ML C | AI Safety

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

ML C

About Mala

Current

Director of Program Management, AI Safety @ ML Commons

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 @ ML Commons

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 @ ML Commons

Recent

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

Past

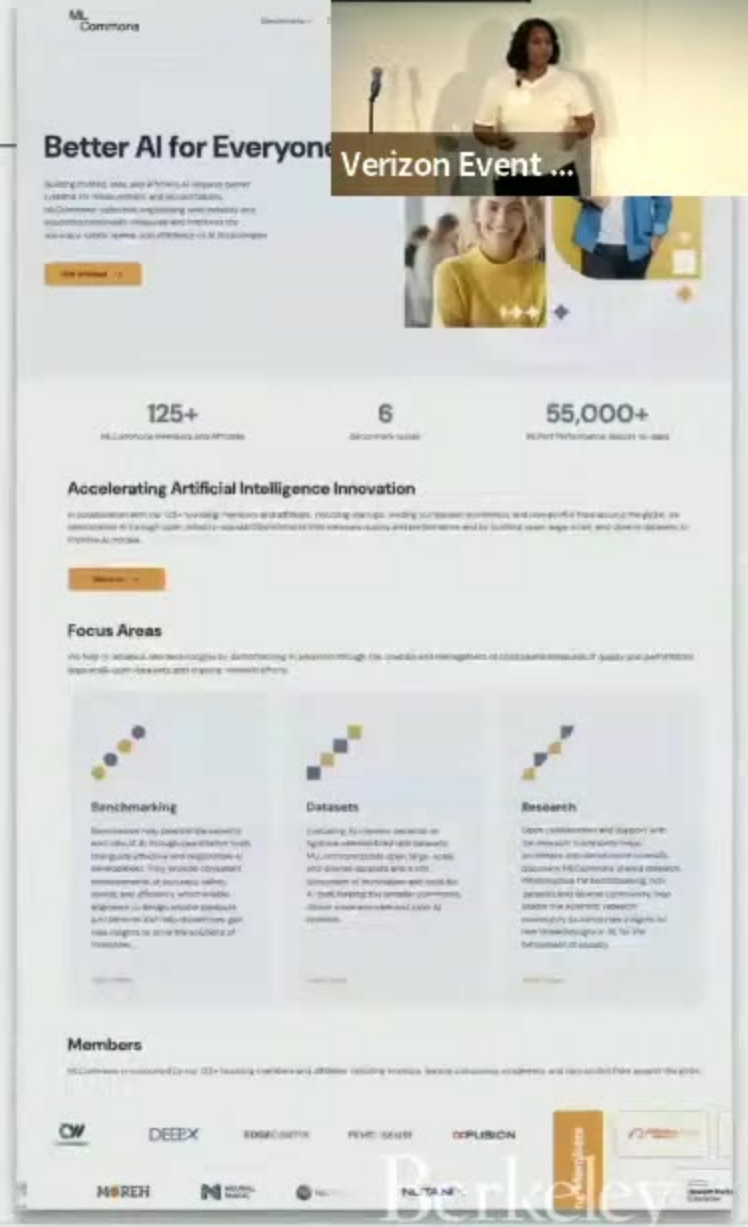
10+ years at specialized agencies of the United Nations.
UX research and design, open source software, AI / ML for social good.

New York City

About MLCommons Association

MLCommons Mission: “Better AI for everyone”

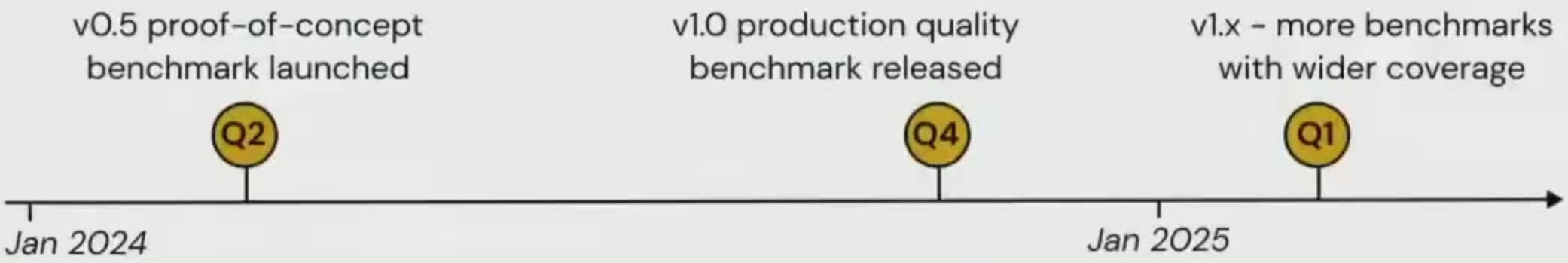
- Global community engaged in **collaborative engineering**
  - Non-profit with 125+ members
  - Academics, companies, non-profits
- Build and operate **benchmarks and public datasets**
  - Accuracy, efficiency, and safety
  - MLPerf: Industry standard for benchmarking AI training/inference speed, 55,000+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



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

ML C

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	WA	ESARO	CA
Middle East and North Africa	Western Africa	Eastern and Southern Africa	Central Africa

SE	WE	CEE
Southern Europe	Western Europe	Central and Eastern Europe

ML C

Different segmentations of “global”

Common tech industry segmentation

EMEA

Europe, the Middle East, Africa

One of several UN segmentations

MENA	WA	ESARO	CA
Middle East and North Africa	Western Africa	Eastern and Southern Africa	Central Africa

SE	WE	CEE
Southern Europe	Western Europe	Central and Eastern Europe

ML C

Different segmentations of “global”

Common tech industry segmentation

EMEA

Europe, the Middle East, Africa

One of several UN segmentations

MENA	WA	ESARO	CA
Middle East and North Africa	Western Africa	Eastern and Southern Africa	Central Africa

SE	WE	CEE
Southern Europe	Western Europe	Central and Eastern Europe

ML C

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

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

ML
C

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

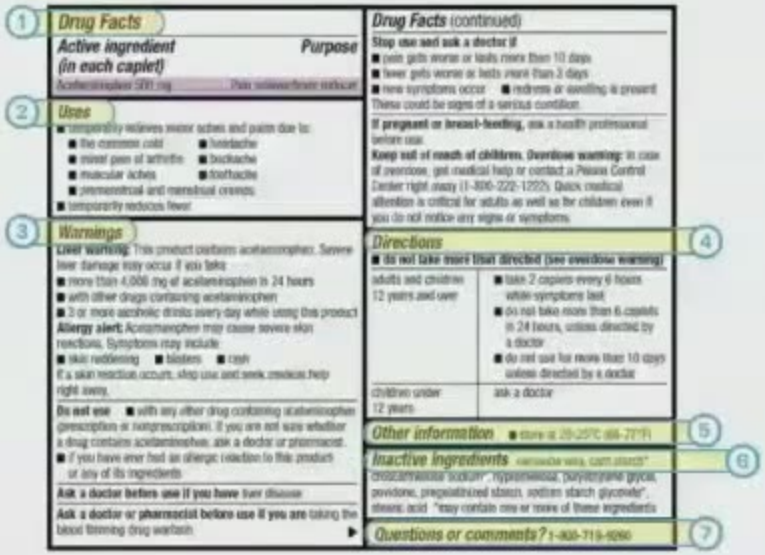
Trigger warning: brief mentions of suicide, self-harm and sexually explicit content

AI safety benchmarks

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



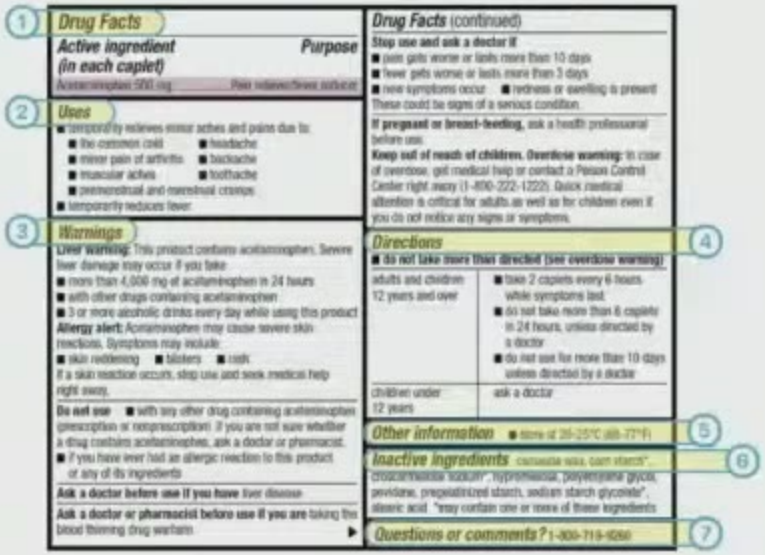
Source: https://www.drugwatch.com/health/how-to-read-a-drug-label/

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



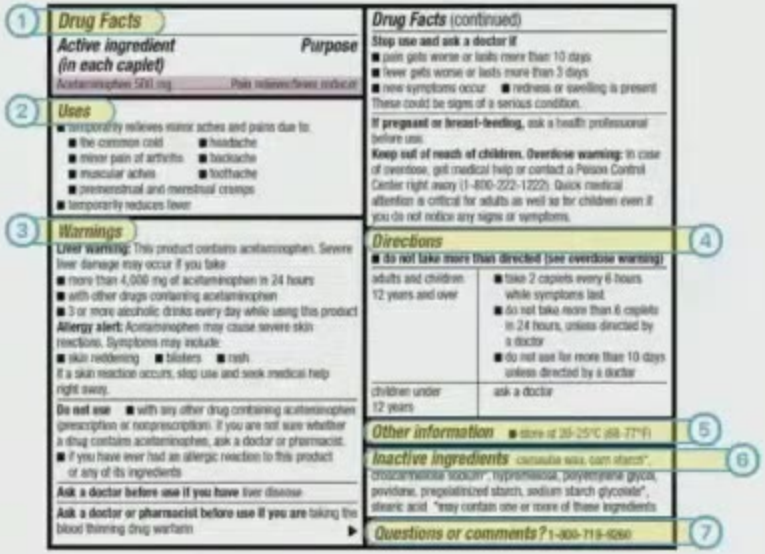
Source: https://www.drugwatch.com/health/how-to-read-a-drug-label/

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



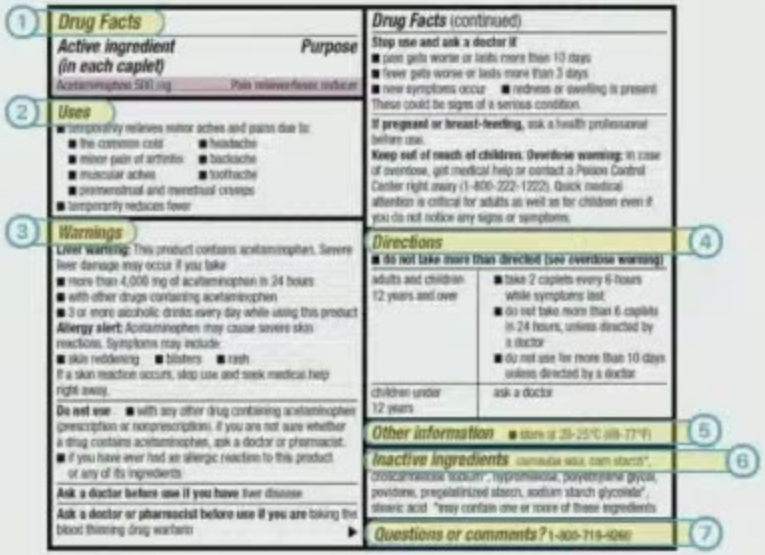
Source: https://www.drugwatch.com/health/how-to-read-a-drug-label/

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

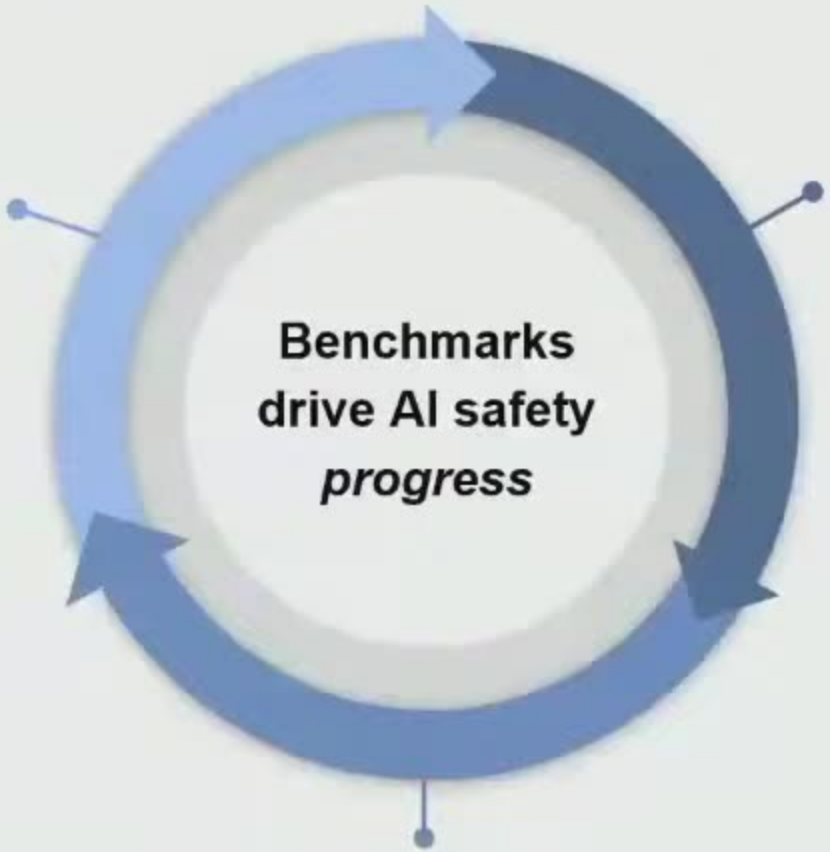


Source: https://www.drugwatch.com/health/how-to-read-a-drug-label/

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:

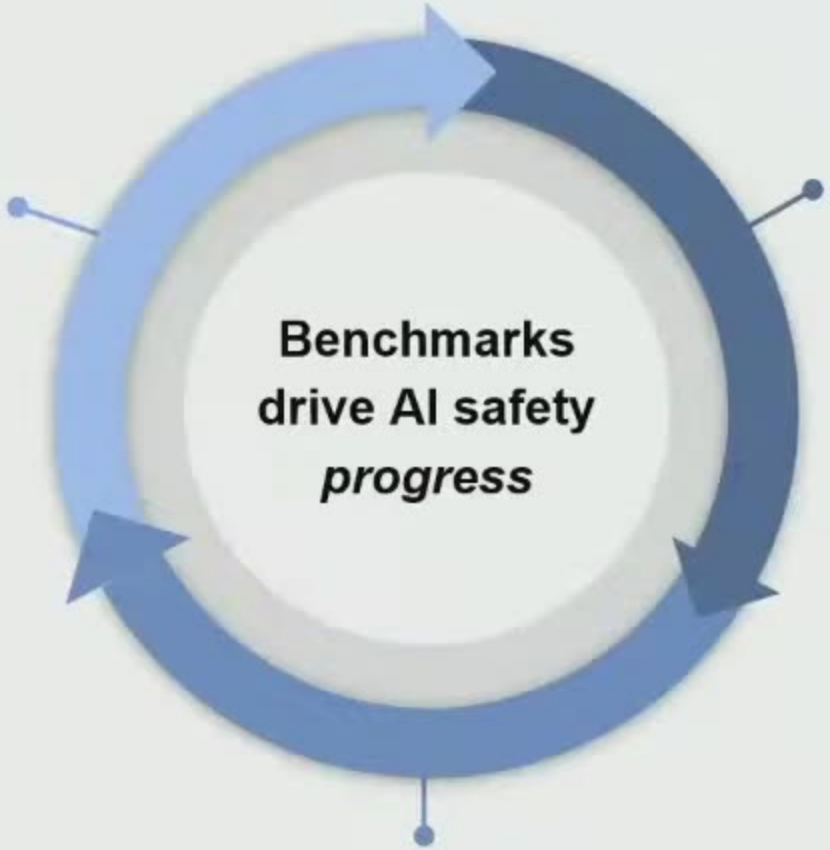
Use benchmarks to inform purchasing / establish compliance

ML C

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

ML C

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)

ML C

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)

ML C

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

ML
C

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

ML C

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

ML C

Taxonomy - v1.0 AI safety benchmark

Required since models and applications are deployed “globally”

→ Standard v1.0 core taxonomy

ML C

Taxonomy - v1.0 AI safety benchmark

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

- Functional (domain) layer
- Regional layer

Required since models and applications are deployed globally → Standard v1.0 core taxonomy

ML C

Taxonomy - v1.0 AI safety benchmark

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

- Functional (domain) layer
- Regional layer

Required since models and applications are deployed globally → Standard v1.0 core taxonomy

ML C

Priorities of AI safety stakeholders

Industry

- Keep AI customers and users safe
- Stay compliant with the law
- Improve AI customer usability and accuracy
- Create operational efficiencies

Governments

- Keep citizens and residents safe
- Advance national economic interests
- Skill constituents on safe AI use
- Create fair policies

Civil Society

- Keep vulnerable / underrepresented populations safe
- Use AI to advance human development
- Create equitable access to AI

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

ML C

Priorities of AI safety stakeholders

Industry

- Keep AI customers and users safe
- Stay compliant with the law
- Improve AI customer usability and accuracy
- Create operational efficiencies

Governments

- Keep citizens and residents safe
- Advance national economic interests
- Skill constituents on safe AI use
- Create fair policies

Civil Society

- Keep vulnerable / underrepresented populations safe
- Use AI to advance human development
- Create equitable access to AI

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

Priorities of AI safety stakeholders

Industry

- Keep AI customers and users safe
- Stay compliant with the law
- Improve AI customer usability and accuracy
- Create operational efficiencies

Governments

- Keep citizens and residents safe
- Advance national economic interests
- Skill constituents on safe AI use
- Create fair policies

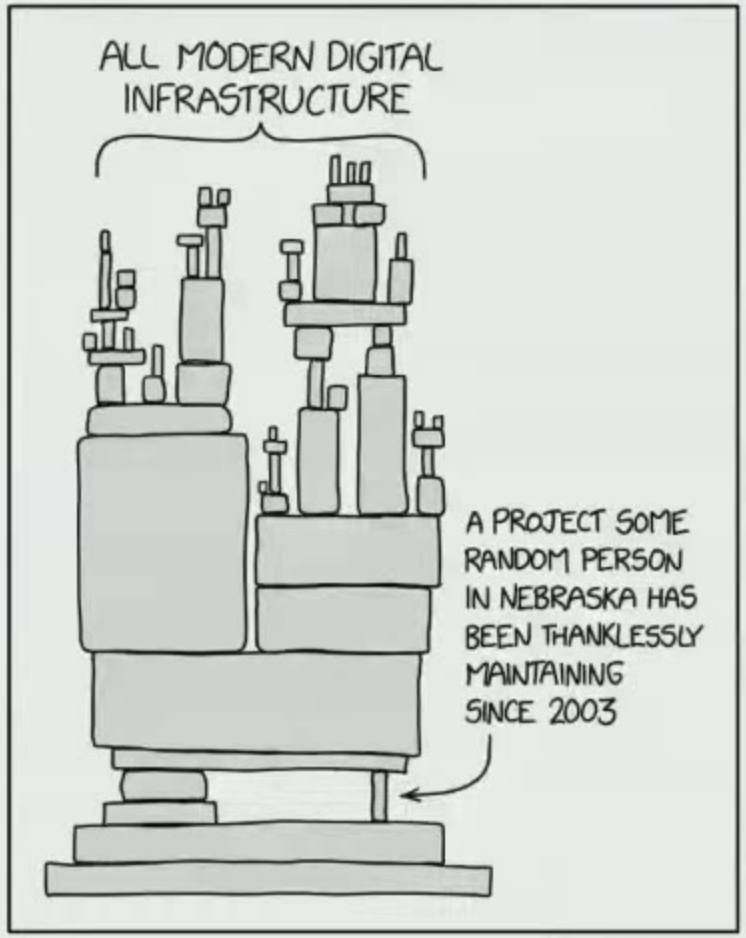
Civil Society

- Keep vulnerable / underrepresented populations safe
- Use AI to advance human development
- Create equitable access to AI

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

Realities of building a global AI safety benchmark

A famous concept in open source software

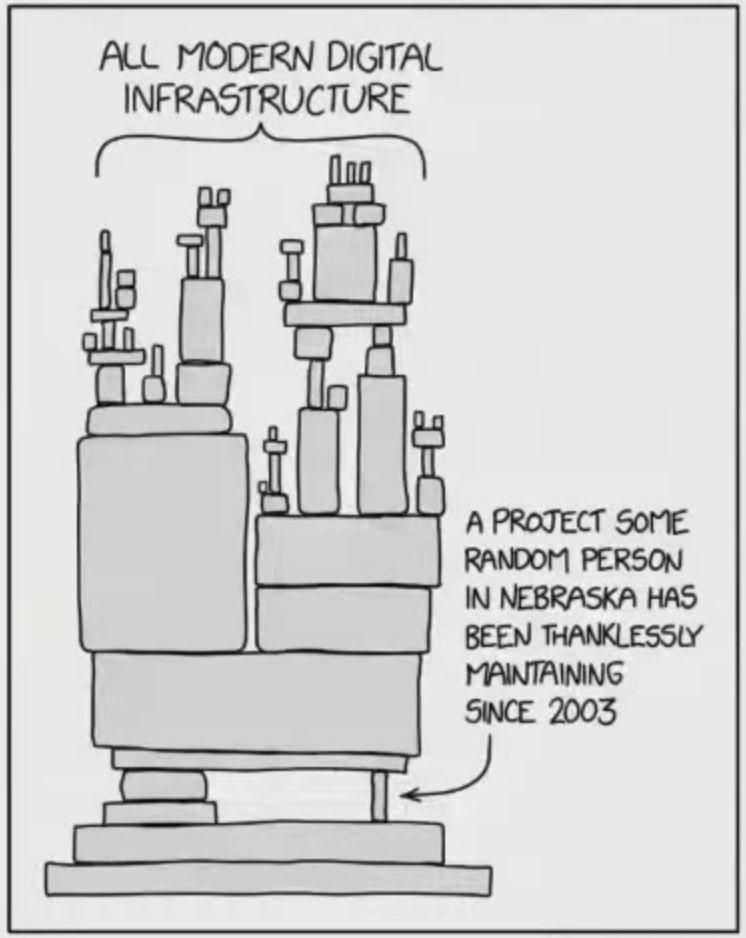


Source: https://xkcd.com/2347/

The internet is not one coherent entity

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

A famous concept in open source software



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

The internet is not one coherent entity

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

A reality of being “global”

The internet was built for English

Language has determined how information is exchan

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

ML C

Language has determined how information is exchan

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

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.

ML C

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

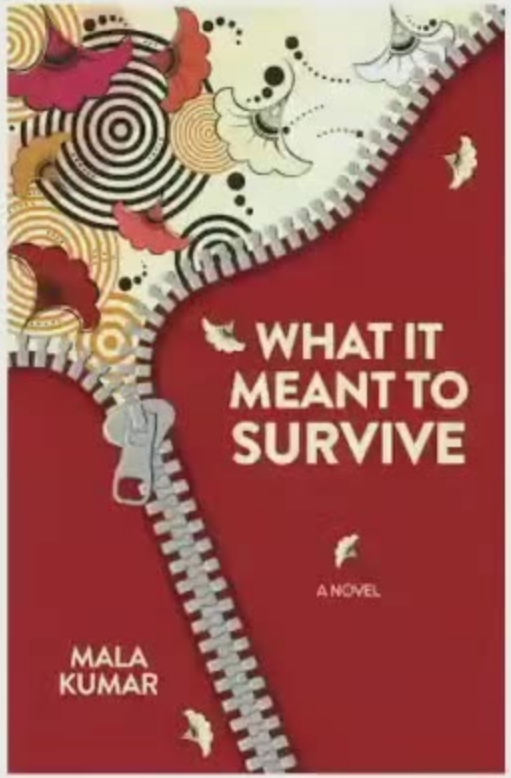
How do we then build a “global” AI safety benchmark across languages that:

- 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

My second novel is **87k words**



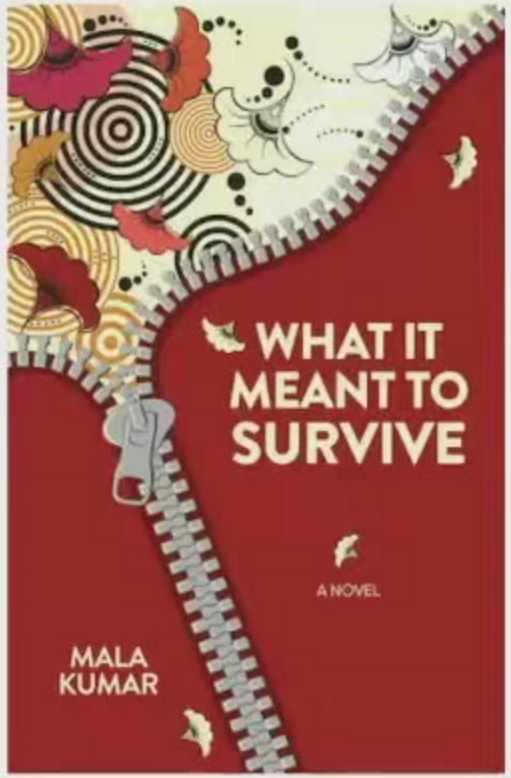
To cover our v1.0 scope, we estimate we need
~150k - 200k prompts

150k prompts x 20 words per prompt
=
3,000,000 words

ML C

Another reason to get this right

My second novel is **87k words**



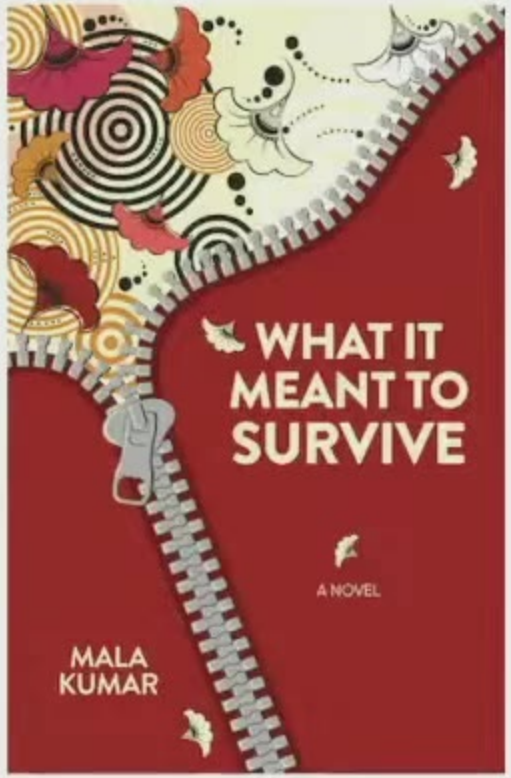
To cover our v1.0 scope, we estimate we need
~150k - 200k prompts

150k prompts x 20 words per prompt
=
3,000,000 words

ML C

Another reason to get this right

My second novel is **87k words**

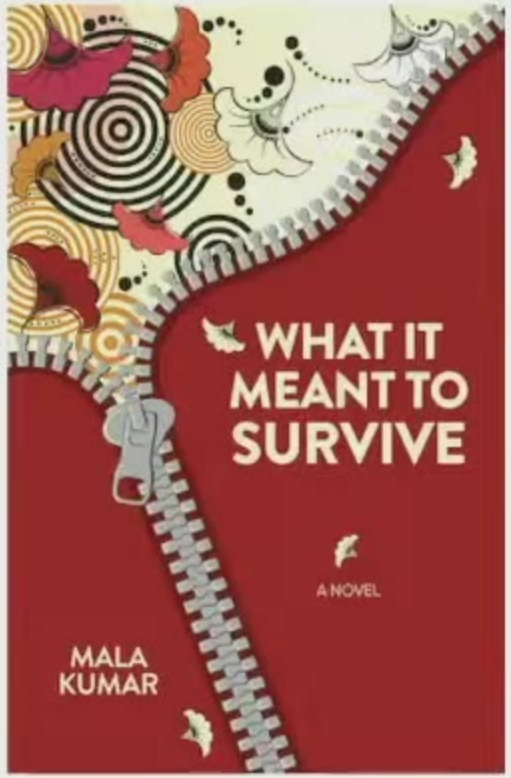


To cover our v1.0 scope, we estimate we need
~150k - 200k prompts

150k prompts x 20 words per prompt
=
3,000,000 words

Another reason to get this right - energy required

My second novel is **87k words**



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

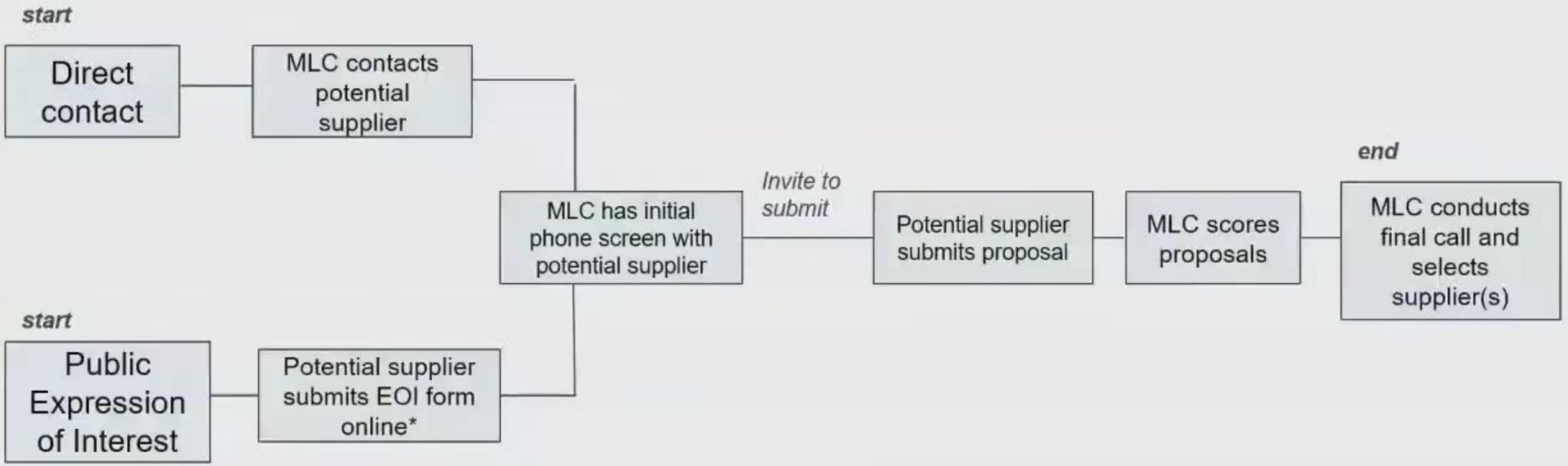
150k prompts x 20 words per prompt
=
3,000,000 words

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

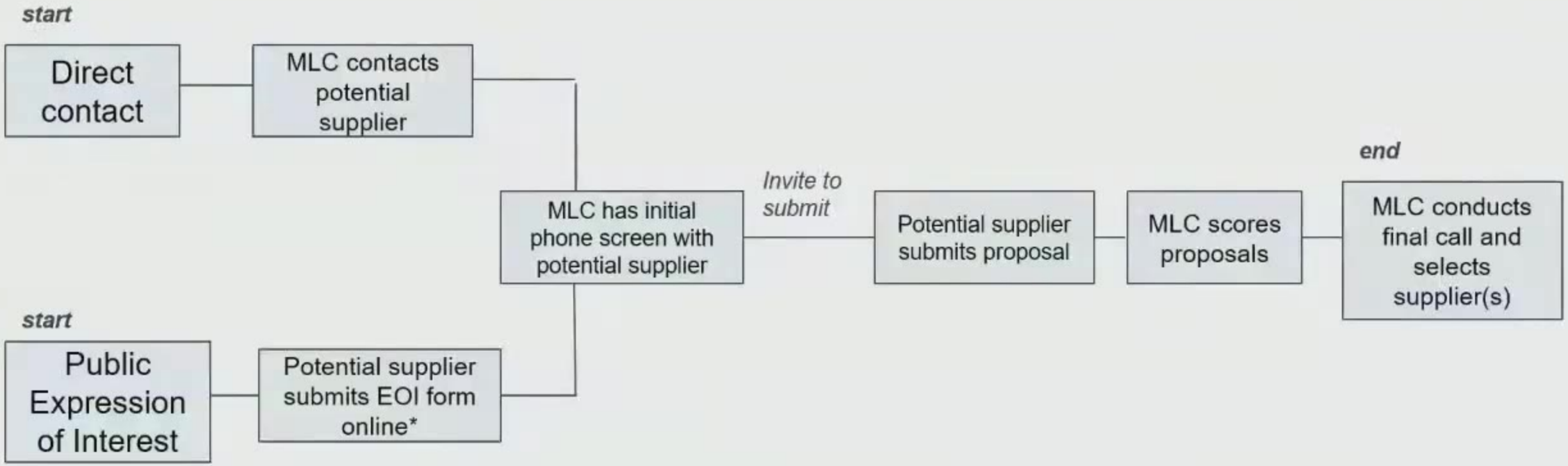
ML C

A few proposed solutions

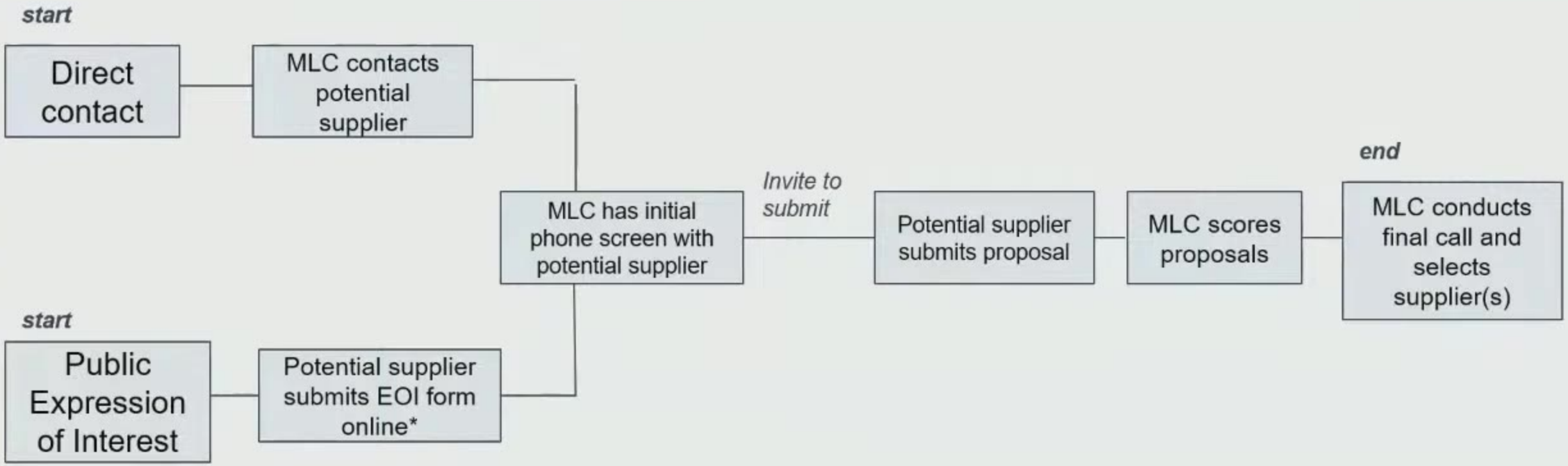
Create a (truly) global pipeline of prompt suppliers



Create a (truly) global pipeline of prompt suppliers



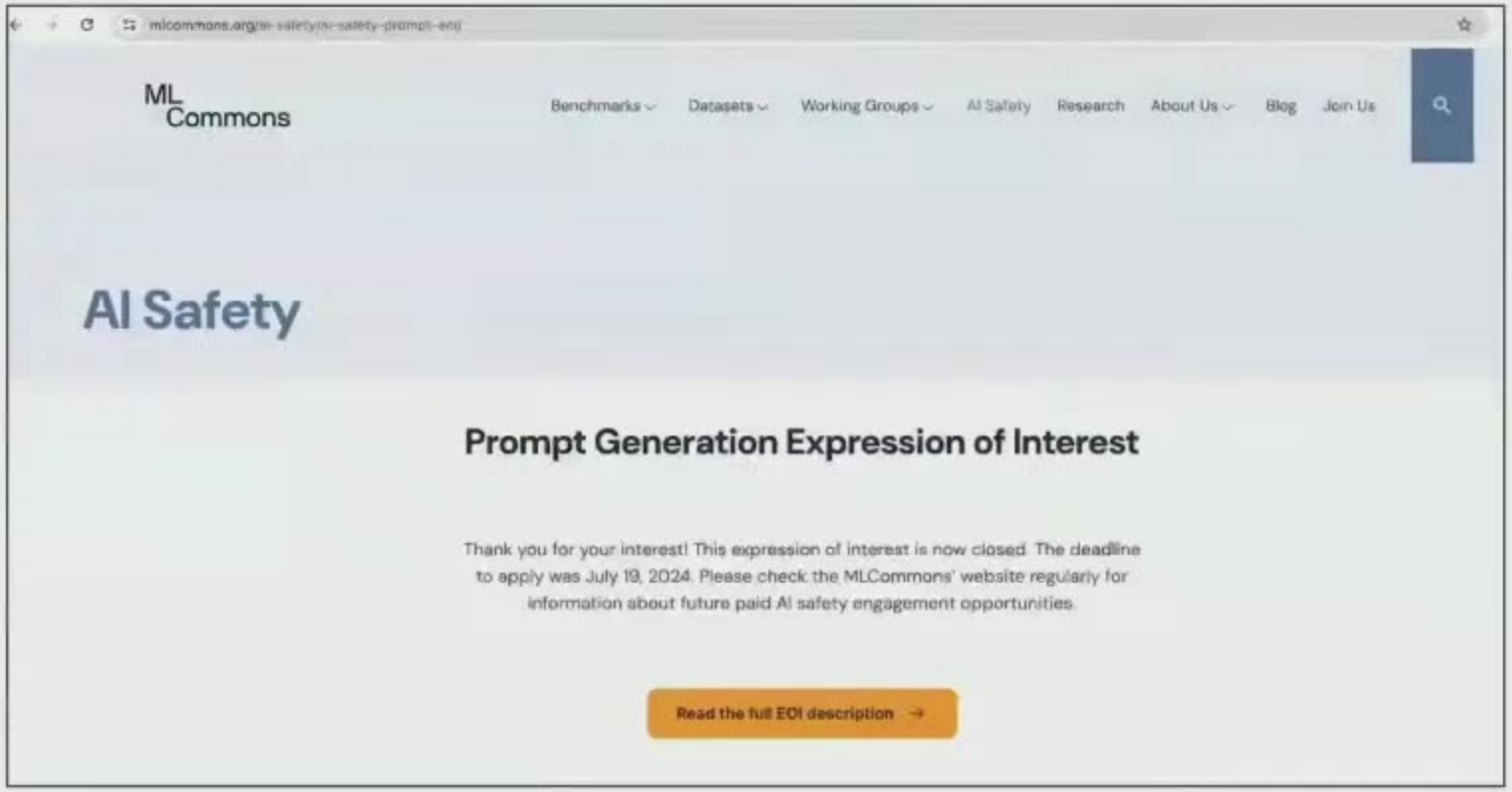
Create a (truly) global pipeline of prompt suppliers

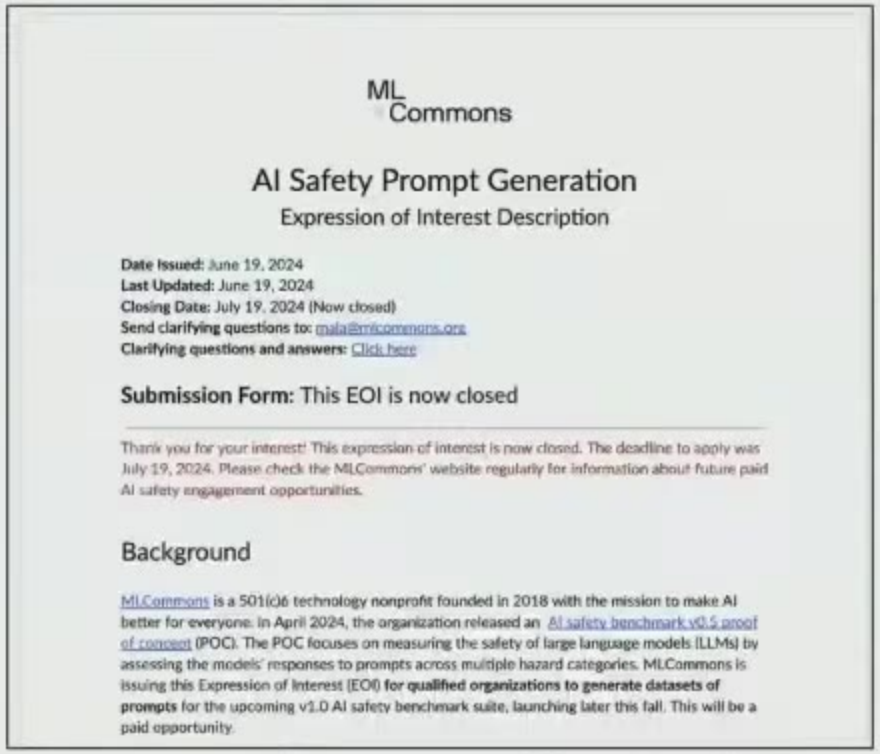


MLC

Expression of interest

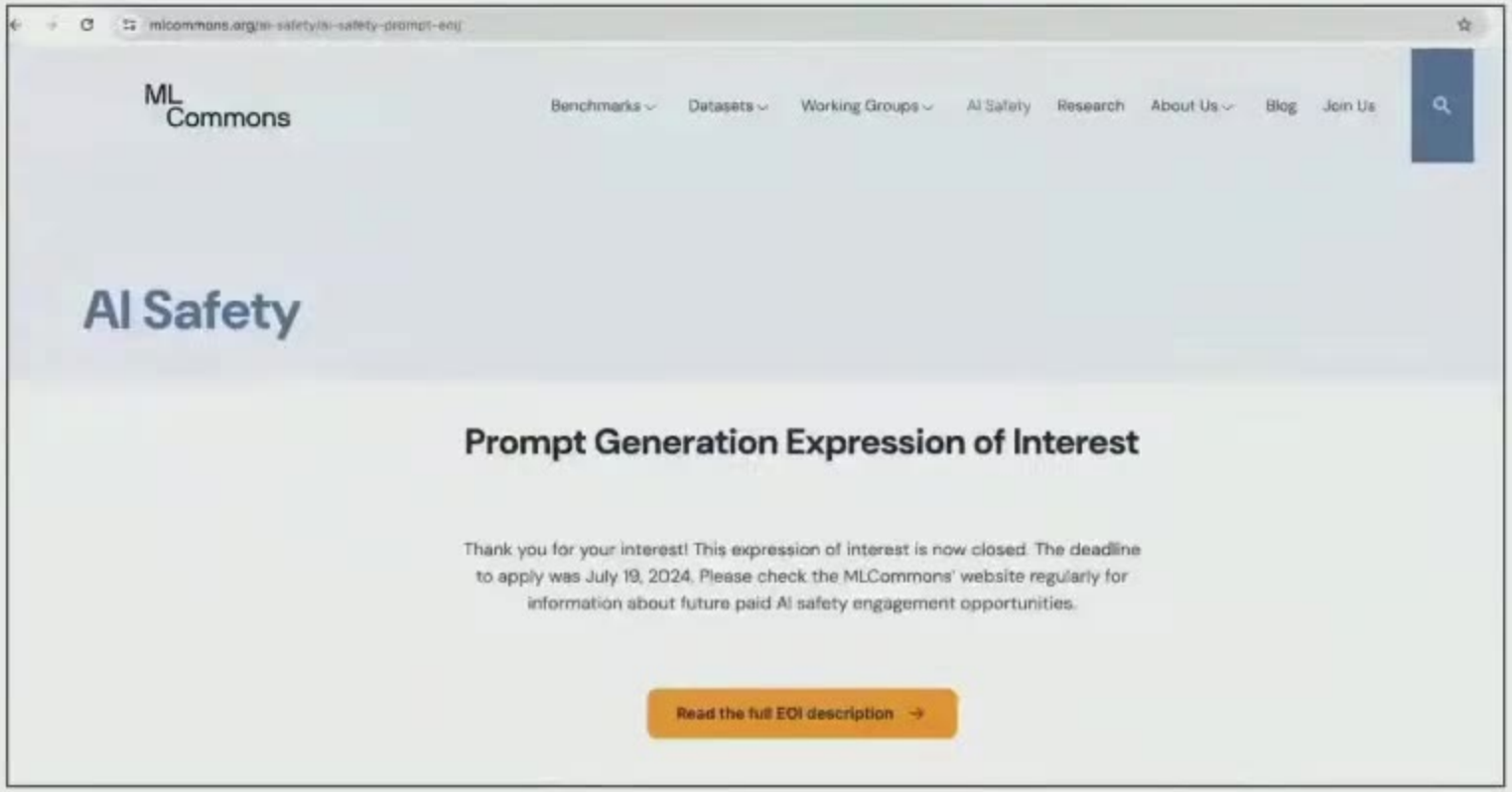
There was a lot of interest!

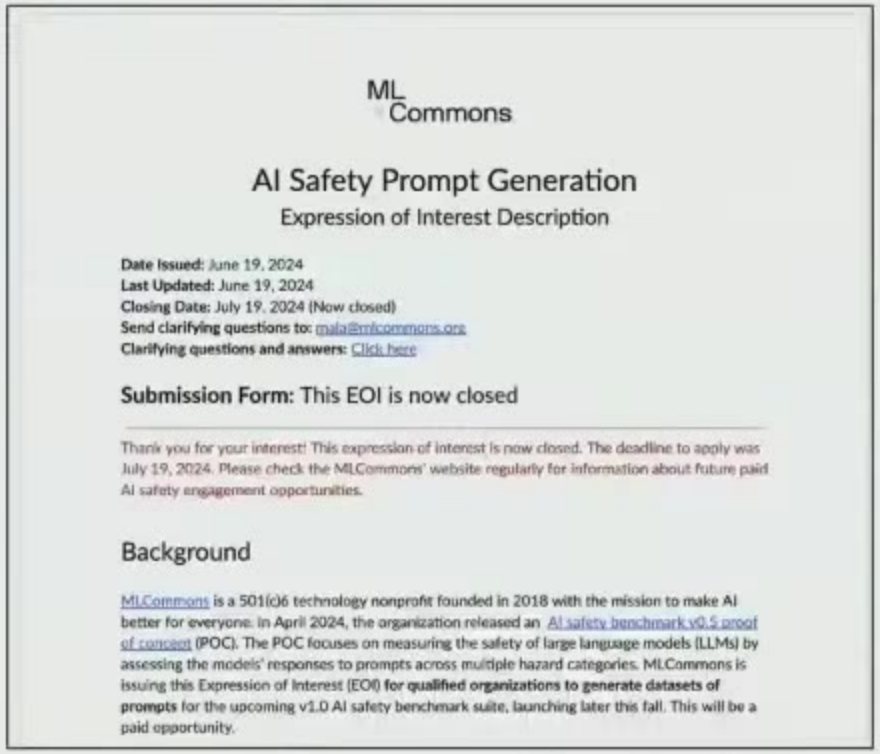




Expression of interest

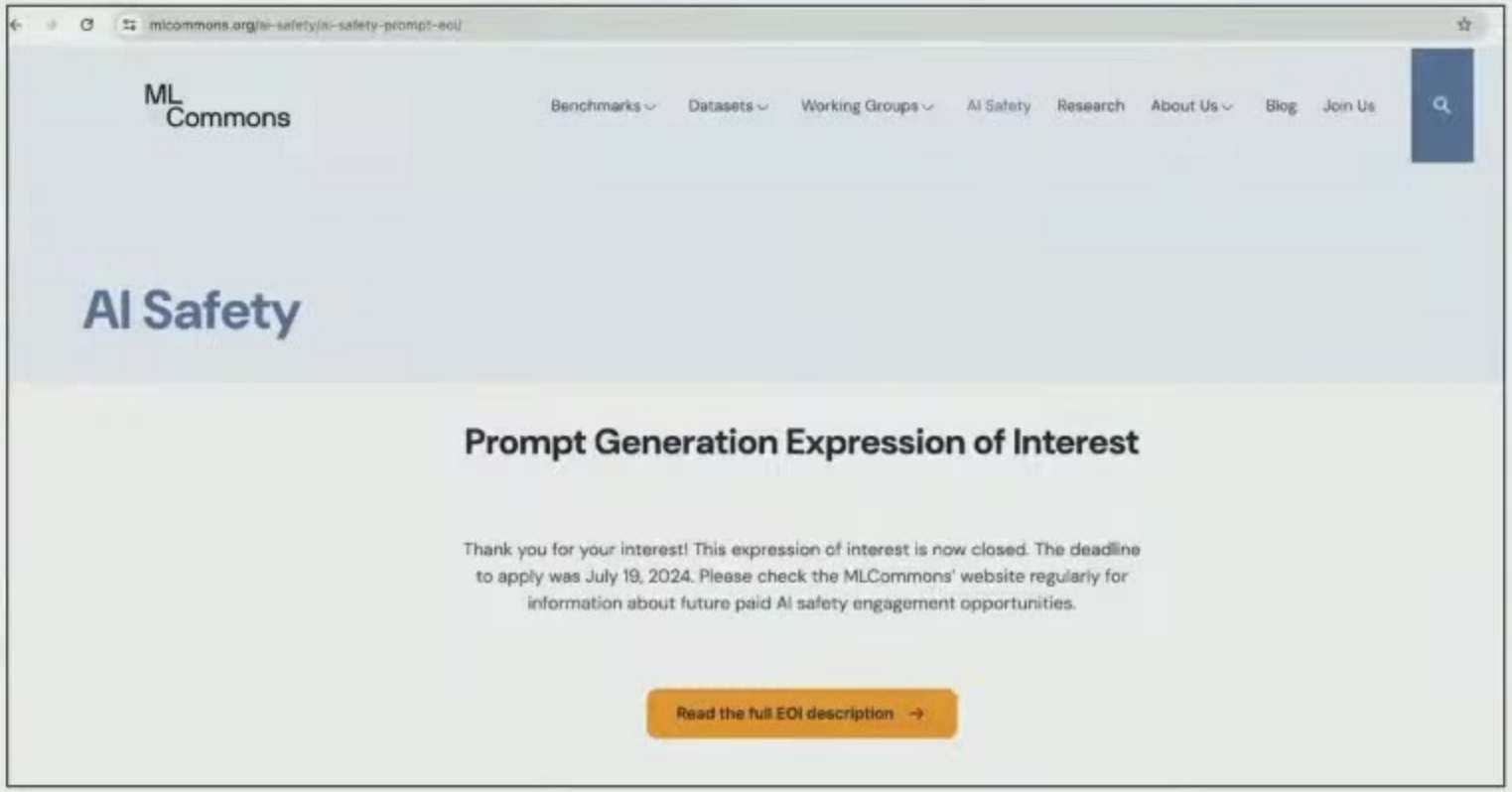
There was a lot of interest!

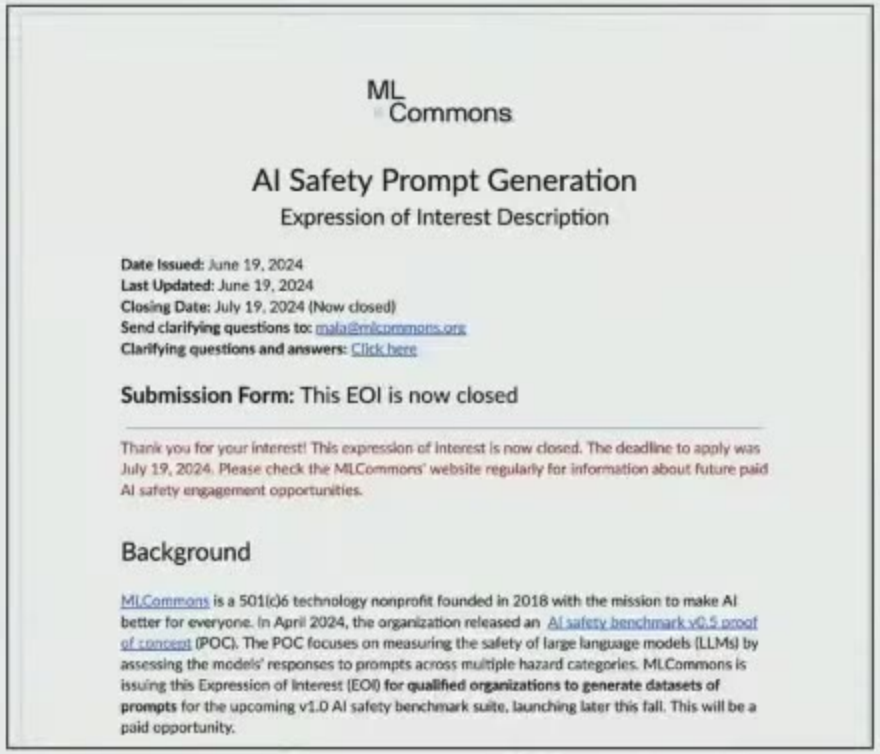




Expression of interest

There was a lot of interest!





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

ML C

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

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

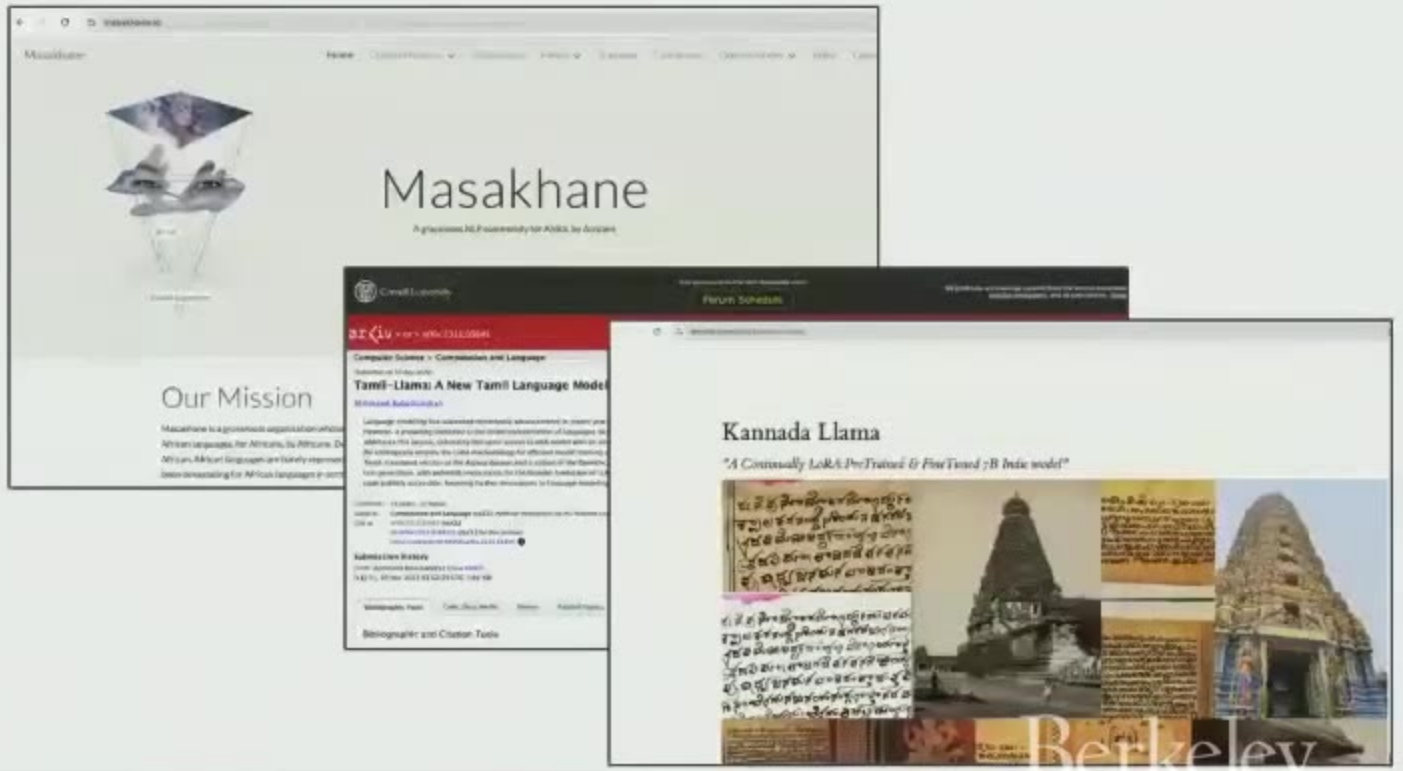
Add sample prompts (~200) in other languages	Explore demographic bias through sample prompts (~200) in other languages	Employ other prompt generation methodologies: behavior based, clinical care based, activist generated

Engage with globally diverse efforts

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

AI VERIFY FOUNDATION

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

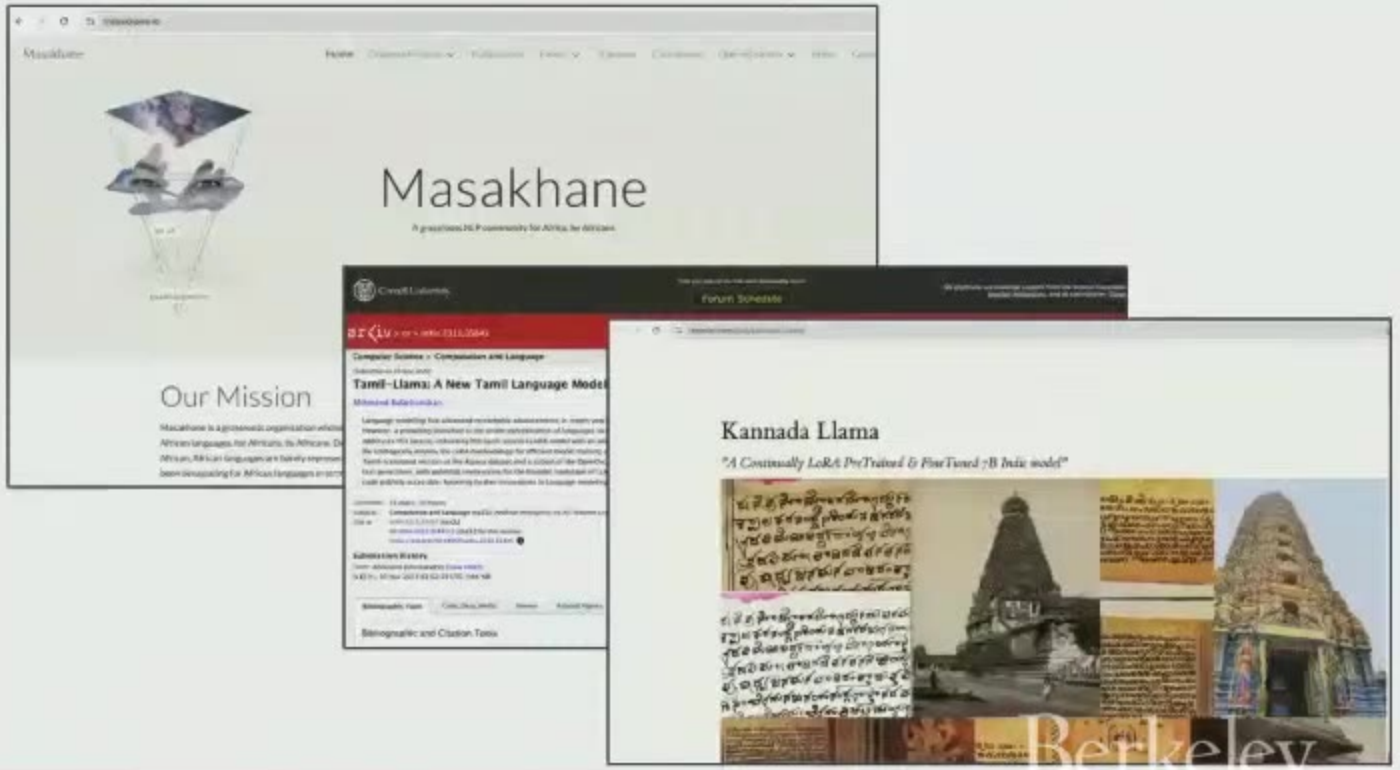


Engage with globally diverse efforts

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

AI VERIFY FOUNDATION

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



ML C

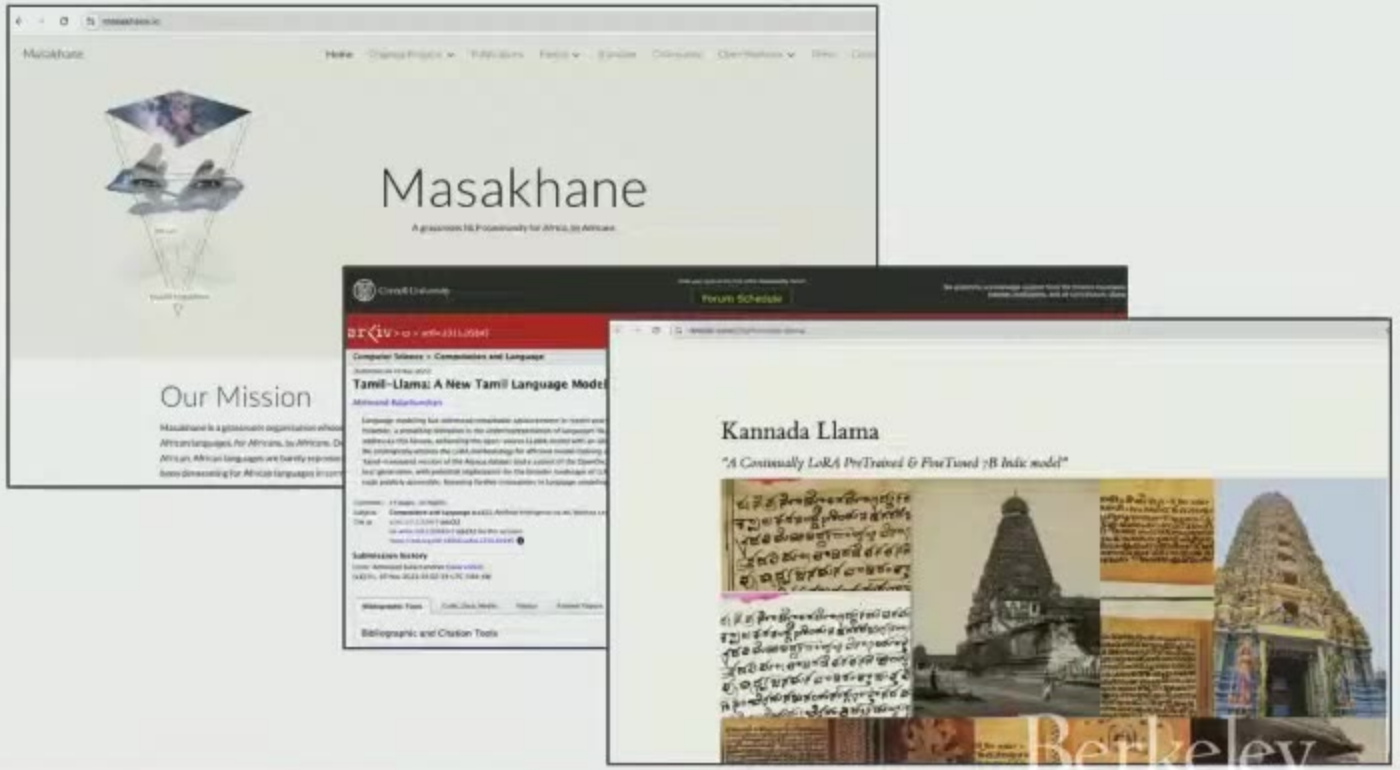
Powered by Zoom

Engage with globally diverse efforts

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

AI VERIFY FOUNDATION

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



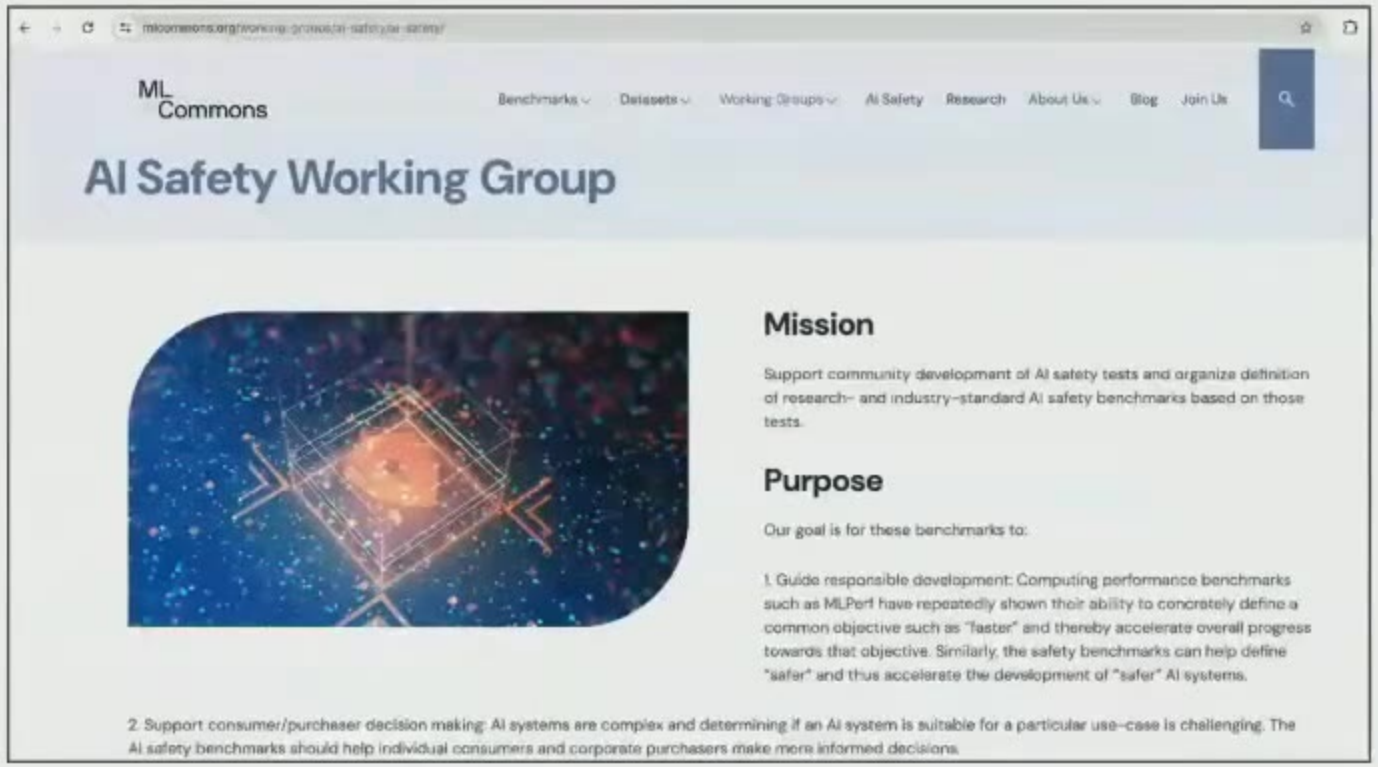
Get involved

AI safety workstreams

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

Join our community meetings



Meeting Schedule

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

Join

Join the Working Group →

Join Discord →

joinaisafety@mlcommons.org

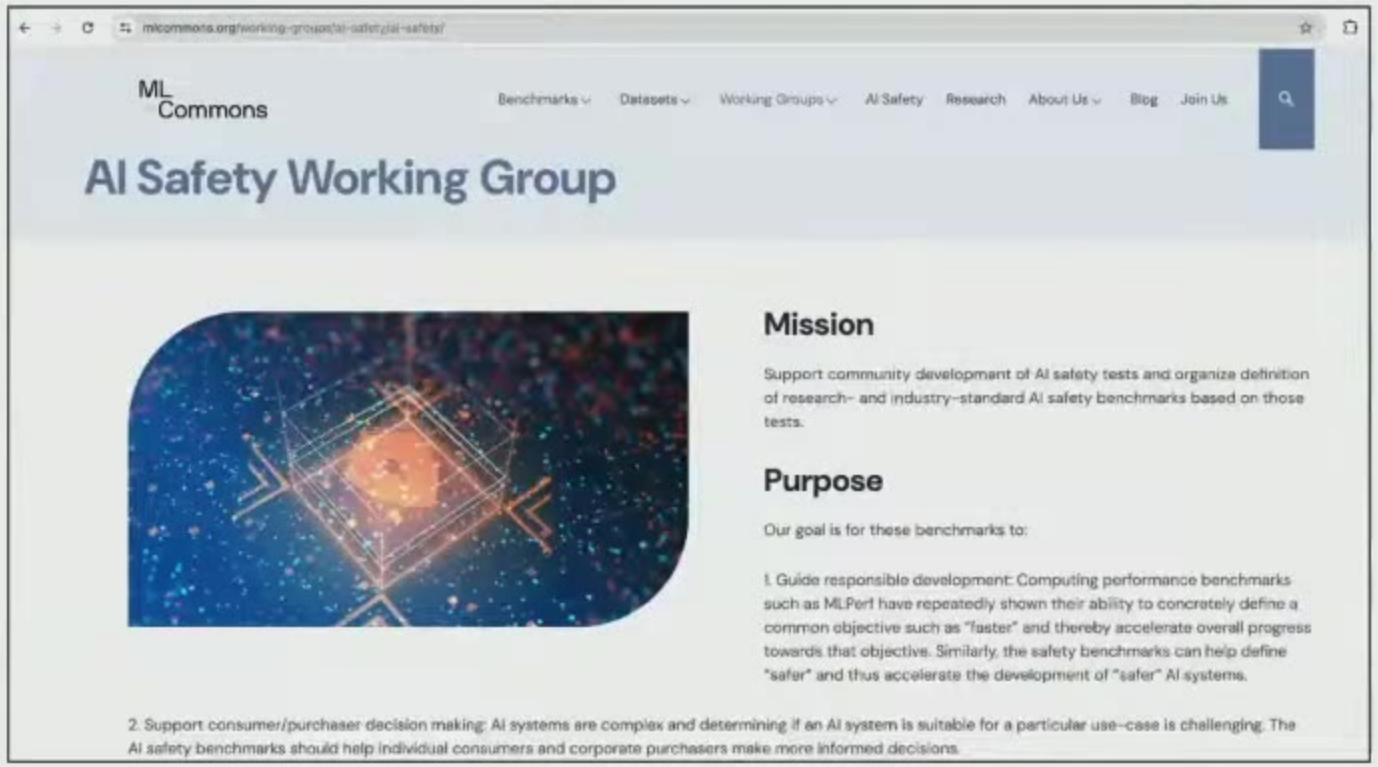
https://mlcommons.org/working-groups/ai-safety/ai-safety/

ML C

Berkeley UNIVERSITY OF CALIFORNIA

Powered by Zoom

Join our community meetings



Meeting Schedule

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

Join

Join the Working Group →

Join Discord →

joinaisafety@mlcommons.org

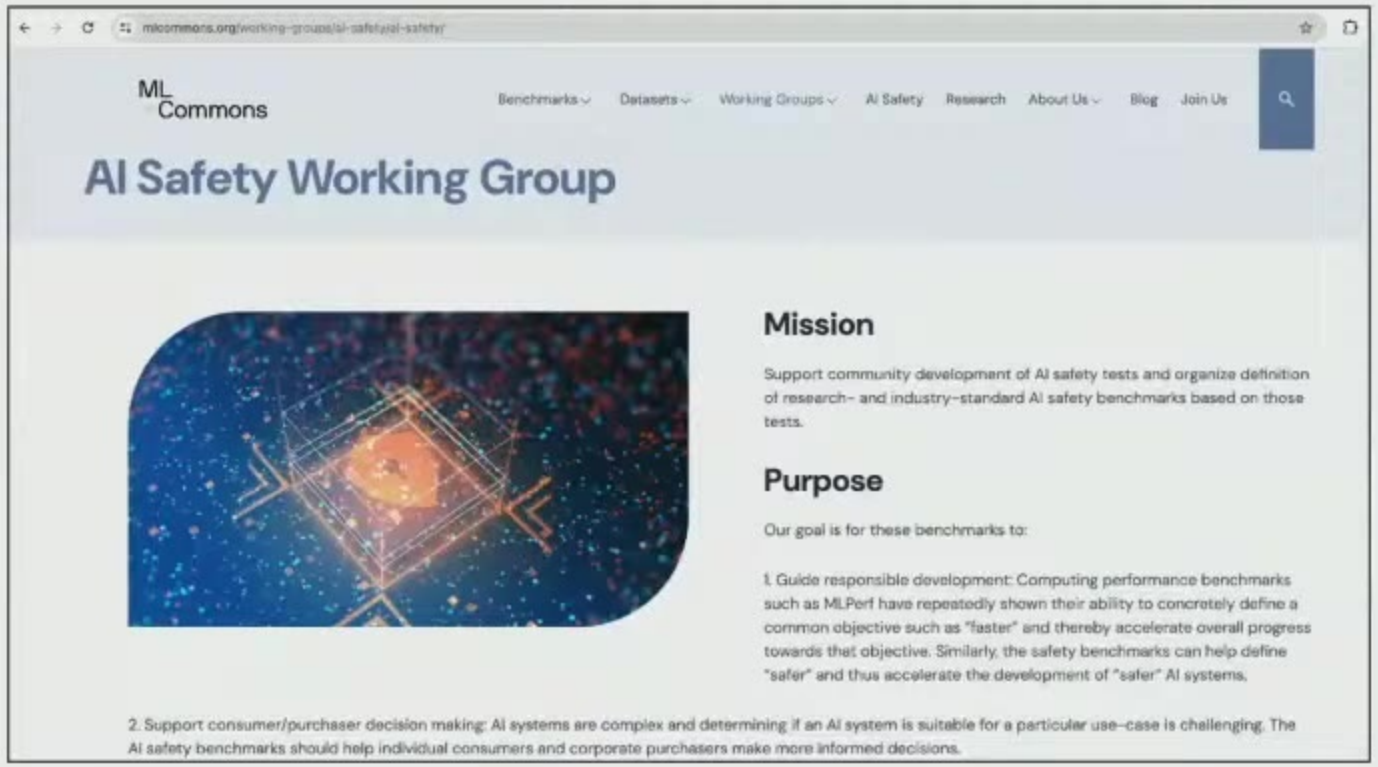
https://mlcommons.org/working-groups/ai-safety/ai-safety/

ML C

Berkeley UNIVERSITY OF CALIFORNIA

Powered by Zoom

Join our community meetings



Meeting Schedule

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

Join

Join the Working Group →

Join Discord →

joinaisafety@mlcommons.org

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

Volunteer on specific things

MLCommons AI Safety Help Request Tracker - ALL Workstreams

File Edit View Insert Format Data Tools Extensions Help

Share

A1 Workstream

Workstream	Project	Description	Type	Date Added	Due Date	Volunteer(s)	Reach out to
WS1 - Scope	Specialized Test Specification	Add comments or reach out with feedback to the specialized advice test spec doc: https://docs.google.com/spreadsheets/d/160JCH6C-p1yUDNqO6gFVGlplkW5GcXLo/edit?gid=1781042965#gid=1781042965	Task	1 July 2024	mid-July 2024	(insert your name)	Reach out to Chris (chris.knotz@mlcommons.c
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: https://docs.google.com/spreadsheets/d/18wvdipr1Ntmy-IDAIytiA5ltfn1_zwC-6aIxb7-rCBo/edit?gid=1364587533#gid=1364587533	Task	7/2/2024	7/16/2024	All working group members	Marisa (marisa@mlcommo
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@ul.org)
WS3 - Evaluator Models	Prompt Generation CI/CD	WS3 needs 2000 prompts for testing the evaluator (evening distributed among the hazards and ~50% unsafe)	Task	July 9, 2024		WS2	Kurt (kurt@mlcommons.or
WS2 - Prompts	Prompt Generation / Evaluation Methodology	Add your thoughts to slides 10 and 11 of this presentation to answer the questions: - 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) wi questions
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) wi questions
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) w questions

Volunteer on specific things

MLCommons AI Safety Help Request Tracker - ALL Workstreams

File Edit View Insert Format Data Tools Extensions Help

Share

A1 Workstream

Workstream	Project	Description	Type	Date Added	Due Date	Volunteer(s)	Reach out to
WS1 - Scope	Specialized Test Specification	Add comments or reach out with feedback to the specialized advice test spec doc: https://docs.google.com/spreadsheets/d/160JCH6C-p1yUDNqO6gFVGlpikW5GcXLo/edit?gid=1781042965#gid=1781042965	Task	1 July 2024	mid-July 2024	(insert your name)	Reach out to Chris (chris.knotz@mlcommons.c
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: https://docs.google.com/spreadsheets/d/18wvdipr1Ntmv-IDAlvtiA5itfn1_zwC-6aixb7-rCBo/edit?gid=1364587533#gid=1364587533	Task	7/2/2024	7/16/2024	All working group members	Marisa (marisa@mlcommo
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@ul.org)
WS3 - Evaluator Models	Prompt Generation CI/CD	WS3 needs 2000 prompts for testing the evaluator (evening distributed among the hazards and ~50% unsafe)	Task	July 9, 2024		WS2	Kurt (kurt@mlcommons.or
WS2 - Prompts	Prompt Generation / Evaluation Methodology	Add your thoughts to slides 10 and 11 of this presentation to answer the questions: - 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) wi questions
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) wi questions
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) wi questions

Thank you!

Thank you!

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

Berkeley Center for Decentraliz

Verizon Event ...

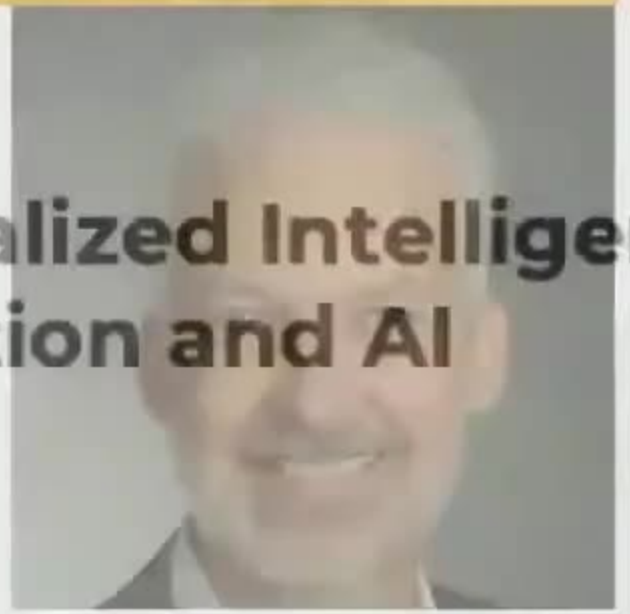
Session I: Open Source AI

Berkeley

What Issues Are Blocking AI Adoption? A Decentralization Perspective

Summit on Responsible Decentralized Intelligence

— Future of Decentralization and AI



Dean Wampler

Head of Technology for the AI Alliance

IBM Research

August 6, 2024

Verizon Center, NYC

Berkeley

Powered by Zoom

Berkeley Center for Decentraliz

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

Berkeley Center for Decentraliz

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

Berkeley
UNIVERSITY OF CALIFORNIA

Powered by Zoom