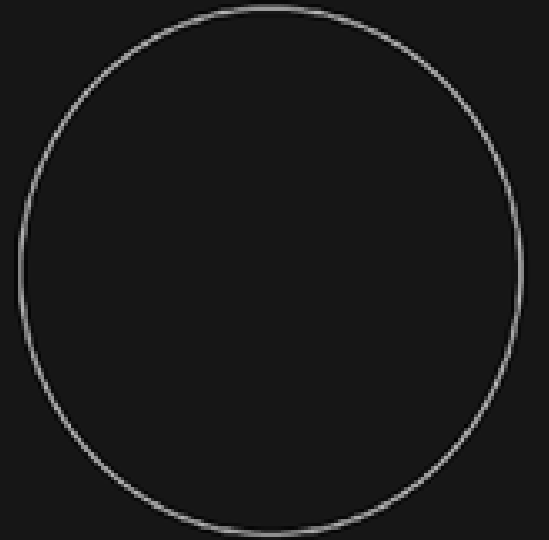


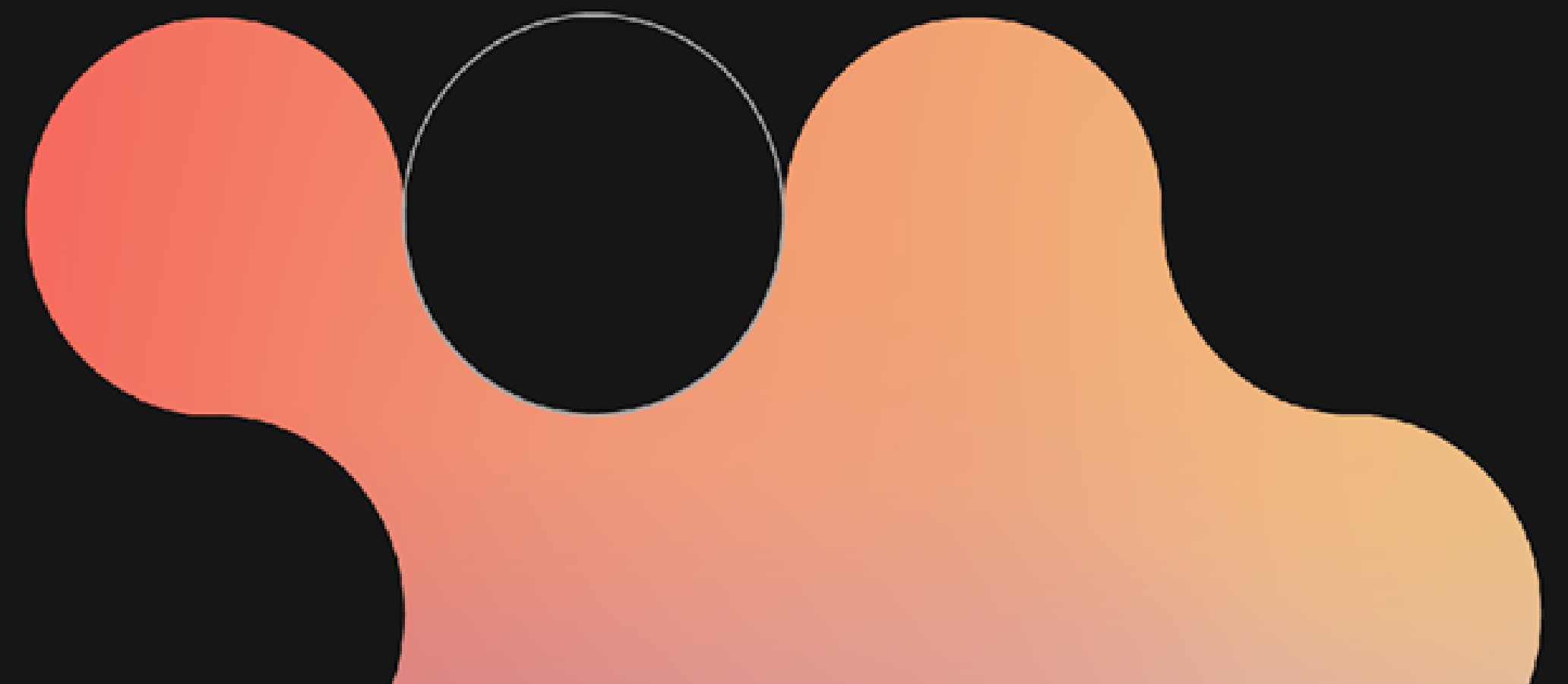
The AI Alliance

Trust and Safety Evaluations Initiative



Overview

March 12, 2025



Focus Areas & Mission

Represents the investment priorities for the AI Alliance

1. Skills & Education

Support global AI skills building, education, and exploratory research.

2. Trust & Safety

Create benchmarks, tools, and methodologies to ensure and evaluate high-quality and safe AI.

3. Applications & Tools

Build and advance efficient and capable software frameworks for model builders and developers.

Member organizations have the choice to take part in one or more of these six focus areas and the agility to shift participation based on their interest and priorities.

4. HW Enablement

Foster a vibrant AI hardware accelerator ecosystem through SW.

5. Foundation Models & Data

Enable an ecosystem of open foundation models and datasets for diverse modalities.

6. Advocacy

Advocate for regulatory policies that create a healthy open ecosystem for AI.

Trust & Safety Evaluations Initiative (TSEI)

Problem Statement

AI builders need to evaluate models and apps for various concerns.

- I am not an expert; where do I start?
- What are the most important concerns for my domain and use cases?
- For those concerns, are there evaluations defined for them?
- How do different open models score for those evaluations?
- How can I run those evaluations on my private tuned models and apps?

Trust & Safety Evaluations Initiative (TSEI)

What We Are Building

- **Taxonomy of evaluations:** Safety, alignment, performance, etc.
- **Evaluators and Benchmarks:** Implementations of *evaluations*
- **Leaderboards:** Find the evaluations for your domains and use-cases.
- **Reference Stack:** Run evaluators offline, during R&D, and online, during inference.

An Initiative of Focus Area 2: AI Trust and Safety

TSEI Website

The screenshot shows the TSEI website interface. On the left is a navigation menu with the following items: Home, Glossary of Terms, User Personae and Their Needs, Taxonomy, Evaluators and Benchmarks, Leaderboards, Evaluation Platform Reference Stack, Contributing, and About Us. The main content area features a search bar at the top right, a hero banner with the text 'The AI Alliance' and abstract colorful shapes, and two buttons: 'Join Our Initiative' and 'GitHub Repo'. Below this is a section titled 'Trust and Safety Evaluations Initiative' containing a table with the following data:

Authors	The AI Alliance Trust and Safety Work Group (see About Us)
Last Update	V0.4.2, 2025-03-07

At the bottom of the page, there is a copyright notice: 'Copyright © 2024-2025, The AI Alliance.' and a license notice: 'This work is licensed under CC BY 4.0'. A footer note states: 'This site uses Jekyll and Just the Docs, a documentation theme.'

Welcome to the **The AI Alliance** initiative for **Trust and Safety Evaluations**.

Unlike traditional software systems that rely on prescribed specifications and application code, *AI systems* based on machine learning *models* depend on training data to map inputs to outputs. Consequently, these systems are inherently non-deterministic and may produce errors due to variability in the training data or the probabilistic nature of the underlying algorithms. To *evaluate*

Trust & Safety Evaluations Initiative (TSEI)

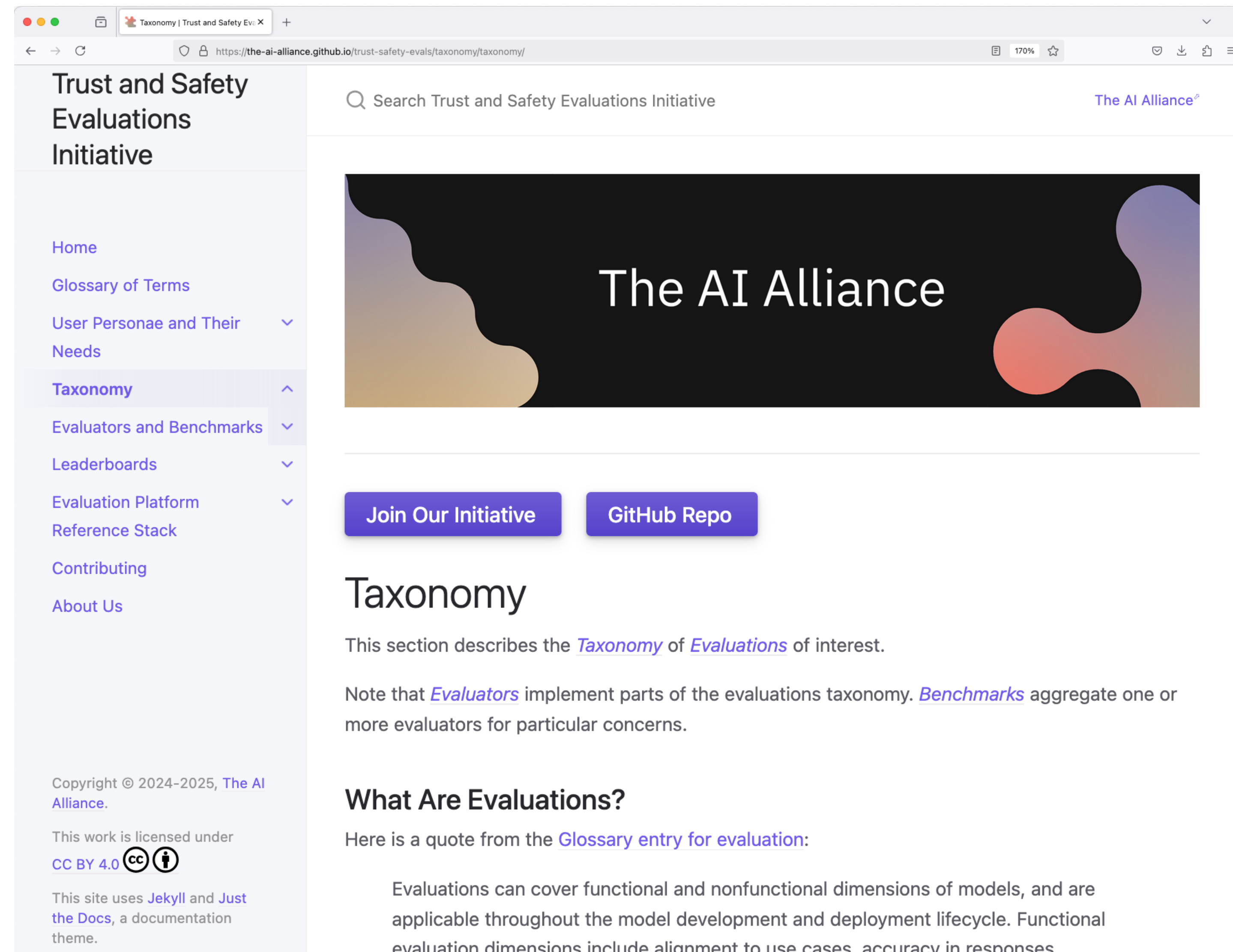
Taxonomy of Evaluations

Unify the work of industry leaders:

- E.g., MLCommons AILuminte, IBM Risk Atlas and Granite Guardian, Meta Llama Guard, ...
- Not just safety, but alignment, performance, domain-specific concerns, and user customization
- Clarify known gaps

An Initiative of Focus Area 2: AI Trust and Safety

TSEI Website



The screenshot shows a web browser displaying the 'Trust and Safety Evaluations Initiative' website. The page title is 'Trust and Safety Evaluations Initiative'. The navigation menu includes: Home, Glossary of Terms, User Personae and Their Needs, Taxonomy (highlighted), Evaluators and Benchmarks, Leaderboards, Evaluation Platform Reference Stack, Contributing, and About Us. The main content area features a search bar, a search button, and a search result for 'The AI Alliance'. Below the search bar, there are two buttons: 'Join Our Initiative' and 'GitHub Repo'. The main heading is 'Taxonomy', followed by a paragraph: 'This section describes the *Taxonomy* of *Evaluations* of interest. Note that *Evaluators* implement parts of the evaluations taxonomy. *Benchmarks* aggregate one or more evaluators for particular concerns.' Below this, there is a section titled 'What Are Evaluations?' with a quote from the 'Glossary entry for evaluation': 'Evaluations can cover functional and nonfunctional dimensions of models, and are applicable throughout the model development and deployment lifecycle. Functional evaluation dimensions include alignment to use cases, accuracy in responses'.

Trust & Safety Evaluations Initiative (TSEI)

Evaluators and Benchmarks

TSEI Website

Implementations of the *evaluations*:

- Aggregate the tools used by MLCommons, IBM, Llama Guard, Meta, and others
- Implement missing evaluators for defined evaluations.
- Make it easy for users to define custom evaluators!

The screenshot shows a web browser window displaying the 'Trust and Safety Evaluations Initiative' website. The page title is 'Evaluators and Benchmarks | Tr'. The URL is 'https://the-ai-alliance.github.io/trust-safety-evals/evaluators/evaluators/'. The page features a search bar, a navigation menu, and a main content area. The navigation menu includes links for Home, Glossary of Terms, User Personae and Their Needs, Taxonomy, Evaluators and Benchmarks (highlighted), Leaderboards, Evaluation Platform, Reference Stack, Contributing, and About Us. The main content area has a header with 'The AI Alliance' logo and two buttons: 'Join Our Initiative' and 'GitHub Repo'. Below this, the section is titled 'Evaluators and Benchmarks' and contains text describing the evaluators and benchmarks. The footer includes copyright information for 2024-2025, The AI Alliance, and mentions that the work is licensed under CC BY 4.0 and the site uses Jekyll and Just the Docs theme.

Trust & Safety Evaluations Initiative (TSEI)

Leaderboards

TSEI Website

Discovery for users:

- Find evaluations (with corresponding evaluators and benchmarks) for their domain, use cases, etc.?
- See how popular, open models perform
- Download deployable configurations

The screenshot shows a web browser displaying the 'Leaderboards' page of the Trust and Safety Evaluations Initiative. The page features a dark header with the 'The AI Alliance' logo and a search bar. A navigation sidebar on the left lists various sections, with 'Leaderboards' highlighted. Below the sidebar, there are two prominent blue buttons: 'Join Our Initiative' and 'GitHub Repo'. The main content area is titled 'Leaderboards' and contains introductory text about the initiative's tools and goals. At the bottom, there is a section titled 'Plans for Leaderboards and Other Tools'.

Trust and Safety Evaluations Initiative

Search Trust and Safety Evaluations Initiative

The AI Alliance

Home

Glossary of Terms

User Personae and Their Needs

Taxonomy

Evaluators and Benchmarks

Leaderboards

SafetyBAT Leaderboard

Risk Atlas Nexus

Evaluation Platform Reference Stack

Contributing

About Us

Copyright © 2024-2025, The AI Alliance.

This work is licensed under CC BY 4.0

This site uses Jekyll and Just the Docs, a documentation theme.

Join Our Initiative

GitHub Repo

Leaderboards

This section describes the leaderboards and related tools that are maintained by this initiative or separately by other AI Alliance members.

The leaderboards provide results from running benchmark suites of the [evaluators](#) against various models and AI systems that use them.

The other tools assist software engineers in identifying important risks for their use cases and finding the evaluators and benchmarks that support testing for those risks.

Plans for Leaderboards and Other Tools

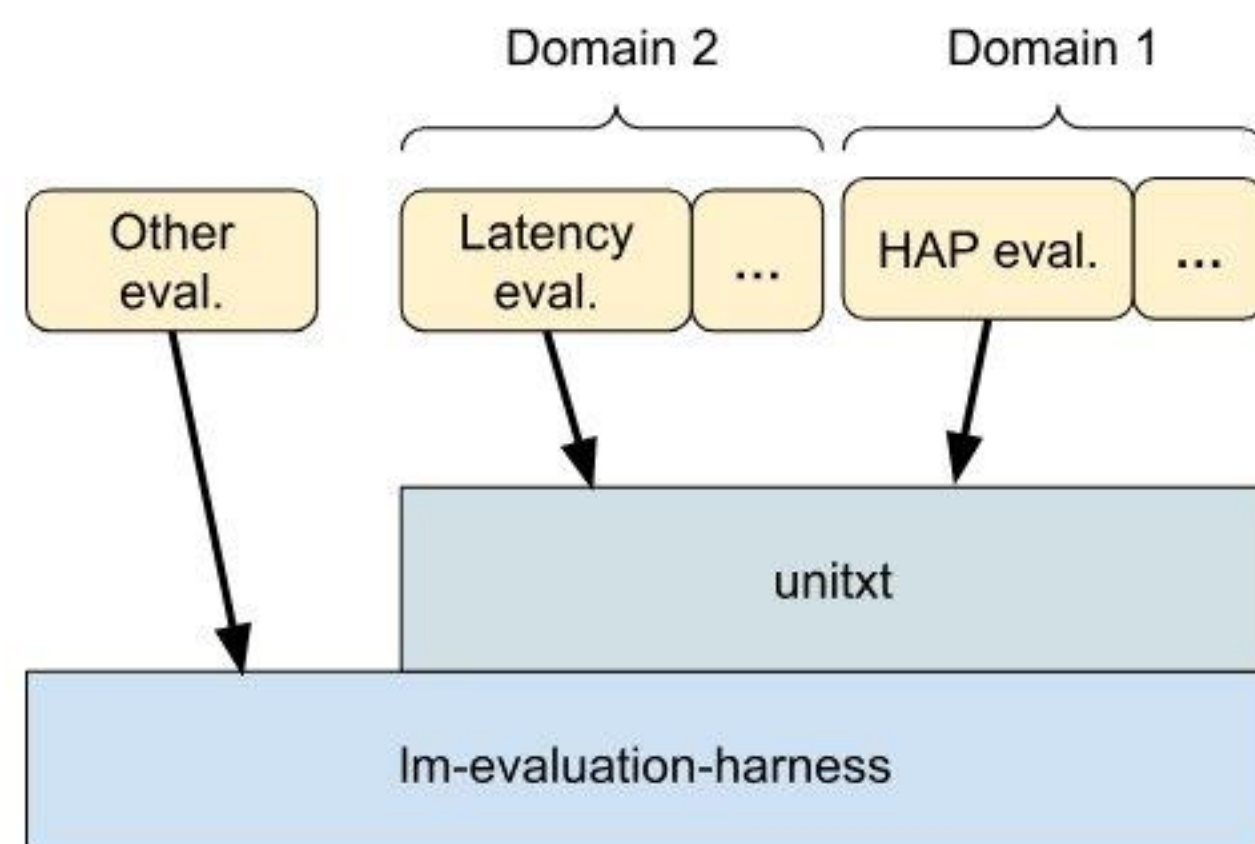
Planned leaderboards will include the leading open-source models to serve as evaluation targets

Trust & Safety Evaluations Initiative (TSEI)

Reference Stack

How to run the evaluators and benchmarks

- Industry-leading stack: lm-evaluation-harness, unitxt, ...
- Download and deploy configurations from the leaderboards



TSEI Website

The screenshot shows the TSEI Website interface. The page title is 'Trust and Safety Evaluations Initiative'. The main content area features a search bar, a navigation menu, and a large banner for 'The AI Alliance'. Below the banner are two buttons: 'Join Our Initiative' and 'GitHub Repo'. The main heading is 'Evaluation Platform Reference Stack', followed by a paragraph describing the reference stack and its purpose. The footer contains copyright information and licensing details.

Trust & Safety Evaluations Initiative (TSEI)

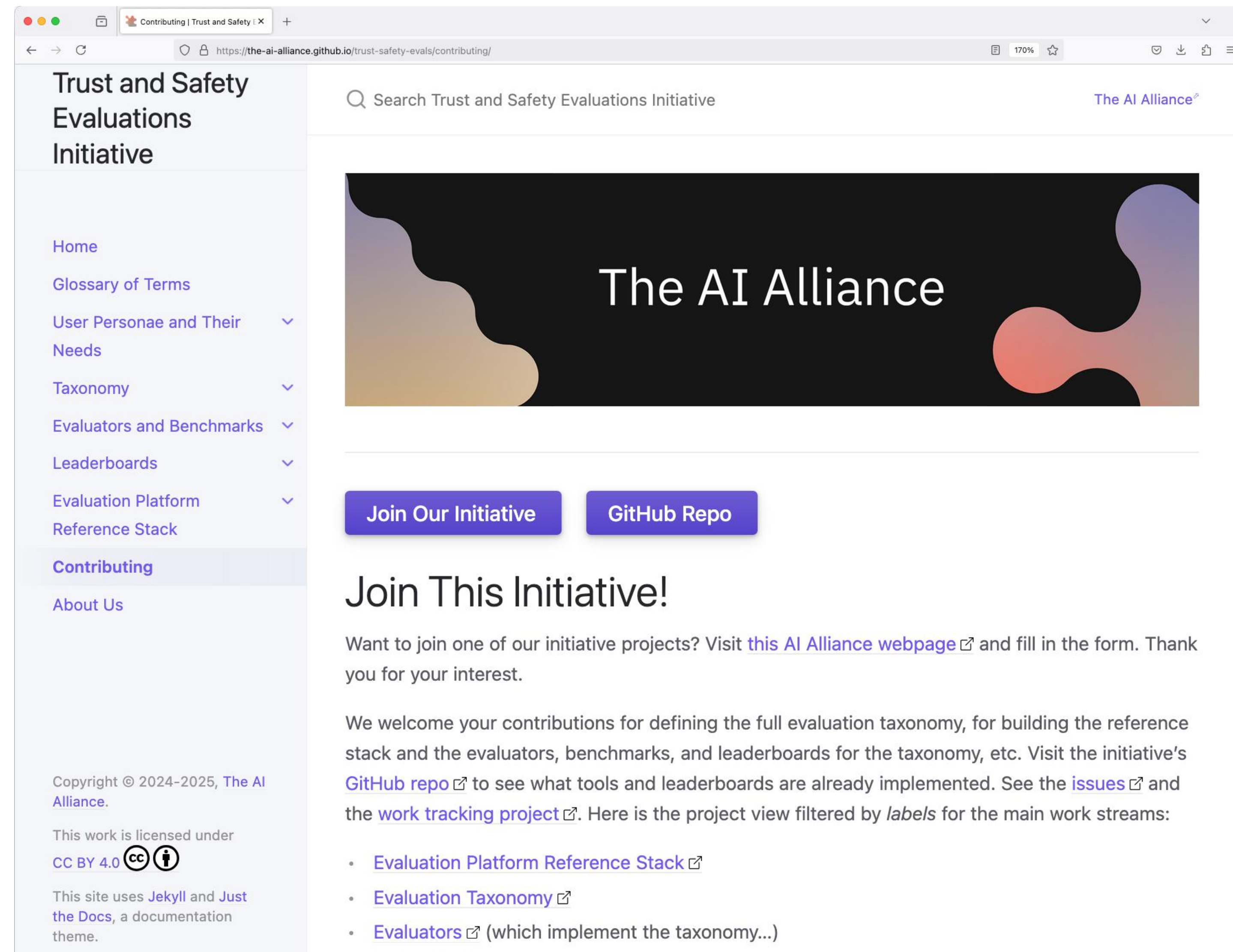
Please Join Us!!

Help us help you...

- Contribute your evaluation expertise
- Contribute your domain expertise
- Help us implement the reference stack, evaluators, and leaderboards

An Initiative of Focus Area 2: AI Trust and Safety

TSEI Website



The screenshot shows a web browser window with the URL <https://the-ai-alliance.github.io/trust-safety-evals/contributing/>. The page title is "Trust and Safety Evaluations Initiative". The navigation menu includes: Home, Glossary of Terms, User Personae and Their Needs, Taxonomy, Evaluators and Benchmarks, Leaderboards, Evaluation Platform Reference Stack, Contributing (highlighted), and About Us. The main content area features a search bar, a search button, and a search results section. The search results section includes a large banner with the text "The AI Alliance" and two buttons: "Join Our Initiative" and "GitHub Repo". Below the banner, there is a heading "Join This Initiative!" followed by a paragraph: "Want to join one of our initiative projects? Visit [this AI Alliance webpage](#) and fill in the form. Thank you for your interest." Below this, there is another paragraph: "We welcome your contributions for defining the full evaluation taxonomy, for building the reference stack and the evaluators, benchmarks, and leaderboards for the taxonomy, etc. Visit the initiative's [GitHub repo](#) to see what tools and leaderboards are already implemented. See the [issues](#) and the [work tracking project](#). Here is the project view filtered by *labels* for the main work streams:" followed by a list of links:

- [Evaluation Platform Reference Stack](#)
- [Evaluation Taxonomy](#)
- [Evaluators](#) (which implement the taxonomy...)

Copyright © 2024-2025, The AI Alliance.

This work is licensed under [CC BY 4.0](#)

This site uses [Jekyll](#) and [Just the Docs](#), a documentation theme.