Eval 3.3 Task Scoring Details

Example Data

Example data, including ArtifactGraphs, input EvidenceGraphs and output EvidenceGraphs can be found in the 3.3 examples on ownCloud

System API

API changes and usage are documented in the Rolling Evaluation 3.3 - Developer Notes

ScoringEngine Documentation:

https://gitlab.semaforprogram.com/semafor/teams/TA3/scoringengine

NOTES:

The EvidenceGraph diagrams shown in this document are not intended to limit/restrict what TA1 algorithms should produce, but rather to provide a reference for what is required for evaluation. If additional evidence (such as image/video/audio localizations) can be provided, please do so

---

3.3.1.1 - Deepfake Video Detection

Hypothesis

The hypothesis of this task is "The video in this MMA is a deepfake"

Task: DETECTION

Scoring behavior

As there is no AnalyticScope in use for this task, the scoring process is done by finding the first ConsistencyCheck node (under the **EvDetectionNode**) whose leaf EvReferenceNodes resolve to a violation in the summary file (in this case the video asset). This remains identical to the RE3.1.2 scoring

LLR Scores

LLR Scores will be taken at the consistency check

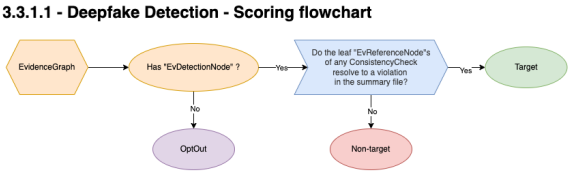
- +LLR indicates that the video is a deepfake
- -LLR indicates that the video is pristine

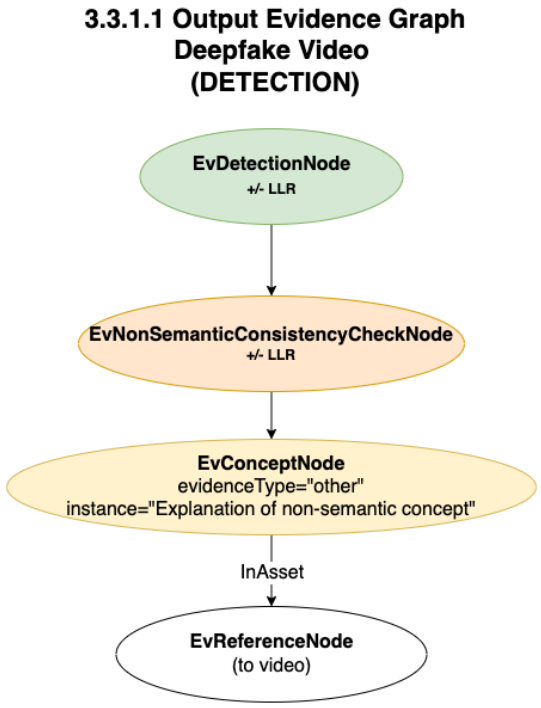
Targets

- Target probes are deepfake videos

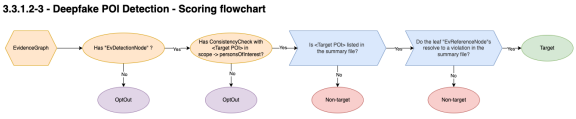
Non-targets

- Non-target probes are pristine videos





3.3.1.2-3 - Deepfake POI Detection



Hypothesis

The hypothesis of this task is "The video in this MMA is a deepfake of <Target POI>"

Task: DETECTION

Persons of Interest:

- 3.3.1.2 - Putin
- 3.3.1.3 - Zelenskyy

NOTE:

This task differs from 3.3.1.1 in that the non-target class is comprised of both deepfake and pristine videos.

Scoring behavior

Scoring behavior is similar to a regular detection task with the addition of an AnalyticScope field containing the target POI. The first consistency check containing the target POI will be evaluated

Targets

- Target probes are deepfake videos of the target POI

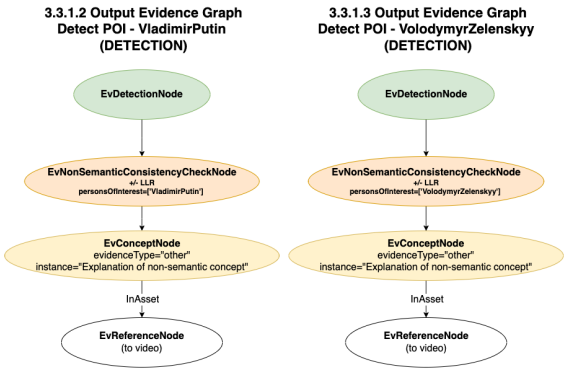
Non-targets

- Pristine videos of the target POI
- Deepfake videos of non-target POIs

LLR Scores

LLR scores will be taken at the consistency check containing the target POI

- +LLR indicates that the video is indeed a deepfake of the target POI
- -LLR indicates the video is not a deepfake of the target POI



3.3.1.4-5 - Generator Attribution (Video)

Hypothesis

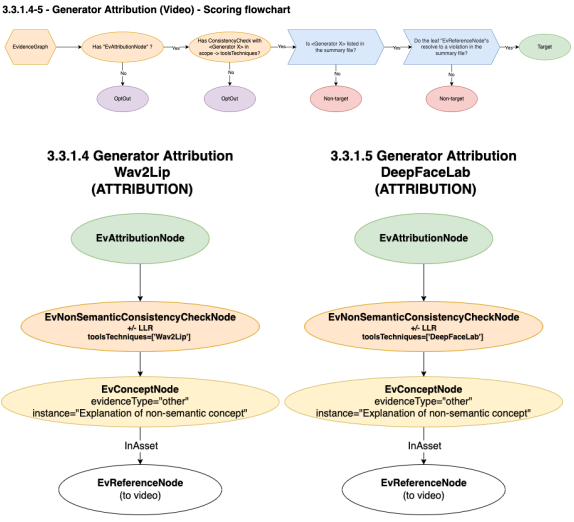
The hypothesis of this task is "The video in this MMA came from generator X"

Task: DETECTION

Generators used:

- 3.3.1.4 - Wav2Lip
- 3.3.1.5 - DeepFaceLab

Scoring behavior



This task will be scored similarly to RE3.2.4. The first consistency check whose scope contains the target generator will be evaluated

Targets

- Target probes are videos that were produced with generator X

Non-targets

- Non-target probes are videos produced with other generator types

LLR Scores

LLR scores will be taken at the consistency check containing the target generator

- +LLR indicates that the video came from generator X
- -LLR indicates that the video did not come from generator X

3.3.2.1 - Generated Audio Detection

Hypothesis

The hypothesis of this task is "The audio in this MMA is falsified"

Task: DETECTION

Scoring behavior

As there is no AnalyticScope in use for this task, the scoring process is done by finding the first ConsistencyCheck node (under the **EvDetectionNode**) whose leaf EvReferenceNodes resolve to a violation in the summary file (in this case the audio asset). This remains identical to the RE3.1.3 scoring

Targets

- Target probes will have generated/synthetic audio

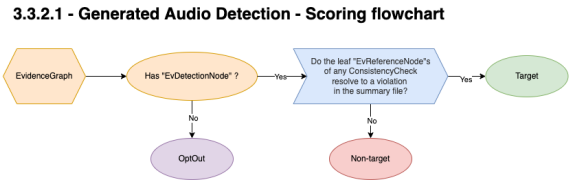
Non-targets

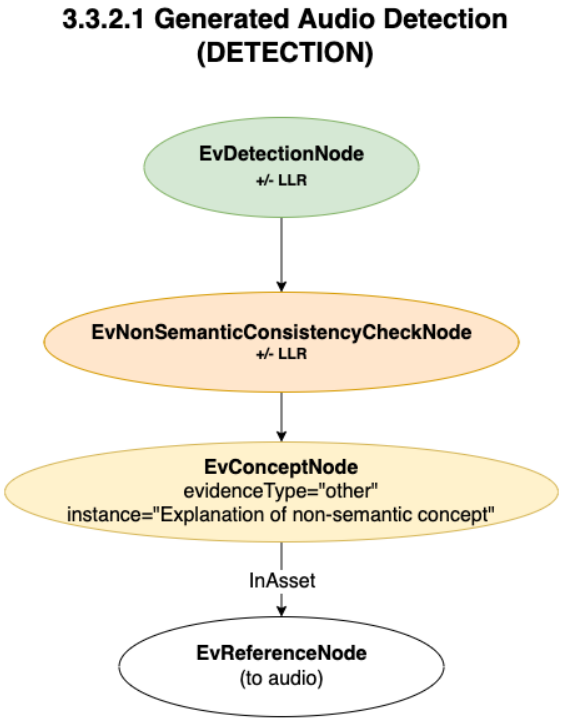
- Non-target probes will have pristine audio

LLR Scores

LLR Scores will be taken at the consistency check

- +LLR indicates the audio is generated/synthetic
- -LLR indicates the audio is pristine





3.3.2.2-3 - Generator Attribution (Audio)

Hypothesis

The hypothesis of these tasks is "The audio in this MMA came from generator X"

Task: ATTRIBUTION

Generators used:

- 3.3.2.2 - Real-Time-Voice-Cloning (RTVC)
- 3.3.2.3 - Multi-Tacotron-Voice-Cloning (MTVC)

Scoring behavior

This task will be scored similarly to RE3. 2.4. The first consistency check whose scope contains the target generator will be evaluated

Targets

- Audio produced with <generator X>

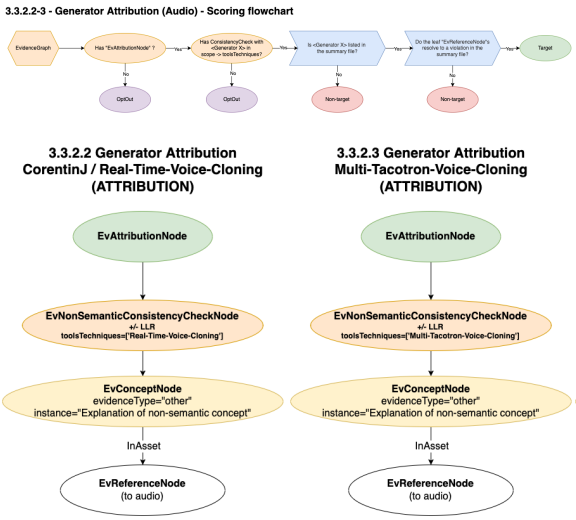
Non-targets

- Audio produced with other generators

LLR Scores

LLR scores will be taken at the consistency check containing the target generator

- +LLR indicates that the audio came from <generator X>
- -LLR indicates that the audio did not come from <generator X>



3.3.3.1a/b - Generated Text Detection

Hypothesis

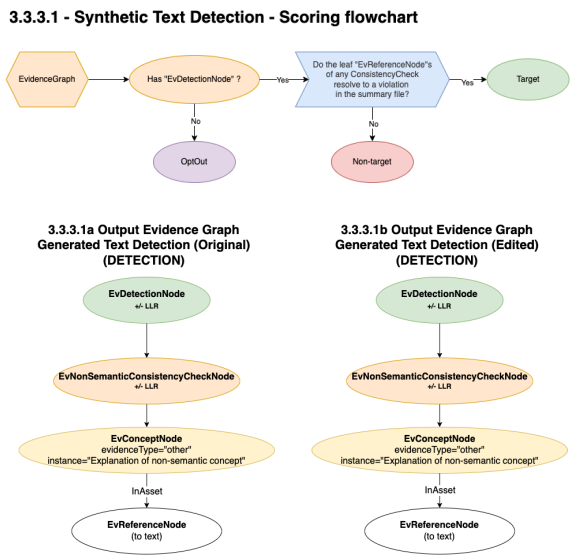
The hypothesis of these tasks is "The text in this MMA is falsified"

Task: DETECTION

Scoring behavior

As there is no AnalyticScope in use for this task, the scoring process is done by finding the first ConsistencyCheck node (under the **EvDetectionNode**) whose leaf EvReferenceNodes resolve to a violation in the summary file (in this case the text asset). This remains identical to the RE3.1.4 scoring

Targets



- 3.3.3.1a - Generated articles
- 3.3.3.1b - Generated+maniulated articles

Non-targets

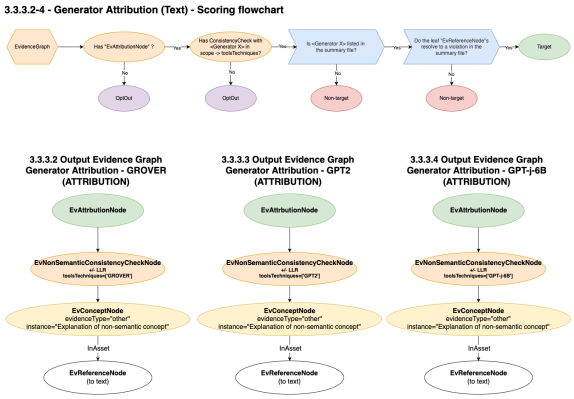
- 3.3.3.1a+b - Pristine articles

LLR Scores

LLR Scores will be taken at the consistency check

- +LLR indicates the article is falsified
- -LLR indicates the article is pristine

3.3.3.2-4 - Generator Attribution (Text)



Hypothesis

The hypothesis for this task is "This MMA came from generator X"

Task: ATTRIBUTION

Generators used:

- 3.3.3.2 - GROVER
- 3.3.3.3 - GPT2
- 3.3.3.4 - GPT-j-6B

Scoring behavior

Scoring for this task is similar to the 3.2. 4 Synthetic Media Attribution task from RE 3.2. An AnalyticScope field of toolsTechniques will be provided to analytics containing the task's generator.

Targets

- Target probes will be articles produced from <generator X>

Non-targets

- Non-target probes will have been produced from other generators (the target probes of other sub-tasks)

LLR Scores

LLR scores will be taken at the consistency check containing the target generator

- +LLR indicates that the MMA came from <generator X>
- -LLR indicates that the MMA did not come from <generator X>

3.3.4.1 - Detect MMA Inconsistencies

Hypothesis

The hypothesis for this task is "There is an inconsistency between the provided AOM structures"

Task: DETECTION

AOM anchors provided:

- 3.3.4.1a - Headline, Image
- 3.3.4.1.b - Headline, Body
- 3.3.4.1.c - Image, Body
- 3.3.4.1.d - Headline, Image, Body

NOTE:

- This task makes use of the new **aomStructures** field of the AnalyticScope. Please refer to the Rolling Evaluation 3.3 - Developer Notes for details and usage.
- Subtasks **a-d** will be combined into a single competition in the GYM. The ScoringEngine will be updated at a later date to produce score breakdowns per-subtask

Scoring behavior

All subtasks will be combined into a single competition in the gym. Analytics will be probed with an MMA and a set of AOM anchors and are asked to identify the inconsistency between them. If an analytic doesn't support the combination of anchors provided, we request that analytics explicitly opt-out. The EG should contain an **EvDetection Node** along with an **EvSemanticConsi stencyCheckNode** containing the analysis. We request that performers report the aomStructures back as follows:

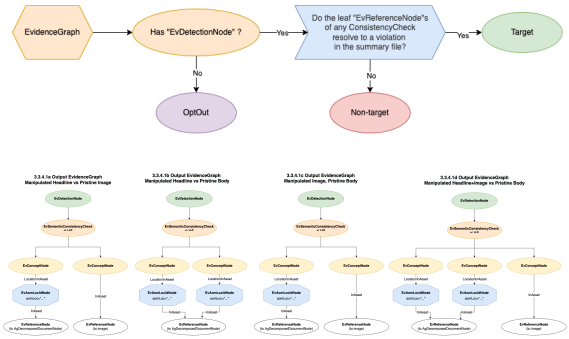
- Headline/Body - report these back in the form of **EvAomLoc IdNode** followed by a reference to the **AgDecompos edDocumentNode**
- Image - Reference the image node in the AG (see the developer notes for how to resolve this using the anchor).
    - Additionally you can provide a localization for the image, this helps later analysis efforts

Targets

- Target probes will have a semantic inconsistency between the provided anchors.

Non-targets

3.3.4.1 - Detect MMA Inconsistencies - Scoring Flowchart



- Non-target probes will be pristine articles. We will still provide the aomStructures, but there will exist a semantic inconsistency between them

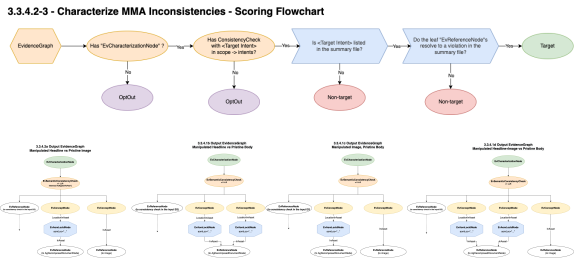
LLR Scores

LLR scores will be taken at the consistency check

- +LLR indicates that there exists an inconsistency between the provided aomStructures
- -LLR indicates that there is no inconsistency between the provided aomStructures

---

3.3.4.2-3 - Characterize MMA Inconsistencies



Hypothesis

The hypothesis for this task is "The MMA was falsified with intent X"

The dataset for this task is comprised of the same manipulated data from 3.3.4.1. Instead of an aomStructures analytic scope, there will be input EvidenceGraphs identifying the inconsistencies. These will be structurally similar to the output EvidenceGraphs for 3.3.4.1.

Because there is 4 different variations of manipulations, input and output EGs for this task will have 4 basic forms

Intents

- 3.3.4.2 - ToAppealToFear
- 3.3.4.3 - ToMinimize

Scoring behavior

Targets

- Target probes will be falsified with <intent X>

Non-targets

- Non-target probes will be falsified with another intent

LLR Scores

LLR scores will be taken at the consistency check containing the target intent

- +LLR indicates the MMA was falsified with <intent X>
- -LLR indicates the MMA was not falsified with <intent X>