TrustyAI community meeting

May 2025

May 2025 updates

- TrustyAI operator
  - Current: 1.37.0 release
    - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.37.0
  - quay.io/trustyai/trustyai-service-operator:v1.37.0
- TrustyAI LMEval
  - Current: 0.48.0 release
    - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.37.0
  - quay.io/trustyai/ta-lmes-job:v1.37.0
- TrustyAI Guardrails
  - Current: 0.9.4 release
    - https://github.com/trustyai-explainability/fms-guardrails-orchestrator
  - quay.io/trustyai/ta-guardrails-orchestrator:v1.37.0

What’s new?

TrustyAI - What's new?

- **TrustyAI operator 1.37.0**
  - Dependabot added
    - Many dependency updates and minor CVE fixes
  - Guardrails gateway
    - Correct sidecar gateway image name [#466]
    - Correct gateway port to 8090 [#472]
  - LMEval driver
    - Enable cgo in driver [#469]
    - cgo permission denied on LMEval driver [#479]
  - Guardrails
    - Rename external guardrails route [#473]
    - Bring orchestrator build in-line with upstream [#478]
    - Decouple vllm-gateway and regex detector [#438]
  - LMEval
    - Add custom YAML task support to LMEval [#435]
    - Allow non-admins to deploy and list LMEval jobs [#440]
    - Support custom metrics and tasks by [#436]
    - Update OAuth proxy image in RHOAI overlay [#447]
- **TrustyAI LMEval 0.4.8**
  - Too many fixes to benchmarks to list!
- **TrustyAI Guardrails Orchestrator 0.9.4**
  - Again, too many fixes!

Current work

TrustyAI - Python Service & LMEval

- TrustyAI service (Python)
  - TLS
  - ModelMesh protobuf parsing
  - Bias/Fairness/Drift metrics
- LMEval
  - Custom system templates
  - LLM-as-Judge support in unitxt

TrustyAI - Llama Stack

- Two new Out-Of-Tree providers for Llama Stack
  - LMEval
    - https://github.com/trustyai-explainability/llama-stack-provider-lmeval
    - Leverage LMEval from the LLS API
  - Guardrails
    - https://github.com/trustyai-explainability/llama-stack-provider-trustyai-fms
    - Leverage Guardrails from the LLS API
- Ready to use in your Llama Stack Distro

TrustyAI - Explainability

- Remote LLM Attributions
  - Enabled black box explainability for remotely hosted (vLLM) LLMs with PyTorch Captum
    - PR merged to Captum - https://github.com/pytorch/captum/pull/1544
- Experimental Llama Stack Implementation
  - A reference implementation introducing an 'explanation' API to Llama Stack using Remote LLM Attributions with Captum
    - https://github.com/saichandrapandraju/lls-explanation-reference

TrustyAI - Community

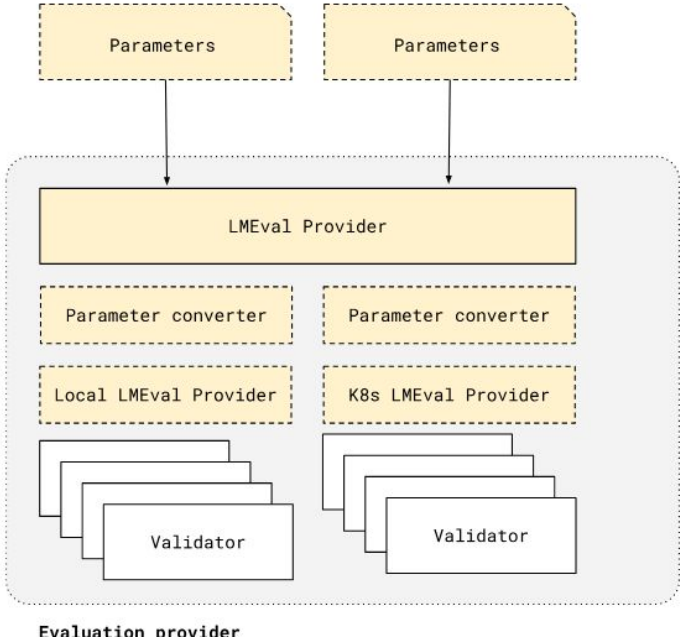
- TrustyAI at the 2025 Dublin AI masterclass
  - Mac Misiura: Evaluation and Guardrails
  - Diego Maniloff: Llama Stack
- Red Hat Summit 2025
  - Christina Xu: TrustyAI, LMEval and Guardrails
- Congratulations to all!

TrustyAI - Python TrustyAI v1

- Migrating to Python 3.12 as minimum requirement

TrustyAI - Next

- Designing the next phase of TrustyAI
  - ADR coming soon
  - Concept of TrustyAI as an universal API for AI safety providers
    - Eval
    - Guardrails
    - Explainability
    - Alignment
    - *etc*
  - Seamless experience in local and Kubernetes



Roadmap

TrustyAI 2024 roadmap

- KServe explainer integration
- Detoxification fine-tuning
- Python TrustyAI service
  - Saliency Explainers
- TrustyAI v2
- Guardrails
  - Orchestrator
- LM-Eval
  - LM-Eval v2 iteration on upstream roadmap (target 6th December)
    - https://github.com/trustyai-explainability/trustyai-service-operator/issues/366

Legend
Not started
In progress
Completed

Other topics