AI Explainability WG

December 2024

December 2024 updates

- TrustyAI core / service
  - Current: 0.24.0 release
    - https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.24.0
    - quay.io/trustyai/trustyai-service:v0.24.0
- TrustyAI operator
  - Current: 1.30.0 release
    - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.30.0
  - quay.io/trustyai/trustyai-service-operator:v1.30.0

What’s new?

TrustyAI - What's new?

- **TrustyAI core/service 0.24.0**
  - Fixes
    - Added tag validation to drift metrics, ensured name mappings are exposed to Prometheus (#660)
  - Documentation
    - Update core readme (#656)

TrustyAI - What's new?

- **TrustyAI operator 1.30.0**
  - **LMEval**
    - Remove unneeded metrics fetch (#382)
    - Filter protected env vars set from CR directly (#381)
    - Add operator online and execution feature flag (#379)
    - Add offline metrics to LMEval job image (#377)
    - Disable online evaluation (#375)
    - Disable remote code execution (#371)
    - Add offline mode as default (#370)
  - **Kueue support**
    - Disallow empty nodeAffinity (#369)

Current work

TrustyAI - LM-Eval

- Available on ODH 2.22
  - Documentation available at https://trustyai-explainability.github.io/trustyai-site/main/lm-eval-tutorial.html
- Kueue support as an overlay
  - This requires Kueue installed and manual changes to Kueue's configuration

TrustyAI - Guardrails

- Work ongoing in
  - Orchestrator support KServe serverless
  - TLS configuration for Serverless
  - gRPC service addressing

Roadmap

TrustyAI 2024 roadmap

- KServe explainer integration
- Detoxification fine-tuning
- Saliency Explainers
- Guardrails
- 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