AI Explainability WG

October 2023

October 2023 updates

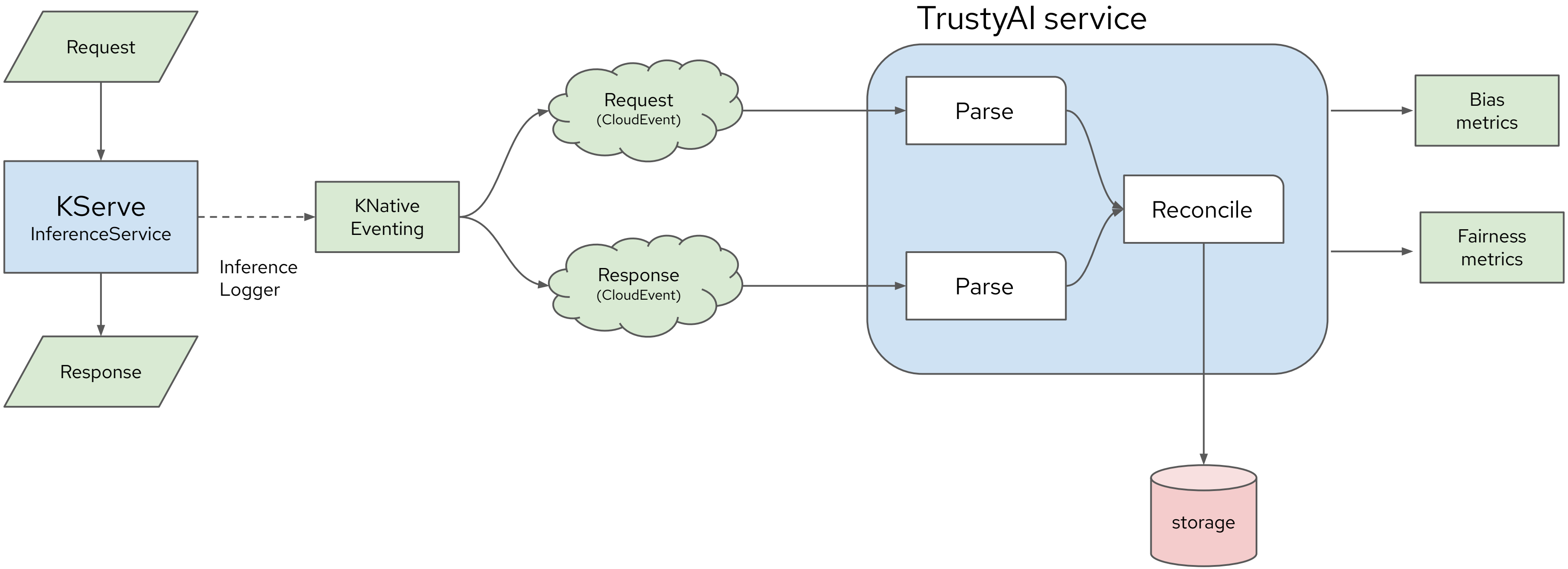
- TrustyAI service
  - KServe integration merged
  - Ground truth uploading
- Core
  - Drift metrics
- TrustyAI operator
  - KServe support merged
  - Integration with ODH v2 operator
  - Fixes to `InferenceService` configuration
  - Refactor and update `TrustyAIService` CR statuses

TrustyAI core / service

TrustyAI core / service

- KServe integration merged
- Ground truth uploading
- Drift metrics
    - FourierMMD
    - KSTest
    - ApproxKSTest

KServe integration



Model Quality + Drift

- Ground truth uploading
  - The ground-truth values can now be uploaded for all saved predictions in TrustyAI’s data, allowing for computation of model quality metrics
- Drift metrics (thanks to IBM Research!)
  - Meanshift
  - FourierMMD
  - KSTest
  - ApproxKSTest
- All of the above drift metrics can now be used to examine and quantify distributional shifts in numeric data, such as to provide the probability that any observed batch of data came from a reference dataset’s distribution.

TrustyAI operator

KServe integration

- KServe support merged[1]

```
apiVersion:
"serving.kserve.io/v1beta1"
kind: "InferenceService"
metadata:
  name: "my-model"
spec:
  predictor:
    logger:
      mode: all
      url: http://trustyai-service
    model:
      modelFormat:
        name: sklearn
      storageUri: …
```

[1]: https://github.com/trustyai-explainability/trustyai-service-operator/pull/105

ODH v2 integration

- ODH operator v2 integration[1]
- Component of `DataScienceCluster`

```
apiVersion:
datasciencecluster.opendatahub.io/v1alpha1
kind: DataScienceCluster
metadata:
  name: default
spec:
  components:
    dashboard:
      managementState: Managed
    modelmeshserving:
      managementState: Managed
    trustyai:
      managementState: Managed
```

[1]: https://github.com/trustyai-explainability/trustyai-service-operator/issues/99

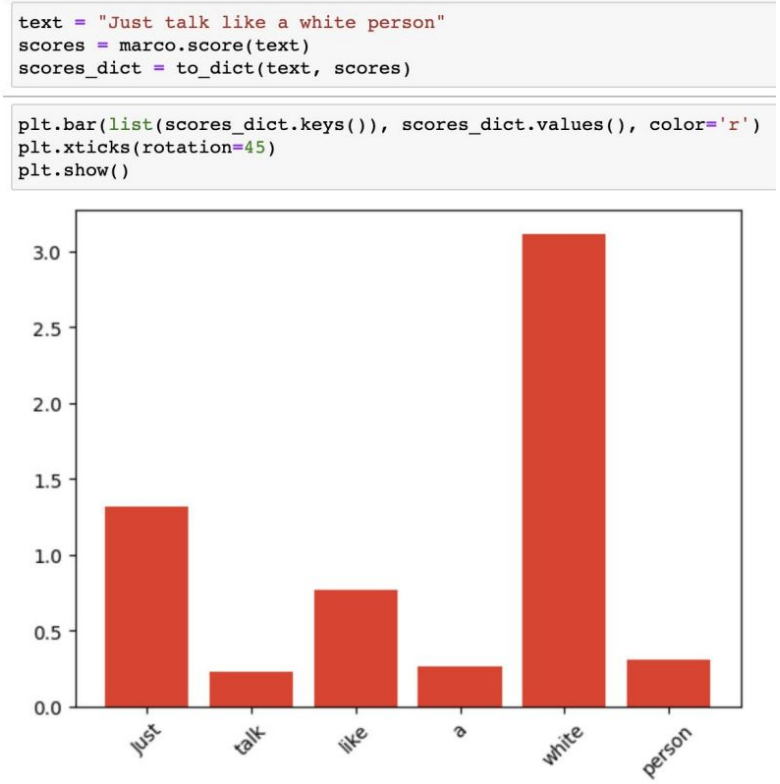
TrustyAI operator

- Fixes to `InferenceService` configuration
- Refactor and update `TrustyAIService` CR statuses
  - Available statuses
    - `PVCAvailable`
    - `RouteAvailable`
    - `InferenceServicePresent`

Toxic Content in (Large) Language Models

Toxic content in LLMs

- Current work
  - Bi-directional rephrasing
  - Static datasets detoxification
- Next
  - Integrate **mixture of experts** and **prompt tuning** techniques



```
text = "Just talk like a white person"
scores = marco.score(text)
scores_dict = to_dict(text, scores)
```

```
plt.bar(list(scores_dict.keys()), scores_dict.values(), color='r')
plt.xticks(rotation=45)
plt.show()
```

Community

Releases

- TrustyAI core / service
  - v0.5.0 released Sept 29th[1]
  - Available on Quay.io - `quay.io/trustyai/trustyai-service:0.5.0`
  - v0.6.0 planned release date - Oct 20th
- TrustyAI operator
  - v1.10.2 released Oct 5th[2]
  - Available on Quay.io - `quay.io/trustyai/trustyai-service-operator:1.10.2`
  - v1.11.0 planned release date - Oct 20th

[1]: https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.5.0
[2]: https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.10.2

Roadmap

TrustyAI 2023 roadmap

July 2023

- *Explainers*
  - Support for explainers LIME, SHAP, CF at service level
- *Metrics*
  - Flexible scheduling/batching
  - Improve service metadata endpoints
    - Include available categories
- *Operator*
  - TrustyAI Operator v1
- *Explainers*
  - Support for external explainability libraries

September 2023

- *Storage*
  - Wider storage support (database backends)
- *Metrics*
  - Additional metrics
  - Metrics statistical tests
- KServe integration
- Drift detection
- ODH v2 onboarding

December 2023

- *Storage*
  - Wider storage support (database backends)
- *Explainers*
  - NLP explainability support
- *Detection*
  - HAP/PII
- *Metrics*
  - Support for user-defined historical windows