TrustyAI community meeting

April 2025

April 2025 updates

- TrustyAI core / service
  - Current: 0.28.0 release
    - https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.28.0
    - quay.io/trustyai/trustyai-service:v0.28.0
- TrustyAI operator
  - Current: 1.35.0 release
    - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.35.0
  - quay.io/trustyai/trustyai-service-operator:v1.35.0

What’s new?

TrustyAI - What's new?

- **TrustyAI core/service 0.28.0**
  - Override netty handler version (#675)
  - Fix `UploadEndpoint nullpointererror` (#676)
  - CI
    - Add PR write permissions for build+push job (#677)
- **TrustyAI operator 1.35.0**
  - Update component_metadata.yaml (#433)

Current work

TrustyAI - LMEval & Guardrails

- Llama Stack LMEval integration
  - https://github.com/trustyai-explainability/llama-stack/tree/release-0.1.9-lmeval
  - https://github.com/ruivieira/lls-lmeval-reference
    - Example Jupyter notebook
- Llama Stack 0.2.2
  - OpenAI Inference
- Llama Stack Guardrails integration
  - https://github.com/meta-llama/llama-stack/pull/1419

TrustyAI - LM-Eval - Custom (YAML) Tasks

```yaml
apiVersion: trustyai.opendatahub.io/v1alpha1
kind: LMEvalJob
metadata:
  name: lmeval-llama-stack-job
  namespace: test
spec:
  logSamples: true
  model: hf
  modelArgs:
  - name: pretrained
      value: "google/flan-t5-base"
  - name: tokenizer
      value: "google/flan-t5-base"
  taskList:
      customTasks:
        source:
          git:
      url: https://github.com/trustyai-explainability/lm-eval-tasks.git
      branch: main
      commit: "b995eeb986daab3c9f77f101af6cc51a34f8f1da"
      path: tasks/
      taskNames:
      - tiny_offline_arithmetic
      - arc_easy
```

TrustyAI - Python TrustyAI service

- TrustyAI service migration Java ↣ Python
  - https://github.com/trustyai-explainability/trustyai-service
  - Enable leveraging community frameworks and algorithms
- Initial drop-in replacement
  - No API changes, no new features
  - Bias/Fairness/Drift metrics

Roadmap

TrustyAI 2024 roadmap

- KServe explainer integration
- Detoxification fine-tuning
- Python TrustyAI service
  - Saliency Explainers
- 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