

AI Alliance

Trust and Safety Evaluations Initiative

the-ai-alliance.github.io/trust-safety-evals/

June 5, 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.

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 applications**:

- I am not an expert; where do I start?
- Good techniques exist for *evaluating* safety, but what about the “*last mile*” of evaluation; how do I evaluate if my app works well for *my* domain and *my* use cases?
- If I have those evaluations, how do different open models score for those evaluations?
- How do I run those evaluations locally on my privately tuned models and applications built with RAG, agents, etc.?

Trust & Safety Evaluations Initiative (TSEI) - Current Picture

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.

the-ai-alliance.github.io/trust-safety-evals/

The screenshot shows a web browser displaying the homepage of the AI Alliance Trust and Safety Evaluations Initiative (TSEI). The URL in the address bar is the-ai-alliance.github.io/trust-safety-evals/. The page features a header with the AI Alliance logo and the text "Trust and Safety Evaluations Initiative". A sidebar on the left lists navigation links: 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 contains a search bar, two blue buttons ("Join Our Work Group" and "Visit Our GitHub Repo"), and a section titled "Trust and Safety Evaluations Initiative (TSEI)". It states: "Our goal is to create the world's most comprehensive and useful tool set and resources for anyone seeking to benchmark, evaluate, and monitor AI systems." Below this is a table with two rows: "Authors" (The AI Alliance Trust and Safety Work Group) and "Last Update" (V0.4.4, 2025-06-03). A welcome message at the bottom says: "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 such systems, benchmarks are commonly used to address user concerns, such as accuracy and bias. However, since benchmarks can be manipulated over time to achieve favorable results, it is essential to establish a flexible evaluation framework that supports rapid updates to evaluation criteria and benchmark selection. Given the critical role of testing and evaluation in deploying AI systems, there is a pressing need for a consistent methodology and robust tool support for these activities."

Trust & Safety Evaluations Initiative (TSEI) – Current Picture

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

the-ai-alliance.github.io/trust-safety-evals/

The screenshot shows a web browser window displaying the AI Alliance Trust and Safety Evaluations Initiative website. The URL in the address bar is the-ai-alliance.github.io/trust-safety-evals/taxonomy/taxonomy/. The page features a navigation sidebar on the left with links like Home, Glossary of Terms, User Personae and Their Needs, Taxonomy (which is expanded), Evaluators and Benchmarks, Leaderboards, Evaluation Platform, Reference Stack, Contributing, and About Us. The main content area has a search bar at the top and two blue buttons: "Join Our Work Group" and "Visit Our GitHub Repo". Below these buttons is a section titled "Taxonomy" which contains text about the taxonomy of evaluations and evaluators. Another section titled "What Are Evaluations?" includes a quote from the glossary entry for evaluation.

AI Alliance

Trust and Safety Evaluations Initiative

Join Our Work Group Visit Our GitHub Repo

Taxonomy

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.

What Are Evaluations?

Here is 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, faithfulness to given context, robustness against perturbations and noise, and adherence to safety and social norms. Nonfunctional evaluation dimensions include latency, throughput, compute efficiency, cost to execute, carbon footprint and other sustainability concerns. Evaluations are applied as regression tests while models are trained and fine-tuned, as benchmarks while GenAI-powered applications are designed and developed, and as guardrails when these applications are deployed in production. They also have a role in compliance, both with specific industry regulations, and with emerging government policies. Lastly, there are numerous techniques used in implementing evaluations. Common techniques are rule-based, machine learning, and hybrid approaches. These techniques can be used to evaluate various aspects of a model, such as its fairness, bias, and explainability.

Trust & Safety Evaluations Initiative (TSEI) – Current Picture

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 it easy for users to define custom evaluators!

the-ai-alliance.github.io/trust-safety-evals/

The screenshot shows a web browser window displaying the AI Alliance Trust and Safety Evaluations Initiative website. The URL in the address bar is the-ai-alliance.github.io/trust-safety-evals/. The page features a navigation menu on the left with options like Home, Glossary of Terms, User Personae and Their Needs, Taxonomy, Evaluators and Benchmarks (which is currently selected), Leaderboards, Evaluation Platform, Reference Stack, Contributing, and About Us. The main content area is titled "Evaluators and Benchmarks" and contains text about the implementations of evaluations, links to related resources like the unitxt catalog and lm-evaluation-harness tasks, and a section for "Evaluators and Benchmarks to Explore". A search bar at the top right says "Search Trust and Safety Evaluations Initiative". Two blue buttons at the top right say "Join Our Work Group" and "Visit Our GitHub Repo". The overall design is clean and modern, using a purple and white color scheme.

Trust & Safety Evaluations Initiative (TSEI) – Current Picture

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

the-ai-alliance.github.io/trust-safety-evals/

The screenshot shows a web browser window displaying the AI Alliance Trust and Safety Evaluations Initiative website. The URL in the address bar is the-ai-alliance.github.io/trust-safety-evals/leaderboards/leaderboards/. The page features a sidebar on the left with the AI Alliance logo and navigation links including Home, Glossary of Terms, User Personae and Their Needs, Taxonomy, Evaluators and Benchmarks, Leaderboards (which is currently selected), SafetyBAT Leaderboard, Risk Atlas Nexus, Evaluation Platform Reference Stack, Contributing, and About Us. The main content area has a search bar at the top and two blue buttons: "Join Our Work Group" and "Visit Our GitHub Repo". Below these buttons is a section titled "Leaderboards" with a sub-section titled "Plans for Leaderboards and Other Tools". The "Leaderboards" section contains text about the initiative's goals and the "Evaluation Platform Reference Stack" section.

AI Alliance

Trust and Safety Evaluations Initiative

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

Join Our Work Group
Visit Our 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 and as evaluation judges. Initially, we are focusing on Meta's [Llama family of models](#) and IBM's [Granite family of models](#), with others to follow.

As we fill in the evaluation [taxonomy](#), we will add corresponding evaluators and benchmarks to the leaderboards, along with search capabilities to find the topics of interest.

Finally, we plan to provide downloadable and deployable configurations of the [Evaluation Platform Reference Stack](#) with the selected evaluators for easy and rapid use.

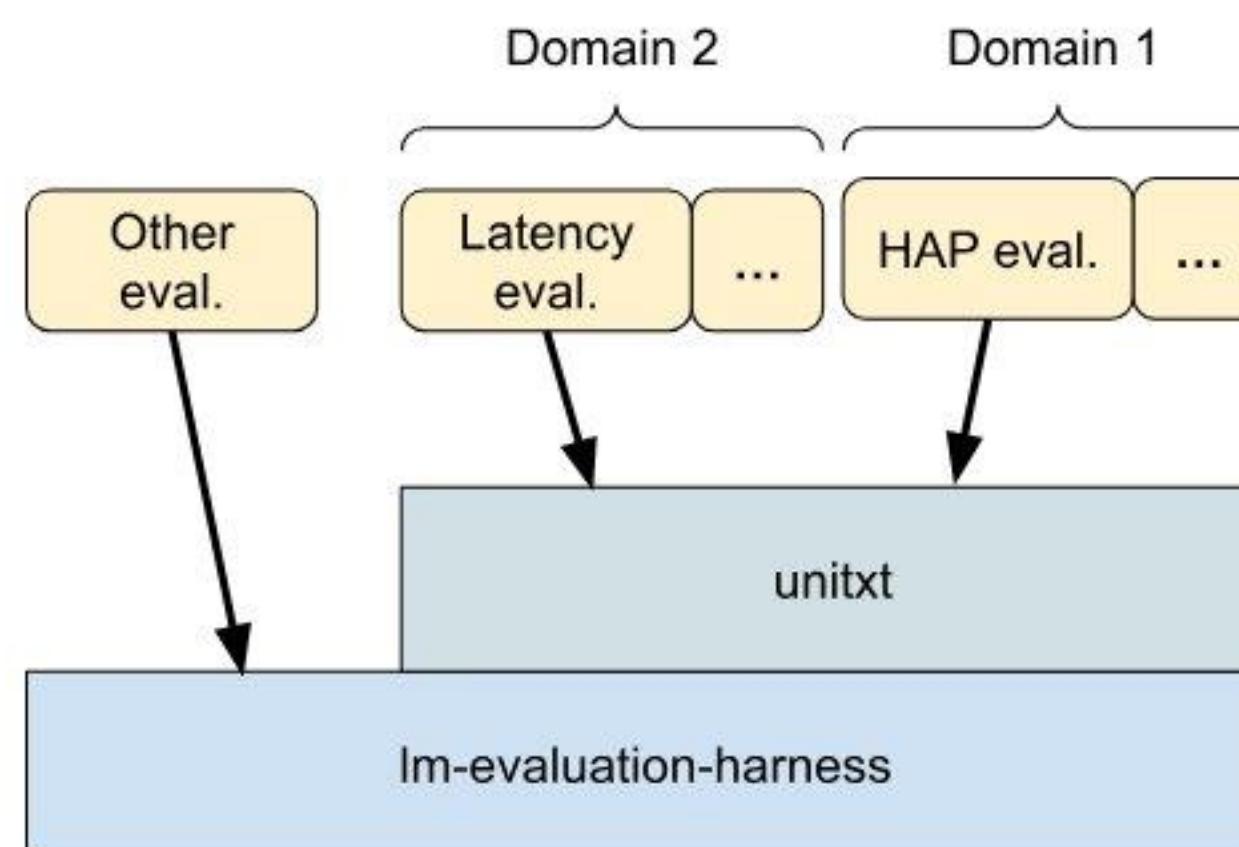
The *child* pages listed next describe the leaderboards and other tools that are currently available.

Trust & Safety Evaluations Initiative (TSEI) – Current Picture

Reference Stack

How to run the evaluators and benchmarks

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



the-ai-alliance.github.io/trust-safety-evals/

The screenshot shows a web browser window titled "Evaluation Platform Reference Stack". The URL in the address bar is "the-ai-alliance.github.io/trust-safety-evals/ref-stack/ref-stack/". The page header features the "AI Alliance" logo and the text "Trust and Safety Evaluations Initiative". On the right side, there are two purple buttons: "Join Our Work Group" and "Visit Our GitHub Repo". Below these buttons, the section title "Evaluation Platform Reference Stack" is displayed. The main content area contains text explaining the reference stack: "This section describes the reference stack that can be used to run the [evaluators](#) and benchmarks aggregated from them. It is important to note the separation between the stack that is agnostic about particular evaluations of interest, and the ‘plug-in’ evaluators themselves. A set of evaluators in a given stack deployment may represent a defined benchmark for particular objectives. The evaluation platform is under development, based on [shared needs](#) of all users. A theme expressed in those needs is the ability to support both running the evaluation platform for public collaborative tasks and leaderboards, as well as support private deployments for evaluating proprietary models and systems. Both offline evaluation, such as for leaderboards and research investigations, and online inference should be able to use the same stack, with appropriate scaling and hardening of the deployments, as required."

Architecture

Schematically, a trust and safety deployment using the reference stack with example evaluators is shown in Figure 1:

The diagram shows three separate "Example Domain" boxes, each with a bracket underneath it. These brackets are positioned below a horizontal line, indicating they are connected to or part of the "Evaluation Platform Reference Stack" shown above.

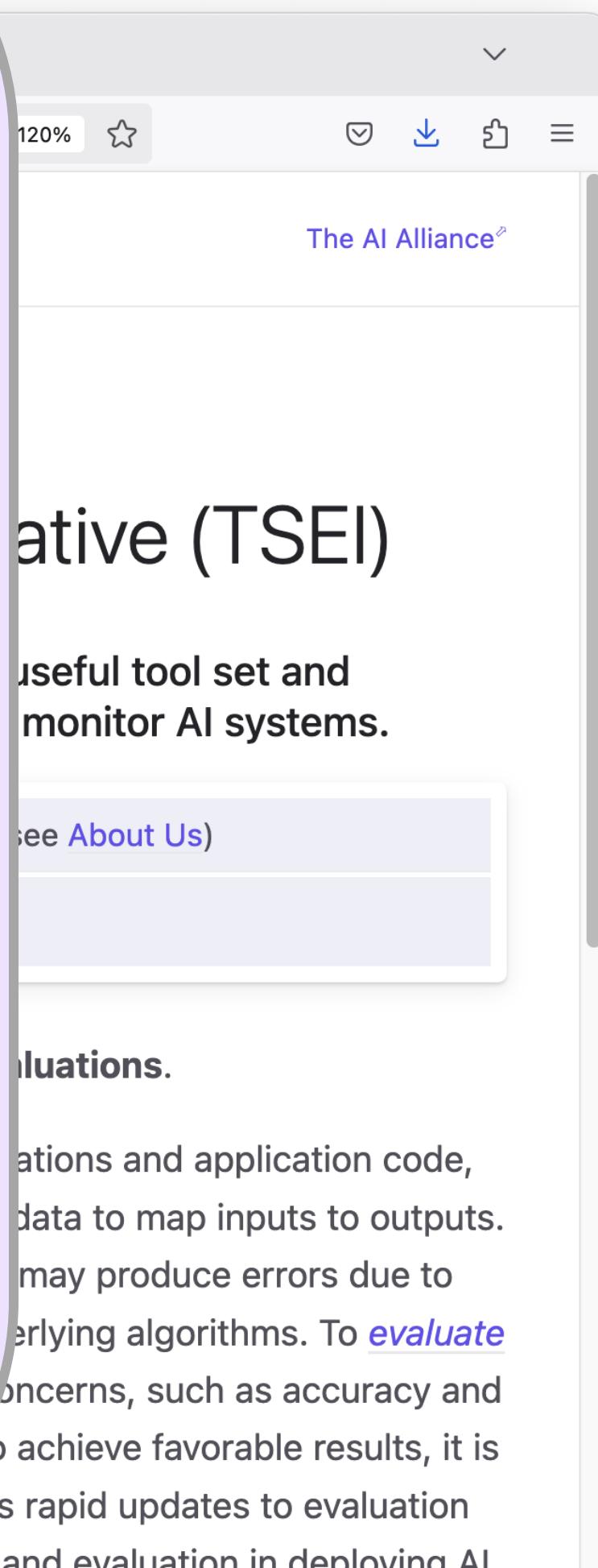
Trust & Safety Evaluations Initiative (TSEI) – Current Picture

What We Are Doing

- Taxonomy and alignment
- Evaluation framework and implementation
- Leadership and resources for your organization
- Reference materials, including offline and online training during the year

But what's not clear is the essential difference between *evaluation*, in general, and the specific area of *trust and safety*, in particular...

[trust-evals/](#)

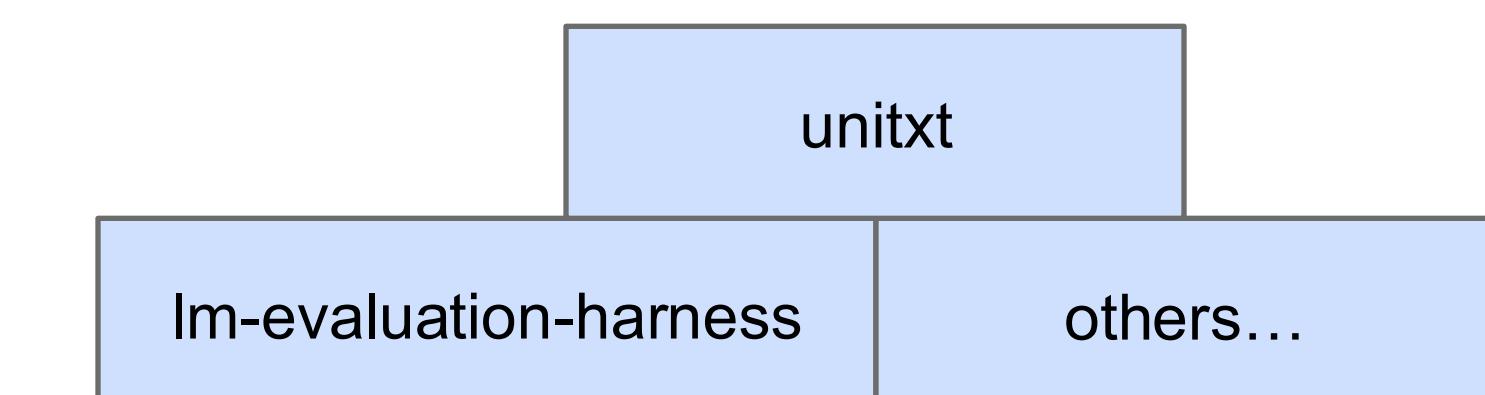


Clearer Separation of Our Efforts

- **Focus on *two key areas* of evaluation:**
- **While still working on aspects common to both:**
 - *The global taxonomy of evaluations.*
 - *Core tools and practices.*
 - *Private deployments for proprietary evaluations*
 - *Public benchmarks and leaderboards.*

Agentic Applications
Highly domain- and use case-specific evaluation required for *alignment* to requirements. This are is a leading-edge challenge.

“Classic” Trust and Safety
Prevent hate speech, harmful behavior, cybersecurity, etc.
More mature techniques and processes, but end users need help using them!



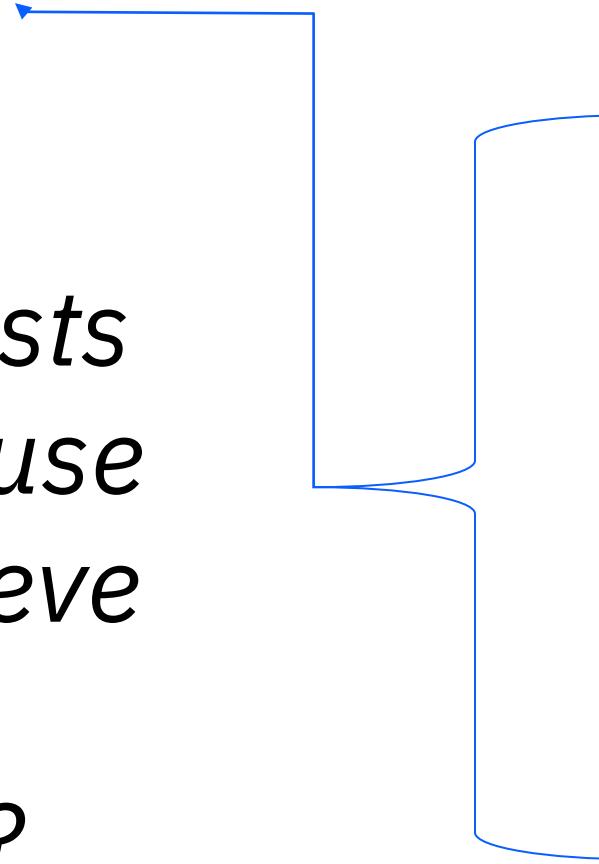
Trust & Safety Evaluations Initiative (TSEI) – Refinement

The Most Pressing Evaluation Challenge of 2025

- **The “last mile” of evaluation¹**

- *As an application developer, I am accustomed to writing deterministic tests to verify my requirements are met, my use cases are implemented. How do I achieve the same assurance when I have probabilistic AI models in the system??*

- *We have two other projects already working on this challenge...*



Agentic Applications

Highly domain- and use case-specific evaluation required for *alignment* to requirements. This are is a leading-edge challenge.

“Classic” Trust and Safety

Prevent hate speech, harmful behavior, cybersecurity, etc. More mature techniques and processes, but end users need help using them!

unitxt

Im-evaluation-harness

others...

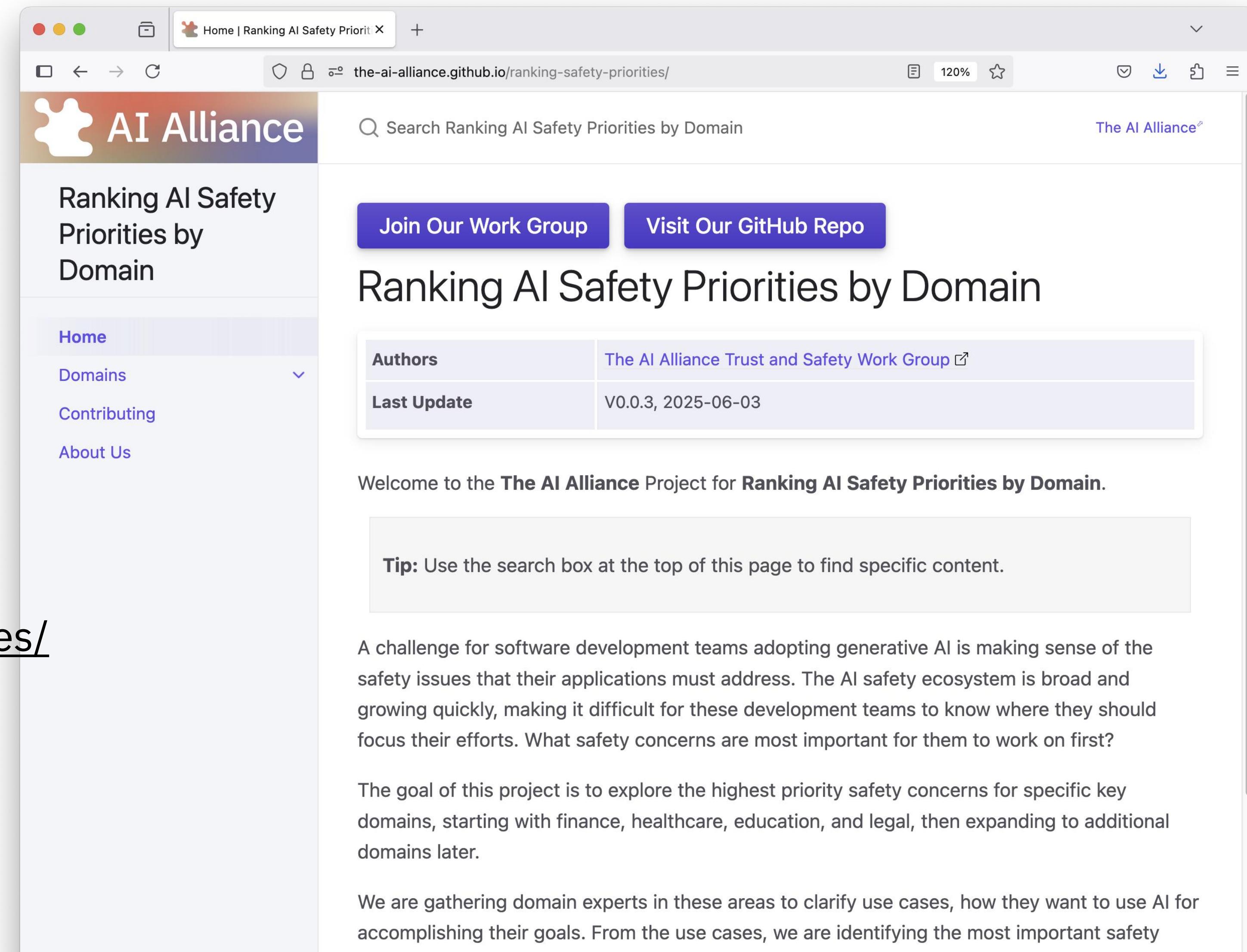
1. A reference to “cable” deployments in the last ~30 years. The last mile was the final optical fiber connection to your home to give you the high bandwidth you need for your TikTok videos...

Trust & Safety Evaluations Initiative (TSEI) – Refinement

#1: Understanding Domain-Specific Risk Concerns

Working with domain experts in healthcare and finance, with other domains planned.

the-ai-alliance.github.io/ranking-safety-priorities/



The screenshot shows a web browser window with the URL the-ai-alliance.github.io/ranking-safety-priorities/. The page has a header with the AI Alliance logo and navigation links for Home, Domains, Contributing, and About Us. The main content area is titled "Ranking AI Safety Priorities by Domain" and includes a search bar, two blue buttons ("Join Our Work Group" and "Visit Our GitHub Repo"), and a table with two rows: "Authors" (The AI Alliance Trust and Safety Work Group) and "Last Update" (V0.0.3, 2025-06-03). Below the table, a tip box says: "Tip: Use the search box at the top of this page to find specific content." To the right, there is a text block about the challenge of making sense of safety issues for software development teams and the goal of the project to explore highest priority safety concerns for specific key domains.

Ranking AI Safety Priorities by Domain

Authors: The AI Alliance Trust and Safety Work Group

Last Update: V0.0.3, 2025-06-03

Welcome to the **The AI Alliance** Project for **Ranking AI Safety Priorities by Domain**.

Tip: Use the search box at the top of this page to find specific content.

A challenge for software development teams adopting generative AI is making sense of the safety issues that their applications must address. The AI safety ecosystem is broad and growing quickly, making it difficult for these development teams to know where they should focus their efforts. What safety concerns are most important for them to work on first?

The goal of this project is to explore the highest priority safety concerns for specific key domains, starting with finance, healthcare, education, and legal, then expanding to additional domains later.

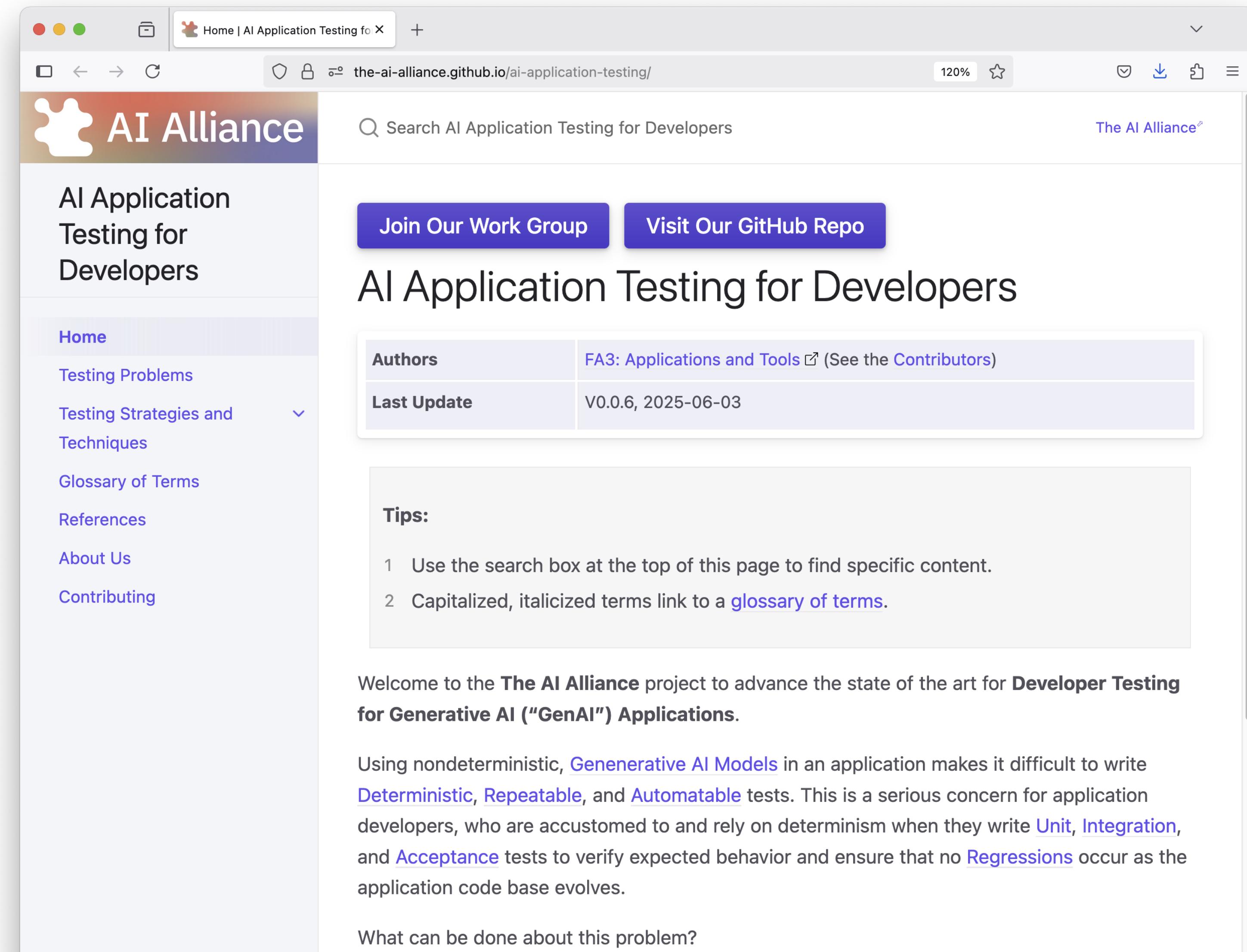
We are gathering domain experts in these areas to clarify use cases, how they want to use AI for accomplishing their goals. From the use cases, we are identifying the most important safety

Trust & Safety Evaluations Initiative (TSEI) – Refinement

#2: Last Year, I Started a Developer-Focused Project...

We are merging the two projects, because they are really tackling the same problem from different points of view...

the-ai-alliance.github.io/ai-application-testing/



The screenshot shows a web browser window displaying the AI Alliance website at the-ai-alliance.github.io/ai-application-testing/. The page has a header with the AI Alliance logo and navigation links for Home, Testing Problems, Testing Strategies and Techniques, Glossary of Terms, References, About Us, and Contributing. The main content area features a search bar, two prominent blue buttons for "Join Our Work Group" and "Visit Our GitHub Repo", and a section titled "AI Application Testing for Developers". This section includes a table with authors (FA3: Applications and Tools) and last update (V0.0.6, 2025-06-03). Below this is a "Tips:" section with two numbered items: "Use the search box at the top of this page to find specific content." and "Capitalized, italicized terms link to a glossary of terms." At the bottom, there is a welcome message about developer testing for Generative AI ("GenAI") applications and a note about nondeterministic models making it difficult to write deterministic, repeatable, and automatable tests.

Home | AI Application Testing for Developers

the-ai-alliance.github.io/ai-application-testing/

AI Alliance

AI Application Testing for Developers

Search AI Application Testing for Developers

The AI Alliance

Join Our Work Group

Visit Our GitHub Repo

AI Application Testing for Developers

Authors: FA3: Applications and Tools (See the Contributors)

Last Update: V0.0.6, 2025-06-03

Tips:

- 1 Use the search box at the top of this page to find specific content.
- 2 Capitalized, italicized terms link to a glossary of terms.

Welcome to the **The AI Alliance** project to advance the state of the art for **Developer Testing for Generative AI ("GenAI") Applications**.

Using nondeterministic, **Generative AI Models** in an application makes it difficult to write **Deterministic**, **Repeatable**, and **Automatable** tests. This is a serious concern for application developers, who are accustomed to and rely on determinism when they write **Unit**, **Integration**, and **Acceptance** tests to verify expected behavior and ensure that no **Regressions** occur as the application code base evolves.

What can be done about this problem?

Trust & Safety Evaluations Initiative (TSEI) – Refinement

Different World Views...

App Developers	AI Researchers
Expect determinism (mostly...)	Embrace probabilistic behavior
Write tests to verify expectations	Write benchmarks to measure expectations
Only measure pass/fail	Measure percentage passing and statistical measures of “confidence”
Expect 100% pass rate	Accept anything they can get... 😊

We need to understand each other's perspective, tools, and techniques...

Trust & Safety Evaluations Initiative (TSEI) – Refinement

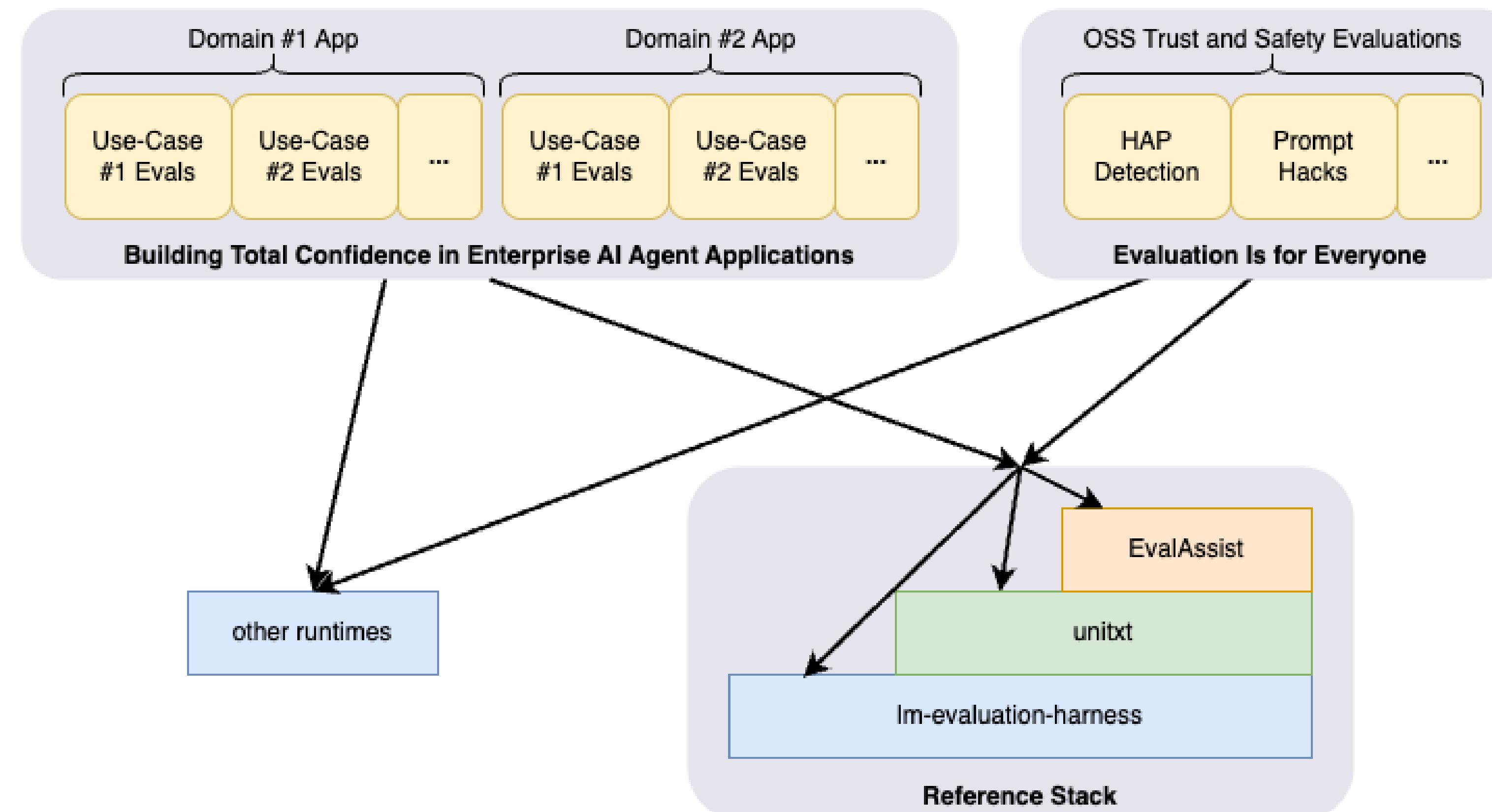
Example: Testing Best Practices... Translated

App Developers	AI Researchers	Explanation
Unit tests	“Unit benchmarks”	Very fine-grained, focused on one “unit” of behavior.
Integration tests	“Integration benchmarks”	E.g., “generic” benchmarks for agent system evaluation.
Acceptance tests	“Acceptance benchmarks”	Per-use case benchmark; does the app implement the whole, end-to-end flow of the user experience

Developers will still need to learn how to work with statistical results, including what % pass rate is “good enough”.

Trust & Safety Evaluations Initiative (TSEI) – Refinement

How the Work Streams Will Be Structured



Provisional work stream names
in **bold** in the *purple* boxes.

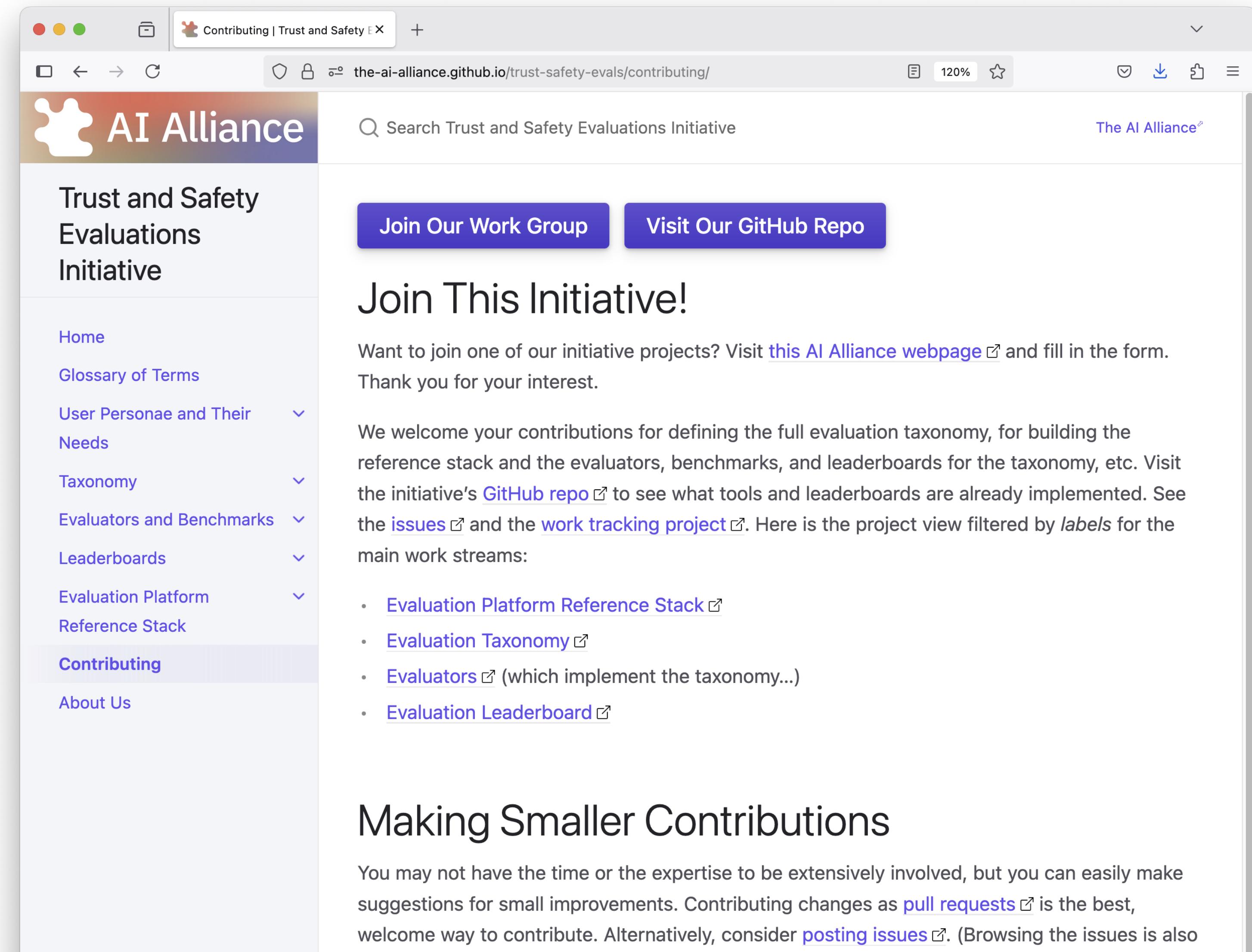
Trust & Safety Evaluations Initiative (TSEI)

Please Join Us!!

Help us help you...

- Contribute your evaluation expertise
- Contribute your domain expertise
- Help us implement the reference stack, evaluators, and leaderboards
- Help us help enterprise developers evaluate their apps

the-ai-alliance.github.io/trust-safety-evals/



The screenshot shows a web browser window with the URL the-ai-alliance.github.io/trust-safety-evals/contributing/. The page has a header with the AI Alliance logo and the text "AI Alliance". Below the header is a sidebar with the following menu items: Home, Glossary of Terms, User Personae and Their Needs, Taxonomy, Evaluators and Benchmarks, Leaderboards, Evaluation Platform Reference Stack, Contributing (which is highlighted), and About Us. The main content area features a search bar, two blue buttons ("Join Our Work Group" and "Visit Our GitHub Repo"), and a section titled "Join This Initiative!". It includes text about joining projects, a GitHub repository, issues, and work tracking. At the bottom, there's a section titled "Making Smaller Contributions" with text about making suggestions via pull requests or posting issues.

Contributing | Trust and Safety

the-ai-alliance.github.io/trust-safety-evals/contributing/

AI Alliance

Search Trust and Safety Evaluations Initiative

The AI Alliance

Join Our Work Group Visit Our GitHub Repo

Join This Initiative!

Want to join one of our initiative projects? Visit [this AI Alliance webpage](#) and fill in the form. Thank you for your interest.

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:

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

Making Smaller Contributions

You may not have the time or the expertise to be extensively involved, but you can easily make suggestions for small improvements. Contributing changes as [pull requests](#) is the best, welcome way to contribute. Alternatively, consider [posting issues](#). (Browsing the issues is also