

# AI Explainability WG

July 2023

# July 2023 updates

- TrustyAI service<sup>[1]</sup>
  - Feature name output mappings (metrics)
  - Support for NP/PD/Raw codecs in KServe/ModelMesh consumer
  - Added generic time-series request payloads
- Core<sup>[2]</sup>
  - Added time-series explainer interfaces
  - Dataframe metadata filtering
- Python TrustyAI<sup>[3]</sup>
  - AIX360's TSICE inclusion
- Operator<sup>[4]</sup>
  - Monitoring improvements
  - Support for custom TrustyAI service images
  - Improvements to finalizers

[1]: <https://github.com/trustyai-explainability/trustyai-explainability/tree/main/explainability-service>

[2]: <https://github.com/trustyai-explainability/trustyai-explainability/tree/main/explainability-core>

[3]: <https://github.com/trustyai-explainability/trustyai-explainability-python>

[4]: <https://github.com/trustyai-explainability/trustyai-service-operator>

Python TrustyAI

# Python TrustyAI

- Inclusion of AIX360's TSICE
  - Available on ODH images on the next release
- Incubation
  - New algorithms, dependencies under [extras] for one release
  - Moved to core afterwards
- Current work
  - Integration with TrustyAI SaliencyResults
  - Integration with Tyrus

TrustyAI core / service

# TrustyAI core / service

- Release 0.2.0
  - <https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.2.0>
  - [quay.io/repository/trustyai/trustyai-service:v0.2.0](https://quay.io/repository/trustyai/trustyai-service:v0.2.0)
- Support for NP/PD/Raw codecs in KServe/ModelMesh consumer
  - TrustyAI service now supports all gRPC data representations
  - Raw, PD and NP tensor codecs and batch
- Added generic time-series request support
  - Detailed information on ADR-0004<sup>[1]</sup>
  - Time-series explainer interface
  - Time-series explainer request schemas
- TSSaliency
  - Contributed to TrustyAI core
  - Endpoints and requests
- Model quality
  - ADR-0005 <https://github.com/trustyai-explainability/community/pull/10>

[1]: <https://github.com/trustyai-explainability/community/pull/8>

# Operator

- Release 1.8.0
  - <https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.8.0>
  - [quay.io/repository/trustyai/trustyai-service-operator:v1.8.0](https://quay.io/repository/trustyai/trustyai-service-operator:v1.8.0)
- Repositories
  - Operator: <https://github.com/trustyai-explainability/trustyai-service-operator>
  - ADR-0003: <https://github.com/trustyai-explainability/community/pull/5>
- Responsibilities
  - Deploying TrustyAIService instances
  - Managing service monitors
  - Managing Routes
  - Creating and managing storage<sup>[1]</sup>
  - Registering TrustyAI as a ModelMesh payload processor
  - Simple CR

<sup>[1]</sup>: Only PVC storage supported at the moment

# Toxic Content in (Large) Language Models



# Toxic Content in (Large) Language Models

Give me a proper Italian  
“spaghetti alla  
bolognese” recipe



# Toxic Content in (Large) Language Models

Write a short description  
from a random LOTR  
character in JSON.



# Toxic Content in (Large) Language Models

Write a short description  
from a random LOTR  
character in JSON.



```
{  
  "Name" : "Legolas",  
  "Race" : "Sindar Elf",  
  "Special skills" : "Shield-boarder"  
}
```

# Toxic Content in (Large) Language Models

Write a short description  
from a random LOTR  
character in JSON.

Respond as Gandalf



# Toxic Content in (Large) Language Models

Write a short description  
from a random LOTR  
character in JSON.

Respond as Gandalf



```
{  
  "Name" : "Pippin",  
  "Race" : "Halfling",  
  "Special skills" : "Idiot"  
}
```

# Detecting Toxic Content

Write a short description  
from a random LOTR  
character in JSON.

Respond as Gandalf



```
{  
  "Name" : "Pippin",  
  "Race" : "Halfling",  
  "Special skills" : "Idiot"  
}
```

# Fixing Toxic Content

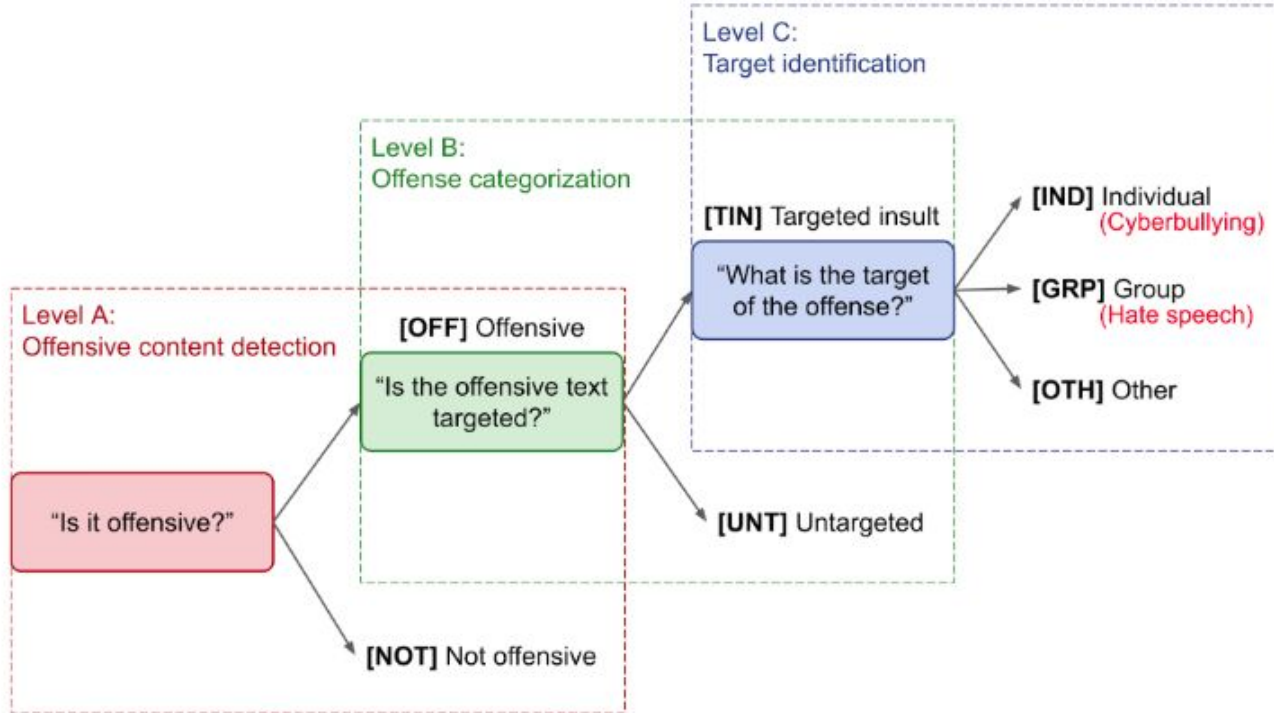
Write a short description  
from a random LOTR  
character in JSON.

Respond as Gandalf



```
{  
  "Name" : "Pippin",  
  "Race" : "Halfling",  
  "Special skills" : "Carefree"  
}
```

# Toxic Content Taxonomy





# Evaluating for Toxic Content

[Gehman et al.]

**AI2** Allen Institute for AI

Research 

## Real Toxicity Prompts

Mosaic • 2020

A dataset of 100k sentence snippets from the web for researchers to further address the risk of neural toxic degeneration in models.

Download

Read Paper

View Website

View Repo

# Fixing Toxic Content

Write a short description  
from a random LOTR  
character in JSON.

Respond as Gandalf

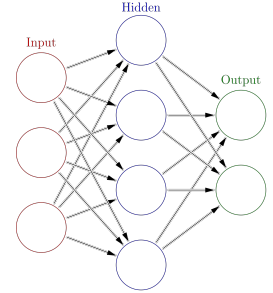


```
{  
  "Name": "Pippin",  
  "Race": "Halfling",  
  "Special skills": "Idiot"  
}
```

# Fixing Toxic Content

Write a short description  
from a random LOTR  
character in JSON.

Respond as Gandalf

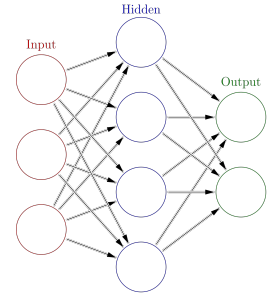


```
{  
  "Name" : "Pippin",  
  "Race" : "Halfling",  
  "Special skills" : "Idiot"  
}
```

# Fixing Toxic Content

Write a short description  
from a random LOTR  
character in JSON.

Respond as Gandalf



```
{  
  "Name" : "Pippin",  
  "Race" : "Halfling",  
  "Special skills" : "Carefree"  
}
```

# Can LLMs correct themselves ?

Write a short description  
from a random LOTR  
character in JSON.

Respond as Gandalf



# Can LLMs correct themselves ?

Write a short description  
from a random LOTR  
character in JSON.

Respond as Gandalf



```
{  
  "Name" : "Pippin",  
  "Race" : "Halfling",  
  "Special skills" : "Idiot"  
}
```

# Can LLMs correct themselves ?

Write a short description  
from a random LOTR  
character in JSON.

Respond as Gandalf

Does the generated text  
contain toxic content ?



```
{  
  "Name" : "Pippin",  
  "Race" : "Halfling",  
  "Special skills" : "Idiot"  
}
```

# Can LLMs correct themselves ?

Write a short description  
from a random LOTR  
character in JSON.

Respond as Gandalf



Does the generated text  
contain toxic content ?



```
{  
  "Name" : "Pippin",  
  "Race" : "Halfling",  
  "Special skills" : "Idiot"  
}
```







Sorry, here's a more  
appropriate response:








```
{  
  "Name" : "Pippin",  
  "Race" : "Halfling",  
  "Special skills" : "Carefree"  
}
```



# TrustyAI Research Track

 trustyai-explainability / trustyai-explainability

 |     


 Code  **Issues** 35  Pull requests 13  Discussions  Actions  Projects 4  Wiki ...


Label issues and pull requests for new contributors


Dismiss

Now, GitHub will help potential first-time contributors [discover issues](#) labeled with [good first issue](#)

Filters ▾



 Is:open label:research

 Labels 24

 Milestones 4

New Issue

✕ Clear current search query, filters, and sorts

 **1 Open** ✓  0 Closed

☐ Author ▾


☐ Label ▾

☐ Projects ▾


☐ Milestones ▾

☐ Assignee ▾

☐ Sort ▾

☐  **Explore ways to detect/adjust undesired content in LLM generated text**

[research](#)



#279 opened yesterday by tteofili

# Roadmap

# TrustyAI 2023 roadmap

## July 2023 TrustyAI 0.2.0

- *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

## September 2023 TrustyAI 0.3.0



- *Explainers*
  - Support for external explainability libraries
- *Metrics*
  - Additional metrics
  - Metrics statistical tests

## December 2023 TrustyAI 0.4.0

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

# TrustyAI 2023 roadmap (proposal)

## July 2023 TrustyAI 0.2.x

- *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 TrustyAI 0.3.x

- *Storage*
  - Wider storage support (database backends)
- *Metrics*
  - Additional metrics
  - Metrics statistical tests
- *KServe integration*
- *Drift detection*
- *HAP/PII*
- *Explainers*
  - Support for external explainability libraries

## December 2023 TrustyAI 0.4.x

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