TrustyAI community meeting

February 2025

February 2025 updates

- TrustyAI core / service
  - Current: 0.25.0 release
    - https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.25.0
    - quay.io/trustyai/trustyai-service:v0.25.0
- TrustyAI operator
  - Current: 1.32.0 release
    - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.32.0
  - quay.io/trustyai/trustyai-service-operator:v1.32.0

What’s new?

TrustyAI - What’s new?

- **TrustyAI core/service 0.25.0**
  - Maintenance release
  - Small fixes to unit tests and CI for 0.26.0

TrustyAI - What's new?

- [TrustyAI operator 1.30.0 … 1.32.0]
  - **LMEval**
    - Add operator online and execution feature flag by [#379]
    - Filter protected env vars set from CR directly [#381]
    - Remove unneeded metrics fetch [#382]
    - Allow unitxt code execution when LMEval code execution enabled [#385]
    - Disable telemetry [#383]
    - Use only local unitxt catalogs when in offline mode [#386]
    - Make lm-evaluation-harness source path consistent [#388]
    - Move LMEval to release 0.4.6 [#398]
    - Run PodExec against the main container [#397]
  - **CVEs**
    - CVE-2024-45338: Update golang.org/x/net to 0.33.0
    - CVE-2024-45337): Enforce x/crypto 0.31.0 [#390]
  - **Manager**
    - Increase resource limits for the TrustyAI operator [#393]

Current work

TrustyAI - LM-Eval

- Available on ODH 2.25 (21st Feb 2025)
- **LMEval**
  - Support for S3 as artifact storage
  - Support for custom templates and system prompts
- **Guardrails**
  - TrustyAI operator support for Guardrails orchestrators[1]

[1] - https://github.com/trustyai-explainability/reference/tree/main/guardrails/orchestrator

TrustyAI - LM-Eval - Custom prompts API

```
apiVersion: trustyai.opendatahub.io/v1alpha1
kind: LMEvalJob
metadata:
  name: custom-card-template
  namespace: tas
spec:
  allowOnline: true
  allowCodeExecution: true
  model: hf
  modelArgs:
  - name: pretrained
      value: google/flan-t5-base
  taskList:
    taskRecipes:
      - template:
          ref: tp_0
        systemPrompt:
          ref: sp_0
        card:
          name: "cards.wnli"
```

```
   custom:
    templates:
     - name: tp_0
      value: |
   {
      "__type__": "input_output_template",
      "input_format": "{text_a_type}:
{text_a}\n{text_b_type}: {text_b}",
      "output_format": "{label}",
      "target_prefix": "The {type_of_relation} class is ",
      "instruction": "Given a {text_a_type} and
{text_b_type} classify the {type_of_relation} of the
{text_b_type} to one of {classes}.",
      "postprocessors": [
         "processors.take_first_non_empty_line",
         "processors.lower_case_till_punc"
      ]
   }
    systemPrompts:
     - name: sp_0
      value: "Be concise. At every point give the shortest
acceptable answer."
 logSamples: true
```

TrustyAI - LM-Eval - S3 support API

```
apiVersion: trustyai.opendatahub.io/v1alpha1
kind: LMEvalJob
metadata:
      name: evaljob-sample
spec:
      allowOnline: false
      model: hf
      modelArgs:
      - name: pretrained
      value: /opt/app-root/src/hf_home/flan
      taskList:
         taskNames:
            - arc_easy
      logSamples: true
```

```
offline:
   storage:
      s3:
      accessKeyId:
              name: s3-secret
              key: AWS_ACCESS_KEY_ID
      secretAccessKey:
              name: s3-secret
              key: AWS_SECRET_ACCESS_KEY
      bucket:
              name: s3-secret
              key: AWS_S3_BUCKET
      endpoint:
              name: s3-secret
              key: AWS_S3_ENDPOINT
      region:
              name: s3-secret
              key: AWS_DEFAULT_REGION
      path: ""
      verifySSL: false
```

TrustyAI - Guardrails - Orchestrator API

```
apiVersion: trustyai.opendatahub.io/v1alpha1
kind: GuardrailsOrchestrator
metadata:
  name: gorch-test
spec:
  orchestratorConfig: "fms-orchestr8-config-nlp"
  vllmGatewayConfig: "fms-orchestr8-config-gateway"
  replicas: 1
```

```
kind: ConfigMap
apiVersion: v1
metadata:
  name: fms-orchestr8-config-nlp
data:
  config.yaml: |
      chat_generation:
        service:
          hostname: llm-predictor.guardrails-test.svc.clust
        port: 8032
        detectors:
          regex:
             type: text_contents
             service:
               hostname: "127.0.0.1"
               port: 8080
             chunker_id: whole_doc_chunker
             default_threshold: 0.5
```

TrustyAI - Guardrails - Orchestrator API

```
apiVersion: trustyai.opendatahub.io/v1alpha1
kind: GuardrailsOrchestrator
metadata:
  name: gorch-test
spec:
  orchestratorConfig: "fms-orchestr8-config-nlp"
  vllmGatewayConfig: "fms-orchestr8-config-gateway"
  replicas: 1
```

```
kind: ConfigMap
apiVersion: v1
metadata:
  name: fms-orchestr8-config-gateway
  labels:
      app: fmstack-nlp
data:
  config.yaml: |
      orchestrator:
        host: "localhost"
        port: 8032
      detectors:
        - name: regex
          detector_params:
      regex:
            - email
            - ssn
        - name: other_detector
      routes:
        - name: pii
          detectors:
      - regex
        - name: passthrough
          detectors:
```

TrustyAI - Work in Progress

- Work ongoing in
  - Explainers gRPC service addressing
  - LLama Stack API support (Guardrails and LMEval)

Roadmap

TrustyAI 2024 roadmap

- KServe explainer integration
- Detoxification fine-tuning
- 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