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.

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.

AI Alliance

Trust & Safety Evaluations Initiative (TSEI)

Problem Statement

AI builders need to **evaluate models and applications**:

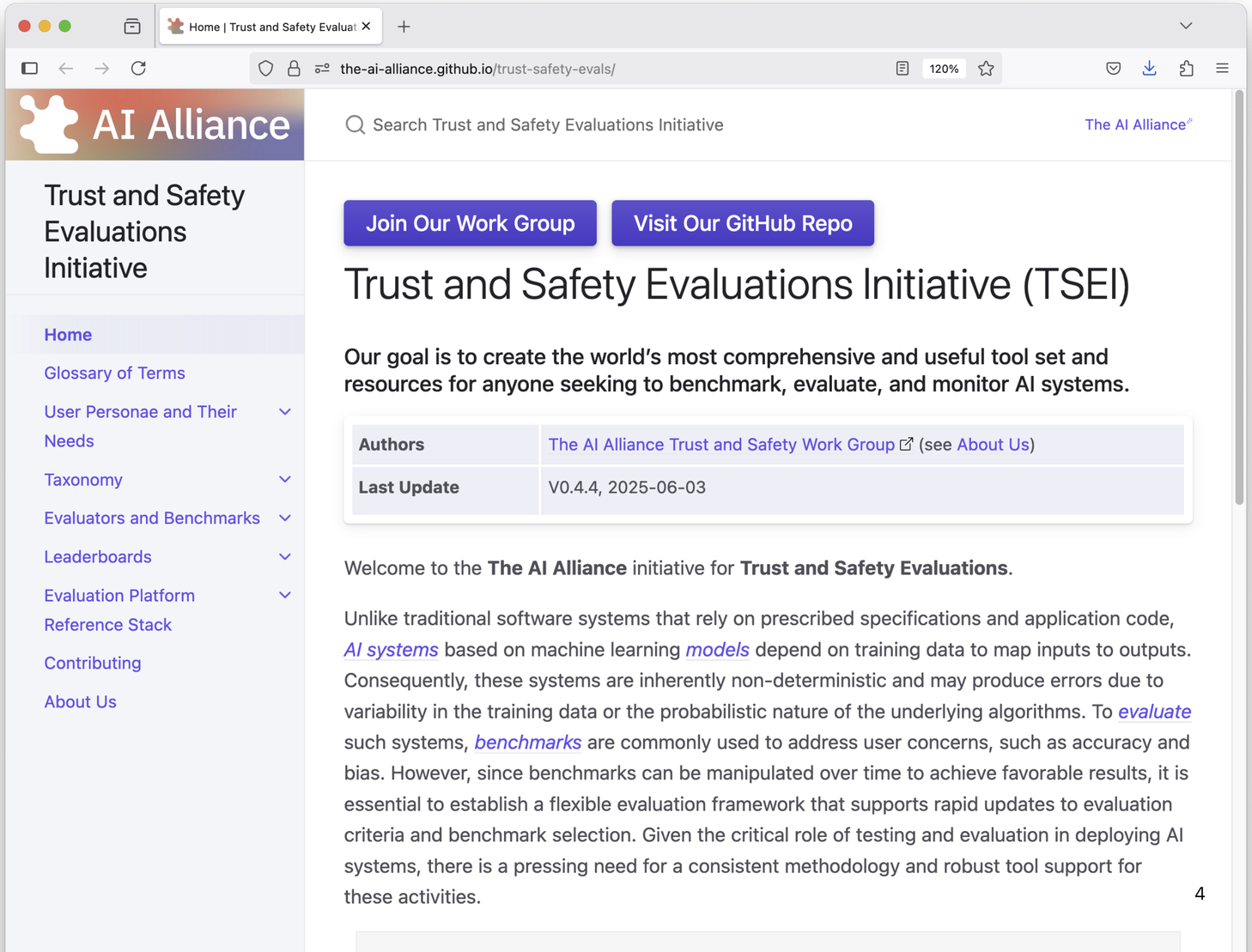
- I am not an expert; where do I start?
- Good techniques exist for *evaluating* safety, but what about the *“last mile”* of evaluation; how do I evaluate if my app works well for *my* domain and *my* use cases?
- If I have those evaluations, how do different open models score for those evaluations?
- 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

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

- **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

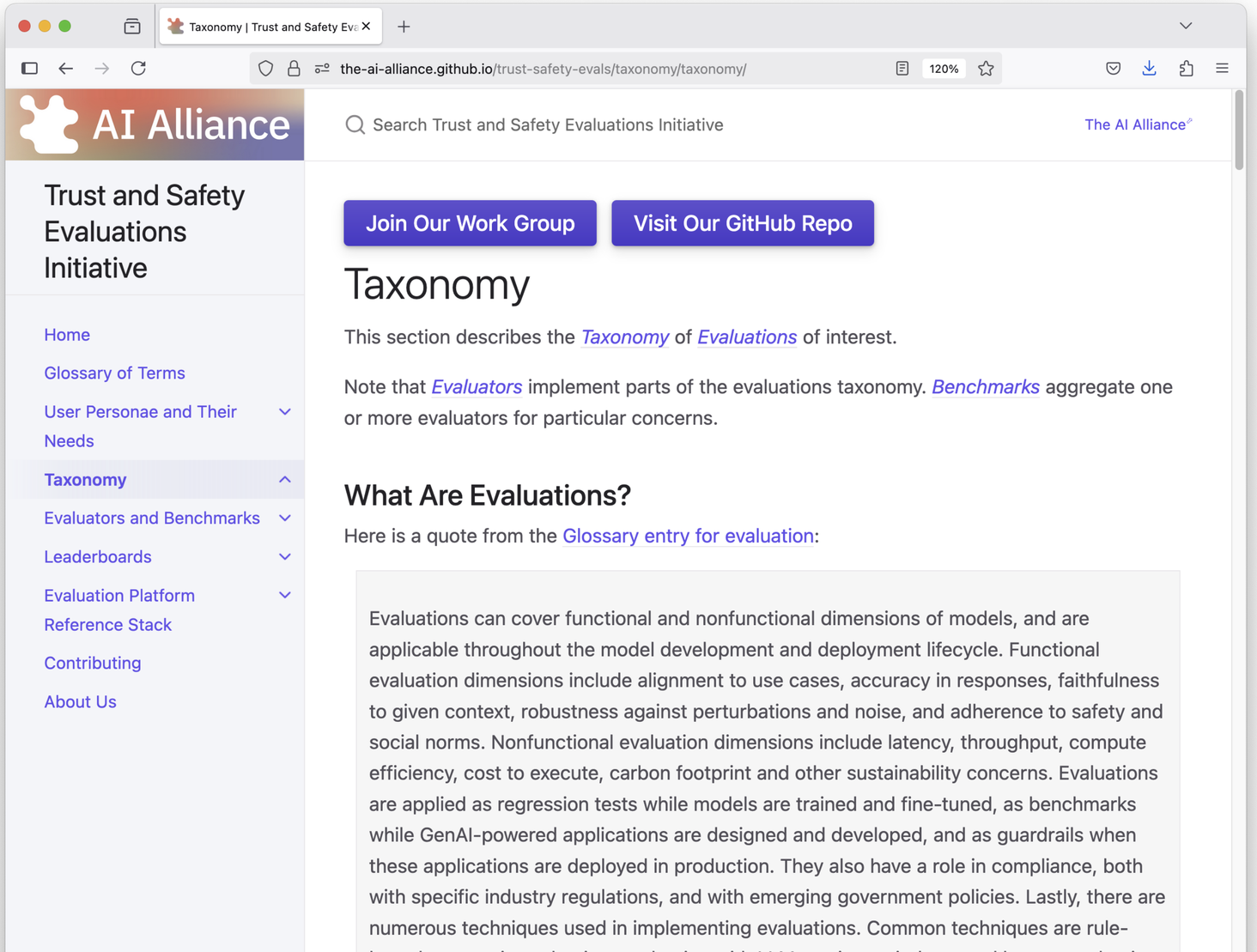
Trust & Safety Evaluations Initiative (TSEI) – Current Picture

Taxonomy of Evaluations

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

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

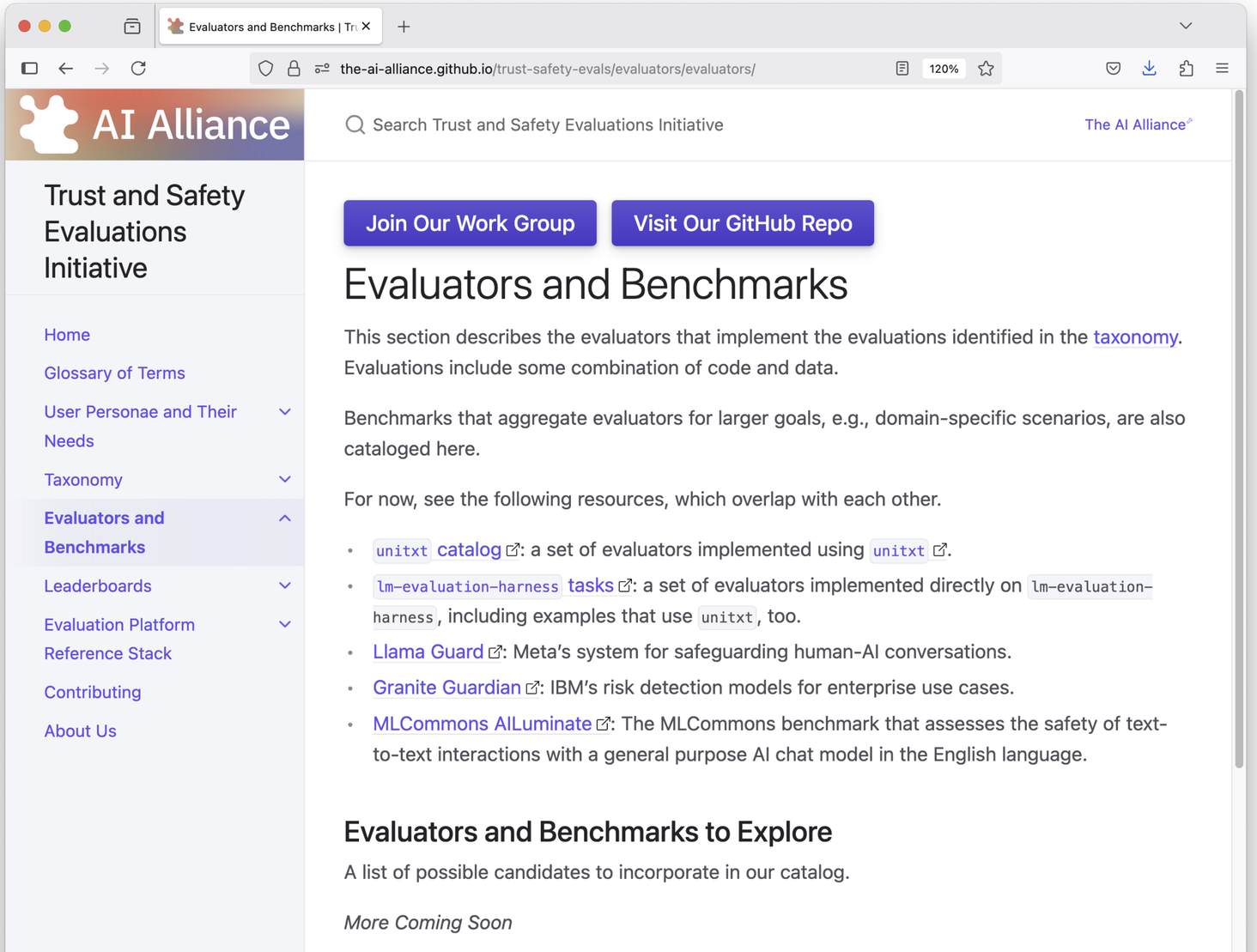
Trust & Safety Evaluations Initiative (TSEI) – Current Picture

Evaluators and Benchmarks

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

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

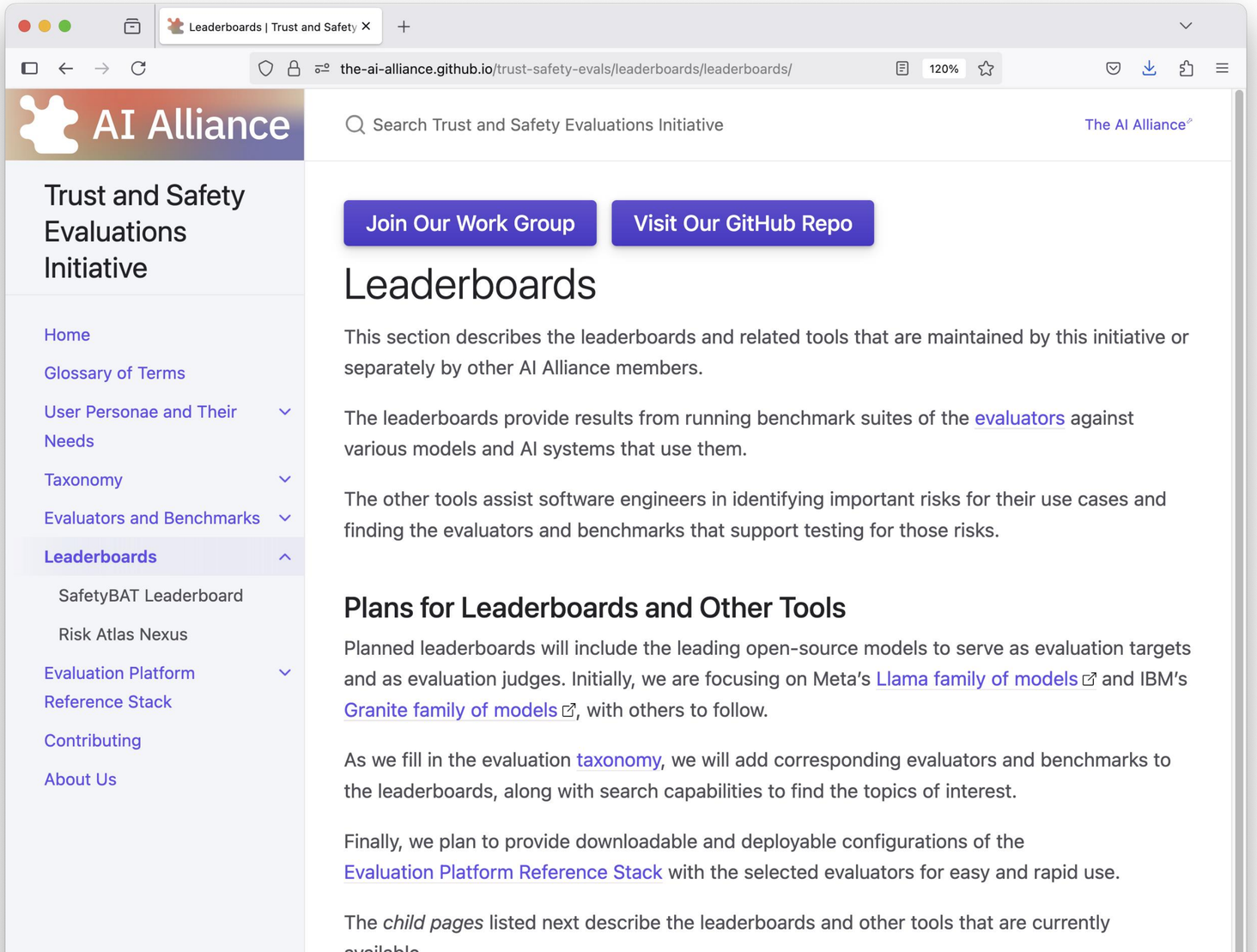
Trust & Safety Evaluations Initiative (TSEI) – Current Picture

Leaderboards

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

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



An Initiative of Focus Area 2: AI Trust and Safety

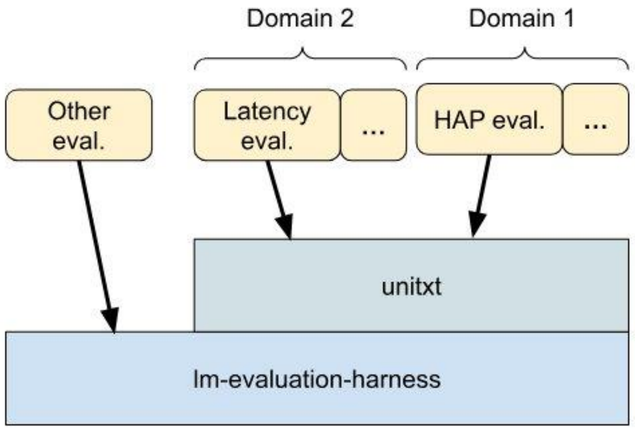
Trust & Safety Evaluations Initiative (TSEI) – Current Picture

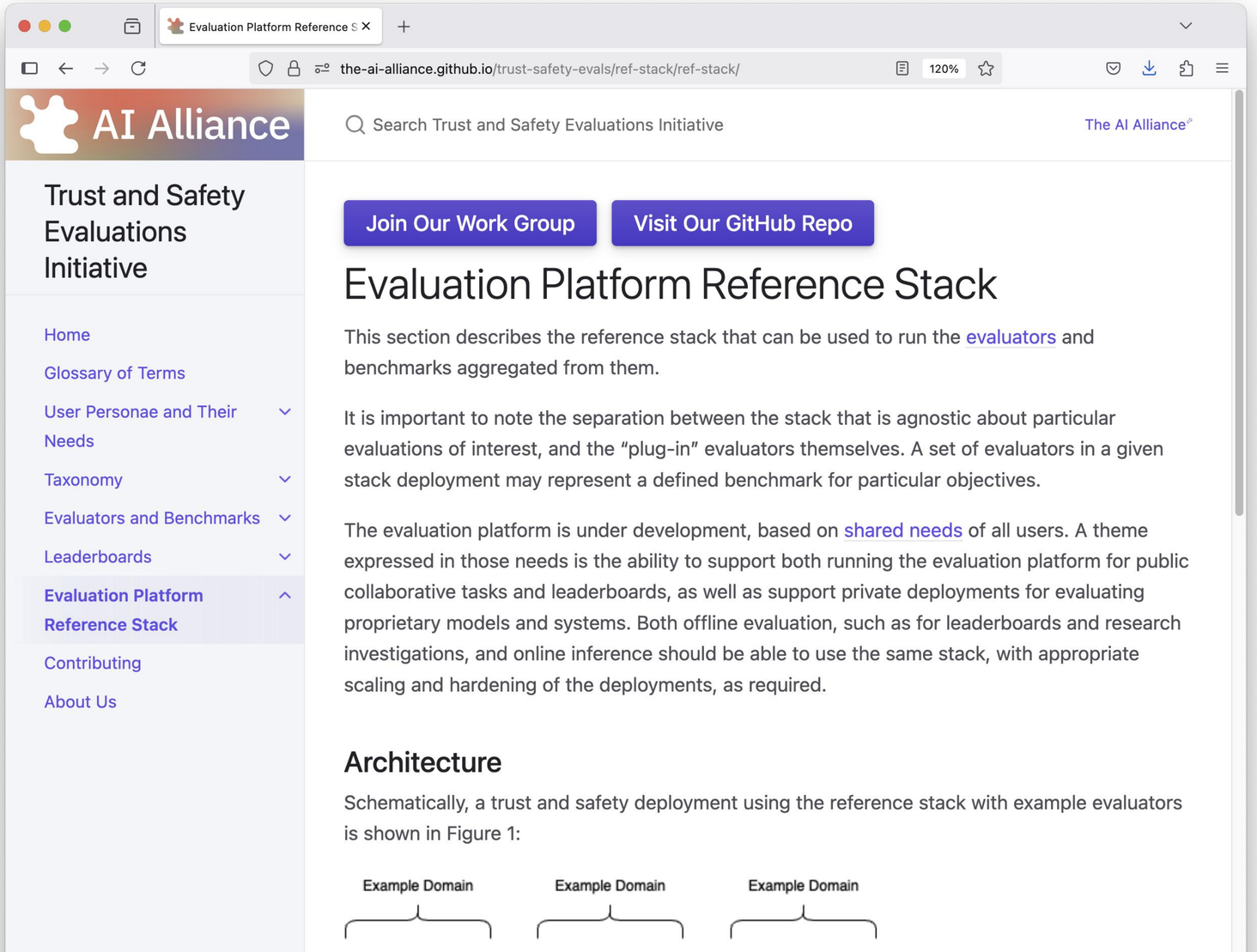
Reference Stack

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

How to run the evaluators and benchmarks

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





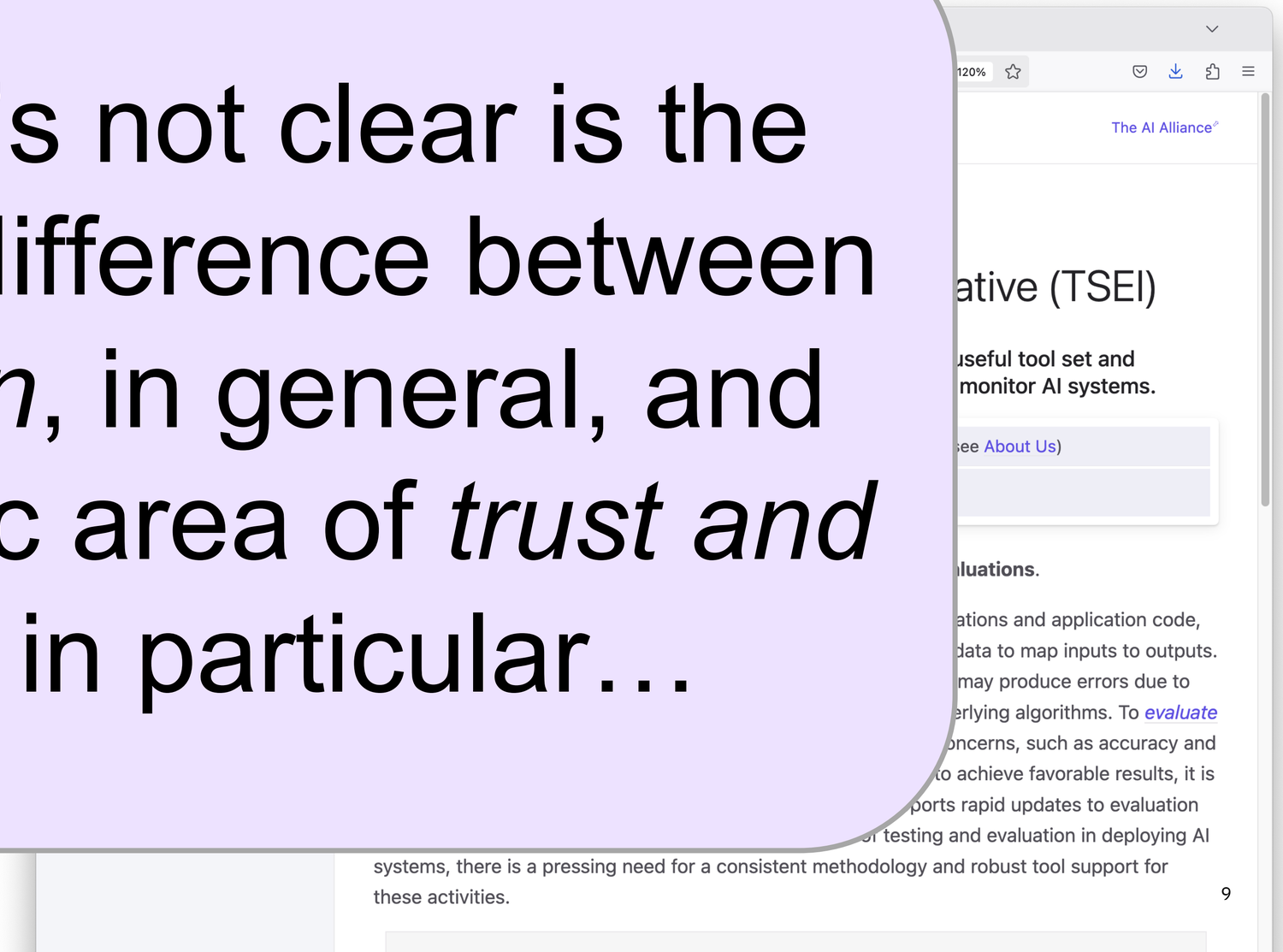
Trust & Safety Evaluations Initiative (TSEI) – Current Picture

What We A

- **Taxon**
  align
- **Evalu**
  Imple
- **Leade**
  for you
- **Refere**
  offline
  during

ty-evals/

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



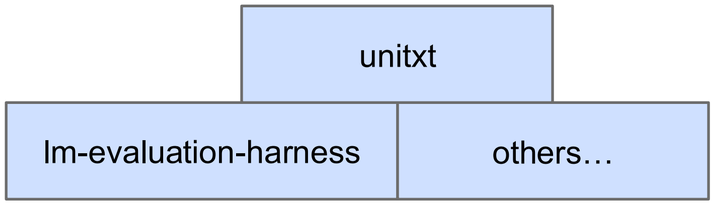
Trust & Safety Evaluations Initiative (TSEI) – Refinement

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 evaluation[1]**
  - *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!

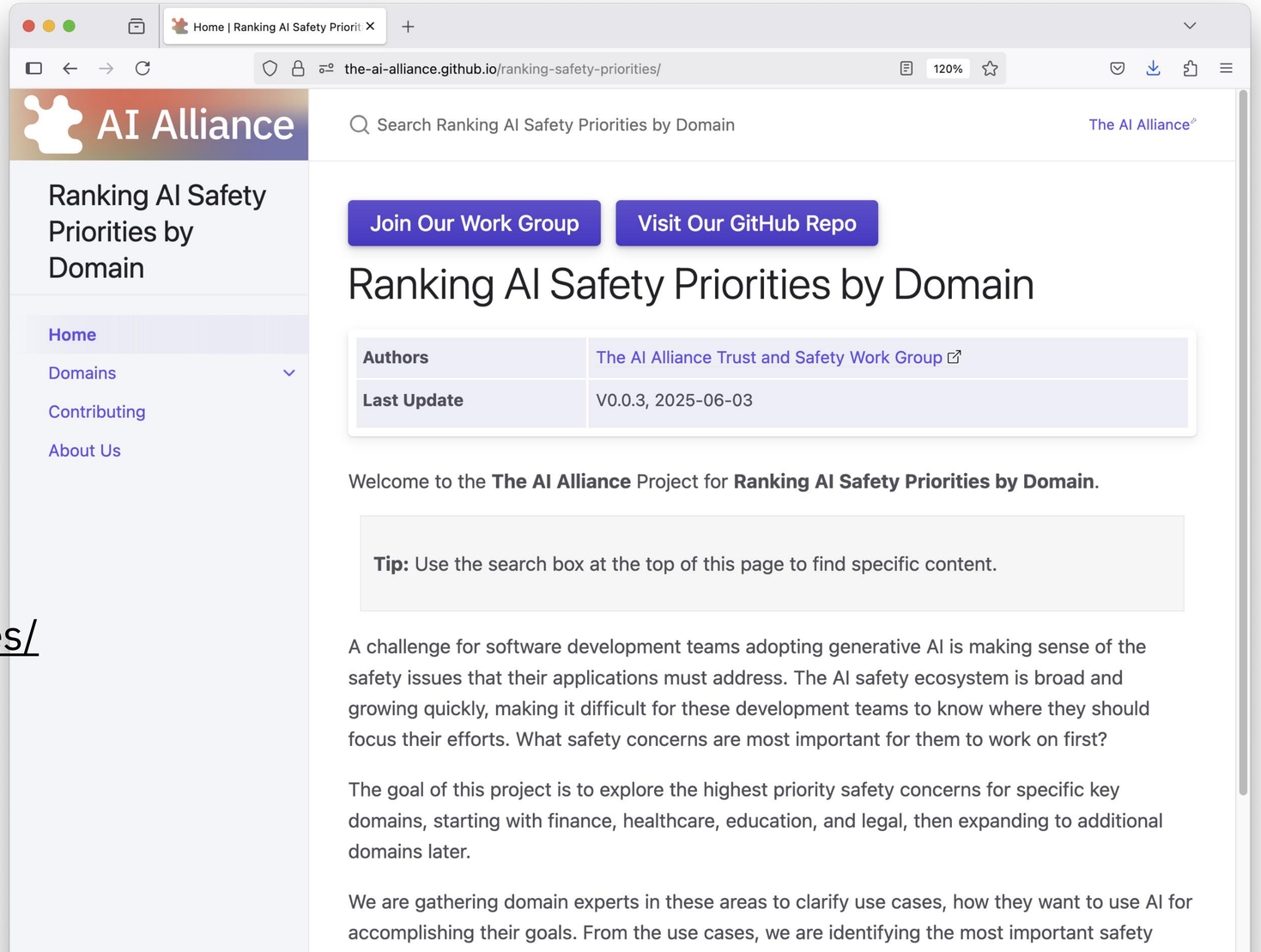
1. A reference to “cable” deployments in the last ~30 years. The last mile was the final optical fiber connection to your home to give you the high bandwidth you need for your TikTok videos…

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/

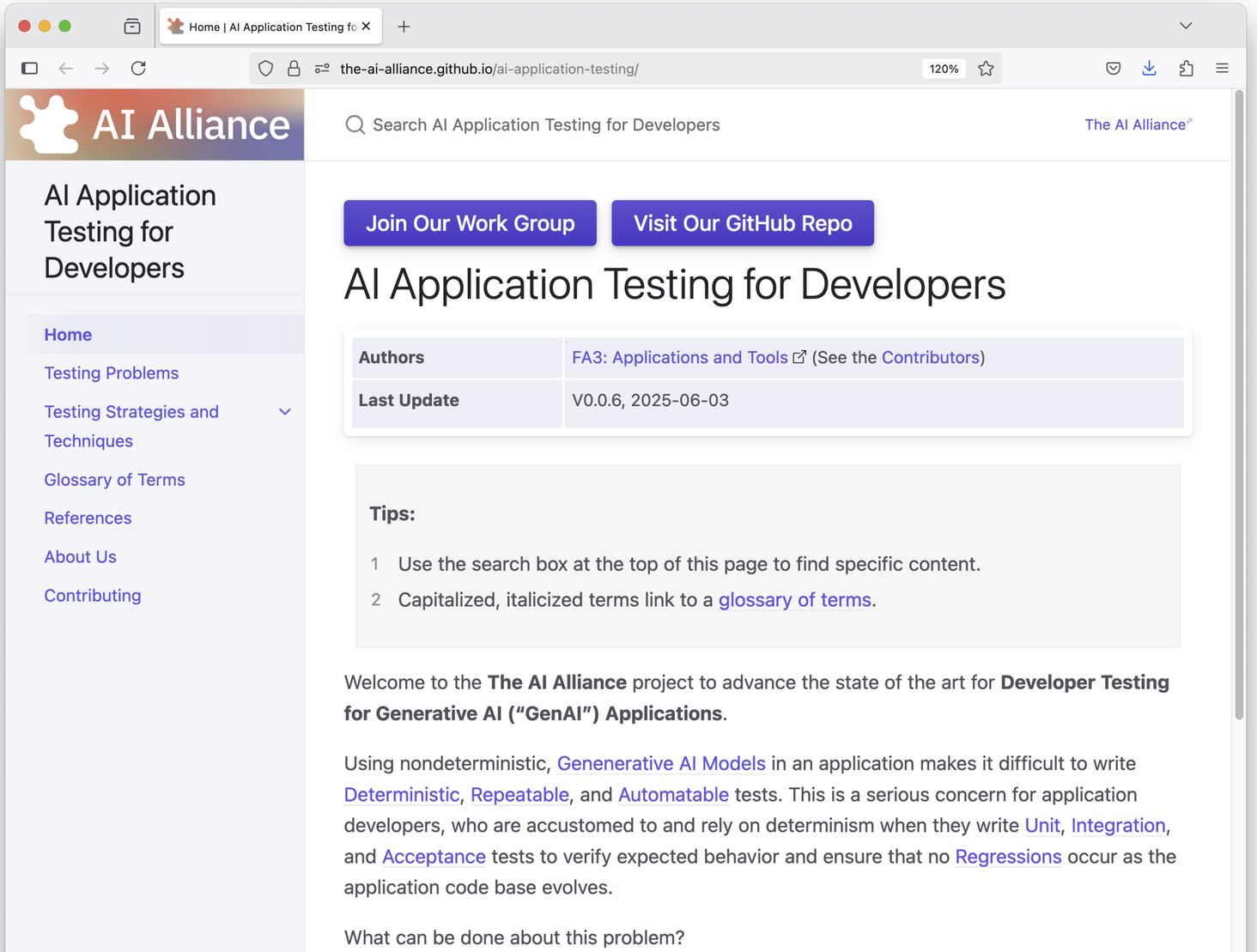


Trust & Safety Evaluations Initiative (TSEI) – Refinement

#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/



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

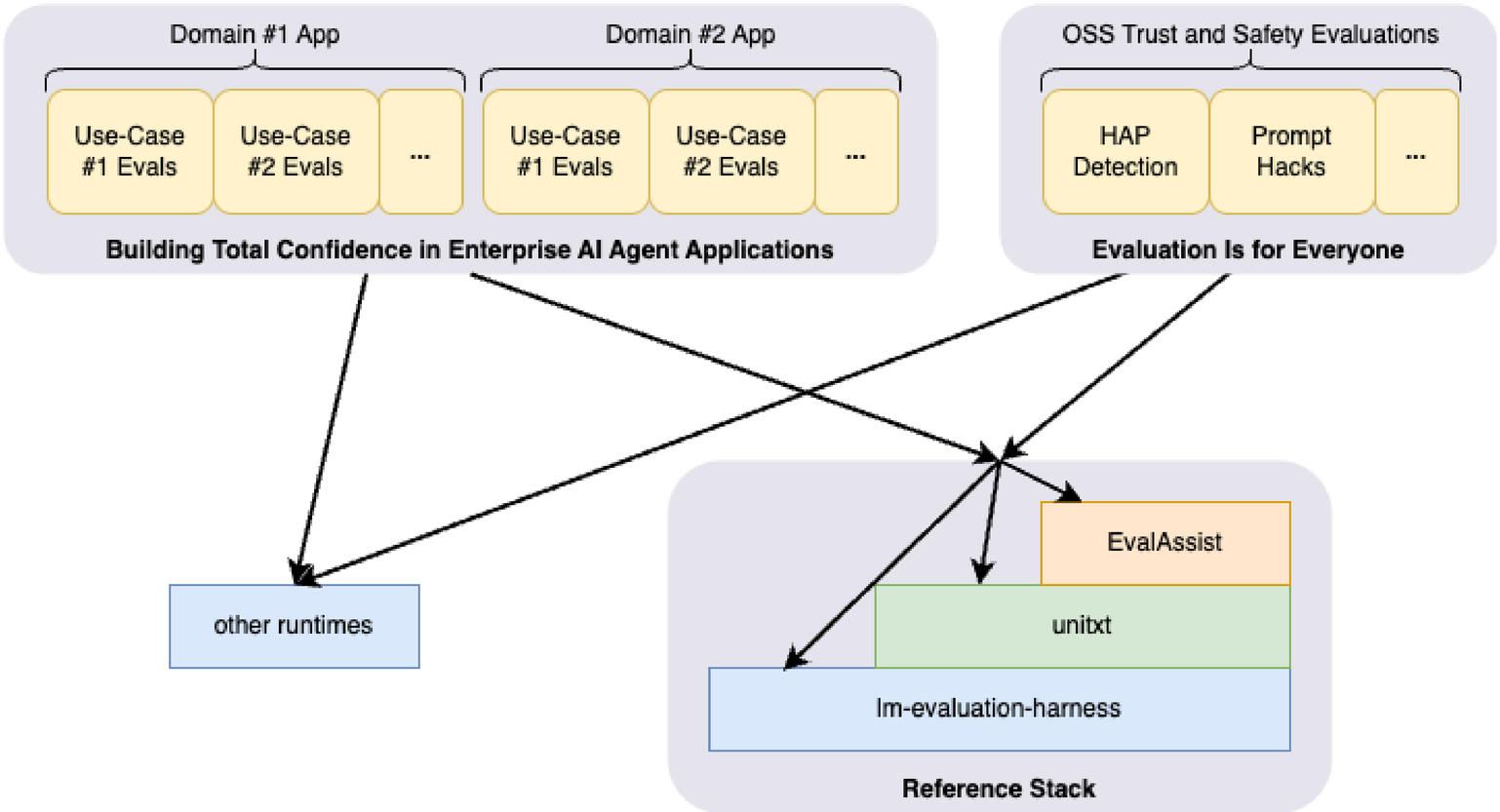
Example: Testing Best Practices… Translated

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!!

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

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