# AI Explainability WG

July 2024

# July 2024 updates

- TrustyAI core / service
  - 0.17.0 release
  - https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.17.0
  - quay.io/trustyai/trustyai-service:v0.17.0
- TrustyAI operator
  - 1.23.0 release
  - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.23.0
  - quay.io/trustyai/trustyai-service-operator:v1.23.0
- TrustyAI KServe explainer
  - 0.1.0 release
  - https://github.com/trustyai-explainability/trustyai-kserve-explainer/releases/tag/v0.1.0
  - quay.io/trustyai/trustyai-kserve-explainer:v0.1.0

# What’s new?

# TrustyAI - What's new?

- **TrustyAI core/service 0.17.0**
  - Add image data drift [#577]
  - Add inference ids metadata endpoint [#582]
  - Support for KServe v2 HTTP protocol as a PredictionProvider [#579]
  - Remove SSL from dev profile [#581]
  - Add explainer configuration to payloads [#584]
  - Small fixes
    - Fix starting index for covariance update [#569]
    - Lower variance of ground-truth distribution in StreamingBackground tests [#578]
    - Add HTTPS connection support [#574]
  - CI
    - Move openshift-ci image to ubi8 [#566]
    - TrustyAI CounterfactualExplainer tests timeouts fixed [#477]
    - Trim installs, set stable stream for Authorino [#572]
- **TrustyAI operator 1.23.0**
  - Update Go builder and toolset to 1.21 [#253]
  - Move test image to ubi [#245]
  - Add internal service TLS [#250]
  - Add overlays [#251]

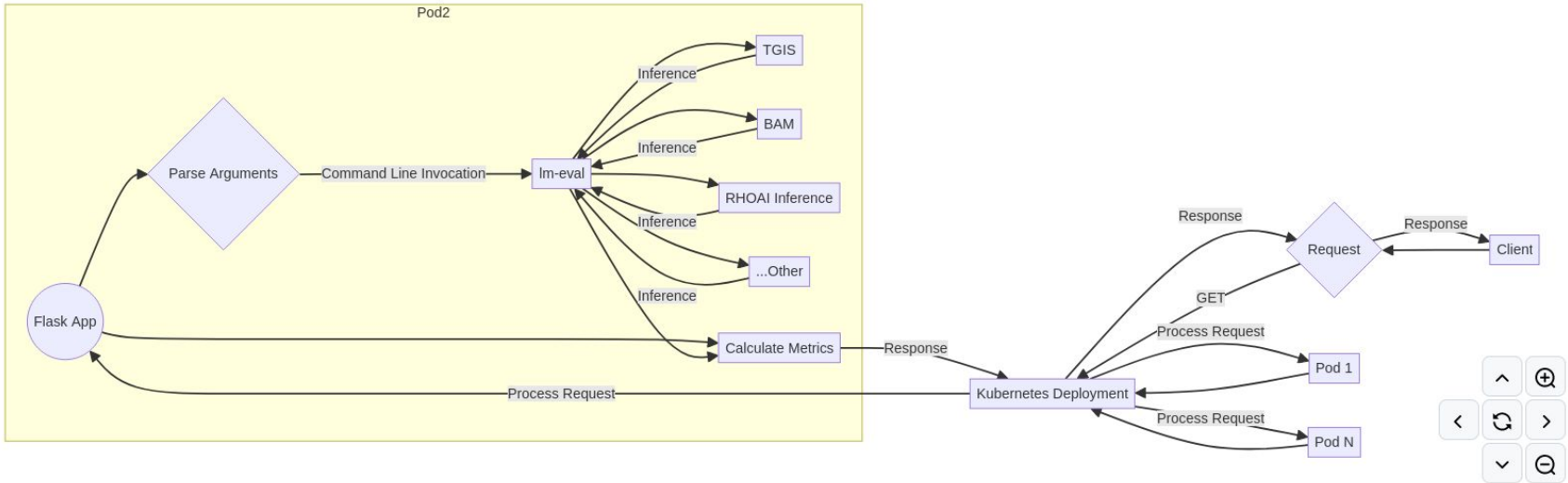
# Current work

# TrustyAI - current work

- **TrustyAI core/service**
  - Support for database storage [merged]
    - Support for PVC- and DB-mode
  - HTTPS endpoint support [merged]
    - ModelMesh HTTPS payload processor support added upstream
    - WIP for ModelMesh certificate support
  - TSSaliency (time-series support)
    - Formalising time-series schema
  - OpenAPI documentation
    - https://trustyai-explainability.github.io/trustyai-site/main/trustyai-service-api-reference.html
- **TrustyAI operator**
  - KServe readiness
- **Testing/CI**
  - Migrating CI to Python-based tests

# LM-Eval

- Discussion
  - https://github.com/foundation-model-stack/fms-lm-eval-service/discussions/20



# TrustyAI - Database support

## PVC-mode

```
apiVersion: trustyai.opendatahub.io/v1alpha1
kind: TrustyAIService
metadata:
  name: trustyai-service
spec:
  storage:
       format: "PVC"
       folder: "/inputs"
       size: "1Gi"
  data:
       filename: "data.csv"
       format: "CSV"
  metrics:
       schedule: "5s"
```

## DB-mode

```
apiVersion: trustyai.opendatahub.io/v1alpha1
kind: TrustyAIService
metadata:
  name: trustyai-service
spec:
  storage:
       format: "DATABASE"
       databaseConfigurations: db-credentials
  metrics:
       schedule: "5s"

—
apiVersion: v1
kind: Secret
metadata:
  name: db-credentials
type: Opaque
stringData:
  databaseKind: mysql
  databaseUsername: "foo"
  databasePassword: "bar"
  databaseService: "mariadb-service"
  databasePort: "3306"
  databaseGeneration: "update"
```

# Roadmap

# TrustyAI 2024 roadmap

- **December 2023 - March 2024 (proposal / discussion)**
  - Out-of-distribution (OOD) metrics completion
    - Finalise OOD metrics, namely data uploading aspect
      - Provide data connection for data upload
  - Explainability service endpoints completion
    - Formalise explainability payloads schema
    - Refining handling of synthetic payloads to avoid interference with metrics
    - Handle potential large computational times in a service setting
  - Detoxification at the library and service level
    - Integrate detoxification with Python TrustyAI (Jupyter main target)
    - Token scoring at the service level
  - Database backend
    - Address scalability
    - Replace PVC with DB?
  - Expand supported types (*eg* image data)
    - Metrics and explainability for non-tabular data
  - Model drift/data drift/anomaly detection
  - Improve handling of unsupported model serving runtimes

**Legend**

Not started

In progress

Completed

# TrustyAI 2024 roadmap

- **June 2024 - August 2024 (proposal / discussion)**
  - KServe explainer integration
  - Detoxification fine-tuning
  - Saliency Explainers

**Legend**
Not started
In progress
Completed