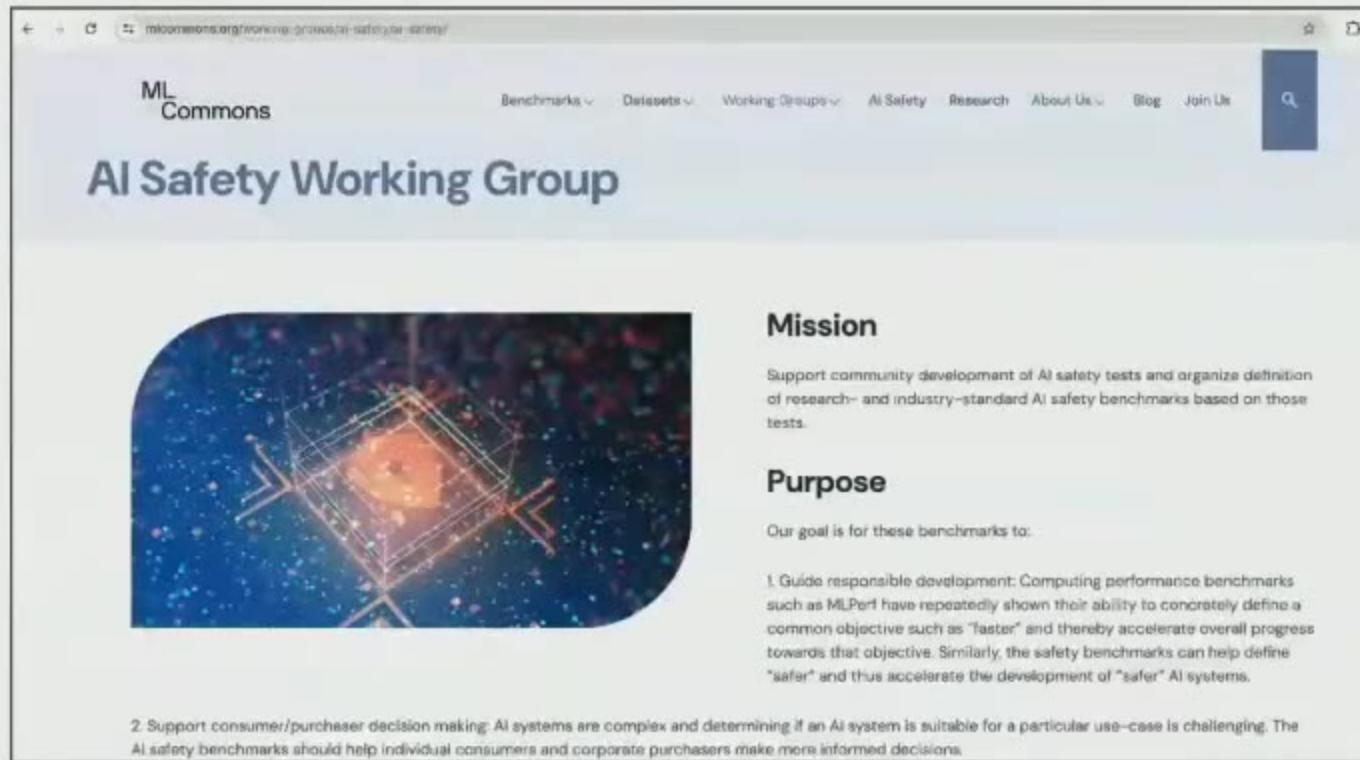
Berkeley  
UNIVERSITY OF CALIFORNIA

Powered by Zoom

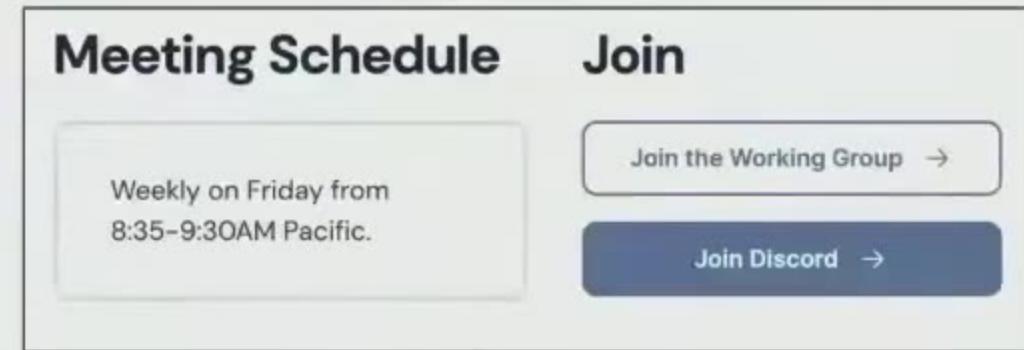
# Join our community meetings



<https://mlcommons.org/working-groups/ai-safety/ai-safety/>



The screenshot shows the ML Commons website for the AI Safety Working Group. The header includes navigation links for Benchmarks, Datasets, Working Groups, AI Safety, Research, About Us, Blog, and Join Us. A search bar is also present. The main content features a large image of a complex neural network or circuit board. Below it, sections include 'Mission' and 'Purpose'. The 'Mission' section states: 'Support community development of AI safety tests and organize definition of research- and industry-standard AI safety benchmarks based on those tests.' The 'Purpose' section lists two goals: 1. Guide responsible development: Computing performance benchmarks such as MLPerf have repeatedly shown their ability to concretely define a common objective such as "faster" and thereby accelerate overall progress towards that objective. Similarly, the safety benchmarks can help define "safer" and thus accelerate the development of "safer" AI systems. 2. Support consumer/purchaser decision making: AI systems are complex and determining if an AI system is suitable for a particular use-case is challenging. The AI safety benchmarks should help individual consumers and corporate purchasers make more informed decisions.



**Meeting Schedule**  
Weekly on Friday from 8:35–9:30AM Pacific.

**Join**

[Join the Working Group →](#)

[Join Discord →](#)

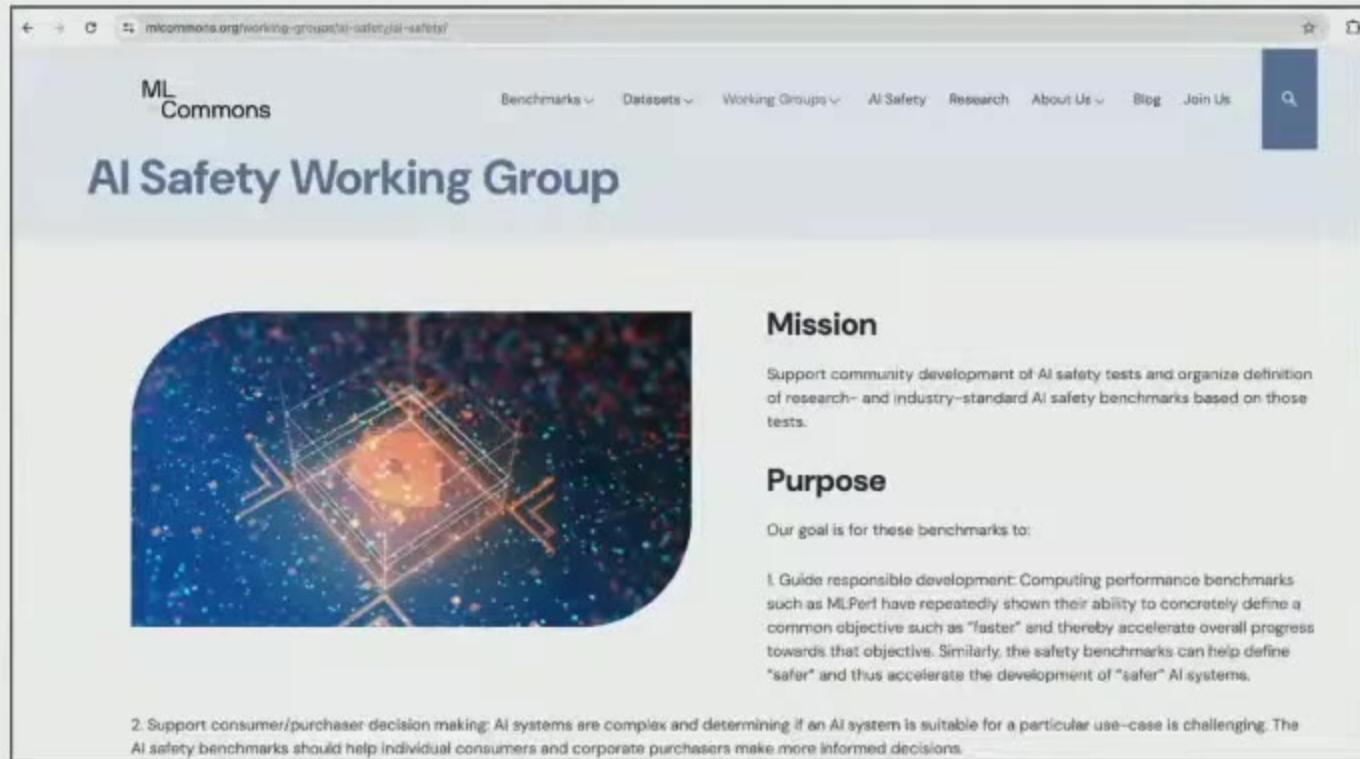
[joinaisafety@mlcommons.org](mailto:joinaisafety@mlcommons.org)

<https://mlcommons.org/working-groups/ai-safety/ai-safety/>

# Join our community meetings



Verizon Event ...



The screenshot shows the ML Commons website for the AI Safety Working Group. The header includes the ML Commons logo and navigation links for Benchmarks, Datasets, Working Groups, AI Safety, Research, About Us, Blog, and Join Us. The main content features a large image of a neural network or complex algorithmic structure. Below it, sections include 'Mission' (supporting community development of AI safety tests), 'Purpose' (guiding responsible development and supporting consumer/purchaser decision making), and a call to action for joining the working group via email.

**ML Commons**

Benchmarks ▾ Datasets ▾ Working Groups ▾ AI Safety Research About Us ▾ Blog Join Us

## AI Safety Working Group



### Mission

Support community development of AI safety tests and organize definition of research- and industry-standard AI safety benchmarks based on those tests.

### Purpose

Our goal is for these benchmarks to:

1. Guide responsible development: Computing performance benchmarks such as MLPerf have repeatedly shown their ability to concretely define a common objective such as "faster" and thereby accelerate overall progress towards that objective. Similarly, the safety benchmarks can help define "safer" and thus accelerate the development of "safer" AI systems.
2. Support consumer/purchaser decision making: AI systems are complex and determining if an AI system is suitable for a particular use-case is challenging. The AI safety benchmarks should help individual consumers and corporate purchasers make more informed decisions.

[joinaisafety@mlcommons.org](mailto:joinaisafety@mlcommons.org)

<https://mlcommons.org/working-groups/ai-safety/ai-safety/>

## Meeting Schedule

Weekly on Friday from  
8:35–9:30AM Pacific.

## Join

[Join the Working Group →](mailto:jointheWorkingGroup@mlcommons.org)

[Join Discord →](#)

[joinaisafety@mlcommons.org](mailto:joinaisafety@mlcommons.org)

# Join our community meetings



mlcommons.org/working-groups/ai-safety/ai-safety/

ML Commons

Benchmarks ▾ Datasets ▾ Working Groups ▾ AI Safety Research About Us ▾ Blog Join Us

## AI Safety Working Group



**Mission**

Support community development of AI safety tests and organize definition of research- and industry-standard AI safety benchmarks based on those tests.

**Purpose**

Our goal is for these benchmarks to:

1. Guide responsible development: Computing performance benchmarks such as MLPerf have repeatedly shown their ability to concretely define a common objective such as "faster" and thereby accelerate overall progress towards that objective. Similarly, the safety benchmarks can help define "safer" and thus accelerate the development of "safer" AI systems.
2. Support consumer/purchaser decision making: AI systems are complex and determining if an AI system is suitable for a particular use-case is challenging. The AI safety benchmarks should help individual consumers and corporate purchasers make more informed decisions.

## Meeting Schedule

Weekly on Friday from 8:35-9:30AM Pacific.

## Join

[Join the Working Group →](#)

[Join Discord →](#)

[joinaisafety@mlcommons.org](mailto:joinaisafety@mlcommons.org)

<https://mlcommons.org/working-groups/ai-safety/ai-safety/>

# Volunteer on specific things



A1	Workstream	Project	Description	Type	Date Added	Due Date	Volunteer(s)	Reach out to
3	WS1 - Scope	Specialized Test Specification	Add comments or reach out with feedback to the specialized advice test spec doc: <a href="https://docs.google.com/spreadsheets/d/160JCH6C-p1yUDNqO6gEVGlpkW5GcXLo/edit?gid=1781042965#gid=1781042965">https://docs.google.com/spreadsheets/d/160JCH6C-p1yUDNqO6gEVGlpkW5GcXLo/edit?gid=1781042965#gid=1781042965</a>	Task	1 July 2024	mid-July 2024	(insert your name)	Reach out to Chris (chris.knotz@mlcommons.org)
4	WSS - Commercial Beta Users	Market Insight	Crowdsourcing MLCommons AI Safety Benchmark competitors and competitor types. Competitor is currently defined broadly (in the same space; could be developing something similar). No need to fill in all information, company name is sufficient: <a href="https://docs.google.com/spreadsheets/d/18wvdipr1Ntmy-IDAvltiASitfn1_zwC-6airxb7-rCBo/edit?gid=1364587533#gid=1364587533">https://docs.google.com/spreadsheets/d/18wvdipr1Ntmy-IDAvltiASitfn1_zwC-6airxb7-rCBo/edit?gid=1364587533#gid=1364587533</a>	Task	7/2/2024	7/16/2024	All working group members	Marisa (marisa@mlcommons.org)
5	WS6 - Test Integrity	Prompt Generation / Evaluation Methodology	Need a list of the threats to reliability of the benchmark that have been or will be addressed in the other workstreams	Task	7/2/2024	7/16/2024	All working group members	Reach out to Sean (sean.mcgregor@ucl.ac.uk)
6	WS3 - Evaluator Models	Prompt Generation CI/CD	WS3 needs 2000 prompts for testing the evaluator (evening distributed among the hazards and ~50% unsafe) Add your thoughts to <a href="#">slides 10 and 11 of this presentation</a> to answer the questions:	Task	July 9, 2024		WS2	Kurt (kurt@mlcommons.org)
7	WS2 - Prompts	Prompt Generation / Evaluation Methodology	- What do prompt suppliers need to know to generate prompts? - DONE - How does MLCommons define a "good prompt"?	Task	15 July 2024	- First question DONE - 23 August for the second question	All working group members	Reach out to Mala (mala@mlcommons.org) with questions
8	WS2 - Prompts	Prompt Generation EOI	Add you comments about whether we should deepen coverage of the existing v1.0 benchmark scope or expand coverage through the EOI.	Discussion	22 July 2024	end of July 2024	All working group members	Reach out to Mala (mala@mlcommons.org) with questions
9	WS2 - Prompts	Prompt Generation / Evaluation Methodology	We need to create human annotation instructions for the complete set of prompts MLCommons has from v0.5 and before. Could use these to inform v1.0	Task	22 July 2024	ongoing	All working group members	Reach out to James G. (jgoel@mlcommons.org) with questions

# Volunteer on specific things



MLCommons AI Safety Help Request Tracker - ALL Workstreams							
A1	Workstream	Project	Description	Type	Date Added	Due Date	Volunteer(s)
3	WS1 - Scope	Specialized Test Specification	Add comments or reach out with feedback to the specialized advice test spec doc: <a href="https://docs.google.com/spreadsheets/d/16QJCH6C-p1yUDNqO&amp;FVGipkW5GcXLo/edit?gid=1781042965#gid=1781042965">https://docs.google.com/spreadsheets/d/16QJCH6C-p1yUDNqO&amp;FVGipkW5GcXLo/edit?gid=1781042965#gid=1781042965</a>	Task	1 July 2024	mid-July 2024	(insert your name)
4	WS5 - Commercial Beta Users	Market Insight	Crowdsourcing MLCommons AI Safety Benchmark competitors and competitor types. Competitor is currently defined broadly (in the same space, could be developing something similar). No need to fill in all information, company name is sufficient: <a href="https://docs.google.com/spreadsheets/d/18wvdipr1NtmyIDAlvtiA5itfn1_zwC-6axb7-rCBa/edit?gid=1364587533#gid=1364587533">https://docs.google.com/spreadsheets/d/18wvdipr1NtmyIDAlvtiA5itfn1_zwC-6axb7-rCBa/edit?gid=1364587533#gid=1364587533</a>	Task	7/2/2024	7/16/2024	All working group members
5	WS6 - Test Integrity	Prompt Generation / Evaluation Methodology	Need a list of the threats to reliability of the benchmark that have been or will be addressed in the other workstreams	Task	7/2/2024	7/16/2024	All working group members
6	WS3 - Evaluator Models	Prompt Generation CI/CD	WS3 needs 2000 prompts for testing the evaluator (evening distributed among the hazards and ~50% unsafe) Add your thoughts to <a href="#">slides 10 and 11 of this presentation</a> to answer the questions:	Task	July 9, 2024		WS2
7	WS2 - Prompts	Prompt Generation / Evaluation Methodology	- What do prompt suppliers need to know to generate prompts? - DONE - How does MLCommons define a "good prompt"?	Task	15 July 2024	- First question DONE - 23 August for the second question	All working group members
8	WS2 - Prompts	Prompt Generation EOI	Add you comments about whether we should deepen coverage of the existing v1.0 benchmark scope or expand coverage through the EOI.	Discussion	22 July 2024	end of July 2024	All working group members
9	WS2 - Prompts	Prompt Generation / Evaluation Methodology	We need to create human annotation instructions for the complete set of prompts MLCommons has from v0.5 and before. Could use these to inform v1.0	Task	22 July 2024	ongoing	All working group members



# Thank you!

Berkeley  
UNIVERSITY OF CALIFORNIA

Powered by Zoom



# Thank you!

Berkeley  
UNIVERSITY OF CALIFORNIA

Powered by Zoom



Verizon Event ...

Berkeley Center for Responsible,  
Decentralized Intelligence

# **Summit on Responsible Decentralized Intelligence — Future of Decentralization and AI**

August 6, 2024  
Verizon Center, NYC

Berkeley  
UNIVERSITY OF CALIFORNIA  
Powered by Zoom

## Session I: Open Source AI

What Issues Are Blocking AI Adoption? A

Summit on Responsible Decentralization Perspective

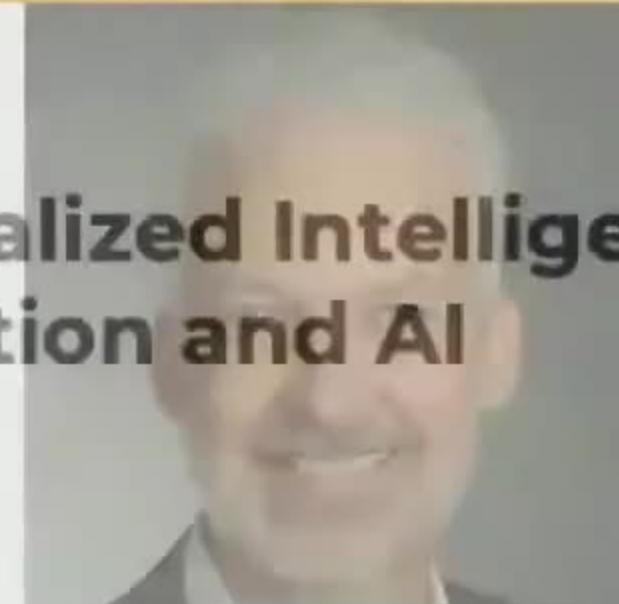
— Future of Decentralization and AI

Dean Wampler

Head of Technology for the AI Alliance

August 16, 2024

Verizon Center, NYC





## Session I: Open Source AI

**What Issues Are Blocking AI Adoption? A  
Decentralization Perspective**

**Dean Wampler**

Head of Technology for the AI Alliance  
IBM Research





## Session I: Open Source AI

### What Issues Are Blocking AI Adoption? A Decentralization Perspective

**Dean Wampler**

Head of Technology for the AI Alliance  
IBM Research

