The AI Alliance

Trust and Safety Evaluations Initiative

Overview

March 12, 2025

thealliance.ai

Focus Areas & Mission

Represents the investment priorities for the AI Alliance

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.

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.

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.

AI Alliance

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?

An Initiative of Focus Area 2: AI Trust and Safety

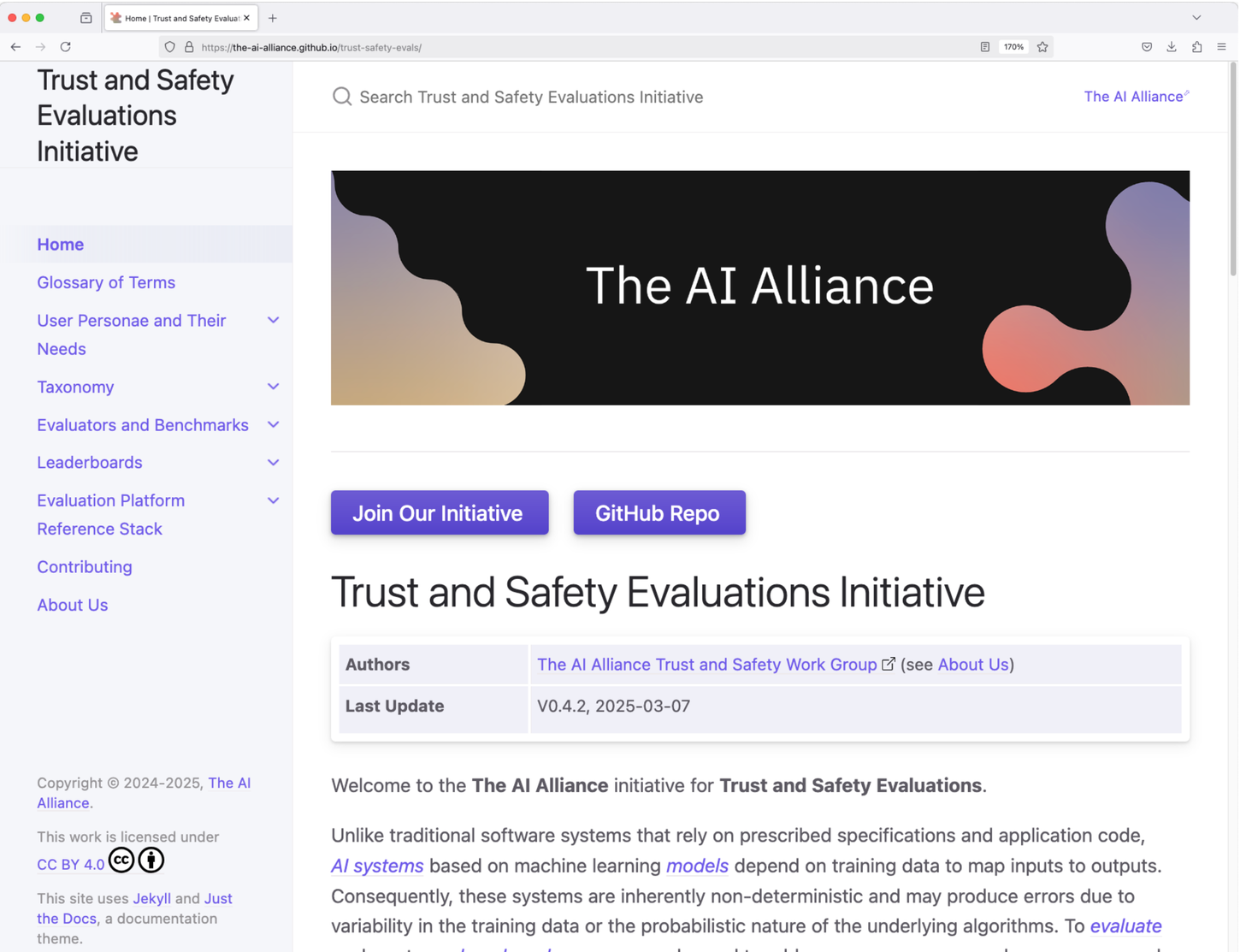
Trust & Safety Evaluations Initiative (TSEI)

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



Trust & Safety Evaluations Initiative (TSEI)

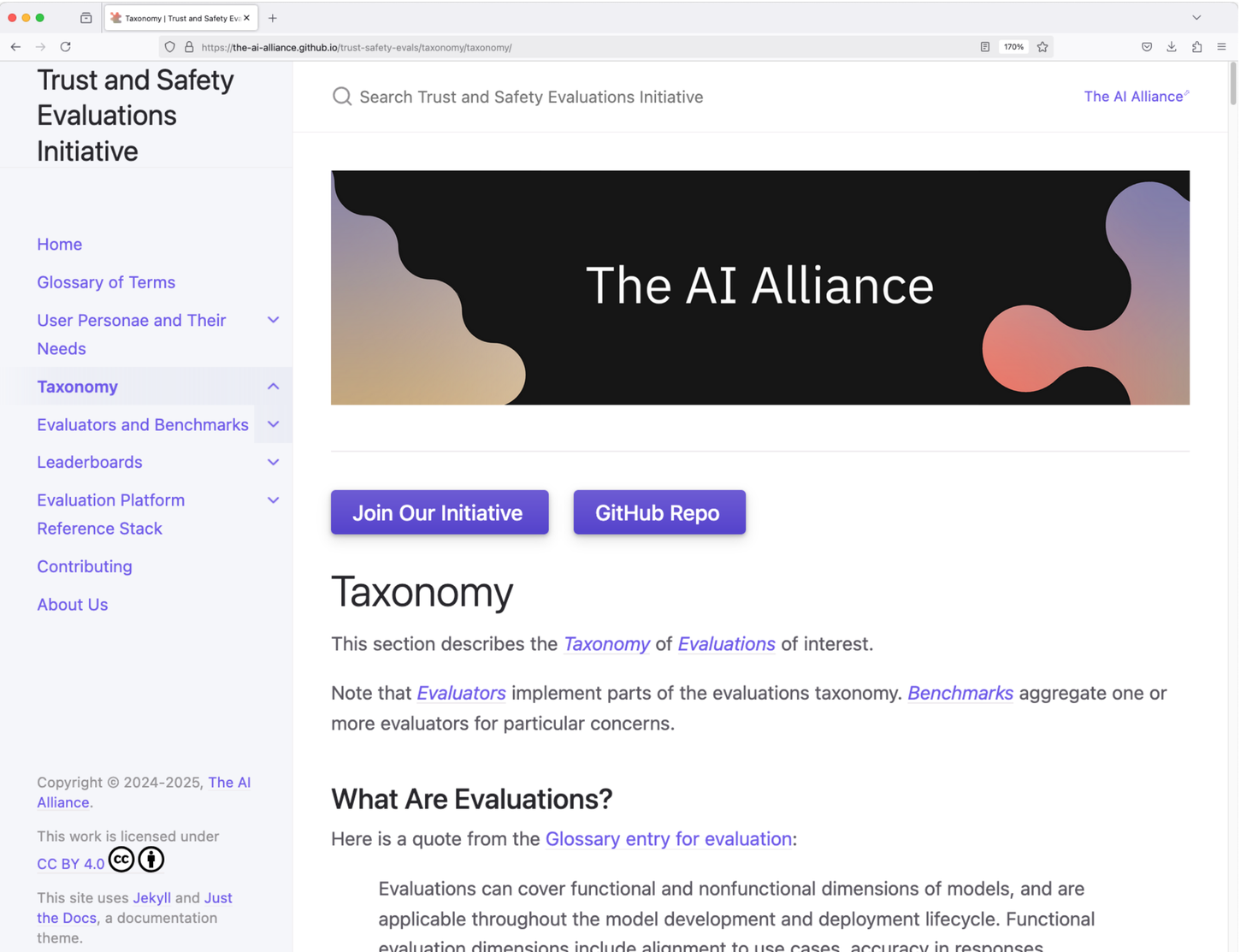
Taxonomy of Evaluations

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



Trust & Safety Evaluations Initiative (TSEI)

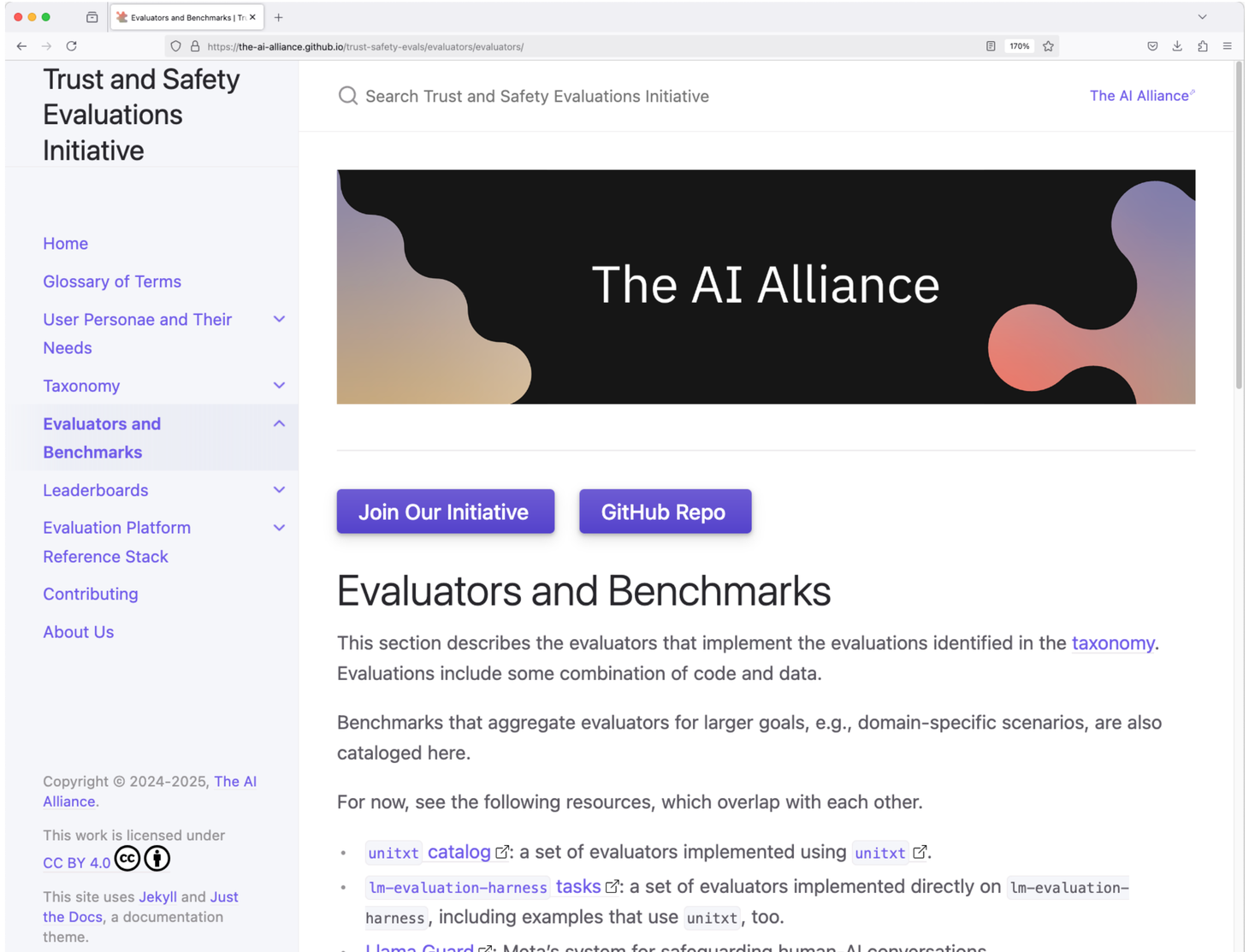
Evaluators and Benchmarks

Implementations of the *evaluations*:

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

An Initiative of Focus Area 2: AI Trust and Safety

TSEI Website



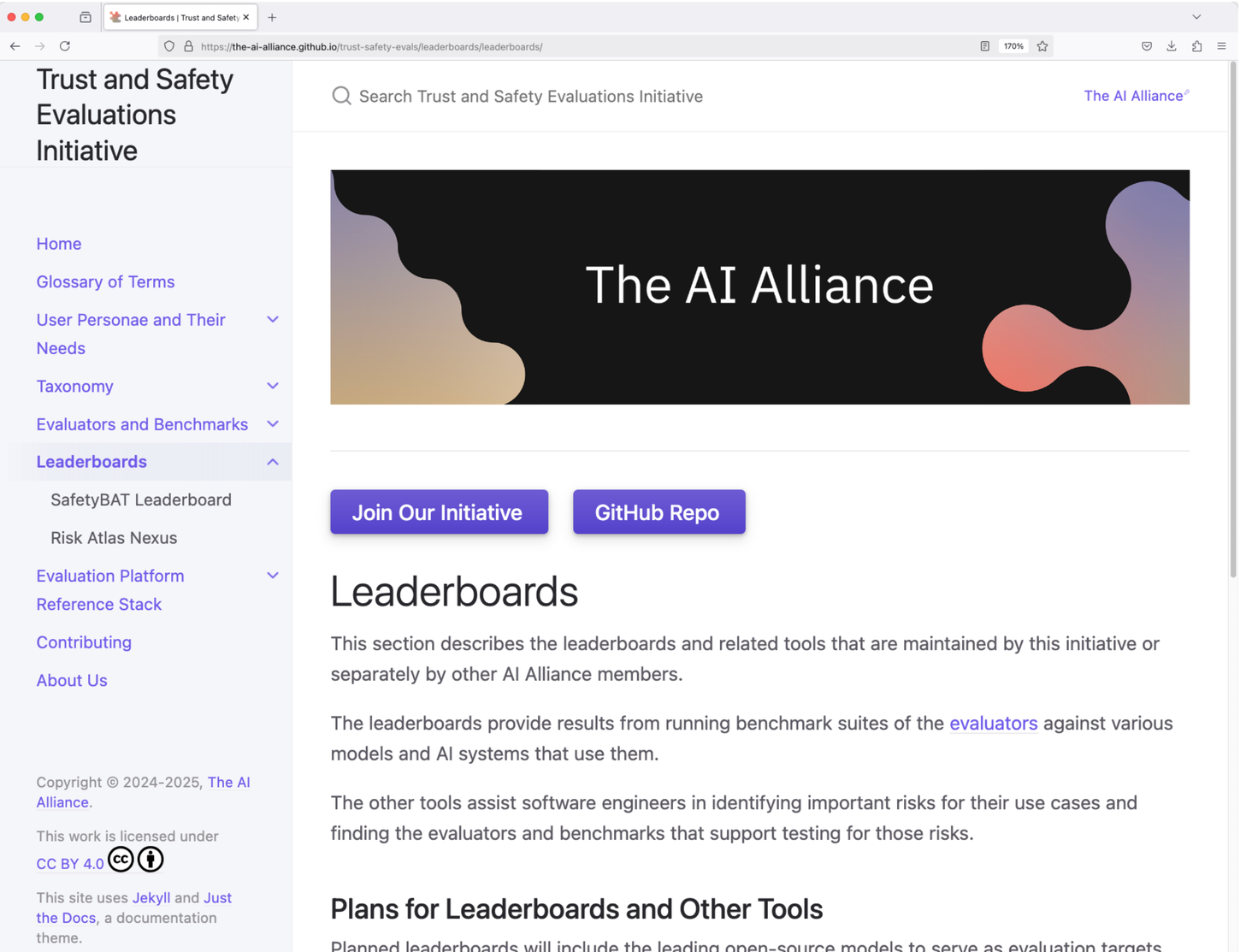
Trust & Safety Evaluations Initiative (TSEI)

Leaderboards

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

TSEI Website



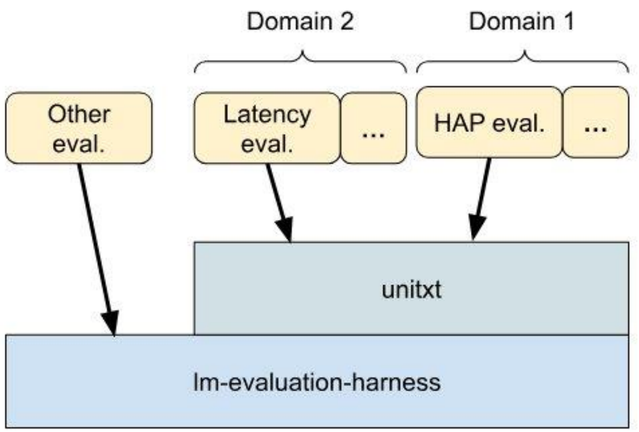
An Initiative of Focus Area 2: AI Trust and Safety

Trust & Safety Evaluations Initiative (TSEI)

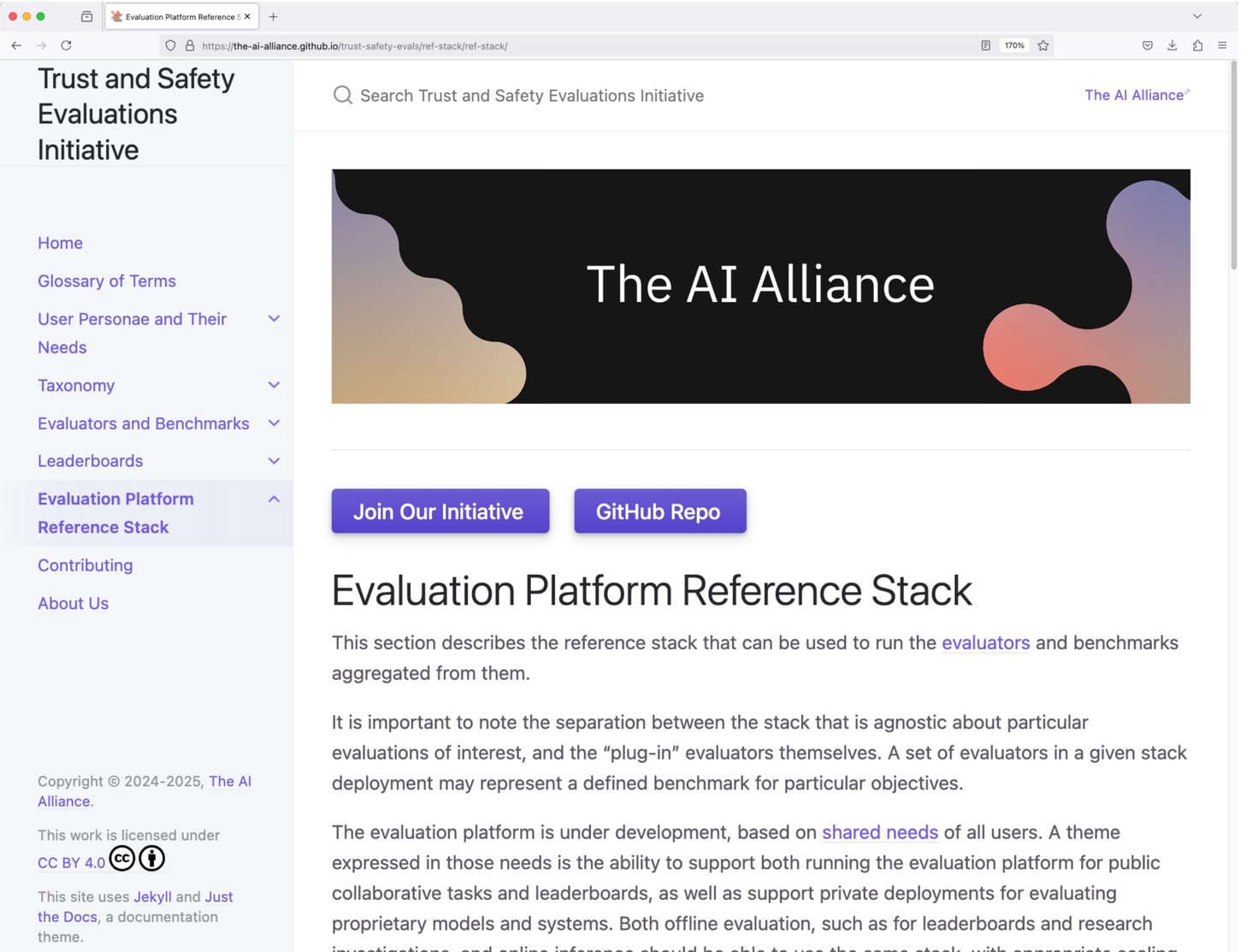
Reference Stack

How to run the evaluators and benchmarks

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



TSEI Website



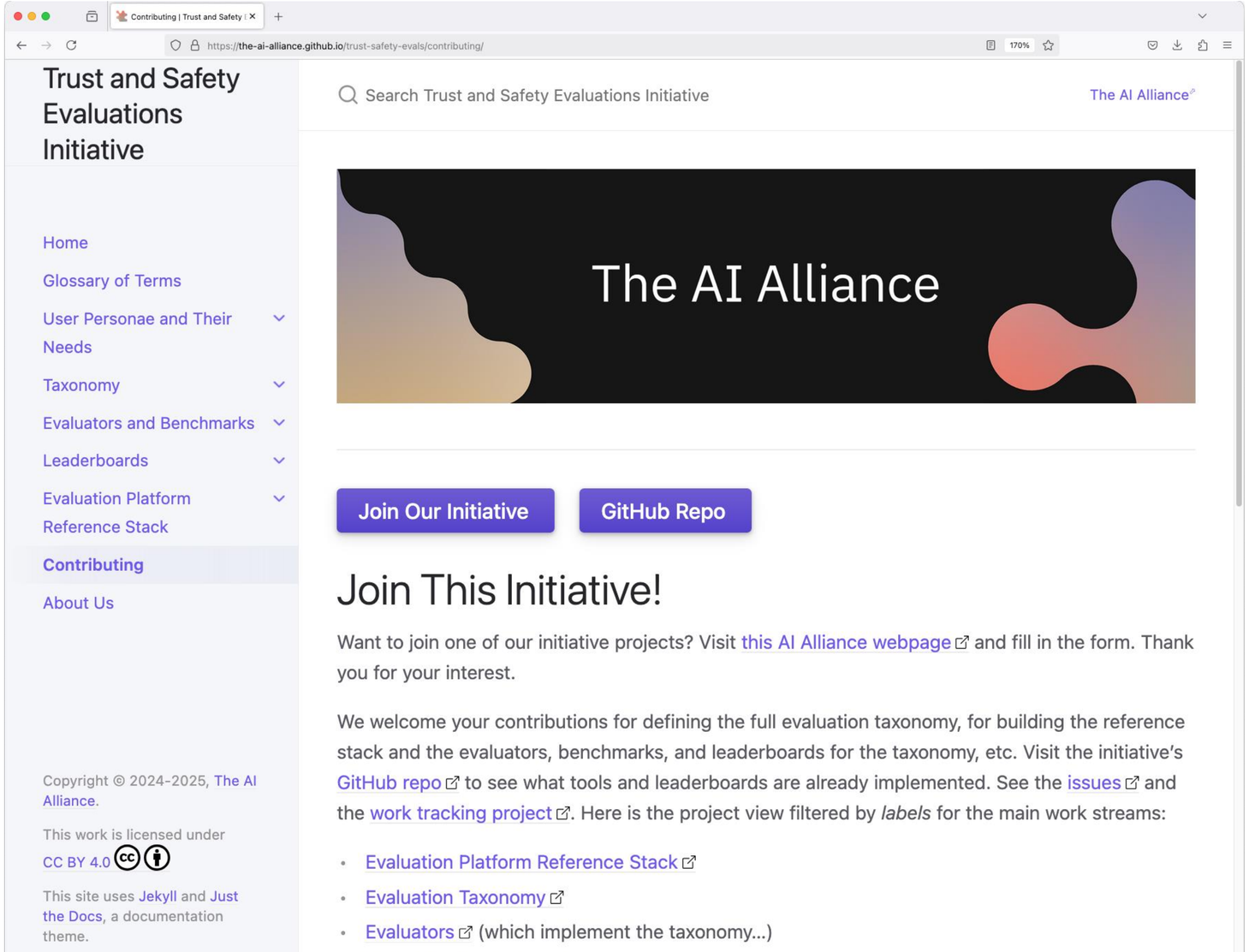
Trust & Safety Evaluations Initiative (TSEI)

Please Join Us!!

Help us help you…

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

TSEI Website



An Initiative of Focus Area 2: AI Trust and Safety