SemaFor
SEMANTIC FORENSICS

SemaFor Evaluation 3 Plan

August 30, 2022

Version 4.2

Change Log

Version	Date	Explanation of Changes
.1	15 March 2022	Initial version of design document for 3.2.
.1.1	23 March 2022	Updated to include current generator information for Evaluation 3.2.
2	15 April 2022	• Added Consistency Check Protocol. • Added probe data examples • Added introduction, timeline, and Phase 2 metrics • Added scoring details • Added Evaluation 3.1 Design • Added Analysis Plan
2.1	19 April 2022	• Update to the Evaluation 3 schedule tables and figure.
3.0	10 May 2022	• Adding generator specifics for Synthetic Media Attribution tasks. • Clarifications added to Evaluation 3.2 tasks and mapping to gym representation. • Updates to localization scoring for tasks 3.2.3. • Updated dates for 3.2 timeline to account for smaller maintenance window and later start date.
4.0	31 May 2022	• Added interrater consistency check results. • Added Evaluation 3.3 Design.
4.1	14 June 2022	• Updated 3.2.4 scoring section to clarify +/- LLR score descriptions
4.2	30 August 2022	• Clarification to Evaluation 3.3 Design. • Added Evaluation 3.4 Design. • Updated 3.2 localization task definitions, design, and scoring approach. • Added examples for Evaluation 3.3 and scoring information. • Add inter rater reliability plan for Evaluation 3.3

Contents

List of Figures

List of Tables

SemaFor Evaluation 3 Plan

Introduction

The TA3 SemaFor team, led by PAR Government, supporting the Defense Advanced Research Projects Agency (DARPA) Semantic Forensics (SemaFor) Broad Agency Announcement (BAA) intends to refine and extend methods for evaluating the identification, attribution, and characterization of semantic inconsistencies in falsified media at scale. This document describes the design for Evaluation 3 of the SemaFor program, including task definitions, evaluation metrics, and scoring procedures.

Drawing upon lessons learned from previous evaluations, the program has moved towards a rolling evaluation framework that will enable more iterative design, development, execution, and evaluation analysis by staging tasks throughout the year to address foundational elements and build to more complexity over the course of the program. In addition to modifying the way evaluations are conducted, the program will be emphasizing several key areas over the course of Evaluation 3. These focus areas include the following:

1. Increase the alignment of evaluation data and tasks to the threat landscape by working with TA4 performers to conduct a mapping of Evaluation 3 tasks to the threat landscape for quantitative representation and review.
2. Move towards a characterization framework approach for evaluating the antecedent tasks necessary to make characterization judgements in a more tractable manner.
3. Focus domains on technical information, technical news articles, and technical propaganda.
4. Leverage real world datasets in experimental tasks to represent transition partner requirements and insights.
5. Expand evaluation tasks to include specific designs for evaluating fusion, prioritization, and explainability within the human-machine interface of the SemaFor prototype.
6. Develop an evaluation plan for assessing LLR calibration.

We note that this is a living document which will be updated with information and specifications for the additional rolling evaluation periods as those are defined by the program.

Evaluation 3 Timeline

Evaluation 3 will be broken into four periods with specific focus areas that build on complexity over time. These include:

1. Rolling Evaluation (RE) 3.1: Bridging from Evaluation 2
2. RE 3.2: Characterization
3. RE 3.3 Audio and Video Modalities, Technical Information
4. RE 3.4: Calibration

An initial schedule is provided below in *Table 1*. Note that this may change slightly over time depending on requirements for maintenance windows, program directions, or resource constraints. To reduce cycle time, TA1 performers are recommended to focus on meaningful updates to the analytics as opposed to incremental updates.

Table 1. Evaluation 3 Schedule.

Rolling Evaluation Period	Initial Open Dates	Reopen Dates
RE 3.1	3/16/22-4/13/22	4/25/22-5/1/22, 5/16/22-6/12/22, 6/15/22-7/26/22, 8/17/22-9/14/22, 10/5/22-10/26/22
RE 3.2 (reopen 3.1)	5/16/22-6/12/22	6/15/22-7/26/22; 8/17/22-9/14/22, 10/5/22-10/26/22
RE 3.3 (reopen 3.1, 3.2)	8/17/22-9/14/22	10/5/22-10/26/22
RE 3.4 (reopen 3.1, 3.2, 3.3)	10/5/22-10/27/22	NA

Each rolling evaluation period is expected to have an initial "open" period of approximately a month to allow for TA1 analytics to iterate and submit over the course of this initial period. TA3 will also be providing 25% of the data across tasks and experimental conditions to performers two weeks in advance of the tasks opening. This data will then be available in the Gate Test to allow for testing of components prior to submission for evaluation. Following the initial open period, tasks will then close to allow for regular system maintenance and upgrades (see *Table 2*). Following these regular system maintenance windows, the rolling evaluation tasks will be reopened to allow for further iteration across components.

Table 2. Initial Scheduled Maintenance Periods.

Initial Scheduled Maintenance Periods
4/14/22-4/24/22
6/12/22-6/14/22
7/27/22-8/16/22
9/15/22-10/4/22

Tasks are anticipated to remain open throughout the duration of Evaluation 3 pending regular scheduled maintenance windows. Evaluation 3.4 will therefore reopen all previous tasks to allow for a final submission for the Evaluation 3 timeframe. TA3 proposes to take the final scores for the formal Evaluation 3 Report at the conclusion of Evaluation 3.4 to allow performers the maximum amount of time for iterating on analytics. Interim analysis on rolling evaluation periods will be provided by TA3 in slide form and will include smaller mini-PI meeting discussions at the Data/Eval WG or Program Scrum to allow performers to share insights across the program on interim results.

SemaFor Evaluation 3 Plan

Phase 2 Metrics

Evaluation 3 falls within Phase 2 of the SemaFor program. The program goals for Phase 2 consist of 85% p(D) @ 0.08 FAR for detection, 85% p(D) @ 0.08 FAR for attribution, and 80% bACC @ EER for characterization, and 70% Precision for prioritization.

Table 3. Phase 2 Program Metrics.

Metrics	Detection	Attribution	Characterization	Prioritization
Probability of Detection, or p(D)	85% p(D)	85% p(D)	N/A	N/A
False Alarm Rate (FAR)	8% FAR	8% FAR	N/A	N/A
Accuracy	N/A	N/A	80% Accuracy	N/A
Precision	N/A	N/A	N/A	70% Precision

SemaFor Evaluation 3.1 Design

Evaluation 3.1 will focus on reopening selected tasks from Evaluation 2 for the purposes of comparing performance over time. These tasks are listed in the below table.

Table 4. Evaluation 3.1 Tasks and Descriptions.

Name	Description	Tasks
3.1.1 Attribution Verification (News Articles) - Author Swaps	Target probes will have authors within the same organization swapped. All articles will purport to be from a known author, but the true source of the content may be unknown.	Attribution
3.1.2 Deepfakes - Social Media (Craig Kelly)	Target probes will have pristine video or deepfaked video (face swap), some will be malicious, and some will not be.	Detection, Attribution, Characterization
3.1.3 Generated Audio - Social Media (Rahul Gandhi)	Target probes will have fully generated audio or interleaved audio, some will be malicious, and some will not be.	Detection, Attribution, Characterization
3.1.4 Generated Text - News Articles	Target probes will have generated text. Some text will be malicious, and some will not be.	Detection, Attribution, Characterization
3.1.5 Text & Image Inconsistencies - News Articles	Target probes will have semantic inconsistencies between text and images, some will be malicious, and some will not be. Note: there will be inconsistencies between modalities.	Detection, Characterization
3.1.6 Text & Image Inconsistencies - Social Media	Target probes will have semantic inconsistencies between text and images, some will be malicious, and some will not be. Note: there will be inconsistencies between modalities.	Detection, Characterization

Feedback from Evaluation 2 as well as new validation techniques were applied to the data for Evaluation 3.1 competitions, which resulted in changes to the data for the following competitions:

- **Task 3.1.1 Attribution Verification (News Articles) - Author Swaps**
  - Probes with content that had been erroneously scraped from the web were dropped.
  - Probes with no authors or authors that did not come from the approved set were dropped.
  - Probes for Authors that had not been included in the original Evaluation 2 set were added.
- **Task 3.1.4 Generated Text - News Articles**
  - Probes with content that had been erroneously scraped from the web were dropped.
  - Probes with caption text duplicated in the text content were modified to repair this issue.

  - Probes where the Author was changed to be from the set of approved authors for the Attribution task were added.
- **Task 3.1.5 Text & Image Inconsistencies - News Articles**
  - Probes with content that had been erroneously scraped from the web were dropped.
  - Probes with caption text duplicated in the text content were modified to repair this issue.
  - Pristine probes with authors from the approved set were added, since this competition shares the pristine set with Generated Text - News Articles.

Task 3.1 Scoring

Scoring information and EG representation for these tasks remains the same as Evaluation 2 and can be found in the Evaluation 2 Report available on Confluence.

SEMAFOR EVALUATION 3.2 DESIGN

This design section details the various definitions, tasks, and conditions for Evaluation 3.2. This evaluation will focus on the tasks and research questions defined in Table 5 below and further described over the course of the design document.

Table 5. Evaluation 3.2 Task Definitions and Research Questions.

Tasks		Research Question
Task 3.2.1 - Characterization Intent Identification (C)	TA1 performers will be given a falsified MMA and must identify the intent of the manipulation across various manipulation tactics.	With manipulation information provided, how does the (1a) manipulation level, (1b) modality manipulated (text, image), and (1c) propaganda tactics impact performers success in identifying the intent of the manipulated MMA?
Task 3.2.2 - Propaganda Tactic Identification (C)	TA1 performers will be given a falsified MMA and must identify the propaganda tactic that was used.	With manipulation information provided, does the (2a) manipulation level, (2b) modality manipulated (text, image), or (2c) type of characterization impact performers' success in identifying specific propaganda tactics in the manipulated MMA?
Task 3.2.3 - Semantic Labeling and Localization (D)	TA1 performers will be given a falsified image and must localize and semantically label the falsification(s).	Does the (3a) type of image manipulation or the (3b) category of semantic content manipulated impact performers' success in successfully localizing and semantically labeling?
Task 3.2.4 - Synthetic Media Attribution (A)	TA1 performers will be given a falsified text or image asset and must detect the falsification and identify the generator or tool that was used to create the falsification.	With no manipulation information provided, is the successful identification of the generator (for text, GPT2, GPT-j-6B, and GROVER; for images, StyleGAN2, StyleGAN3, taming-transformers, latent-diffusion, LSGM, and CLIP-guided-diffusion) more or less difficult due to (4a) modality manipulated (text, image) or (4b) which generator is used?

In the below table we will summarize the tasks as they will appear in the gym competitions.

Table 6. Evaluation 3.2 Tasks and Descriptions.

Name	Description	Tasks
3.2.1.1 Intent Identification – ToDiscreditEntity	Determine if the intent of the manipulation is to Discredit Entity.	Characterization
3.2.1.2 Intent Identification - ToCalltoAction	Determine if the intent of the manipulation is to Call to Action.	Characterization
3.2.2.1 Tactic Identification - HateSpeech	Determine if the propaganda tactic used in the manipulation is Hate Speech.	Characterization
3.2.2.2 Tactic Identification - Dictat	Determine if the propaganda tactic used in the manipulation is Dictat.	Characterization
3.2.2.3 Tactic Identification - Scapegoating	Determine if the propaganda tactic used in the manipulation is Scapegoating.	Characterization
3.2.2.4 Tactic Identification - Bandwagoning	Determine if the propaganda tactic used in the manipulation is Bandwagoning.	Characterization
3.2.3.1 Semantic Label Detection – People	Semantically label the manipulation as being a Person or People.	Detection
3.2.3.2 Semantic Label Detection – Fire	Semantically label the manipulation as being a Fire.	Detection
3.2.3.3 Semantic Label Detection - Symbol	Semantically label the manipulation as being a Symbol.	Detection
3.2.3.4 Semantic Label Detection - Firearms	Semantically label the manipulation as being a Firearm.	Detection
3.2.3.5 Semantic Label Detection – Signs	Semantically label the manipulation as being a Sign.	Detection
3.2.3.6 Semantic Label Detection - Vehicles	Semantically label the manipulation as being a Vehicle.	Detection
3.2.3.7 Semantic Label Localization	Semantically localize the manipulation.	Detection
3.2.4.1 Synthetic Media Attribution – GPT-j-6B	Determine if the manipulation was generated with GPT-j-6B.	Attribution
3.2.4.2 Synthetic Media Attribution – GROVER	Determine if the manipulation was generated with Grover.	Attribution
3.2.4.3 Synthetic Media Attribution – GPT2	Determine if the manipulation was generated with GPT2.	Attribution
3.2.4.4 Synthetic Media Attribution – StyleGAN2	Determine if the manipulation was generated with StyleGAN2.	Attribution
3.2.4.5 Synthetic Media Attribution – StyleGAN3	Determine if the manipulation was generated with StyleGAN3.	Attribution
3.2.4.6 Synthetic Media Attribution – Taming-transformers	Determine if the manipulation was generated with Taming-transformers.	Attribution
3.2.4.7 Synthetic Media Attribution – Latent-diffusion	Determine if the manipulation was generated with Latent-diffusion.	Attribution
3.2.4.8 Synthetic Media Attribution – LSGM	Determine if the manipulation was generated with LSGM.	Attribution
3.2.4.9 Synthetic Media Attribution – Guided-diffusion	Determine if the manipulation was generated with Guided-diffusion.	Attribution

NEWS ARTICLE INTENT CHARACTERIZATION AND IMAGE FALSIFICATION LOCALIZATION CAMPAIGN DEFINITIONS

Characterization

In collaboration with TA1 and TA4, TA3 worked to refine a framework (see *Figure 1. Characterization Framework*) to break down the characterization task into smaller subtasks that could be approached in Evaluation 3.2. This framework aims to look at the subcomponents that would inform a decision-making assessment about malicious or benign designations for an MMA. In this framework we imagine the lowest level of analysis to focus on the detection layer, which consists of first identifying a falsification. The semantic layer then aims to provide context to the identified falsification by providing a label of what was manipulated and where it occurs in the MMA. The attribution layer provides a further level of analysis by aiming to attribute the falsification to particular tools, generators, or actors, and placing that within the threat landscape of the capabilities available to various adversaries. The intent layer builds on the previous layers and aims to identify the intent behind the falsification and the impact it aims to produce. We note that this framework is evolving as the program continues, and we appreciate performers feedback and continued refinement of this framework approach.

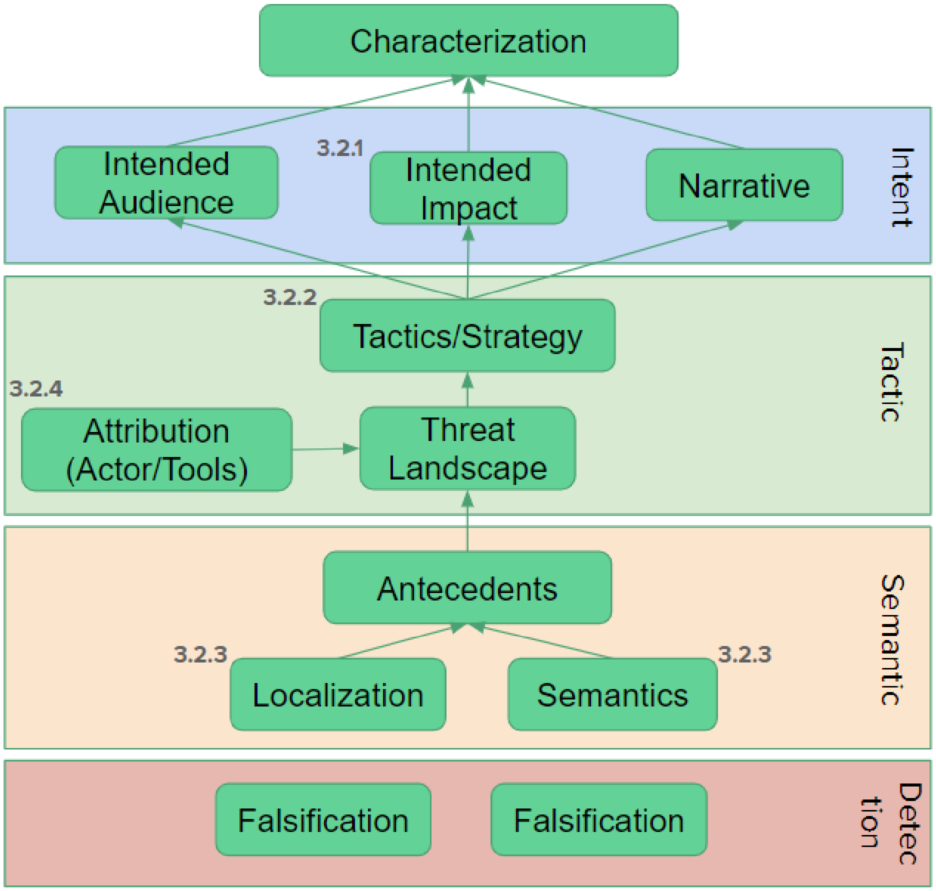


Figure 1. Characterization Framework

In Evaluation 3.2, we have devised four high level tasks, that perform against different elements of the characterization framework and at various levels of analysis:

- **3.2.1: Intent Identification**
- **3.2.2: Propaganda Tactic Identification**
- **3.2.3: Semantic Labeling and Localization**
- **3.2.4: Synthetic Media Attribution**

Evaluation 3.2 will be confined to two characterization intents: **Discredit Entity** and **Call to Action** and the minimum number of manipulation tactics to adequately represent this (four in total).

Discredit Entity

Discredit Entity refers to MMAs that are manipulated with the intention of harming the reputation, trustworthiness, or competence of an organization, nation-state, or individual. Generally, this characterization is negative in tone. This may include disparaging language towards the entity (e.g., slurs, comparisons to semantic categories with negative connotations). Discredit Entity MMAs are pieces of content/media intended to lower an audience's view of an entity (person, government, policy, vaccine, technology, etc.). This can be done by altering or omitting facts and information as well as by adding false information or by adding in suggestive commentary to make emotional appeals to a particular viewpoint. Additionally, this may also include questioning the reputation, trustworthiness, or competence rather than declaration.

Call to Action

Call to Action refers to MMAs that are manipulated with the intention of an exhortation or stimulus to do something to achieve an aim or deal with a problem. This may include content intended to induce a viewer, reader, or listener to perform a specific act, typically taking the form of an instruction or directive. This typically takes the form of an instruction or directive to the audience and is designed to provoke an immediate response. In terms of text, *calls to action* use action verbs and specific dates, times, and numbers ("Tuesday," "Freedom Square," "ten thousand people") and usually use an imperative verb. In terms of the broader contextual environment, *calls to action* contain an issue (over which to take action) and a specific judgment on the status of issue to support the reason for the call to action.

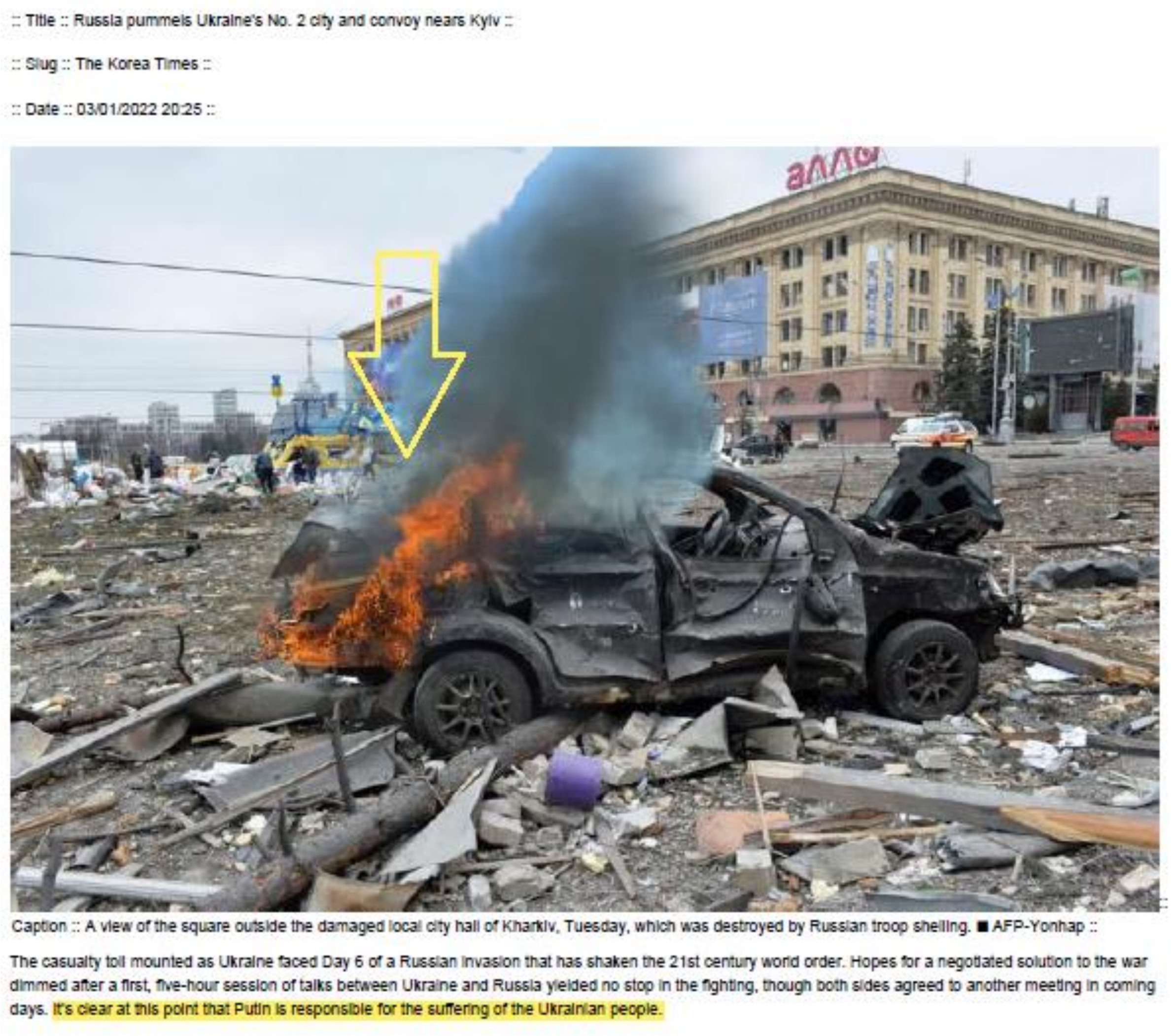
Manipulation Techniques

Embedded within the two characterization intentions, probes will contain four manipulation techniques. Probes will contain *one tactic* per characterization intent. Four techniques, *Scapegoating*, *Hate Speech*, *Bandwagoning*, and *Dictat*, (defined below) were used and embedded under the two characterization types. Specifically, *Scapegoating* and *Hate Speech* were nested under **Discredit Entity**, whereas *Bandwagoning* and *Dictat* were nested under **Call to Action**. Tactics were not used across intents, but fully nested under their characterization types.

The proposed degree of manipulation—regardless of technique—is detailed in the Manipulation Level section below.

Scapegoating

Scapegoating is used to assign blame to an individual or group for a problem, and thus alleviating feelings of guilt from responsible parties and/or distracting attention from the need to fix the problem. Attempts to assign blame to an individual, group, idea, or entity, thus alleviate feelings of guilt from responsible parties and/or distract attention from the need to fix the problem for which blame is being assigned. This manipulation technique transfers blame to one person or group of people without investigating the complexities of the issue. For an example of scapegoating, see *Figure 2. Example Data Probe of Scapegoating with one-sentence text and image manipulation (highlighted)* below.



:: Title :: Russia pummels Ukraine's No. 2 city and convoy nears Kyiv ::

:: Slug :: The Korea Times ::

:: Date :: 03/01/2022 20:25 ::

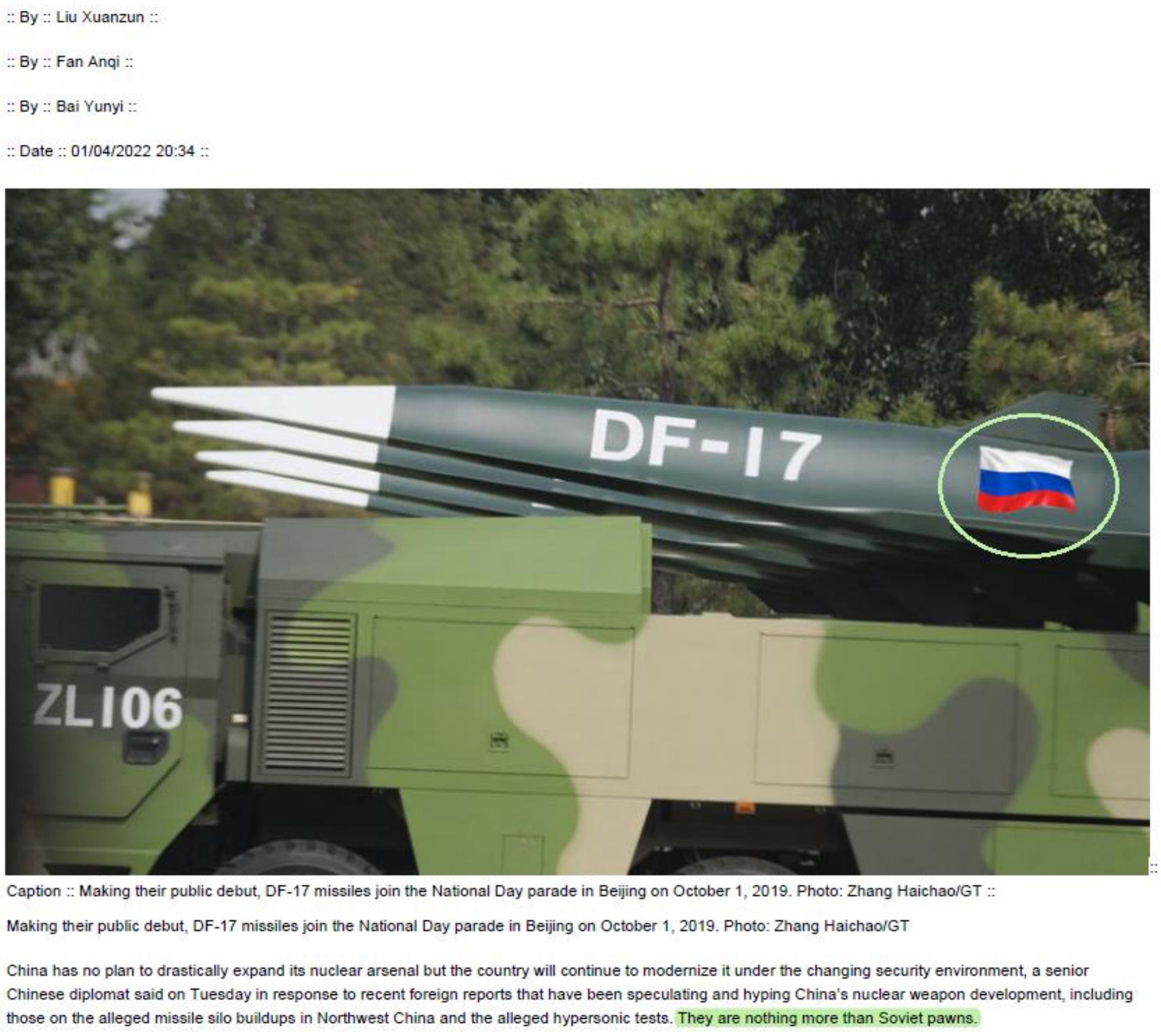
Caption :: A view of the square outside the damaged local city hall of Kharkiv, Tuesday, which was destroyed by Russian troop shelling. ■ AFP-Yonhap ::

The casualty toll mounted as Ukraine faced Day 6 of a Russian invasion that has shaken the 21st century world order. Hopes for a negotiated solution to the war dimmed after a first, five-hour session of talks between Ukraine and Russia yielded no stop in the fighting, though both sides agreed to another meeting in coming days. It's clear at this point that Putin is responsible for the suffering of the Ukrainian people.

Figure 2. Example Data Probe of Scapegoating with one-sentence text and image manipulation (highlighted)

Hate Speech

Hate speech is used to incite fears and arouse prejudices for the purposes of constructing a negative opinion about a group or set of beliefs or ideas. This includes abusive or threatening speech or writing that expresses prejudice against a particular group, especially on the basis of race, religion, or sexual orientation. It attempts to persuade an audience by using abusive, threatening, or derogatory commentary about a group, idea, or policy, with the intent to lessen the audience’s view of that group, idea, or policy. For an example of hate speech, see



:: By :: Liu Xuanzun ::

:: By :: Fan Anqi ::

:: By :: Bai Yunyi ::

:: Date :: 01/04/2022 20:34 ::

::

Caption :: Making their public debut, DF-17 missiles join the National Day parade in Beijing on October 1, 2019. Photo: Zhang Haichao/GT ::

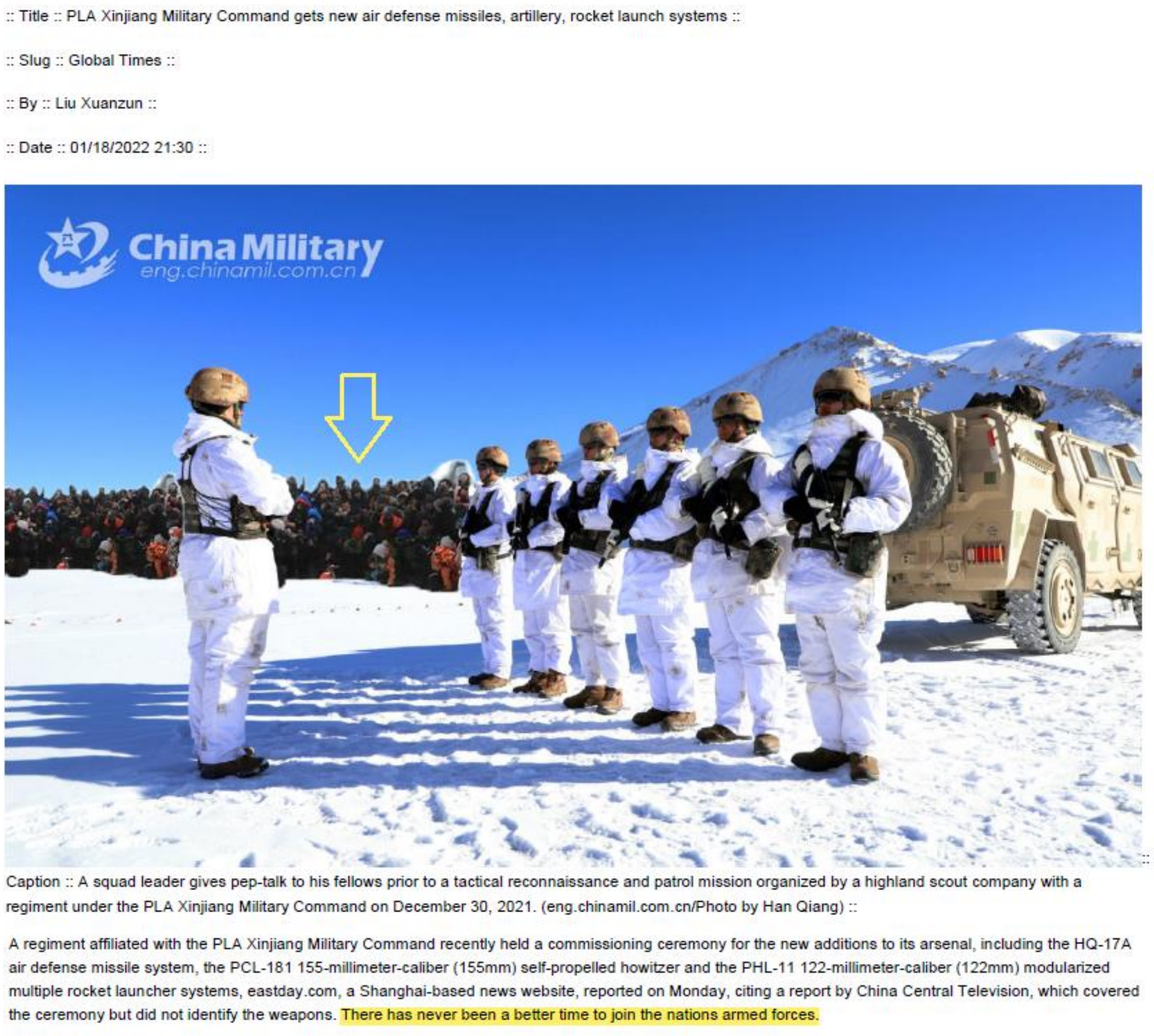
Making their public debut, DF-17 missiles join the National Day parade in Beijing on October 1, 2019. Photo: Zhang Haichao/GT

China has no plan to drastically expand its nuclear arsenal but the country will continue to modernize it under the changing security environment, a senior Chinese diplomat said on Tuesday in response to recent foreign reports that have been speculating and hyping China's nuclear weapon development, including those on the alleged missile silo buildups in Northwest China and the alleged hypersonic tests. They are nothing more than Soviet pawns.

Figure 3. Example Data Probe of Hate Speech with one-sentence text and image manipulation (highlighted)

Bandwagon

Bandwagon techniques attempt to persuade the audience to join an in-group and take the course of action that everyone else is already taking. This technique implies an irresistible or inevitably successful mass movement. Bandwagoning techniques aim to persuade the audience to join in and take the course of action that "everyone else is taking" or uses a "fear of missing out" approach. It asserts and appeals to popularity or strengthens by appeal to authority. For an example of bandwagoning, see *Figure 4. Example Data Probe of Bandwagoning with one-sentence text and image manipulation (highlighted)* below.



:: Title :: PLA Xinjiang Military Command gets new air defense missiles, artillery, rocket launch systems ::

:: Slug :: Global Times ::

:: By :: Liu Xuanzun ::

:: Date :: 01/18/2022 21:30 ::

::

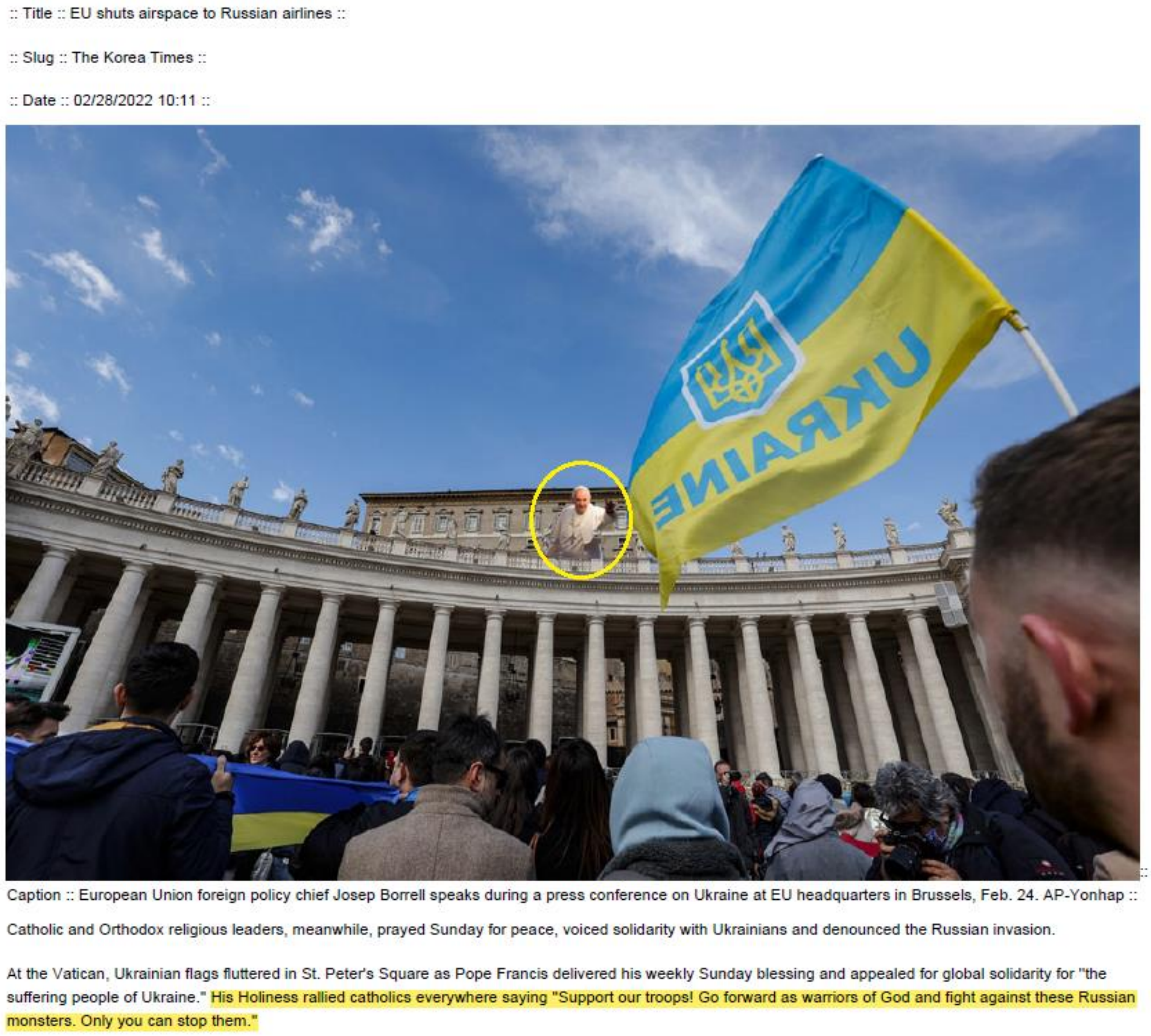
Caption :: A squad leader gives pep-talk to his fellows prior to a tactical reconnaissance and patrol mission organized by a highland scout company with a regiment under the PLA Xinjiang Military Command on December 30, 2021. (eng.chinamil.com.cn/Photo by Han Qiang) ::

A regiment affiliated with the PLA Xinjiang Military Command recently held a commissioning ceremony for the new additions to its arsenal, including the HQ-17A air defense missile system, the PCL-181 155-millimeter-caliber (155mm) self-propelled howitzer and the PHL-11 122-millimeter-caliber (122mm) modularized multiple rocket launcher systems, eastday.com, a Shanghai-based news website, reported on Monday, citing a report by China Central Television, which covered the ceremony but did not identify the weapons. There has never been a better time to join the nations armed forces.

Figure 4. Example Data Probe of Bandwagoning with one-sentence text and image manipulation (highlighted)

Dictat

Dictat seeks to attempt to simplify the decision-making process by using images and words to tell the audience exactly what actions to take, eliminating any other possible choices. This can often come from authority figures supporting a position, idea, argument, or course of action. Authority figures can be used in an appeal to authority. For an example of dictat, see the below example.



:: Title :: EU shuts airspace to Russian airlines ::

:: Slug :: The Korea Times ::

:: Date :: 02/28/2022 10:11 ::

::

Caption :: European Union foreign policy chief Josep Borrell speaks during a press conference on Ukraine at EU headquarters in Brussels, Feb. 24. AP-Yonhap ::

Catholic and Orthodox religious leaders, meanwhile, prayed Sunday for peace, voiced solidarity with Ukrainians and denounced the Russian invasion.

At the Vatican, Ukrainian flags fluttered in St. Peter's Square as Pope Francis delivered his weekly Sunday blessing and appealed for global solidarity for "the suffering people of Ukraine." His Holiness rallied catholics everywhere saying "Support our troops! Go forward as warriors of God and fight against these Russian monsters. Only you can stop them."

Figure 5. Example Data Probe of Dictat with three-sentence text and image manipulation (highlighted)

Manipulation Level

Across tasks, there will be variations in the level or degree of manipulations across characterization intents and techniques. The characterizations, manipulation techniques, and level of manipulation will be tracked in the Journaling Tool, which will also capture the ground truth of the annotation.

Manipulations are done at four different levels: 1) a single sentence insertion in the text, 2) a three sentence insertion in the text, 3) a single sentence insertion and a corresponding image manipulation, or 4) a three sentences insertion and a corresponding image manipulation. Image manipulations can include modifications to vehicles, flags, persons or people for example. The image manipulations are meant to bolster the narrative of the changes that were inserted in the text (e.g., impetus for what is calling the reader to action).

PROBE DISTRIBUTION FOR TASK 3.2.1, 3.2.2, AND 3.2.3

Table 7. Data Generation Plan for Task 3.2.1, 3.2.2, 3.2.3 details the data generation distribution and plan for Tasks 3.2.1 through 3.2.3. This provides representative stimuli to support generating ROC curves for the research questions below.

Table 7. Data Generation Plan for Task 3.2.1, 3.2.2, 3.2.3

	Discredit Entity		**Call to Action**		**Total**
Level of Manipulation	**Hate Speech**	**Scapegoat**	**Bandwagon**	**Dictat**	
1 Sentence	30	30	30	30	120
3 Sentences	30	30	30	30	120
1 Sentence + 1 Image	30	30	30	30	120
3 Sentences + 1 Image	30	30	30	30	120
Total	120	120	120	120	***480***

TOPICS

Data for Evaluation 3.2 will have specific topical and content constraints, as summarized in *Table 8. Evaluation 3.2 Constraints for Assets and Tasks* below. The only topic for Evaluation 3.2 will be Military Capabilities, other topics will follow in future rolling evaluations. Additional sub-topics may be added in future evaluations.

Table 8. Evaluation 3.2 Constraints for Assets and Tasks

Military Capabilities
Capability
Conflict
Parade
Protest
Technology
Vehicles
Other

EVALUATION 3.2.1 CHARACTERIZATION INTENT IDENTIFICATION (C)

Research Question 3.2.1

With manipulation information provided, how does the (1a) manipulation level, (1b) modality manipulated (text, image), and (1c) propaganda tactics impact performers' success in identifying the intent of the manipulated MMA?

Task 3.2.1 – Intent Characterization

For Task 1, TA1 performers will be given a falsified MMA, and must successfully identify the intent of the manipulation across various manipulation tactics. Task 1 will be broken down into two subtasks for the successful verification of the two characterization intents.

For Subtask 1a, TA1 performers must verify which probes are intended to **Discredit Entity** among all available probes. Negative cases will be **Call to Action** probes. For Subtask 1b, TA1 performers must verify which probes are intended as a **Call to Action** among all available probes. Negative cases will be **Discredit Entity** probes. These conditions will be balanced.

For this task, TA1 performers can assume that detection of what was manipulated is already completed. Thus, the purpose of the task is to determine *why* by inferring intent. Accordingly, manipulation information will be provided, including what was manipulated and where the manipulation occurred. TA1 performers will receive evidence graphs with this information. During probe generation, the Journaling Tool will include a way to record characterization type, manipulation technique, level of manipulation, and where manipulation techniques are located in the MMA.

The level of manipulation, particular manipulation tactic implemented, and the modalities may differentially impact performers' success in categorization of characterization intents. TA3 will coordinate to ensure that probe generation is completed with an inventory of those characteristics of the probes.

For Evaluation 3.2, Task 3.2.1 are constrained around the following parameters. Probes in Evaluation 3.2 will focus to only comprise Military Capabilities related topics. MMA content is News Articles and contains only text and images. Manipulation tactics are confined to text and image manipulations as outlined in Table 7.

Task 3.2.1 Scoring

TA1 Performers will be scored on their ability to successfully characterize the type of characterization intended with the program goal of 80% bACC @ EER. For scoring purposes, each intent subtask will be separated into its own competition in the SemaFor Gym and scored individually as a binary classification problem. Within a given competition, target probes will be those falsified with the specified intent. Non-target probes will also be falsified, but with a different intent. Target and non-target classes will be balanced—each will have equal numbers of probes.

The EvCharacterizationNode must be present in the EvidenceGraph and there must be a child consistency check node containing the intent being scored in that node's AnalyticScope. The LLR score of that consistency check will be used for scoring. Just as with target probes, the same nodes must be present. The expectation here is that the consistency check with relevant scope will have a low LLR score for non-target probes (i.e., "not intent X").

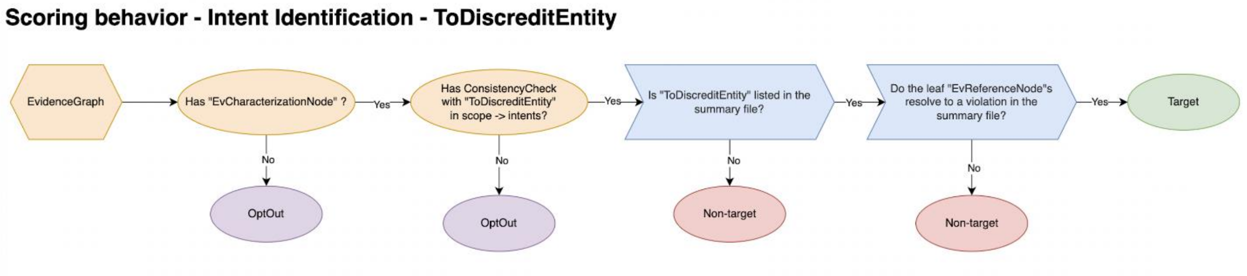
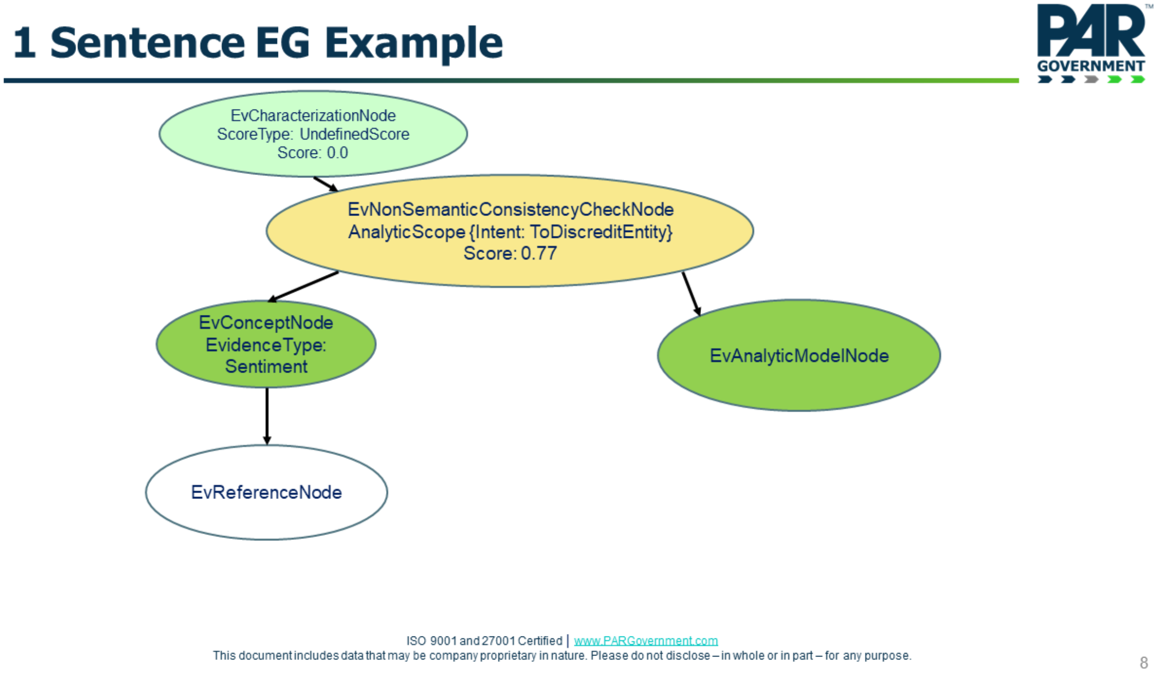

Scoring behavior - Intent Identification - ToDiscreditEntity
EvidenceGraph
Has "EvCharacterizationNode" ?
Yes
Has ConsistencyCheck with "ToDiscreditEntity" in scope -> intents?
Yes
Is "ToDiscreditEntity" listed in the summary file?
Yes
Do the leaf "EvReferenceNode"s resolve to a violation in the summary file?
Yes
Target
No
OptOut
No
OptOut
No
Non-target
No
Non-target

Figure 6. Example Evidence Graph Structure for Task 3.2.2

As part of each probe, analytics will be expected to read in a partial Evidence Graph (EG) containing an already completed EvCharacterizationNode. Analytics should output an as results an EvNonSemanticConsistencyCheckNode pointing to the relevant input from the EG. This should include an intent matching the competition target intent. An example of this EG structure for subtask *1b–Call to Action* is shown in Figure 7 below.



1 Sentence EG Example

EvCharacterizationNode
ScoreType: UndefinedScore
Score: 0.0

EvNonSemanticConsistencyCheckNode
AnalyticScope {Intent: ToDiscreditEntity}
Score: 0.77

EvConceptNode
EvidenceType:
Sentiment

EvAnalyticModelNode

EvReferenceNode

ISO 9001 and 27001 Certified | www.PARGovernment.com
This document includes data that may be company proprietary in nature. Please do not disclose – in whole or in part – for any purpose.

8

Figure 7. Example Evidence Graph Structure for Task 3.2.1

Analytics will be scored based on the LLR scores contained within the EvNonSemanticConsistencyCheckNode for all probes within a competition. The LLR score should represent the algorithm's confidence in the probe containing the target intent. TA3 will use the established approach, described in Appendix A, for dynamically determining the LLR thresholds used for scoring each analytic within a competition. Since the LLR score is being taken at the consistency check instead of the task node for 3.2, if the ScoringEngine is unable to find a consistency check with the matching AnalyticScope, it is considered an opt-out. Confusion matrix values for a given threshold will be assigned to results as described in Table 9.

Table 9. Task 3.2.1 Confusion Matrix Value Assignment

System Response	**Discredit Entity**	**Not Discredit Entity**
ConsistencyCheck LLR above threshold	TP	FP
ConsistencyCheck LLR below threshold	FN	TN
System Response	**Call to Action**	**Not Call to Action**
ConsistencyCheck LLR above threshold	TP	FP
ConsistencyCheck LLR below threshold	FN	TN

Although performers will not be scored on their ability to characterize across manipulation techniques, modalities, and manipulation level, TA3 will track this information to answer Research Question 3.2.1.

EVALUATION 3.2.2 PROPAGANDA TACTIC IDENTIFICATION (C)

Research Question 3.2.2

With manipulation information provided, does the (2a) manipulation level, (2b) modality manipulated (text, image), or (2c) type of characterization impact performers’ success in identifying specific propaganda tactics in the manipulated MMA?

Task 3.2.2

For Task 3.2.2, TA1 performers will be given a falsified MMA, and must successfully identify the propaganda tactic that was used in creating the falsification. Again, TA1 performers will receive evidence graphs that include what was manipulated and where in the MMA. Task 2 will be broken down into four subtasks for the successful verification of each of the four manipulation tactics individually.

For Subtask 2a, TA1 performers must verify which probes are intended as *Hate Speech* among all available probes. For Subtask 2b, TA1 performers must verify which probes are intended to *Scapegoat* among all available probes. For Subtask 2c, TA1 performers must verify which probes are intended to *Bandwagon* among all available probes. For Subtask 2d, TA1 performers must verify which probes are intended as a *Dictat.* To ensure a balanced sample of probes between the targets and negative cases for each subtask, a stratified random sample of probes will be selected from the remaining classes. The sample will be stratified based on the parameters of the research question. That is, the negative cases will be sampled equally from the other manipulation tactic classes to ensure a balanced representation of positive to negative cases as well as manipulation level, modality manipulated, and type of characterization. The LLR threshold will be identified based off the down-selected sample of probes for the evaluation.

Just as for Task 1, Task 2 is constrained around the following parameters. Probes only comprise Military Capabilities related topics. MMA content is News Articles and contains only text and images. Manipulation tactics are confined to text and image manipulations. The intention is for TA1 analytics to reason about the tactic in the context of the article for the purpose of the intent. As these data comprise the same probes as Task 1, they will be documented to include the level of manipulation and the modalities. This information will serve to score Subtasks 3.2.2a-2d.

Task 3.2.2 Scoring

TA1 Performers will be scored on their ability to successfully characterize the type of characterization intended with the program goal of 80% bACC @ EER. For scoring purposes, each tactic subtask will be separated into its own competition in the SemaFor Gym and scored individually as a binary classification problem. Within a given competition, target probes will be those falsified with the specified tactic. Non-target probes will also be falsified, but with a different tactic (i.e., “not tactic X”). Each competition will have an equal number of target and non-target probes. Further, non-target probes for each competition will be equally distributed across the three non-target classes.

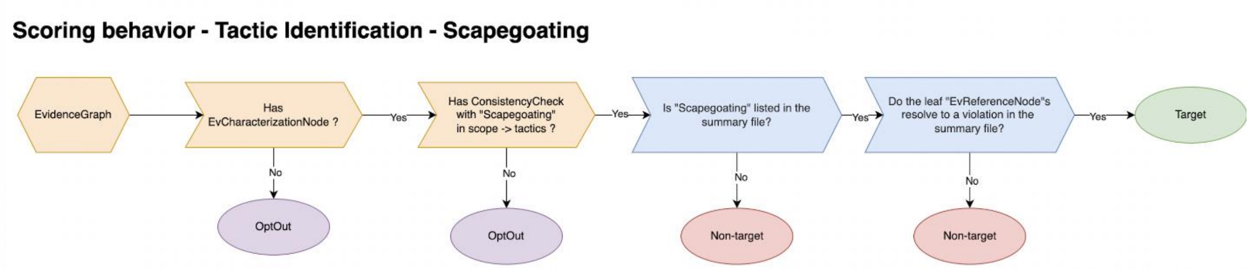
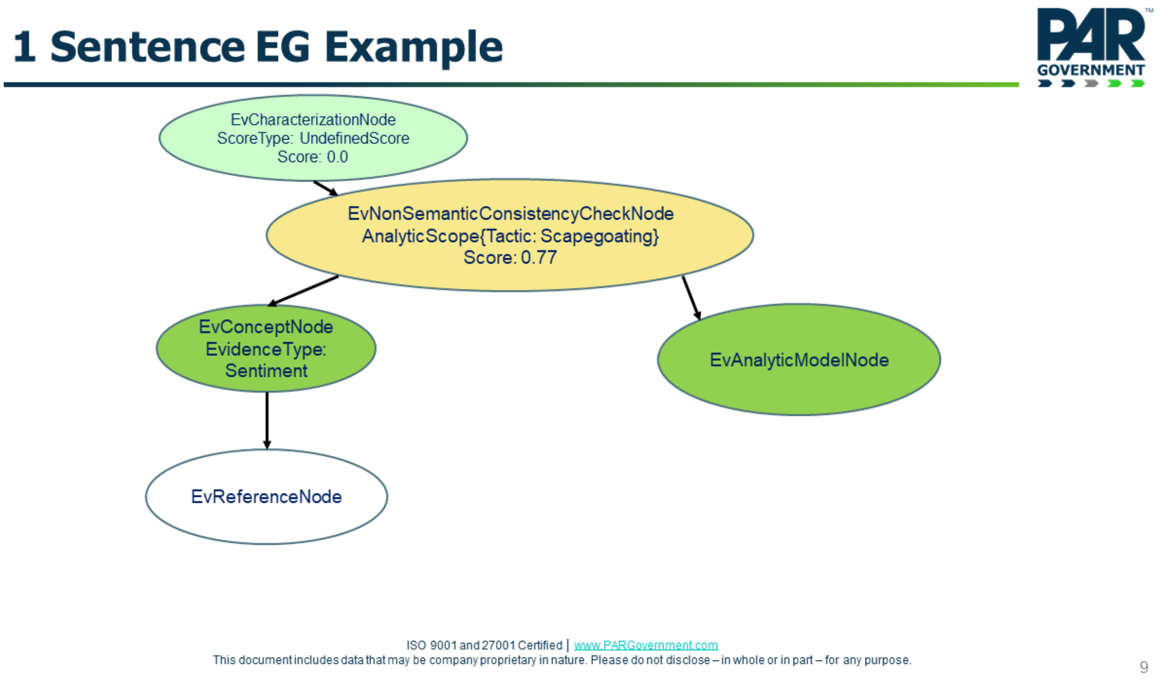

Scoring behavior - Tactic Identification - Scapegoating
EvidenceGraph
Has EvCharacterizationNode ?
Yes
Has ConsistencyCheck with "Scapegoating" in scope -> tactics ?
Yes
Is "Scapegoating" listed in the summary file?
Yes
Do the leaf "EvReferenceNode"s resolve to a violation in the summary file?
Yes
Target
No
OptOut
No
OptOut
No
Non-target
No
Non-target

Figure 8. Example Evidence Graph Structure for Task 3.2.2

As part of each probe, analytics will be expected to read in a partial Evidence Graph (EG) containing an already completed EvCharacterizationNode. The EvCharacterizationNode must be present in the EvidenceGraph and there must be a child consistency check node containing the intent being scored in that node's AnalyticScope. Just as with target probes, the same nodes must be present in non-targets. The expectation here is that the consistency check with relevant scope will have a low LLR score for non-target probes (i.e., "not Intent X"). Since the LLR score is being taken at the consistency check instead of the task node for 3.2, if the ScoringEngine is unable to find a consistency check with the matching AnalyticScope, it is considered an opt-out. The LLR score of that consistency check will be used for scoring.

Analytics should output as results an augmented version of that EG containing an EvNonSemanticConsistencyCheckNode with a tactic matching the competition target tactic. An example of this EG structure for subtask *2c–Bandwagoning* is shown in Figure 9 below.



1 Sentence EG Example

ISO 9001 and 27001 Certified | www.PARGovernment.com
This document includes data that may be company proprietary in nature. Please do not disclose – in whole or in part – for any purpose.

9

Figure 9. Example Evidence Graph Structure for Task 3.2.2

Analytics will be scored based on the LLR scores contained within the EvNonSemanticConsistencyCheckNode for all probes within a competition. TA3 will use the established approach, described in Appendix A, for dynamically determining the LLR thresholds used for scoring each analytic within a competition. Confusion matrix values for a given threshold will be assigned to results as described in Table 10.

Table 10. Task 3.2.2a Confusion Matrix Value Assignment

System Response	"Hate Speech"	"Not Hate Speech"
ConsistencyCheck LLR above threshold	TP	FP

System Response	"Hate Speech"	"Not Hate Speech"
ConsistencyCheck LLR below threshold	FN	TN
System Response	**"Bandwagoning"**	**"Not Bandwagoning"**
ConsistencyCheck LLR above threshold	TP	FP
ConsistencyCheck LLR below threshold	FN	TN
System Response	**"Scapegoating"**	**"Not Scapegoating"**
ConsistencyCheck LLR above threshold	TP	FP
ConsistencyCheck LLR below threshold	FN	TN
System Response	**"Dictat"**	**"Not Dictat"**
ConsistencyCheck LLR above threshold	TP	FP
ConsistencyCheck LLR below threshold	FN	TN

Although performers will not be scored on their ability to identify manipulation techniques across characterization intent, modalities, and manipulation level, TA3 will track this information to answer Research Question 3.2.2.

EVALUATION 3.2.3 SEMANTIC LABELING AND LOCALIZATION (D)

Research Question 3.2.3

Does the (3a) type of image manipulation or the (3b) category of semantic content manipulated impact performers' success in successfully localizing and semantically labeling?

Task 3.2.3

For Task 3, TA1 performers will be given a falsified image, and must localize and semantically label the falsification(s). For Subtask 3a, TA1 performers must verify where manipulations are implemented in the probes among all available probes. For Subtask 3b, TA1 performers must semantically label the manipulations that are implemented in the probes among all available probes. Note that in existing task designs, probes will only contain one instance of a concept to be localized. This task will be limited to only probes containing images, and only image localizations will be scored.

Just as for Tasks 3.2.1 and 3.2.2, Task 3.2.3 is constrained around the following parameters. Probes will only comprise Military Capabilities related topics. Probes will only contain images. A taxonomy of types of manipulated objects includes the following: crowds/people, explosions/fire, flag/emblem/national symbol, firearms, text/sign, vehicles. Task 3.2.3 will be broken down into two subtasks: semantic labeling and localization. *Figure 10* below illustrates an example of the probe data for Bandwagoning with the image manipulation highlighted and its associated mask data used for localization for comparison.

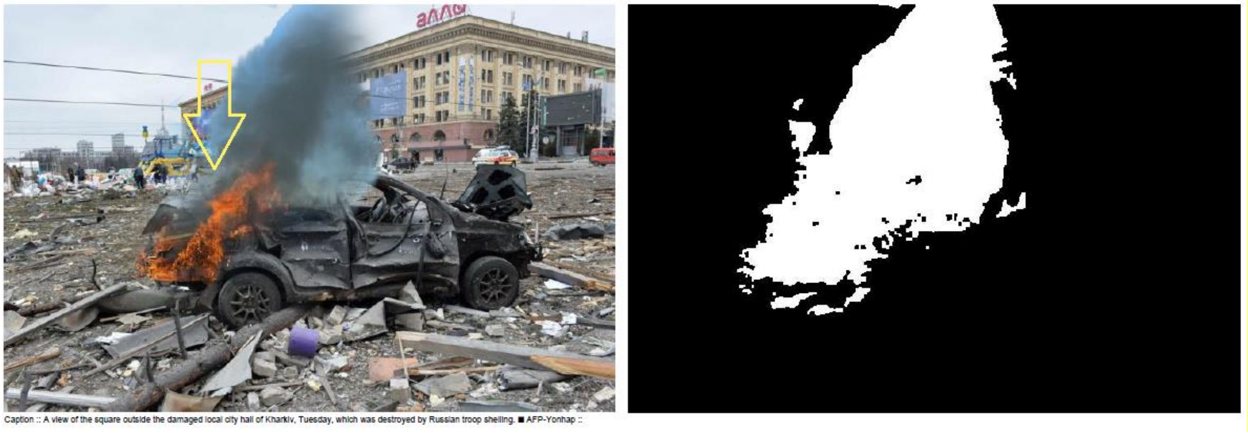

Caption :: A view of the square outside the damaged local city hall of Kharkiv, Tuesday, which was destroyed by Russian troop shelling. ■ AFP-Yonhap ::

Figure 10. Example Data Probe (Bandwagoning) with image manipulation and associated mask.

Task 3.2.3 Scoring

TA1 performers will be scored on their ability to successfully localize and semantically label the manipulation implemented in the MMA. The subtasks of localization and semantic labeling will not be combined or necessarily considered sequentially.

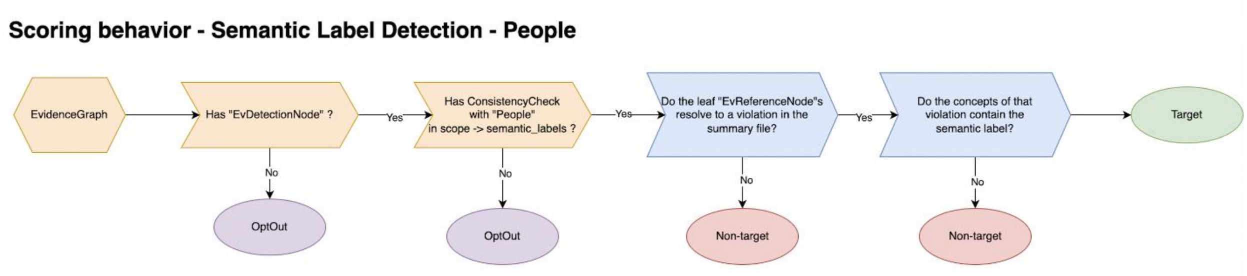
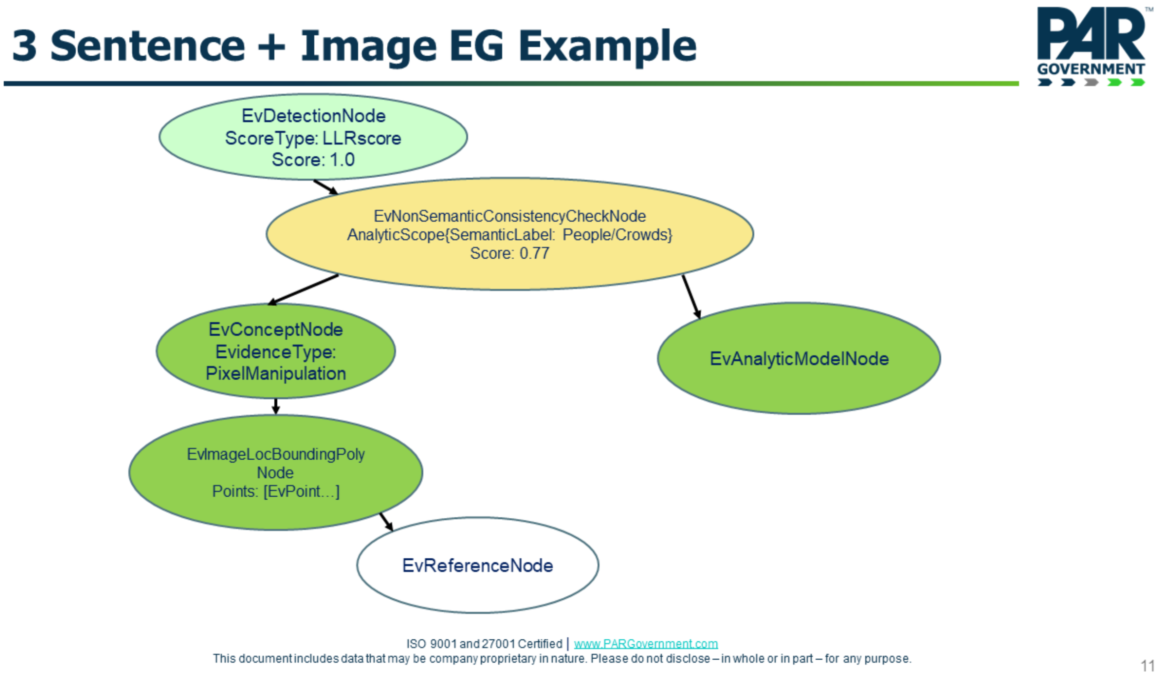

Scoring behavior - Semantic Label Detection - People
EvidenceGraph
Has "EvDetectionNode" ?
Yes
Has ConsistencyCheck with "People" in scope -> semantic_labels ?
Yes
Do the leaf "EvReferenceNode"s resolve to a violation in the summary file?
Yes
Do the concepts of that violation contain the semantic label?
Target
No
OptOut
No
OptOut
No
Non-target
No
Non-target

Figure 11. Example Scoring Behavior Flowchart for Task 3.2.3a

Each semantic label will be separated into its own competition in the GYM and scored individually as a binary classification problem. The positive class (target probes) will be probes falsified with “Semantic label X”. The negative class (non-target probes) will be probes falsified with “Not semantic label X”. The **EvDetectionNode** must be present in the EvidenceGraph and there must be a child **ConsistencyCheckNode** containing the semantic label being scored in that node's AnalyticScope. The LLR score of that consistency check will be used for scoring. Just as with target probes, the same nodes must be present for non-targets. The expectation here is that the consistency check with relevant scope will have a low LLR score for non-target probes.

As part of each probe, analytics will be expected to read in a partial Evidence Graph (EG) containing an already completed EvDetectionNode. Analytics should output as results an EvReferenceNode pointing to the relevant input EG consistency check. This includes containing at least one EvNonSemanticConsistencyCheckNode with a semantic label matching one of the labels in the task taxonomy, as well as an EvImageLocBoundingPolyNode containing a localization of the corresponding object. An example of this EG structure is shown in Figure 12 below.



3 Sentence + Image EG Example

EvDetectionNode
ScoreType: LLRscore
Score: 1.0

EvNonSemanticConsistencyCheckNode
AnalyticScope{SemanticLabel: People/Crowds}
Score: 0.77

EvConceptNode
EvidenceType:
PixelManipulation

EvAnalyticModelNode

EvImageLocBoundingPoly
Node
Points: [EvPoint…]

EvReferenceNode

ISO 9001 and 27001 Certified | www.PARGovernment.com
This document includes data that may be company proprietary in nature. Please do not disclose – in whole or in part – for any purpose.

11

Figure 12. Example Evidence Graph Structure for Task 3.2.2

Subtask 3.2.3.1-6

The ScoringEngine will produce a subset of results for each semantic label present in the EvidenceGraph produced by an analytic. For example, if an EvidenceGraph contains analysis for Crowds/People and Explosions/Fire, they will appear as two subsets of scores—one against Crowds/People, and the other against Explosions/Fire. For scoring purposes, each semantic label will be scored individually as a binary classification problem. For each subset of results, target probes will be those with the specified semantic label, while non-target probes will be without the specified semantic label. All probes will be falsified in some way. Probes for each will be equally distributed across the semantic classes.

Analytics will be scored for each label based on the LLR scores contained within all EvNonSemanticConsistencyCheckNodes with that label. The LLR score should represent the algorithm's confidence in the probe containing the target semantic label. TA3 will use the established approach, described in Appendix A, for dynamically determining the LLR thresholds used for scoring each analytic within a competition. Since the LLR score is being taken at the consistency check instead of the task node for 3.2, if the ScoringEngine is unable to find either the EvDetectionNode node or consistency check with the matching AnalyticScope, it is considered an opt-out as it is not score-able. Confusion matrix values for a given threshold will be assigned to results following the example shown in Table 11.

Table 11. Task 3.2.3.1-6 Value Assignment for Semantic Label Subscore

System Response	**“People”**	**“Not People”**
ConsistencyCheck LLR above threshold	TP	FP
ConsistencyCheck LLR below threshold	FN	TN
System Response	**“Fire”**	**Not “Fire”**
ConsistencyCheck LLR above threshold	TP	FP
ConsistencyCheck LLR below threshold	FN	TN
System Response	**“Symbol”**	**“Not Symbol”**
ConsistencyCheck LLR above threshold	TP	FP
ConsistencyCheck LLR below threshold	FN	TN
System Response	**“Firearms”**	**“Not Firearms”**
ConsistencyCheck LLR above threshold	TP	FP
ConsistencyCheck LLR below threshold	FN	TN
System Response	**“Sign”**	**“Not Sign”**
ConsistencyCheck LLR above threshold	TP	FP
ConsistencyCheck LLR below threshold	FN	TN
System Response	**“Vehicles”**	**“Not Vehicles”**
ConsistencyCheck LLR above threshold	TP	FP
ConsistencyCheck LLR below threshold	FN	TN

Subtask 3.2.3.7

This task will be limited to only probes containing images, and only image localizations will be scored. Analytics are expected to localize the manipulated pixels and report that localization in the form of an EvImageLocBoundingPolyNode in the EG. A **localization IoU (Intersection over Union)** will be computed for each probe against the corresponding ground truth localization. The flowchart for scoring this task is shown below in Figure 13.

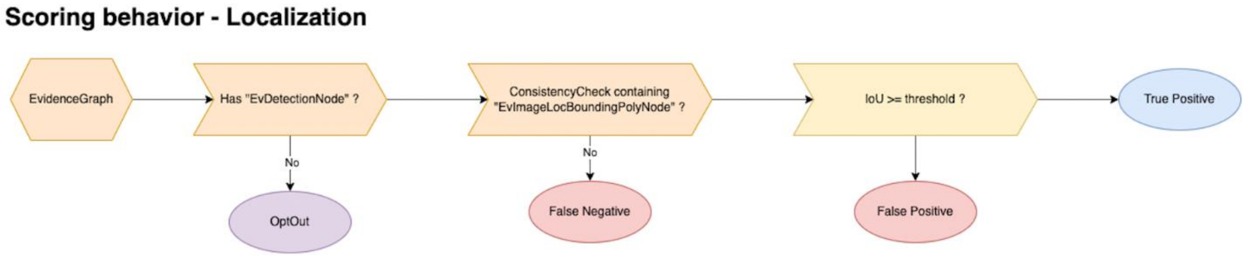


Figure 13. Example Scoring Behavior Flowchart for Task 3.2.3.7

Localization will be scored by comparing the Intersection over Union (IoU) of the localizations on the EvImageLocBoundingPolyNode of the outputted EG and summary to some threshold t. A probe is considered a target if IoU >= t. The following IoU thresholds will be used in evaluation:

- 0.0001
- 0.1
- 0.2
- 0.4

Each IoU threshold is used to bin the analytic's responses into the following categories:

- True Positive (correct detection) = IoU >= t
- False Positive (incorrect detection) = IoU < t
- False Negative = no localization provided

The following metrics are computed for each IoU threshold:

- Probability of Detection (#TP / All responses)
- False Alarm Rate (#FP / All responses)
- False Negative Rate (#FN / All responses) (constant across thresholds)

It is expected that as the IoU threshold increases, the p(D) will decrease and the FAR will increase. The FNR will remain constant.

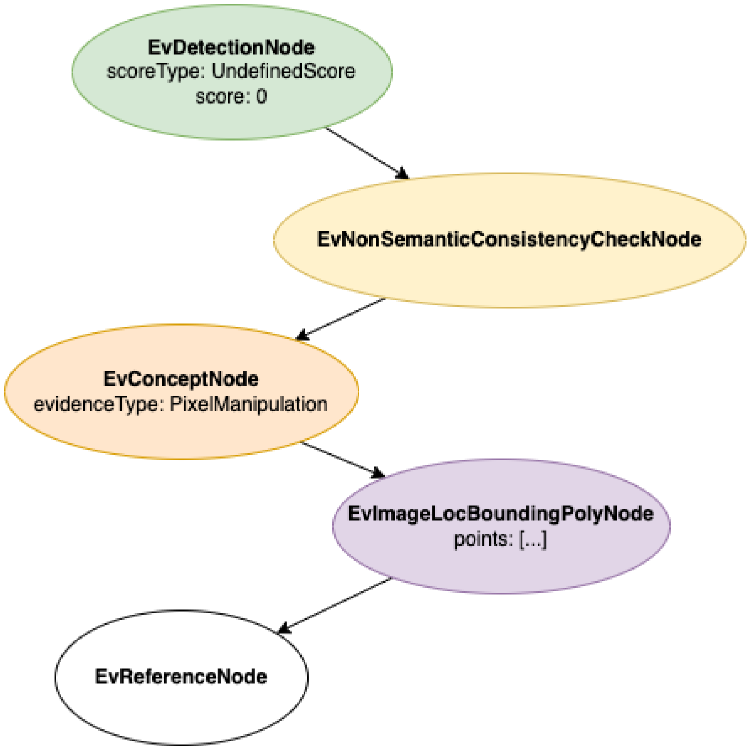

EvDetectionNode
scoreType: UndefinedScore
score: 0
EvNonSemanticConsistencyCheckNode
EvConceptNode
evidenceType: PixelManipulation
EvImageLocBoundingPolyNode
points: [...]
EvReferenceNode

Figure 14. Example Evidence Graph Structure for Task 3.2.3.7

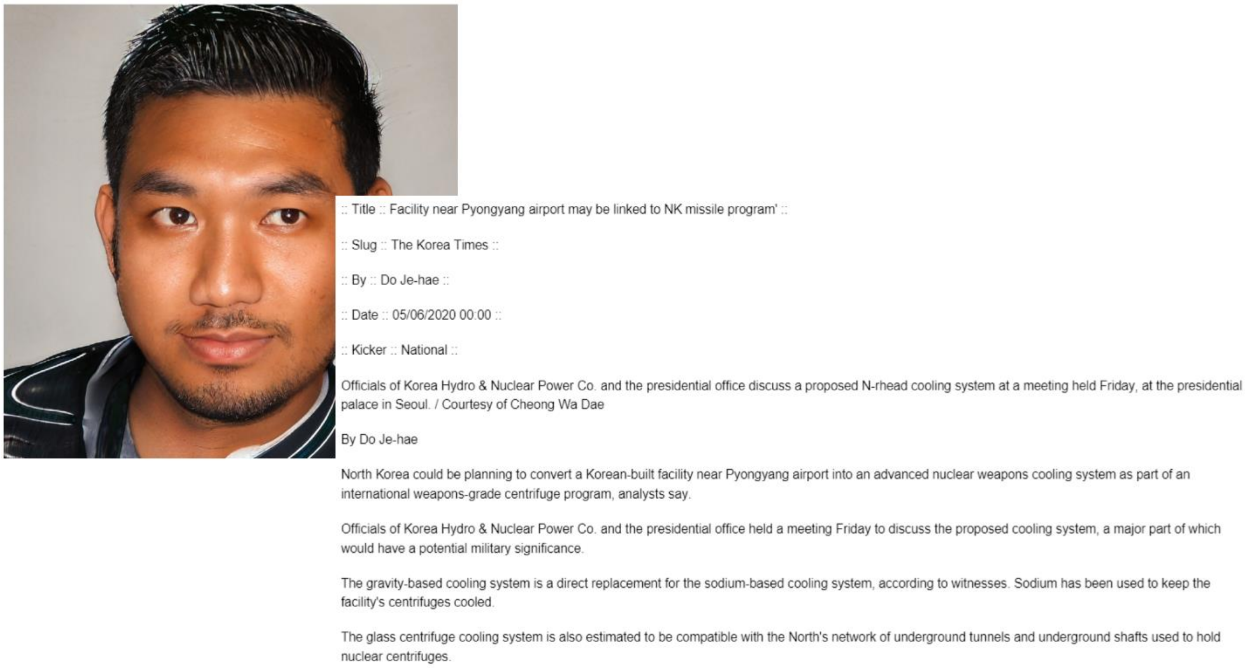
EVALUATION 3.2.4 SYNTHETIC MEDIA ATTRIBUTION (A)

Research Question 3.2.4

With no manipulation information provided, is the successful identification of the generator (e.g., for text, GPT2, GPT-j-6B, and GROVER; for images, StyleGAN2, StyleGAN3, taming-transformers, latent-diffusion, LSGM, and CLIP-guided-diffusion) more or less difficult due to (4a) modality manipulated (text, image) or (4b) which generator is used?

Task 3.2.4 Synthetic Media Attribution – Text & Images

For Task 3.2.4, TA1 performers will be given a falsified text or image asset and must identify of the generator or tool that was used to create the falsification. Initial discussion of the synthetic media tools includes the following proposed generators: Three different generators will be used for text generation (GPT2, GPT-j-6B, and GROVER) and six different generators will be used for image generation (StyleGAN2, StyleGAN3, taming-transformers, latent-diffusion, LSGM, and CLIP-guided-diffusion). *Figure 15. Synthetically generated image and text probe* illustrate two examples of synthetically generated media produced by StyleGAN3 and GROVER, respectively.



:: Title :: Facility near Pyongyang airport may be linked to NK missile program' ::

:: Slug :: The Korea Times ::

:: By :: Do Je-hae ::

:: Date :: 05/06/2020 00:00 ::

:: Kicker :: National ::

Officials of Korea Hydro & Nuclear Power Co. and the presidential office discuss a proposed N-rhead cooling system at a meeting held Friday, at the presidential palace in Seoul. / Courtesy of Cheong Wa Dae

By Do Je-hae

North Korea could be planning to convert a Korean-built facility near Pyongyang airport into an advanced nuclear weapons cooling system as part of an international weapons-grade centrifuge program, analysts say.

Officials of Korea Hydro & Nuclear Power Co. and the presidential office held a meeting Friday to discuss the proposed cooling system, a major part of which would have a potential military significance.

The gravity-based cooling system is a direct replacement for the sodium-based cooling system, according to witnesses. Sodium has been used to keep the facility's centrifuges cooled.

The glass centrifuge cooling system is also estimated to be compatible with the North's network of underground tunnels and underground shafts used to hold nuclear centrifuges.

Figure 15. Synthetically generated image and text probes

Task 3.2.4 will be broken down into three subtasks for text probes (3.2.4a - 3.2.4c) and six subtasks for image probes (3.2.4d - 3.2.4i) based on the type of generator used to create the asset. For this task, text and image modalities will be independent as the data probes will contain *only* generated text or images and not integrated into a full MMA. Below we list the specific generator types, repository links, and topics.

Table 12. Generator Types, References, and Topics.

Generator	Link	Topic
Latent-diffusion	https://github.com/CompVis/latent-diffusion	Military Vehicles
Image-LSGM	https://github.com/NVlabs/LSGM	Faces
Guided-diffusion	https://github.com/openai/guided-diffusion	Military Vehicles
StyleGAN3	https://github.com/NVlabs/stylegan3	Faces
Taming-transformers	https://github.com/CompVis/taming-transformers	Faces
StyleGAN2	https://github.com/NVlabs/stylegan2-ada-pytorch	Military Vehicles
Grover	https://github.com/rowanz/grover	N/A
GPT2	https://github.com/openai/gpt-2	N/A
GPT-j-6B	https://huggingface.co/EleutherAI/gpt-j-6B	N/A

To ensure a balanced sample of probes between the targets and negative cases for the text generation subtask, a stratified random sample of probes will be selected from the remaining classes of text generators (i.e., among GPT2, GPT-j-6B, and GROVER). For a balanced sample of probes between the targets and negative cases for the image generation subtask, a stratified random sample of probes will be selected from the remaining classes of image generators (i.e., among StyleGAN2, StyleGAN3, taming-transformers, latent-diffusion, LSGM, and CLIP-guided-diffusion). The LLR threshold will be identified based off the down-selected sample of probes for the evaluation.

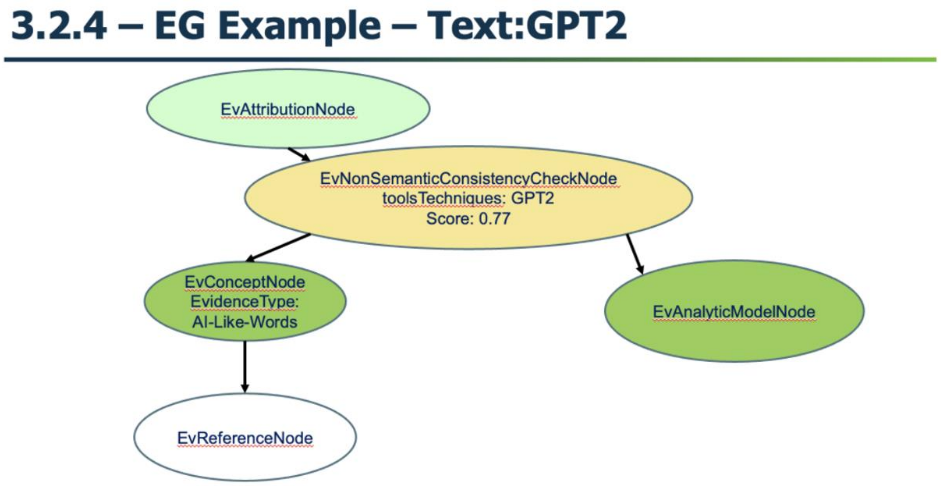
Future evaluations will continue to build these assets into a full MMA, and then add additional layers of human manipulation. See *Table 13. Data Generation Plan for Task 3.2.4* for the data generation plan and asset distribution. Note that for this task, it is not specific to the topic of Military Capabilities and includes generators that focus on faces for example.

Table 13. Data Generation Plan for Task 3.2.4

Text Generators	Text Assets	Image Generators	Image Assets
GPT-2	120	StyleGAN2	120
GPT-j-6B	120	StyleGAN3	120
Grover	120	taming-transformers	120
Total	*360*	latent-diffusion	120
		LSGM	120
		CLIP-guided-diffusion	120
		Total	*720*

Task 3.2.4 Scoring

This task will be scored against the program goal of 85% pD @ 0.08 FAR across probes. TA3 will use this performance information to answer Research Question 3.2.4a and b. For scoring purposes, each subtask will be separated into its own competition in the SemaFor Gym and scored individually as a binary classification problem. Within a given competition, target probes will be those falsified using the specified generator. Non-target probes will also be falsified but come from a different generator. Each competition will have an equal number of target and non-target probes. Further, non-target probes for each competition will be equally distributed across all other generator types.



3.2.4 – EG Example – Text:GPT2

Figure 16. Example Evidence Graph Structure for Task 3.2.4, applicable to all text generator subtasks

As part of each probe, analytics will be expected to read in a partial Evidence Graph (EG) containing an already completed EvAttributionNode. The EvAttributionNode must be present in the EvidenceGraph and there must be a child ConsistencyCheckNode containing the generator being scored in the toolsTechniques of that node's AnalyticScope. Just as with target probes, the same nodes must be present. The expectation here is that the consistency check with relevant scope will have a negative LLR score for non-target probes. The LLR score of that consistency check will be used for scoring. Analytics should output as results an augmented version of that EG containing an EvNonSemanticConsistencyCheckNode with a toolsTechniques value matching the competition target tactic. Since the LLR score is being taken at the consistency check instead of the task node for 3.2, if the ScoringEngine is unable to find either the EvAttributionNode node or consistency check with the matching AnalyticScope, it is considered an opt-out as it is not score-able. An example of this EG structure for text generator subtasks is shown in Figure 16. Figure 17 shows an example EG for image generator subtasks.

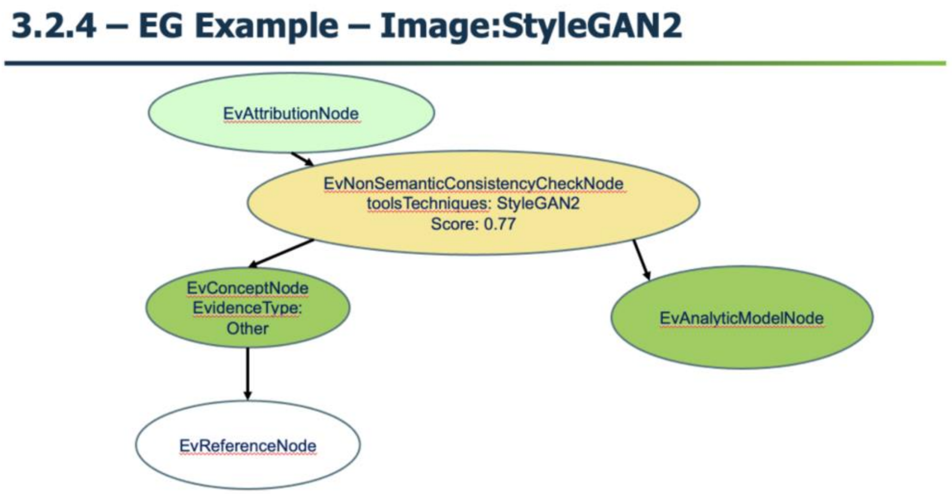

3.2.4 – EG Example – Image:StyleGAN2
EvAttributionNode
EvNonSemanticConsistencyCheckNode
toolsTechniques: StyleGAN2
Score: 0.77
EvConceptNode
EvidenceType:
Other
EvAnalyticModelNode
EvReferenceNode

Figure 17. Example Evidence Graph Structure for Task 3.2.4, applicable to all image generator subtasks

Analytics will be scored based on the LLR scores contained within the EvNonSemanticConsistencyCheckNode for all probes within a competition. The LLR score should represent the algorithm's confidence in the probe being attributed to the target generator, where positive scores are for correctly attributing the generator and negative scores for non-target generators. TA3 will use the established approach, described in Appendix A, for dynamically determining the LLR thresholds used for scoring each analytic within a competition. Confusion matrix values for a given threshold will be assigned to results as described in Table 14.

Table 14 Task 3.2.4 Confusion Matrix for Result Value Assignment and Generator Enumerations

System Response	Target Generator	Not Target Generator
ConsistencyCheck LLR above threshold	TP	FP
ConsistencyCheck LLR below threshold	FN	TN

Generator Enumerations		
Modality	**Generator**	**Enumeration**
Text	GPT-j-6B	GPT-j-6B

Generator Enumerations		
Modality	**Generator**	**Enumeration**
Text	GROVER	GROVER
Text	GPT2	GPT2
Image	StyleGAN2	StyleGAN2
Image	StyleGAN3	StyleGAN3
Image	Taming-transformers	Taming-transformers
Image	Latent-diffusion	Latent-diffusion
Image	LSGM	LSGM
Image	guided-diffusion	guided-diffusion

SEMAFOR EVALUATION 3.3 DESIGN

This design section details the various definitions, tasks, and conditions for Evaluation 3.3. Evaluation 3.3 will include 15 sub-tasks in total, nested under 4 primary tasks. This will comprise 6 detection, 7 attribution, and 2 characterization tasks. The four high level tasks include:

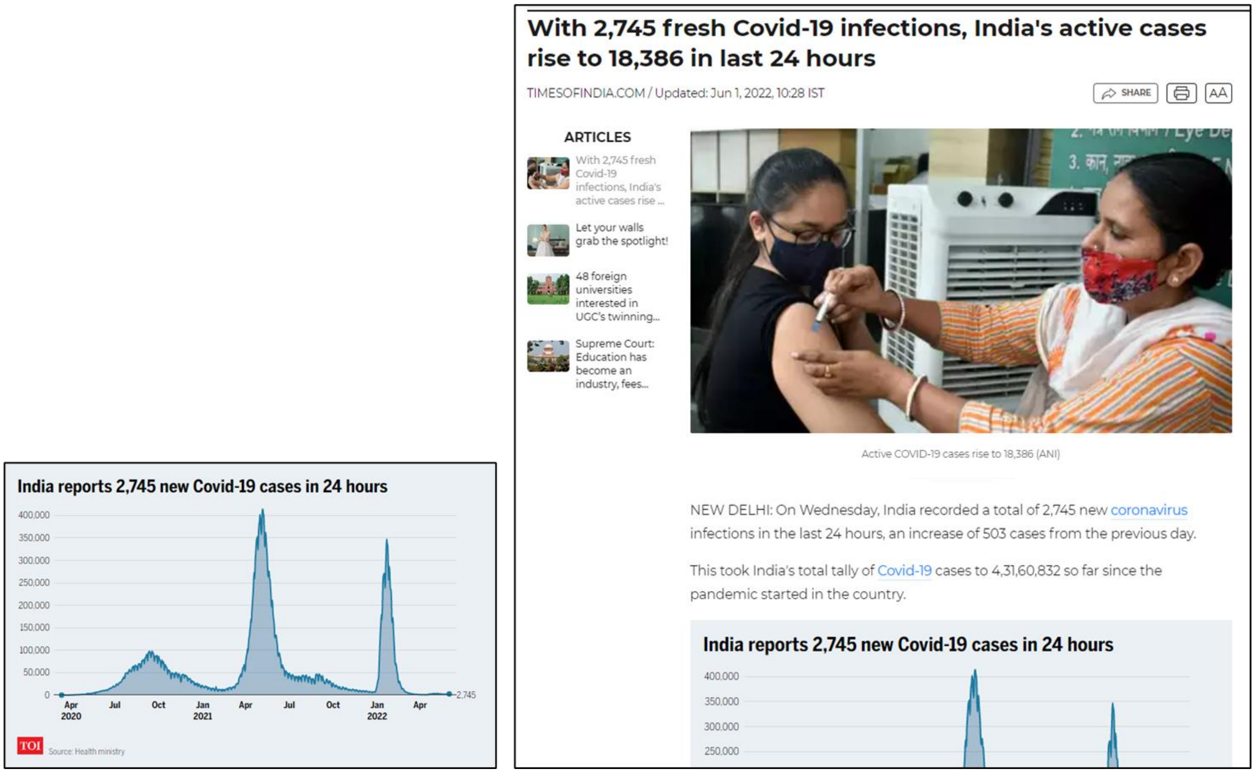
- **3.3.1: Deepfake Videos (D, A)**
- **3.3.2: Synthetic Audio (D, A)**
- **3.3.3: Synthetic Media – Text (D, A)**
- **3.3.4: Multi-Modal Inconsistencies – News Articles with Technical Information (D, C)**

Evaluation Task 3.3 will introduce two new characterization intents—specific to the topic of COVID—*Minimization* and *Appeal to Fear*.

Relevant Evaluation 3.3. Definitions

Technical Information

Evaluation 3.3 will introduce news articles that contain detailed technical information. For Evaluation 3.3 and future evaluations using this material, technical information is defined as news articles that contain detailed information that is referenced as fact. Information will be verifiable between assets within a single multi-modal asset. Assets include the article body, headline, or image. That is, the focus of tasks using technical information articles will *not* be on fact checking but in identifying inconsistencies within the MMA (see more information on inconsistencies in *Task 3.3.4*). See Figure 18. below for an example of a technical information article (original article linked here).



With 2,745 fresh Covid-19 infections, India's active cases rise to 18,386 in last 24 hours

TIMESOFINDIA.COM / Updated: Jun 1, 2022, 10:28 IST

ARTICLES

With 2,745 fresh Covid-19 infections, India's active cases rise ...

Let your walls grab the spotlight!

48 foreign universities interested in UGC's twinning...

Supreme Court: Education has become an industry, fees...

Active COVID-19 cases rise to 18,386 (ANI)

NEW DELHI: On Wednesday, India recorded a total of 2,745 new coronavirus infections in the last 24 hours, an increase of 503 cases from the previous day.

This took India's total tally of Covid-19 cases to 4,31,60,832 so far since the pandemic started in the country.

Figure 18. Example Article for Probe Generation representing technical information.

Images within these news articles will be in a form that conveys information such as a graph, chart, or infographic. The information in the image will directly correspond to information stated within the article headline or body. An inconsistency would occur if the information in the image does *not* correspond to information stated within the article headline or body. This may include an infographic or heading stating one statistic while the article body references a different value. For example, an infographic stating a vaccine is 35% effective and the article body stating that the vaccine is 95% effective. Images may be original to the article, original but manipulated, or completely fabricated (a chart is created and inserted into the article).

When inconsistent, images may be original to the article, original but manipulated, or completely fabricated (e.g., a chart is created and inserted into the article). This task requires performers to identify the inconsistency, not simply what was manipulated. This will provide for probes that contain inconsistencies across modalities. Technical Information for Evaluation 3.3 will only apply to *Task 3.3.4 MMA Inconsistencies*.

COVID Propaganda Technique: Minimization

Minimization refers to manipulating a technical news article with the intent to reduce, moderate, or reject the negative elements of the original article. Minimization is the opposite of exaggeration in that it moderates or downplays the extremity of a real situation. This type of deception involves denial coupled with rationalization in situations where a more complete or overt denial is implausible. This tactic is an intentional manipulation of the media asset to minimize concerns, reduce fear, or otherwise dispel apprehension that would be considered reasonable about an issue. For example, this may include manipulating an article heading to state that long-haul COVID only lasts a few weeks when the original article body text states that long-haul COVID tends to last a few months. This manipulation technique may be used across headings, images, or both in multi-media assets. Body text will not be manipulated. Manipulations may include fabricating information.

COVID Propaganda Technique: Appeal to Fear

Appealing to Fear refers to manipulating a technical news article with the intent to exaggerate or amplify the negative elements of the original article. This tactic is an intentional manipulation of the media asset to induce fear or raise concerns when it is not necessary or reasonable. This tactic aims to instill anxieties, concerns, and panic in the typical reader. This approach may also be used to influence public perception by disseminating negative, dubious, or false information designed to undermine the credibility of a particular belief. For example, this may include manipulating an infographic to state that a vaccine is 35% effective whereas the original article body text states that the vaccine is 95% effective. This manipulation technique may be used across headings, images, or both in multi-media assets. Body text will not be manipulated. Manipulations may include fabricating information. Recall, that this task is not about fact checking claims, but examining inconsistencies about information within an MMA. It is out of scope of the effort to examine if the manipulated data is true or false.

EVALUATION 3.3.1 DEEPFAKE VIDEO (D, A)

Evaluation *Task 3.3.1* will be broken down into five subtasks: 3.3.1.1-3.3.1.5. For Evaluation *Task 3.3.1.1* through *Task 3.3.1.5*, the topic will be constrained to media related to the Russia/Ukraine conflict. Given its sensitivity, deepfake and puppeteering videos and their associated audio will *not* be shared outside of the program.

Two generators to be used to produce these videos, DeepFaceLab (https://github.com/iperov/DeepFaceLab) and Wav2Lip https://github.com/Rudrabha/Wav2Lip). All new deepfakes will be of one of the two POIs. Videos may contain other other pristine faces, but each video will only have a single deepfaked face or head. Videos will range from one to three minutes.

Videos will be in the MP4 file format and may be either landscape or portrait orientation. All videos will be color (RGB). Two different resolutions will be produced: low with 640 x 360 resolution and high with 1920 x 1080. Similarly, two frame rates will be produced: low with 15fps and high with 30fps. Finally, videos may have additional post-production editing done such as background replacement. These edits will be done in Adobe Premiere Pro or Adobe After Effects.

All deepfakes will have audio. This may be the real POI's voice or voice of a different person. Additionally, voices may be altered using software such as Adobe Premiere Pro. However, there will be no synthetically generated voices in this round of deepfakes. The final audio format will be acc audio and embedded within the associated MMA videos. Deepfake videos may be in three different languages: English, Russian or Ukrainian. Some tasks will use pristine videos/audio of the targets POIs, and are noted as such below. Subtasks will include detection and attribution elements. The generated audio tasks are examined separately in *3.3.2.1*, and *3.3.2.2* and *3.3.2.3* which are focused on generated audio detection and generator attribution.

In the below table, we summarize the tasks as they will appear in the gym competitions.

Table 15. Evaluation 3.3.1 Tasks and Descriptions.

Deepfake Video, Puppeteering Tasks			
Task	**Title**	**Description**	**Focus Area**
3.3.1.1	*Detect Deepfake Video*	Detection whether a video is real or a deepfake of a person. Targets will be deepfakes and non-targets will be pristine videos. The dataset will use all deepfakes from previous evals in addition to the newly created deepfakes for this eval.	Detection
3.3.1.2	*Detect POI - Putin*	Using specific traits or biometrics, determine if a video is authentic video of the specific POI. The non-targets will be pristine videos of the specific POI and deepfake videos of previous POI from Evaluation 2.	Detection
3.3.1.3	*Detect POI - Zelenskyy*	Using specific traits or biometrics, determine if a video is authentic video of the specific POI. The non-targets will be pristine videos of the specific POI. and deepfake videos of previous POI from Evaluation 2.	Detection
3.3.1.4	*Generator Attribution (Wav2Lip)*	Given a set of manipulated videos, determine if the generator used is Wav2Lip. Non-targets will leverage other generator types.	Attribution
3.3.1.5	*Generator Attribution (DFL)*	Given a set of manipulated videos, determine if the generator used is DFL. Non-targets will leverage other generator types.	Attribution

Probe Distribution for Task 3.3.1

For this task, a single data set will be created comprising deepfakes, puppeteering, and non-target videos. Some tasks will use pristine videos/audio of the targets POIs. Non-targets for the deepfake detection task will include pristine videos. The non-targets for the detect POI tasks will include pristine videos of the specific POI, as well as deepfake videos of the previous POI from Evaluation 2, that would serve as "imposter" videos. Some videos will have additional post-processing such as green screen replacement. This distribution is described in Table 16 and further elaborated on in the task definitions. This dataset provides representative stimuli to support generating ROC curves for the research questions. This dataset will be used to answer Research Questions 3.3.1a-c.

Table 16. Data Generation Plan for Task 3.3.1

TOPIC: Military Conflict	Probe Generation Table for Deepfakes, Puppeteering Tasks				Total
Person-of- Interest	**Vladimir Putin**		**Volodymyr Zelenskyy**		
Type of Manipulation	*Call to Action*	*Discredit Entity*	*Call to Action*	*Discredit Entity*	
Generator Attribution (Wav2Lip)	15	15	15	15	60
Generator Attribution (DFL)	15	15	15	15	60
Total Manipulated Probes	*30*	*30*	*30*	*30*	*120*

Note. An additional 60 pristine probes per POI will be used in this task as non-targets.

Research Question 3.3.1a-c

3.3.1a. Are TA1 performers better able to detect deepfakes versus puppeteered videos, and is this moderated by the degree or type of video manipulation added?

3.3.1b. Are TA1 performers better able to successfully detect POI between Putin and Zelenskyy, and is this moderated by the availability of training data? Does the ability to detect POI perform better than the ability to detect deepfake and puppeteered falsification broadly?

3.3.1c. Is the successful identification of the generator more or less difficult depending on which generator is used (i.e., Face2Vid and DeepFaceLab), and is this moderated by the degree or type of video manipulation added?

Tasks 3.3.1.1-3.3.1.5

MMAs in this task will include deepfakes, puppeteered, and pristine data and will be specific to two persons-of-interest: Vladimir Putin and Volodymyr Zelenskyy. These probes have been developed using the latest version of DeepFaceLab (DFL) for deepfakes and Face2Vid for puppeteered videos. Training models may vary but will be restricted to models available in latest DFL version. It is most likely that the same model will be used. The final format of the deepfake videos provided to TA1 performers will be .mp4/acc audio. In addition to the deepfaked and puppeteered POIs, some videos may have backgrounds altered or replaced. Some videos will have additional post-processing such as green screen replacement. Final videos may come directly from DFL, Face2Vid, or from video editing software such Adobe Premiere Pro or Adobe After Effects. Deepfake videos may have English, Russian, or Ukrainian language included. All deepfakes will have audio that may be the real POI’s voice or a voice of a different person. Voices may be altered using software such as Adobe Premiere Pro. There will be no synthetically generated voices in this set of data for these tasks.

Task 3.3.1 will be broken down into five subtasks for the successful detection, POI attribution, and generator attribution of the MMA probe in the ways outlined below. These will include different subsets of pristine or manipulated videos depending on the nature of the task.

For *Task 3.3.1.1 Detect Deepfake Video*, TA1 performers will be provided with pristine or deepfaked videos, and must successfully detect whether the video includes falsification. Thus, the task is to tell if the person in a video is the real POI or a deepfake/puppeteer. The dataset will consist of videos of real people or videos where the POI has been deepfaked/puppeteered. TA1 performers may use techniques related to video artifacts, compression, or the like to determine if a video is a deepfaked, puppeteered, or real. Performers do not need to attribute how the video was generated in this task.

For *Tasks 3.3.1.2 and 3.3.1.3 Detect POI – Putin and Zelenskyy*, TA1 performers will be provided with pristine or manipulated MMAs (deepfakes or puppeteered). The dataset will consist of real videos of the POI and deepfake/puppeteering videos of the POI. TA1 performers must successfully attribute the person-of-interest portrayed by determining whether the video contains either Vladimir Putin or Volodymyr Zelenskyy. TA1 performers may use techniques to attribute or classify specific traits of an individual, such as use visual elements or biometrics (e.g., facial features, physical movement) to identify if the person in the video is actually the POI or a deepfake of the POI. See Figure 19. Example Probe for Task 3.3.2: POI – PutinFigure 19 below for an example probe for Task 3.3.2. In this example, performers would be expected to determine if the video is an authentic video of Vladamir Putin rather than a deepfaked video of Putin.

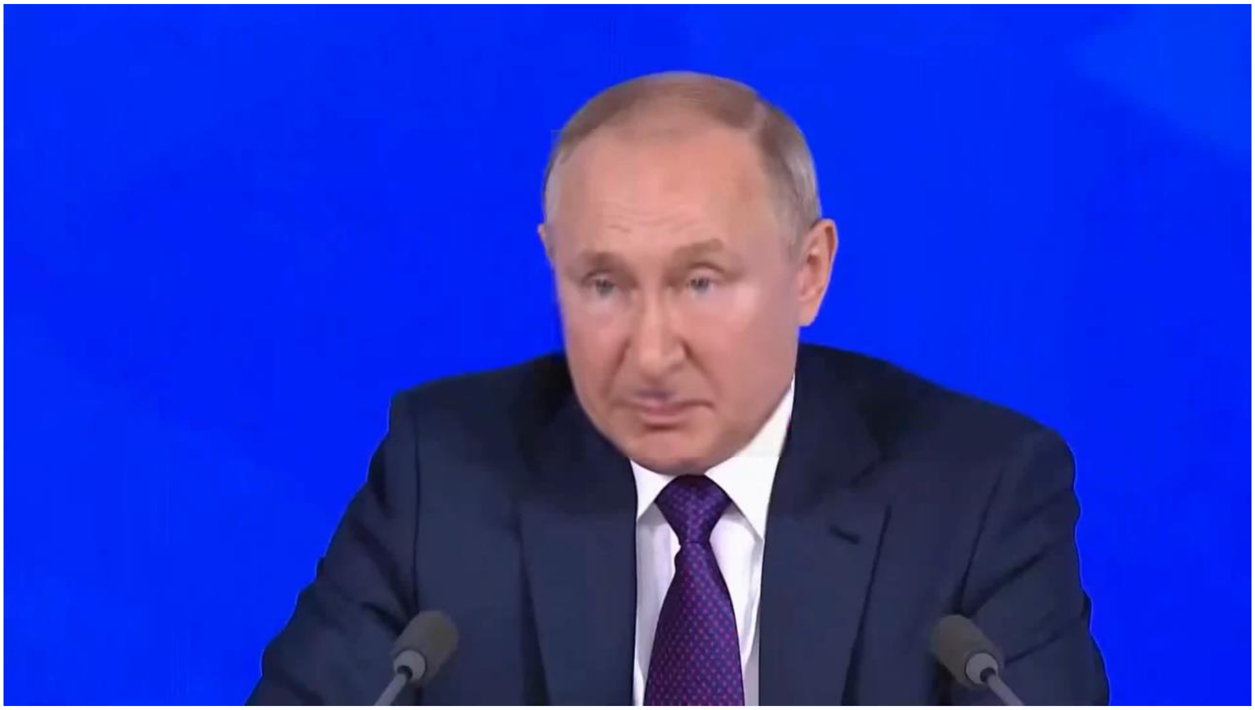


Figure 19. Example Probe for Task 3.3.2: POI – Putin

See Figure 20 below for an example probe for Task 3.3.3. Similarly, in this example, performers would be expected to determine if the video is an authentic video of Volodomyr Zelenskyy rather than a deepfaked video of Zelenskyy. For Task 3.3.2 and 3.3.3, the particular generator that was used need not be identified.

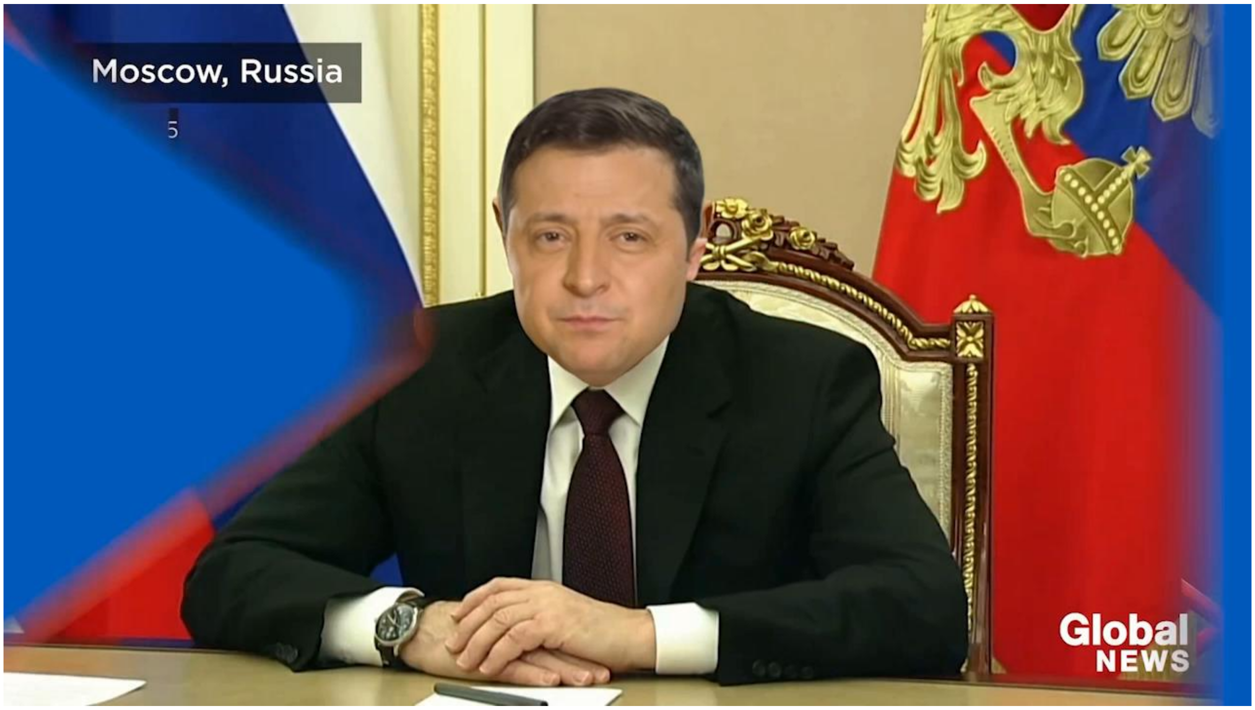


Figure 20. Example Probe for Task 3.3.1: POI – Zelenskyy.

For *Tasks 3.3.1.4 and Tasks 3.3.1.5 Generator Attribution – Wav2Lip, DeepFaceLab,* TA1 performers will be provided with only manipulated MMAs (deepfakes or puppeteered). The dataset will consist of all deepfake or puppeteering videos created with DeepFaceLab or Face2Vid. TA1 performers must successfully attribute which generator was used to create the deepfaked or puppeteered video. The task is to determine if the video was created using DeepFaceLab or Face2Vid. TA1 performers may use techniques to attribute what software/model was used to create the deepfake or puppeteering video. For DeepFaceLab, two different models may be used. See Figure 21 below for an example probe for Task 3.3.4. In this example, performers would be expected to correctly attribute that the generator used was Wav2Lip.

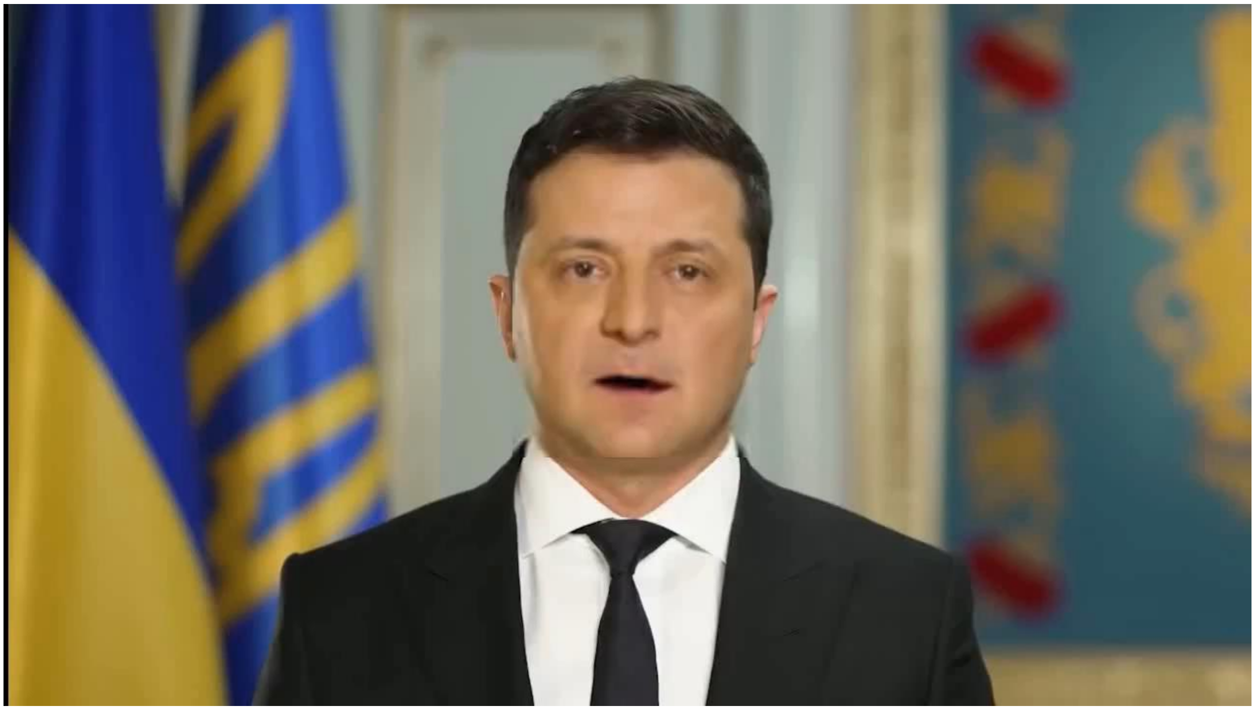


Figure 21. Example Probe for Task 3.3.1: POI – Zelenskyy Wav2Lip.

See Figure 22 below for an example probe for Task 3.3.1. In this example, performers would be expected to correctly attribute that the generator was DeepFaceLab.

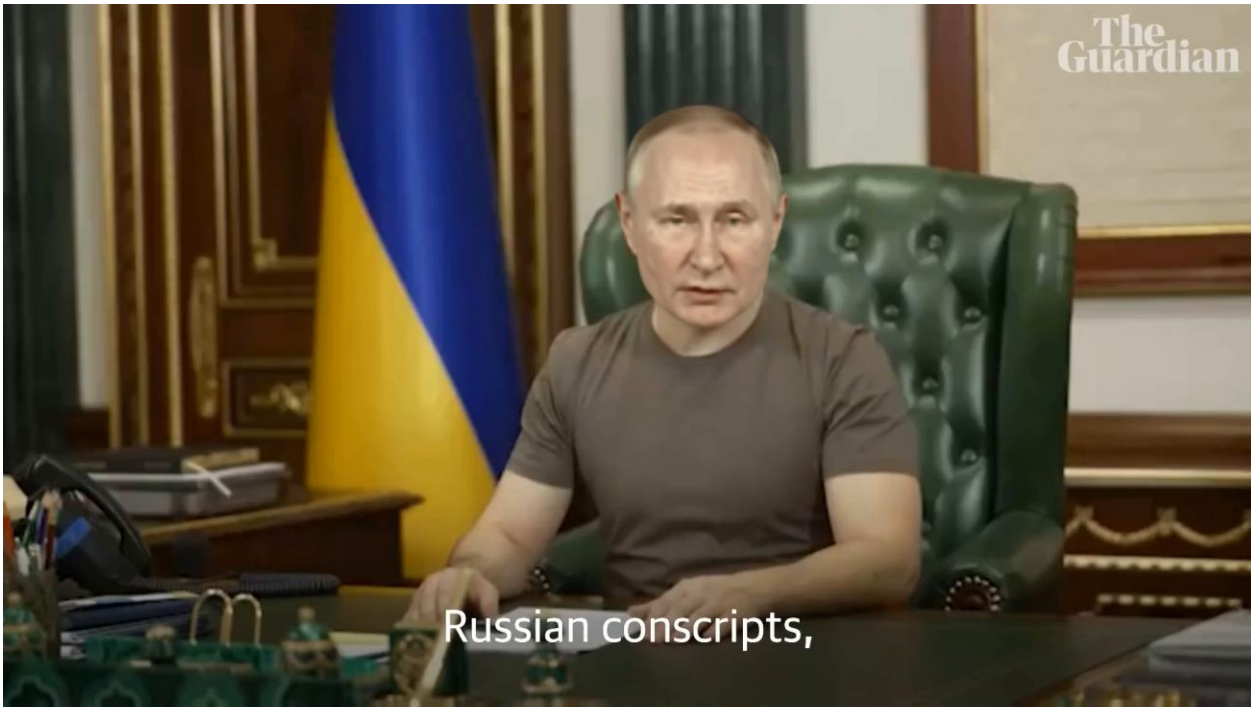


Figure 22. Example Probe for Task 3.3.1: POI – Putin DeepFaceLab.

Task 3.3.1 Scoring

TA1 Performers will be scored on their ability to successfully detect deepfake and puppeteering videos, detect the two POIs, and attribute the generator used with the program goal of 85% p(D) at 8% FAR. For scoring purposes, each intent subtask will be separated into its own competition in the SemaFor Gym and scored individually as a binary classification problem across *Tasks 3.3.1.1*-*3.3.1.5*. The number of probes across target and non-target classes will be balanced in each task.

EVALUATION 3.3.2 SYNTHETIC AUDIO (D, A)

For *Tasks 3.3.2.1*-*3.3.2.3*, TA1 performers will be provided with generated audio files. The task is to determine if an audio sample is generated or real using techniques related to audio artifacts. Two generators will be used to create synthetic audio: Either generator may be used for POI voice cloning. CorentinJ / Real-Time-Voice-Cloning (https://github.com/CorentinJ/Real-Time-Voice-Cloning) and vlomme / Multi-Tacotron-Voice-Cloning (https://github.com/vlomme/Multi-Tacotron-Voice-Cloning). The dataset will consist of synthetically generated audio and non-targets taken from pristine audio. For Task 3.3.2, only English language will be used.

All synthetic audio will be of one speaker, completely synthetic. Audio will range from 1 minute to 3 minutes. Audio will be standalone files of completely generated audio. Final format will be aac audio in MP3 format. Two different bit rate ranges will be produced: low with 8 kbps and high with 1411 kbps. Three bit depth ranges will be produced: 16, 24, 32. Two sampling rate ranges will be produced: low at 8 kHz and high at 48 kHz. Finally, audio will not have any additional editing done to the files that are generated. Audio will not be POI-specific for Evaluation 3.3.

In the below table, we summarize the tasks as they will appear in the gym competitions.

Table 17. Evaluation 3.3.2 Tasks and Descriptions.

Audio Tasks			
Task	**Title**	**Description**	**Focus Area**
3.3.2.1	*Generated Audio Detection*	Detect if the audio is generated/synthetic or a real voice. Non-targets will be pristine audio.	Detection
3.3.2.2	*Generator Attribution (CorentinJ / Real-Time-Voice-Cloning)*	Given a set of generated audio, determine if the generator used is CorentinJ / Real-Time-Voice-Cloning. Non-targets will leverage the other generator type	Attribution
3.3.2.3	*Generator Attribution (https://github.com/vlomme/Multi-Tacotron-Voice-Cloning)*	Given a set of generated audio, determine if the generator used is vlomme/Multi-Tacotron-Voice-Cloning. Non-targets will leverage the other generator type	Attribution

Research Question 3.3.2

3.3.2 Are performers better able to detect synthetically generated audio, and is this moderated by the type of generator used?

Tasks 3.3.2.1-3.3.2.3

For *Task 3.3.2.1 Generated Audio Detection*, TA1 performers will be provided with pristine or artificially generated acc audio and must successfully detect whether the audio is synthetically generated audio. TA1 performers may use techniques related to audio artifacts, compression, or the like to determine if an audio sample is generated or real. Non-targets will be pristine audio.

For *Task 3.3.2.2 Attribute Synthetic Audio Generator (CorentinJ / Real-Time-Voice-Cloning)*, TA1 performers will be provided with only synthetically generated audio. TA1 performers must successfully determine that CorentinJ / Real-Time-Voice-Cloning generated the audio sample. Non-targets will be generated using the alternate audio generator.

For *Task 3.3.2.3 Attribute Synthetic Audio Generator (vlomme / Multi-Tacotron-Voice-Cloning)*,TA1 performers will be provided with only synthetically generated audio. TA1 performers must successfully determine that vlomme/Multi-Tacotron-Voice-Cloning generated the audio sample. Non-targets will be generated using the alternate audio generator.

Task 3.3.2 Scoring

Task 3.3.2.1 Generated Audio Detection will be scored against the program goal of 85% pD at 0.08 FAR across probes. TA3 will use this performance information across the three tasks to answer Research Question 3.3.3.

For *Tasks 3.3.2.2 and 3.3.2.3*, TA1 Performers will be scored on their ability to successfully attribute the generator that was used to produce the generated audio, with the program goal of 85% pD at 0.08 FAR across probes. Although there are only two generators, for scoring purposes each task will be separated into its own competition in the SemaFor Gym and scored individually as a binary classification problem. Within a given competition, target probes will be those that were generated using one of the two audio generators. Non-target probes will also be generated, but by using the alternate audio generator. Target and non-target classes will be balanced—each will have equal numbers of probes.

EVALUATION 3.3.3 SYNTHETIC MEDIA (D, A)

For *Tasks 3.3.3.1-3.3.3.4*, TA1 performers will focus on detecting synthetic media in both text and images, as well as attributing to the appropriate generator. This task builds on the 3.2.4 task by applying a further level of human manipulation to increase the complexity of the probes for both text and images. This task will then examine how the additional level of manipulation influences the ability to both detect and attribute synthetic media following the baseline analysis that was the focus of *Task 3.2.4*.

For the synthetic text task, generated news articles will be produced and then be edited to correct grammar, punctuation, spelling and formatting. Articles will focus on the topic of military capabilities. The two primary tasks in this evaluation will be detecting generated from pristine real world news article text (D) and attributing the generator that was used to generate text (A). Three generators will be used to produce the synthetic text GROVER, GPT2, and GPT-J-6B.

In the below table, we summarize the tasks as they will appear in the gym competitions.

Table 18. Evaluation 3.3.3 Tasks and Descriptions.

3.3.3 Synthetic Media Task Definitions			
Task	**Title**	**Description**	**Focus Area**
3.3.3.1 **3.3.3.1a** **3.3.3.1b**	*Synthetic Text Detection* Synthetic Text Detection – Original Generated Article Synthetic Text Detection – Edited Generated Article	Detect whether the text is synthetically generated or pristine. Targets will consist of an equal number of generator types and articles with be both original output from the generator and edited version of the original output. Non-targets will be pristine text articles.	Detection
3.3.3.2	*Synthetic Text Generator Attribution (Grover)*	Detect whether the text that is generated is from a particular generator. Non-targets will be synthetic text generated from an alternative generator.	Attribution
3.3.3.3	*Synthetic Text Generator Attribution (GPT2)*	Detect whether the image that is generated is from a particular generator. Non-targets will be synthetic text generated from an alternative generator.	Attribution
3.3.3.4	*Synthetic Text Generator Attribution (GPT-j-6B)*	Detect whether the image that is generated is from a particular generator. Non-targets will be synthetic text generated from an alternative generator.	Attribution

* This will be one competition that contains both the original generated article and the edited generated article.

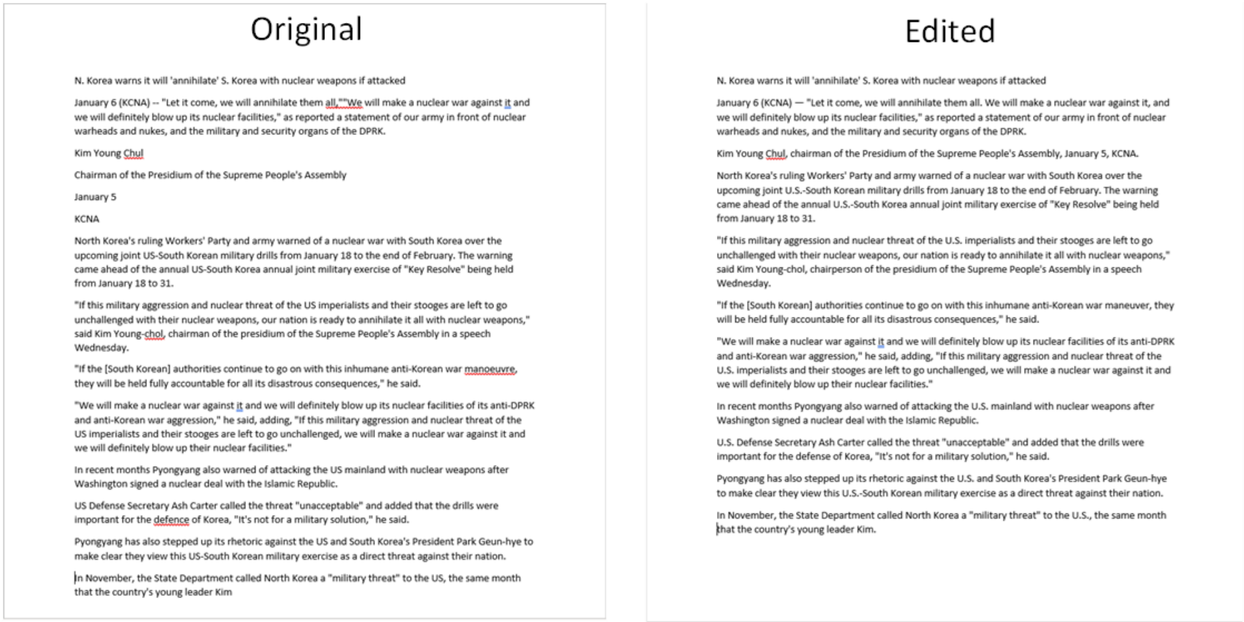
Research Questions 3.3.3

3.3.3a Is the successful detection of synthetic text more or less difficult when text has an additional layer of editing?

3.3.3b Is it more difficult to detect synthetic synthetic text based on the type of generator?

Tasks 3.3.3.1-3.3.3.4

Evaluation Task 3.3.3 will be a continuation of efforts from *Evaluation 3.2* on the attribution of synthetic media. For this task, a dataset of news articles will be created using the same three generators and the same articles produced for *Evaluation 3.2*. Topic(s) for this task will be the same as *Evaluation 3.2*. However, these articles will be edited by humans. For *Evaluation 3.3*, edits will be confined to grammar and punctuation corrections of generator-produced articles. Grammar and punctuation editing will be done manually using a tool such as MS Word. Figure 23 below provides an example of the original and edited documents that will be generated.



Original

N. Korea warns it will 'annihilate' S. Korea with nuclear weapons if attacked

January 6 (KCNA) -- "Let it come, we will annihilate them all,""We will make a nuclear war against it and we will definitely blow up its nuclear facilities," as reported a statement of our army in front of nuclear warheads and nukes, and the military and security organs of the DPRK.

Kim Young Chul

Chairman of the Presidium of the Supreme People's Assembly

January 5

KCNA

North Korea's ruling Workers' Party and army warned of a nuclear war with South Korea over the upcoming joint US-South Korean military drills from January 18 to the end of February. The warning came ahead of the annual US-South Korea annual joint military exercise of "Key Resolve" being held from January 18 to 31.

"If this military aggression and nuclear threat of the US imperialists and their stooges are left to go unchallenged with their nuclear weapons, our nation is ready to annihilate it all with nuclear weapons," said Kim Young-chol, chairman of the presidium of the Supreme People's Assembly in a speech Wednesday.

"If the [South Korean] authorities continue to go on with this inhumane anti-Korean war manoeuvre, they will be held fully accountable for all its disastrous consequences," he said.

"We will make a nuclear war against it and we will definitely blow up its nuclear facilities of its anti-DPRK and anti-Korean war aggression," he said, adding, "If this military aggression and nuclear threat of the US imperialists and their stooges are left to go unchallenged, we will make a nuclear war against it and we will definitely blow up their nuclear facilities."

In recent months Pyongyang also warned of attacking the US mainland with nuclear weapons after Washington signed a nuclear deal with the Islamic Republic.

US Defense Secretary Ash Carter called the threat "unacceptable" and added that the drills were important for the defence of Korea, "It's not for a military solution," he said.

Pyongyang has also stepped up its rhetoric against the US and South Korea's President Park Geun-hye to make clear they view this US-South Korean military exercise as a direct threat against their nation.

In November, the State Department called North Korea a "military threat" to the US, the same month that the country's young leader Kim

Edited

N. Korea warns it will 'annihilate' S. Korea with nuclear weapons if attacked

January 6 (KCNA) — "Let it come, we will annihilate them all. We will make a nuclear war against it, and we will definitely blow up its nuclear facilities," as reported a statement of our army in front of nuclear warheads and nukes, and the military and security organs of the DPRK.

Kim Young Chul, chairman of the Presidium of the Supreme People's Assembly, January 5, KCNA.

North Korea's ruling Workers' Party and army warned of a nuclear war with South Korea over the upcoming joint U.S.-South Korean military drills from January 18 to the end of February. The warning came ahead of the annual U.S.-South Korea annual joint military exercise of "Key Resolve" being held from January 18 to 31.

"If this military aggression and nuclear threat of the U.S. imperialists and their stooges are left to go unchallenged with their nuclear weapons, our nation is ready to annihilate it all with nuclear weapons," said Kim Young-chol, chairperson of the presidium of the Supreme People's Assembly in a speech Wednesday.

"If the [South Korean] authorities continue to go on with this inhumane anti-Korean war maneuver, they will be held fully accountable for all its disastrous consequences," he said.

"We will make a nuclear war against it and we will definitely blow up its nuclear facilities of its anti-DPRK and anti-Korean war aggression," he said, adding, "If this military aggression and nuclear threat of the U.S. imperialists and their stooges are left to go unchallenged, we will make a nuclear war against it and we will definitely blow up their nuclear facilities."

In recent months Pyongyang also warned of attacking the U.S. mainland with nuclear weapons after Washington signed a nuclear deal with the Islamic Republic.

U.S. Defense Secretary Ash Carter called the threat "unacceptable" and added that the drills were important for the defense of Korea, "It's not for a military solution," he said.

Pyongyang has also stepped up its rhetoric against the U.S. and South Korea's President Park Geun-hye to make clear they view this U.S.-South Korean military exercise as a direct threat against their nation.

In November, the State Department called North Korea a "military threat" to the U.S., the same month that the country's young leader Kim.

Figure 23. Example Probe for Task 3.3.3: Synthetic Text.

Table 19. Probe Distribution for Task 3.3.3 Synthetic Media

Topic: Military	**Total**
Level of Manipulation	**Text:** Grammar/ Punctuation Corrections
GROVER	40
GPT-2	40
GPT-j-6B	40
Totals	*120*

Task 3.3.3 Scoring

Tasks 3.3.3.1-4 will be scored against the program goal of 85% pD @ 0.08 FAR across probes. TA3 will use this performance information to answer Research Questions for 3.3.3. For scoring purposes, text subtasks will be separated into competitions in the SemaFor Gym and scored individually as a binary classification problem.

In detection *Task 3.3.3.1*, performers must detect whether the text is synthetically generated or pristine. Targets will consist of an equal number of generator types across text generators. Non-targets will be pristine text articles. Targets and non-targets will be balanced and representative across a subset of generators for each text sub-task.

Within attribution *Tasks 3.3.3.2-4*, target probes will be those falsified using the specified generator. Non-target probes will also be falsified but come from a representative subset of the remaining two alternative generators. Each competition will have an equal number of target and non-target probes.

EVALUATION 3.3.4 MULTI-MODAL ASSET INCONSISTENCIES – NEWS ARTICLES WITH TECHNICAL INFORMATION (D, C)

Probes for Evaluation Task 3.3.4 will comprise news articles that contain detailed technical information (see *Relevant Evaluation 3.3 Definitions above*). MMAs for *Evaluation Task 3.3.4* will comprise news articles focusing on technical information that is available within the news article body. Successful performance in these tasks will be defined by the ability to determine if the elements in the news article are consistent in meaning. Only the headline and image elements in the news article will be altered or manipulated. The text in the news article body will never be altered or manipulated. For example, analytics in this task will attempt to determine if the headline of a news article states/confers/suggests the same thing that the image in the news article conveys, or if the image in a news article conveys the same information as stated in the body of the article.

The topic for these probes will be COVID, and the technical articles will include discussion related to the virus and resulting pandemic. This includes but is not limited to topics of biology, policy, technology, and governmental responses to COVID over the course of the pandemic. That is, the topic of COVID is not confined simply to vaccinations. Again, fact checking is out of scope of this effort; the purpose is to identify *inconsistent technical information* within the same MMA.

Two new propaganda intents will be introduced in this task: *Appeal to Fear* and *Minimization*. Manipulations will be done to represent these intent and are defined previously in this document (see *Relevant Evaluation 3.3 Definitions above*). Only one intent will be present in each MMA. In *Evaluation Task 3.3.4 MMA Inconsistencies*, two propaganda intents related to generating inconsistencies will be implemented: *Appeal to Fear* and *Minimization.*

In the below table, we summarize the tasks as they will appear in the gym competitions.

Table 20. Evaluation 3.3.3 Tasks and Descriptions.

3.3.4 MMA Inconsistency Task Definitions			
Task	**Title**	**Description**	**Focus Area**
3.3.4.1	*Detect MMA Inconsistency*	Detect whether there is a semantic inconsistency between the headline, image, and body in an MMA.	Detection
3.3.4.1a	*Manipulated Headline vs. Original (True) Image*	Detect whether there is a semantic inconsistency between the headline and an image. Targets will be articles that have an inconsistency between the headline and the image. Non-targets will be pristine articles where these assets are consistent.	Detection
3.3.4.1b	*Manipulated Headline vs. Original (True) Body*	Detect whether there is a semantic inconsistency between the headline and the article body. Targets will be articles that have an inconsistency between the headline and the article body. Non-targets will be pristine articles where these assets are consistent.	Detection
3.3.4.1c	*Manipulated Image vs. Original (True) Body*	Detect whether there is a semantic inconsistency between an image and the article body. Targets will be articles that have an inconsistency between the image and the article body. Non-targets will be pristine articles where these assets are consistent.	Detection
3.3.4.1d	*Manipulated Headline and Manipulated Image vs. Original (True) Body*	Detect whether there is a semantic inconsistency between a headline and an image that are consistent with each other but are inconsistent with the information within the article body. Non-targets will be pristine articles where these assets are consistent.	Detection
3.3.4.2	*Characterize Inconsistency: Minimization (COVID)*	Characterize whether the manipulation was done with the propaganda intent of *minimization*. Non-targets will be MMAs of the opposite tactic.	Characterization
3.3.4.3	*Characterize Inconsistency: Appeal to Fear (COVID)*	Given a manipulated article, characterize whether the manipulation was done with the propaganda intent of *appealing to fear*. Non-targets will be MMAs of the opposite tactic.	Characterization

Probe Distribution for Task 3.3.4

Data for *Evaluation 3.3.4* will be constrained on the topic of COVID.

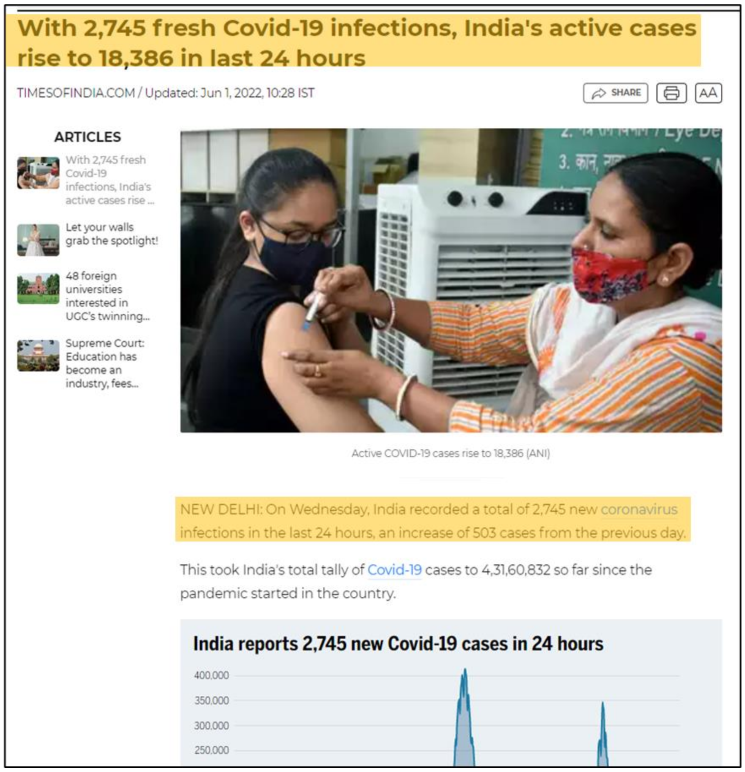
Table 21. Evaluation 3.3.4 Tasks and Descriptions.

Level of Manipulation	**Propaganda Intent**		**Total**
	COVID: Minimization	**COVID:** Appeal To Fear	
Headline and Image Inconsistency	30	30	60
Headline and Body Inconsistency	30	30	60
Image and Body Inconsistency	30	30	60
Headline/Image and Body Inconsistency	30	30	60
Total	120	120	*240*

Note. An additional 120 pristine probes will be used in this task as non-targets. These probes will be the original articles.

Not all articles will have technical information in the headline but they will always have technical information in the article body. Articles will be manipulated so that the technical information within the article body is inconsistent with technical information either in the headline or in an article image. The technical information in the article body will never be manipulated. Manipulations done to the headline or image will be directly inconsistent with the technical information in the article body.

Figure 24below provides an example inconsistent multi-media asset and its original. This article contains technical information regarding COVID infection numbers. The highlighted areas show where the technical information is, where manipulations are highlighted in red and original information in yellow.



With 2,745 fresh Covid-19 infections, India's active cases rise to 18,386 in last 24 hours

TIMESOFINDIA.COM / Updated: Jun 1, 2022, 10:28 IST

ARTICLES

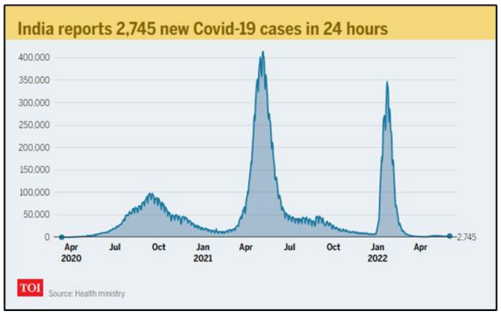
- With 2,745 fresh Covid-19 infections, India's active cases rise ...
- Let your walls grab the spotlight!
- 48 foreign universities interested in UGC's twinning...
- Supreme Court: Education has become an industry, fees...

Active COVID-19 cases rise to 18,386 (ANI)

NEW DELHI: On Wednesday, India recorded a total of 2,745 new coronavirus infections in the last 24 hours, an increase of 503 cases from the previous day.

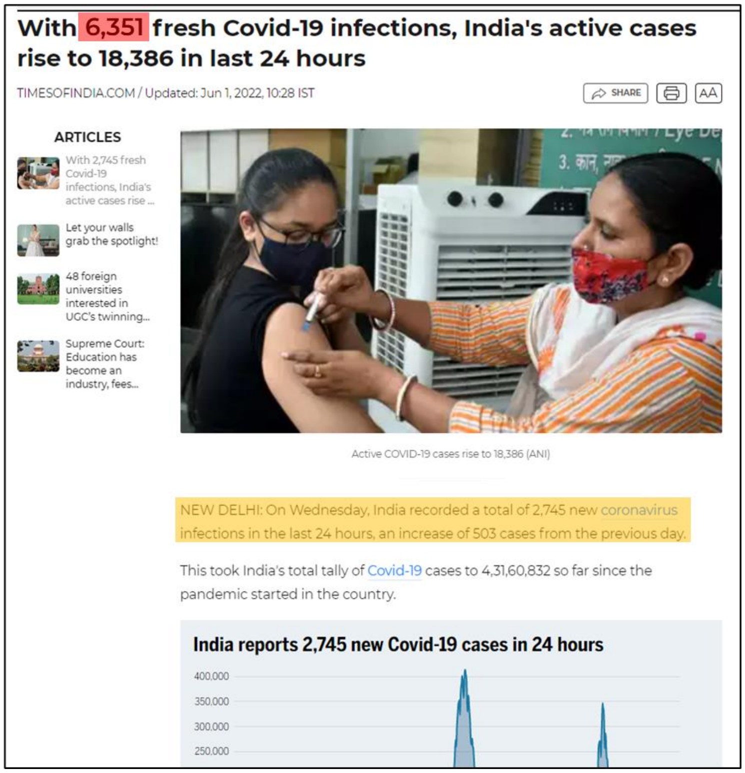
This took India's total tally of Covid-19 cases to 4,31,60,832 so far since the pandemic started in the country.

India reports 2,745 new Covid-19 cases in 24 hours



India reports 2,745 new Covid-19 cases in 24 hours

Source: Health ministry



With 6,351 fresh Covid-19 infections, India's active cases rise to 18,386 in last 24 hours

TIMESOFINDIA.COM / Updated: Jun 1, 2022, 10:28 IST

ARTICLES

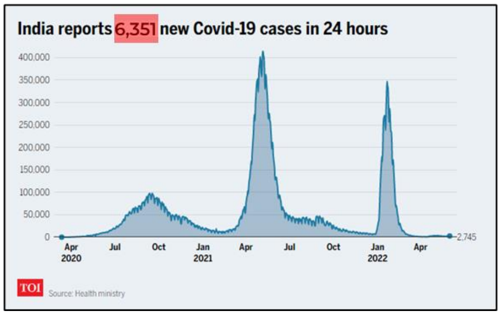
- With 2,745 fresh Covid-19 infections, India's active cases rise ...
- Let your walls grab the spotlight!
- 48 foreign universities interested in UGC's twinning...
- Supreme Court: Education has become an industry, fees...

Active COVID-19 cases rise to 18,386 (ANI)

NEW DELHI: On Wednesday, India recorded a total of 2,745 new coronavirus infections in the last 24 hours, an increase of 503 cases from the previous day.

This took India's total tally of Covid-19 cases to 4,31,60,832 so far since the pandemic started in the country.

India reports 2,745 new Covid-19 cases in 24 hours



India reports 6,351 new Covid-19 cases in 24 hours

Source: Health ministry

Figure 24. Example Probe for Task 3.3.4.1a Detect Inconsistency: Manipulated Headline vs. Original (True) Image.

Research Questions 3.3.4

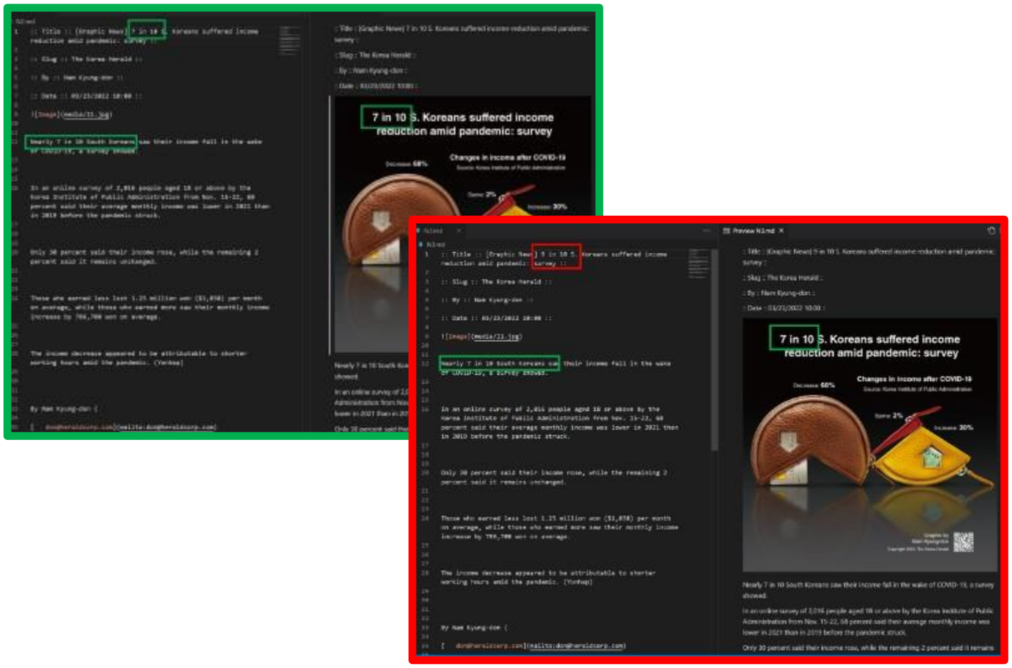
3.3.4 Does the manipulation modality affect the ability to identify an inconsistency? Does the combination of both text and image inconsistency with body make detection easier?

Tasks 3.3.4.1-3.3.4.3

For *Task 3.3.4.1 MMA Inconsistencies Detect Inconsistency: Inconsistencies between different article assets (Headline, Image and Body)*, TA1 performers will be provided with MMAs that are news articles containing technical information. They comprise of a headline, image and article body. The article body will *never* be manipulated and will contain the technical information to be considered true. Images and headlines will be manipulated to be inconsistent with the technical information in the article body. Manipulations to the headline will be text-based. Manipulation to images may be statistical (numbers), graphical, or text-based.

The task is to use the MMA to determine if the headline, image and body are consistent with each other. Targets will be articles that have an inconsistency; non-target will be pristine articles that the information is consistent. The details provided here will be true for all subtasks under this general task. Each subtask will focus on specific combinations of the inconsistent assets underlying each MMA. Analytics will be able to opt-out of probes that contain assets or media that they cannot or are not made to handle.

For *Task 3.3.4.1a Detect Inconsistency: Manipulated Headline vs. Original (True) Image*, TA1 Performers will be provided with a manipulated MMA (technical news article) and must detect whether there is a semantic inconsistency relating to technical information within the headline and an image. Targets will be articles that have an inconsistency between the headline and the image. For this sub-task, only headlines will be manipulated. Non-targets will be pristine articles with information that is consistent. Thus, in this case, the modality inconsistency is between short text and image assets.



```
N1.md
:: Title :: [Graphic News] 7 in 10 S. Koreans suffered income
reduction amid pandemic: survey ::

:: Slug :: The Korea Herald ::

:: By :: Nam Kyung-don ::

:: Date :: 03/23/2022 10:00 ::

![Image](media/11.jpg)

Nearly 7 in 10 South Koreans saw their income fall in the wake
of COVID-19, a survey showed.

In an online survey of 2,016 people aged 18 or above by the
Korea Institute of Public Administration from Nov. 15-22, 68
percent said their average monthly income was lower in 2021 than
in 2019 before the pandemic struck.

Only 30 percent said their income rose, while the remaining 2
percent said it remains unchanged.

Those who earned less lost 1.25 million won ($1,030) per month
on average, while those who earned more saw their monthly income
increase by 786,700 won on average.

The income decrease appeared to be attributable to shorter
working hours amid the pandemic. (Yonhap)

By Nam Kyung-don (
[   don@heraldcorp.com](mailto:don@heraldcorp.com)
```

: Title : [Graphic News] 7 in 10 S. Koreans suffered income reduction amid pandemic: survey :

: Slug : The Korea Herald :

: By : Nam Kyung-don :

: Date : 03/23/2022 10:00 :

7 in 10 S. Koreans suffered income reduction amid pandemic: survey
Changes in income after COVID-19
Source: Korea Institute of Public Administration
Decrease 68%
Same 2%
Increase 30%

Nearly 7 in 10 South Ko

showed.

In an online survey of 2,0
Administration from Nov
lower in 2021 than in 20

Only 30 percent said thei

```
N1.md
:: Title :: [Graphic News] 5 in 10 S. Koreans suffered income
reduction amid pandemic: survey ::

:: Slug :: The Korea Herald ::

:: By :: Nam Kyung-don ::

:: Date :: 03/23/2022 10:00 ::

![Image](media/11.jpg)

Nearly 7 in 10 South Koreans saw their income fall in the wake
of COVID-19, a survey showed.

In an online survey of 2,016 people aged 18 or above by the
Korea Institute of Public Administration from Nov. 15-22, 68
percent said their average monthly income was lower in 2021 than
in 2019 before the pandemic struck.

Only 30 percent said their income rose, while the remaining 2
percent said it remains unchanged.

Those who earned less lost 1.25 million won ($1,030) per month
on average, while those who earned more saw their monthly income
increase by 786,700 won on average.

The income decrease appeared to be attributable to shorter
working hours amid the pandemic. (Yonhap)

By Nam Kyung-don (
[   don@heraldcorp.com](mailto:don@heraldcorp.com)
```

: Title : [Graphic News] 5 in 10 S. Koreans suffered income reduction amid pandemic: survey :

: Slug : The Korea Herald :

: By : Nam Kyung-don :

: Date : 03/23/2022 10:00 :

7 in 10 S. Koreans suffered income reduction amid pandemic: survey
Changes in income after COVID-19
Source: Korea Institute of Public Administration
Decrease 68%
Same 2%
Increase 30%
Graphic by
Nam Kyung-don
Copyright 2022. The Korea Herald

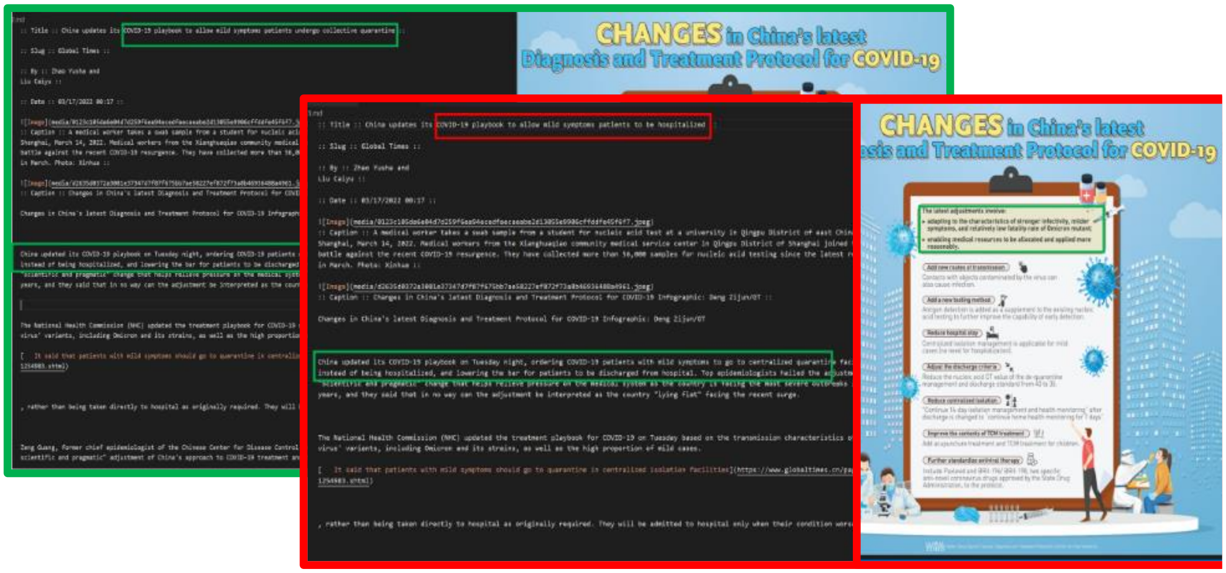
Nearly 7 in 10 South Koreans saw their income fall in the wake of COVID-19, a survey showed.

In an online survey of 2,016 people aged 18 or above by the Korea Institute of Public Administration from Nov. 15-22, 68 percent said their average monthly income was lower in 2021 than in 2019 before the pandemic struck.

Only 30 percent said their income rose, while the remaining 2 percent said it remains

Figure 25. Example Probe for Task 3.3.4.1a Detect Inconsistency: Manipulated Headline vs. Original (True) Image

For *Task 3.3.4.1b Detect Inconsistency - Manipulated Headline vs. Original (True) Body*, TA1 Performers will be provided with a manipulated MMA and must detect whether there is a semantic inconsistency relating to technical information within the headline and the article body. Accordingly, targets will be articles that have an inconsistency between the headline and the article body. For this sub-task, only headlines will be manipulated. Non-targets will be pristine articles with information that is consistent between headline and body. Thus, in this case, the modality inconsistency is between short text versus long text.



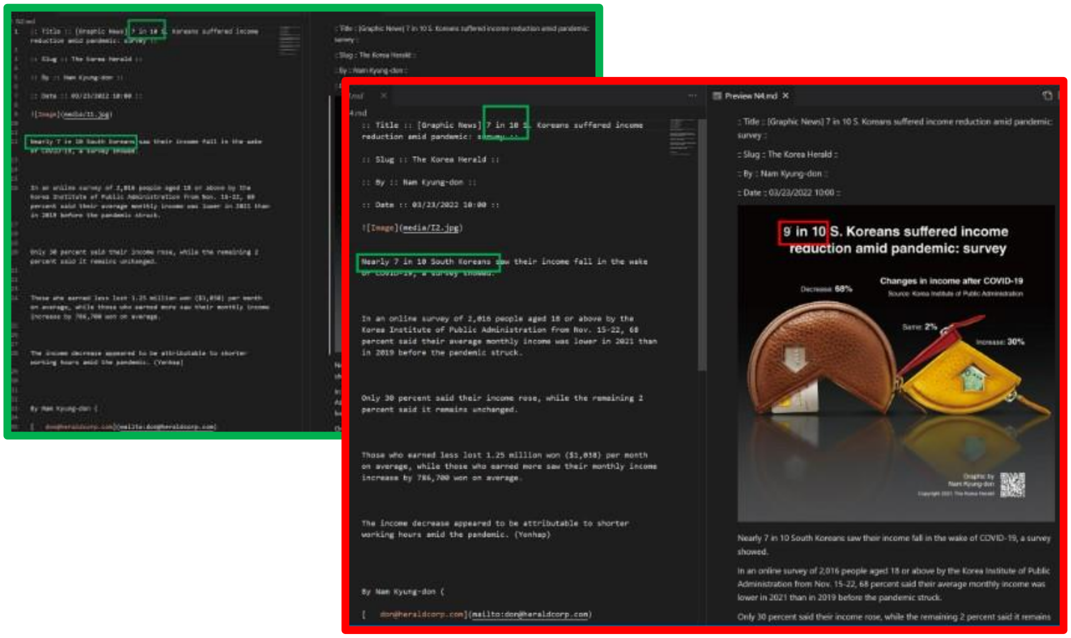
```
:: Title :: China updates its COVID-19 playbook to allow mild symptoms patients undergo collective quarantine ::
```

```
:: Title :: China updates its COVID-19 playbook to allow mild symptoms patients to be hospitalized

:: Slug :: Global Times ::

:: By :: Zhan Yusha and
Liu Caiyu ::

:: Date :: 03/17/2022 00:17 ::

![Image](media/0123c105da6e04d7d259f6ea94ecedfaecaeabe2d13055e9906cffddfe45f6f7.jpeg)
:: Caption :: A medical worker takes a swab sample from a student for nucleic acid test at a university in Qingpu District of east Chin
Shanghai, March 14, 2022. Medical workers from the Xianghuaqiao community medical service center in Qingpu District of Shanghai joined
battle against the recent COVID-19 resurgence. They have collected more than 50,000 samples for nucleic acid testing since the latest r
in March. Photo: Xinhua ::

![Image](media/d2635d0372a1001e37347d7f87f675bb7ae58227ef872f73a8b4693648Bad961.jpeg)
:: Caption :: Changes in China's latest Diagnosis and Treatment Protocol for COVID-19 Infographic: Deng Zijun/GT ::

Changes in China's latest Diagnosis and Treatment Protocol for COVID-19 Infographic: Deng Zijun/GT


China updated its COVID-19 playbook on Tuesday night, ordering COVID-19 patients with mild symptoms to go to centralized quarantine fac
instead of being hospitalized, and lowering the bar for patients to be discharged from hospital. Top epidemiologists hailed the adjustm
"scientific and pragmatic" change that helps relieve pressure on the medical system as the country is facing the most severe outbreaks
years, and they said that in no way can the adjustment be interpreted as the country "lying flat" facing the recent surge.


The National Health Commission (NHC) updated the treatment playbook for COVID-19 on Tuesday based on the transmission characteristics o
virus' variants, including Omicron and its strains, as well as the high proportion of mild cases.

[   It said that patients with mild symptoms should go to quarantine in centralized isolation facilities](https://www.globaltimes.cn/pa
1254983.shtml)


, rather than being taken directly to hospital as originally required. They will be admitted to hospital only when their condition wors
```

CHANGES in China's latest
Diagnosis and Treatment Protocol for COVID-19

Figure 26. Example Probe for Task 3.3.4.1b Detect Inconsistency - Manipulated Headline vs. Original (True) Body

For *Task 3.3.4.1c Detect Inconsistency: Manipulated Image vs. Original (True) Body*, TA1 performers will be provided with a manipulated MMA (technical news article) and must detect whether there is a semantic inconsistency relating to technical information within an image and the article body. Accordingly, targets will be articles that have an inconsistency between the image and the article body. For this sub-task, only images will be manipulated. Non-targets will be pristine articles with information that is consistent image and body. In this case, the modality inconsistency is between image versus long text.



```
:: Title :: [Graphic News] 7 in 10 S. Koreans suffered income
reduction amid pandemic: survey ::

:: Slug :: The Korea Herald ::

:: By :: Nam Kyung-don ::

:: Date :: 03/23/2022 10:00 ::

![Image](media/I1.jpg)

Nearly 7 in 10 South Koreans saw their income fall in the wake
of COVID-19, a survey showed.

In an online survey of 2,016 people aged 18 or above by the
Korea Institute of Public Administration from Nov. 15-22, 68
percent said their average monthly income was lower in 2021 than
in 2019 before the pandemic struck.

Only 30 percent said their income rose, while the remaining 2
percent said it remains unchanged.

Those who earned less lost 1.25 million won ($1,038) per month
on average, while those who earned more saw their monthly income
increase by 786,700 won on average.

The income decrease appeared to be attributable to shorter
working hours amid the pandemic. (Yonhap)

By Nam Kyung-don (
[   don@heraldcorp.com](mailto:don@heraldcorp.com)
```

```
:: Title :: [Graphic News] 7 in 10 S. Koreans suffered income
reduction amid pandemic: survey ::

:: Slug :: The Korea Herald ::

:: By :: Nam Kyung-don ::

:: Date :: 03/23/2022 10:00 ::

![Image](media/I2.jpg)

Nearly 7 in 10 South Koreans saw their income fall in the wake
of COVID-19, a survey showed.

In an online survey of 2,016 people aged 18 or above by the
Korea Institute of Public Administration from Nov. 15-22, 68
percent said their average monthly income was lower in 2021 than
in 2019 before the pandemic struck.

Only 30 percent said their income rose, while the remaining 2
percent said it remains unchanged.

Those who earned less lost 1.25 million won ($1,038) per month
on average, while those who earned more saw their monthly income
increase by 786,700 won on average.

The income decrease appeared to be attributable to shorter
working hours amid the pandemic. (Yonhap)

By Nam Kyung-don (
[   don@heraldcorp.com](mailto:don@heraldcorp.com)
```

:: Title :: [Graphic News] 7 in 10 S. Koreans suffered income reduction amid pandemic: survey ::

:: Slug :: The Korea Herald ::

:: By :: Nam Kyung-don ::

:: Date :: 03/23/2022 10:00 ::

9 in 10 S. Koreans suffered income reduction amid pandemic: survey
Changes in income after COVID-19
Source: Korea Institute of Public Administration
Decrease: 68%
Same: 2%
Increase: 30%
Graphic by Nam Kyung-don
Copyright 2021 The Korea Herald

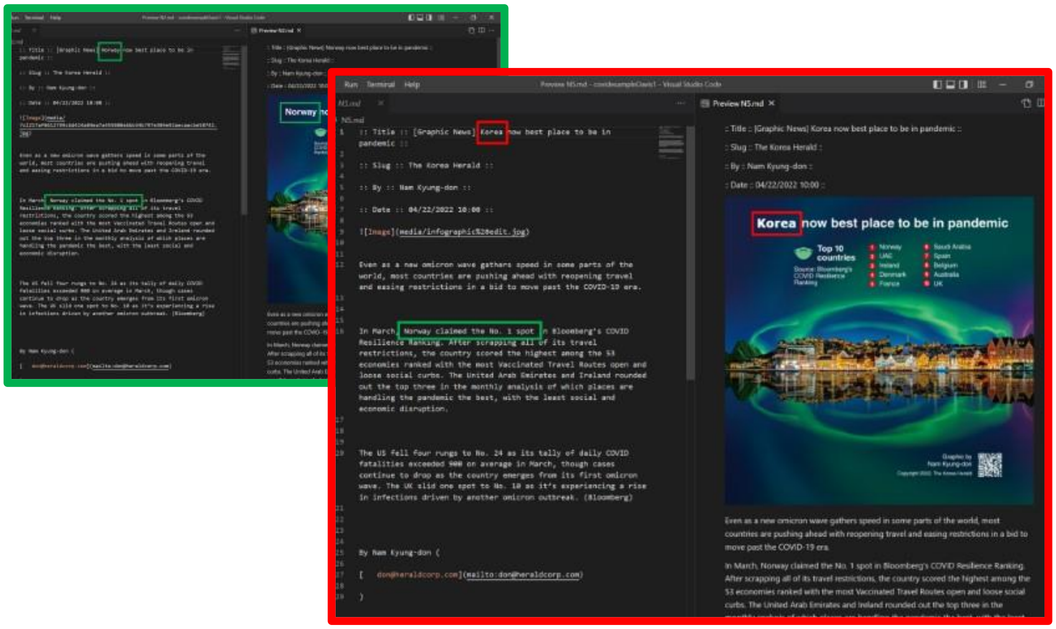
Nearly 7 in 10 South Koreans saw their income fall in the wake of COVID-19, a survey showed.

In an online survey of 2,016 people aged 18 or above by the Korea Institute of Public Administration from Nov. 15-22, 68 percent said their average monthly income was lower in 2021 than in 2019 before the pandemic struck.

Only 30 percent said their income rose, while the remaining 2 percent said it remains

Figure 27. Example Probe for Task 3.3.4.1c Detect Inconsistency: Manipulated Image vs. Original (True) Body

For *Task 3.3.4.1d Detect Inconsistency Manipulated Headline & Manipulated Image vs. Original (True) Body*, TA1 performers will be provided with a manipulated MMA (technical news article) and must detect whether there is a semantic inconsistency relating to technical information within a headline *and* an image that are consistent with each other, but are inconsistent with the information within the article body. For this sub-task, both headlines and images will be manipulated and manipulated in similar ways. Non-targets will be pristine articles with information that is consistent across assets. Thus, in this case, the modality inconsistency is between image *and* short text versus long text.



:: Title :: [Graphic News] Norway now best place to be in pandemic ::

```
:: Title :: [Graphic News] Korea now best place to be in
pandemic ::

:: Slug :: The Korea Herald ::

:: By :: Nam Kyung-don ::

:: Date :: 04/22/2022 10:00 ::

![Image](media/infographic%20edit.jpg)


Even as a new omicron wave gathers speed in some parts of the
world, most countries are pushing ahead with reopening travel
and easing restrictions in a bid to move past the COVID-19 era.



In March, Norway claimed the No. 1 spot in Bloomberg's COVID
Resilience Ranking. After scrapping all of its travel
restrictions, the country scored the highest among the 53
economies ranked with the most Vaccinated Travel Routes open and
loose social curbs. The United Arab Emirates and Ireland rounded
out the top three in the monthly analysis of which places are
handling the pandemic the best, with the least social and
economic disruption.



The US fell four rungs to No. 24 as its tally of daily COVID
fatalities exceeded 900 on average in March, though cases
continue to drop as the country emerges from its first omicron
wave. The UK slid one spot to No. 10 as it's experiencing a rise
in infections driven by another omicron outbreak. (Bloomberg)




By Nam Kyung-don (

[   don@heraldcorp.com](mailto:don@heraldcorp.com)

)
```

:: Title :: [Graphic News] Korea now best place to be in pandemic ::

:: Slug :: The Korea Herald ::

:: By :: Nam Kyung-don ::

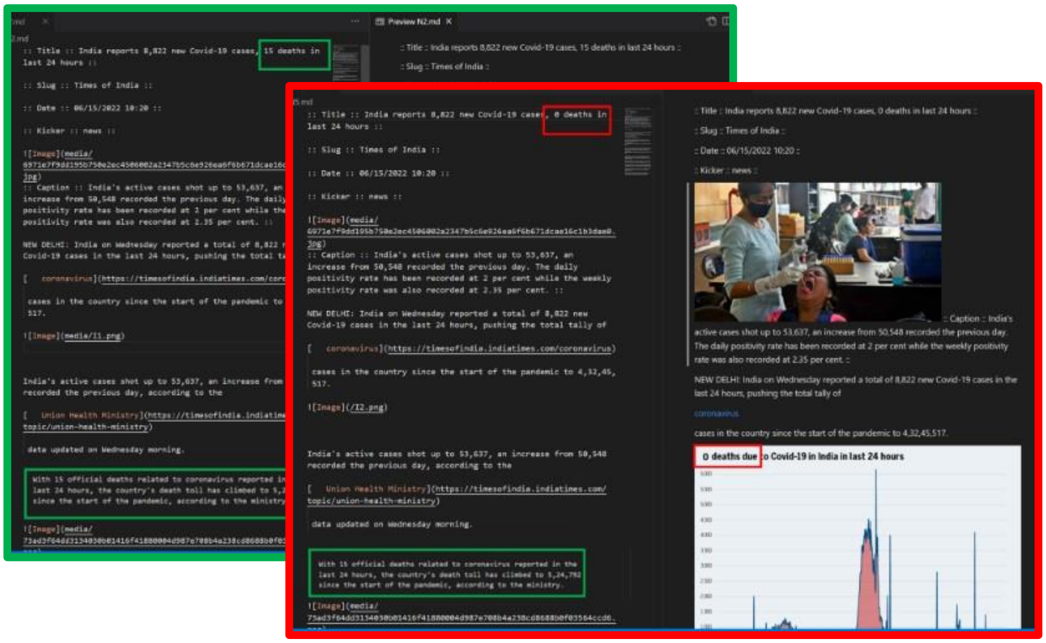
:: Date :: 04/22/2022 10:00 ::

Even as a new omicron wave gathers speed in some parts of the world, most countries are pushing ahead with reopening travel and easing restrictions in a bid to move past the COVID-19 era.

In March, Norway claimed the No. 1 spot in Bloomberg's COVID Resilience Ranking. After scrapping all of its travel restrictions, the country scored the highest among the 53 economies ranked with the most Vaccinated Travel Routes open and loose social curbs. The United Arab Emirates and Ireland rounded out the top three in the

Figure 28. Example Probe for Task 3.3.4.1d Detect Inconsistency Manipulated Headline & Manipulated Image vs. Original (True) Body

For *Task 3.3.4.2 Characterize Inconsistency - Minimization (COVID)*, TA1 performers will be provided with a manipulated MMA (technical news article) and must characterize whether the manipulation was done with the propaganda technique of minimization. Non-targets will be MMAs of the opposite tactic (Appeal to Fear). The dataset for this competition will include all probe variations from 3.3.4.1. Analytics will be expected to opt-out of probes containing media or assets that they are not made to ingest or cannot analyze.



```
:: Title :: India reports 8,822 new Covid-19 cases, 15 deaths in
last 24 hours ::

:: Slug :: Times of India ::

:: Date :: 06/15/2022 10:20 ::

:: Kicker :: news ::

![Image](media/
6971e7f9dd195b750e2ec4506002a2347b5c6e926ea6f6b671dcae16c
jpg)
:: Caption :: India's active cases shot up to 53,637, an
increase from 50,548 recorded the previous day. The daily
positivity rate has been recorded at 2 per cent while the
positivity rate was also recorded at 2.35 per cent. ::

NEW DELHI: India on Wednesday reported a total of 8,822 n
Covid-19 cases in the last 24 hours, pushing the total ta

[   coronavirus](https://timesofindia.indiatimes.com/cor

 cases in the country since the start of the pandemic to
 517.

![Image](media/I1.png)


India's active cases shot up to 53,637, an increase from
recorded the previous day, according to the

[   Union Health Ministry](https://timesofindia.indiatime
topic/union-health-ministry)

 data updated on Wednesday morning.

 With 15 official deaths related to coronavirus reported in
 last 24 hours, the country's death toll has climbed to 5,2
 since the start of the pandemic, according to the ministry

![Image](media/
73ad3f64dd3134030b01416f41880004d987e708b4a238cd8688b0f0
```

Preview N2.md

:: Title :: India reports 8,822 new Covid-19 cases, 15 deaths in last 24 hours ::

:: Slug :: Times of India ::

```
:: Title :: India reports 8,822 new Covid-19 cases, 0 deaths in
last 24 hours ::

:: Slug :: Times of India ::

:: Date :: 06/15/2022 10:20 ::

:: Kicker :: news ::

![Image](media/
6971e7f9dd195b750e2ec4506002a2347b5c6e926ea6f6b671dcae16c1b3dae0.
jpg)
:: Caption :: India's active cases shot up to 53,637, an
increase from 50,548 recorded the previous day. The daily
positivity rate has been recorded at 2 per cent while the weekly
positivity rate was also recorded at 2.35 per cent. ::

NEW DELHI: India on Wednesday reported a total of 8,822 new
Covid-19 cases in the last 24 hours, pushing the total tally of

[   coronavirus](https://timesofindia.indiatimes.com/coronavirus)

 cases in the country since the start of the pandemic to 4,32,45,
 517.

![Image](/I2.png)


India's active cases shot up to 53,637, an increase from 50,548
recorded the previous day, according to the

[   Union Health Ministry](https://timesofindia.indiatimes.com/
topic/union-health-ministry)

 data updated on Wednesday morning.

 With 15 official deaths related to coronavirus reported in the
 last 24 hours, the country's death toll has climbed to 5,24,792
 since the start of the pandemic, according to the ministry.

![Image](media/
73ad3f64dd3134030b01416f41880004d987e708b4a238cd8688b0f03564ccd6.
```

:: Title :: India reports 8,822 new Covid-19 cases, 0 deaths in last 24 hours ::

:: Slug :: Times of India ::

:: Date :: 06/15/2022 10:20 ::

:: Kicker :: news ::

:: Caption :: India's active cases shot up to 53,637, an increase from 50,548 recorded the previous day. The daily positivity rate has been recorded at 2 per cent while the weekly positivity rate was also recorded at 2.35 per cent. ::

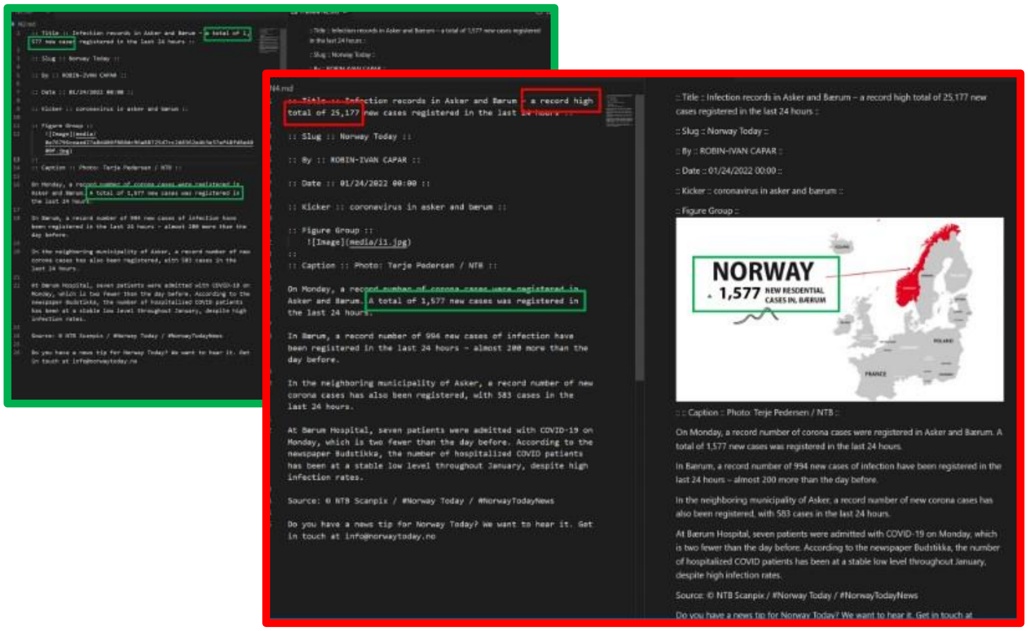
NEW DELHI: India on Wednesday reported a total of 8,822 new Covid-19 cases in the last 24 hours, pushing the total tally of

coronavirus

cases in the country since the start of the pandemic to 4,32,45,517.

Figure 29. Example Probe for Task 3.3.4.2 Characterize Inconsistency - Minimization (COVID)

For *Task 3.3.4.3 Characterize Inconsistency - Appeal to Fear (COVID)*, TA1 performers will be provided with a manipulated MMA (technical news article) and will be asked to characterize whether the manipulation was done with the propaganda technique of appealing to fear. Non-targets will be MMAs of the opposite tactic (Minimization). The dataset for this competition will include all probe variations from 3.3.4.1. Analytics will be expected to opt-out of probes containing media or assets that they are not made to ingest or cannot analyze.



```
N4.md
:: Title :: Infection records in Asker and Barum - a record high
total of 25,177 new cases registered in the last 24 hours ::

:: Slug :: Norway Today ::

:: By :: ROBIN-IVAN CAPAR ::

:: Date :: 01/24/2022 00:00 ::

:: Kicker :: coronevirus in asker and berum ::

:: Figure Group ::
   ![Image](media/i1.jpg)
::
:: Caption :: Photo: Terje Pedersen / NTB ::

On Monday, a record number of corona cases were registered in
Asker and Barum. A total of 1,577 new cases was registered in
the last 24 hours.

In Barum, a record number of 994 new cases of infection have
been registered in the last 24 hours - almost 200 more than the
day before.

In the neighboring municipality of Asker, a record number of new
corona cases has also been registered, with 583 cases in the
last 24 hours.

At Barum Hospital, seven patients were admitted with COVID-19 on
Monday, which is two fewer than the day before. According to the
newspaper Budstikka, the number of hospitalized COVID patients
has been at a stable low level throughout January, despite high
infection rates.

Source: © NTB Scanpix / #Norway Today / #NorwayTodayNews

Do you have a news tip for Norway Today? We want to hear it. Get
in touch at info@norwaytoday.no
```

:: Title :: Infection records in Asker and Bærum – a record high total of 25,177 new cases registered in the last 24 hours ::

:: Slug :: Norway Today ::

:: By :: ROBIN-IVAN CAPAR ::

:: Date :: 01/24/2022 00:00 ::

:: Kicker :: coronavirus in asker and bærum ::

:: Figure Group ::

: : Caption :: Photo: Terje Pedersen / NTB ::

On Monday, a record number of corona cases were registered in Asker and Bærum. A total of 1,577 new cases was registered in the last 24 hours.

In Bærum, a record number of 994 new cases of infection have been registered in the last 24 hours – almost 200 more than the day before.

In the neighboring municipality of Asker, a record number of new corona cases has also been registered, with 583 cases in the last 24 hours.

At Bærum Hospital, seven patients were admitted with COVID-19 on Monday, which is two fewer than the day before. According to the newspaper Budstikka, the number of hospitalized COVID patients has been at a stable low level throughout January, despite high infection rates.

Source: © NTB Scanpix / #Norway Today / #NorwayTodayNews

Do you have a news tip for Norway Today? We want to hear it. Get in touch at

Figure 30. Example Probe for Task 3.3.4.3 Characterize Inconsistency - Appeal to Fear (COVID)

Task 3.3.4 Scoring

All Evaluation Task 3.3.4 subtasks will be combined into a single task and dataset. Each probe will include an AnalyticScope containing the AOM anchors to identify where the analytic should look. These will be labeled according to article headline, body, and image. If an analytic does *not* support the anchors provided, it is expected that the analytic opts-out of the sub-task.

Each subtask uses the same non-target (pristine) data. Thus, for each probe there will be four copies: each pristine AG, each with a different set of AOM anchors, and resulting in an equal distribution of pristine/manipulated for each type of manipulation. The ScoringEngine will be updated to support querying based on manipulation type. This will also allow viewing score breakdowns for each manipulation[1].

Tasks 3.3.4.1a-1d Detect Inconsistencies will be scored against the program goal of 85% pD @ 0.08 FAR across probes. TA3 will use this performance information to answer Research Question 3.3.4.

For *Tasks 3.3.4.2 Characterize Inconsistency: Minimization (COVID) and 3.3.4.3 Characterize Inconsistency: Appeal to Fear (COVID)* will be scored against the characterization program goal of 80% accuracy. TA1 Performers will be scored on their ability to successfully characterize the type of intent within the inconsistent MMA probes. For scoring purposes, each propaganda intent will be separated into its own competition in the SemaFor Gym and scored individually as a

[1] *Note.* This update may come after the opening of Evaluation 3.3.

binary classification problem. Each competition will have an equal number of target and non-target probes. Within a given competition, target probes will be those falsified with the specified tactic (*Appeal to Fear* or *Minimization*). Non-targets will be MMAs generated with the alternative tactic. Target and non-target classes will be balanced—each will have equal number of probes and representative across inconsistency types.

For the characterization tasks, analytics will be expected to read in a partial Evidence Graph (EG) containing an already completed EvCharacterizationNode. The EvCharacterizationNode must be present in the EvidenceGraph and there must be a child consistency check node containing the tactic being scored in that node's AnalyticScope. Just as with target probes, the same nodes must be present in non-targets. The expectation here is that the consistency check with relevant scope will have a low LLR score for non-target probes (i.e., "not Intent X"). Since the LLR score is being taken at the consistency check instead of the task node, if the ScoringEngine is unable to find a consistency check with the matching AnalyticScope, it is considered an opt-out. The LLR score of that consistency check will be used for scoring.

Analytics should output as results an augmented version of that EG containing an EvNonSemanticConsistencyCheckNode with a tactic matching the competition target tactic.

EVALUATION 3 ANALYSIS PLAN

In addition to reporting results at the individual analytic level, TA3 plans to conduct aggregate analysis across all performers and tasks. The goal of this analysis is to answer the research questions described in this document, as well as to provide greater insight into evaluation results, gauge general program strengths and weaknesses, and inform future research directions. Whereas leaderboard scores in the Gym allow comparison of performance across individual analytics, analysis will aim to provide a unified summary of the program capabilities.

Research Questions versus Task Leaderboards

Tasks refer to the specific challenge that individual TA1 performer's analytics will perform against—including multiple variations of conditions as outlined below. For all participating analytics (and performers), performance on each task will be rank ordered against one another. In contrast, research questions will be explored and investigated by aggregating performance across analytics for each task. Findings from this analysis will not be used to compare or rank individual analytics, but rather to identify shared strengths and challenges for program technologies as a whole and provide a better understanding of our niche within the problem space.

TA3 will work throughout this analysis to maintain transparency and ensure that each performer and the program as a whole are represented fairly and accurately. For the purposes of understanding program capabilities to the fullest, analyses conducted by TA3 may involve assessment of an analytic's performance on a task for which it was not expressly designed. Given that, TA3 will ensure that: (1) no analysis results will be reported at the individual analytic level; and (2) all results calculated for analysis purposes will be kept out of the Gym leaderboards and will not be used for official task scoring.

Methods

Although performers will not be scored on their ability to perform across subsets of data, such as across different modalities or manipulation techniques, TA3 will track this information to answer the Evaluation 3 research questions. Generally, analysis will involve recalculation of scores across contrasting subsets of probe data to understand the difference in analytics' ability to perform across conditions. This recalculation will occur by the following procedure:

1. **Identification of interest variable:** For a given analysis, the first step will be to identify the variable of interest and how it is represented in task results. For example, in order to answer *Research Question 3.2.2* we will need to investigate performance across manipulation techniques. Starting at subtask 3.2.2a, analysis will then investigate analytics' ability to discriminate between hate speech and each of the other manipulation techniques.
2. **Balancing dataset:** Where necessary, probes will be either added to or removed from the positive and negative classes to ensure a balanced representation of each. For example, in subtask 3.2.2a, participating analytics will evaluate 120 hate speech probes, but only 40 probes for each other manipulation technique. If we are looking at discrimination

between hate speech and each other class individually, we will want to balance them by either randomly removing 80 of the positive hate speech probes, or by running the analytic over a full set of 120 scapegoating probes. Once the set of probes is adjusted for analysis, LLR thresholds will be recalculated for each unique analytic competition entry. In any cases where this is necessary, the change would not be carried over into any Gym results or official scores. The goal of this analysis would be to compare differences across classes as opposed to individual performance in any one class, so the change in individual scores from those shown in the Gym should not affect the validity of results.

(3) **Recalculating metrics of interest across analytics:** After any rebalancing, the next step will be to assign confusion matrix values *(true positive*, *true negative*, *false positive*, *false negative*) to scores at the individual probe level using the predetermined LLR threshold of interest for each analytic, and then to separate those scores into buckets based on the variable of interest and calculate metrics of interest (pD, FAR, bACC) at relevant thresholds.

(4) **Comparison across conditions:** Lastly, results from each condition will be compared and differences used to draw inferences about program strengths and challenges. In the example above, this might look like comparing the difference in discrimination between hate speech and scapegoating vs hate speech and dictat. Discrepancies in these across analytics on the program could point to areas where analytics shine or problems that are still technically difficult for the program. TA3 hopes these insights will help to inform future evaluations and program directions.

Preliminary results from analysis will be shared out in slide form throughout the evaluation period. A full report of analysis will be included in the Evaluation 3 Report, with results on each of the research questions, as well as general cross-program performance analysis and an emerging areas of research interest.

PROBE GENERATION CONSTRUCT CONSISTENCY

Concurrent with probe generation, members of the TA3 evaluation team also supported ensuring consistency of the characterizations and manipulation techniques intended. The purpose of the inter-rater consistency process was *not* to act as quality control, but to ensure that the generated probes reflect the intended constructs of characterization types and manipulation techniques as defined. Accordingly, this process only pertained to the manipulated constructs, and not to levels or modalities of manipulation.

The approach and methods described here were implemented across Evaluations 3.2-3.3. This document details this process across evaluations. When methods deviated between evaluations, this is highlighted.

Probe Generation Construct Consistency Procedures

Providing Construct Definitions

Definitions for every characterization type and manipulation technique were provided to raters for reference. These definitions were vetted by the TA3 and TA4 teams. Before rating, raters participated in a frame-of-reference training session. This included reviewing the construct definitions of the characterization and manipulation technique types and if needed, discuss these definitions further with the broader research team. In addition to the construct definitions, non-rated probes were provided as examples of each characterization and manipulation technique. This process was repeated for each evaluation when unique tasks and their construct definitions were added.

Rating Instructions

During the rating process, raters were instructed to carefully read and review the probe in its entirety and respond in a provided electronic rating form. Raters were informed of the characterization and manipulation technique intended for all probes before they provided their rating. Raters also had access to the construct definitions during their review of the probes.

The electronic rating form contains two ratings for rating each data probe: one for characterizations and one for manipulation techniques. Raters used drop-down menus to select from the two characterizations and the four manipulation techniques. Raters could also select, "Cannont determine/Unsure" to indicate that the intended characterization or manipulation technique was not adequately or appropriately represented in the probe. If any notes were needed, an additional blank were provided for comments.

Rating Methods

Two raters completed the construct consistency review of a subset of data probes. Agreement was considered achieved if the two raters agreed with one another that the intended characterization and manipulation technique was appropriately represented in the probe. If raters agreed with one another about the characterization and manipulation technique employed, but it did not match the *intent* of the data generation team, this information was reported back to the data generation team to suggest a rescoring of the probe. If consensus could *not* be achieved, this too was reported back to the data generation team to suggest changing or eliminating the probe. Inter-rater agreement for each factor and aggregated notes from raters were provided to the probe generation team after each review phase.

To ensure independence, members of the probe generation team did *not* participate in the inter-rater reliability ratings. Additionally, raters independently coded the characterization and manipulation technique for each probe and were not able to see one another's responses.

Inter-Rater Reliability Analysis Approach

Inter-rater agreement was measured using the *Fleiss kappa*. The Fleiss kappa is an extension of the usual Cohen's kappa for evaluating the degree of agreement between two or more raters[2], and

[2] Note that the Fleiss Kappa does not assume that raters are the same across trials, allowing more flexibility with personnel (Fleiss, Levin, & Paik, 2013).

when the agreement is on a categorical scale containing more than one category (i.e., two characterizations and four possible manipulation techniques). No weighting is used and the categories are considered to be unordered. This measure expresses the degree to which the observed proportion of agreement among raters exceeds what would be expected if ratings were completely randomly. This will allow the TA3 research team to assess the agreement between multiple raters about the intended constructs being appropriately represented in the probes. If acceptable agreement is not found, this information was provided that information to the probe generation team across each phase of probe generation. Upon the completion of probe generation, a total score will be provided to show overall consistency of the intent of probe's characterization type of manipulation technique.

Evaluation 3.2 Construct Consistency and Inter-Rater Reliability Plan for Evaluation 3.2

Probe Data Sampling and Schedule

Ultimately, the raters' goal is to review at least 25% of the sample of generated probes. Because probe generation produced manipulations roughly equally in proportion across the four tactics over time, the raters were able to follow the probe generation phases of approximately four two-week cycles of review. However, two factors led to front-loading the review of more probes earlier. First, this provided the probe generation team with information about the consistency of the probes produced as early as possible if there was any disagreement among raters. After each construct consistency review phase, the results were provided to TA3 Probe Generation Team (see Table 22 below). Second, the final phase of probe generation was immediately deployed for THE Evaluations, leaving little time for review before Evaluation 3.2 began.

The probe generation team reported that production across proportion of characterizations and manipulation techniques may be slightly skewed by characterization and manipulation technique over phases. To ensure that a representative sample of the produced probes was used, a stratified random sample of intended characterizations and manipulation techniques was pulled from the total probes produced by each phase to be coded. In the case of new Evaluation 3.3 tasks, this process was repeated.

Table 22. Timetable for Data Generation and Construct Consistency for Evaluation 3.2

Data Generation Timeline for Rolling Evaluation 3.2					
Probe Data Generation Progress		**Construct Consistency Checks Completed**			
Date Completed	*Probes Produced*	*Date Completed*	*Probes Reviewed by Phase*	*Running Tally of Probes Reviewed*	*Percentage of Total Probes Reviewed*
8-Apr-22	25% (120)	*20-Apr-22*	45	45	~9%
13-Apr-22	50% (240)	*22-Apr-22*	30	75	~16%
27-Apr-22	75% (360)	*2-May-22*	30	105	~22%
4-May-22	100% (480)	*6-May-22*	15	120	25%

Interrater Consistency Results

Two raters independently reviewed 80 probes to determine consistency with the established manipulation technique definitions. These 80 probes represented 27.03% of the available, non-sequestered data and 16.67% of the total probes produced. These probes were a stratified random sample across all characterization types, manipulation techniques, and degrees of manipulation (1 vs 3 sentences manipulated, image manipulations). Because a single original news article was used modified to produce these tasks, reviewed probes were also staggered to ensure that as many unique base articles were reviewed as well.

Of the 80 probes, the two raters reached a Fleiss' Kappa of .746 which is considered a good level of inter-rater agreement. Operationalized another way, the two raters concurred that 62 of the 80 probes well-represented the established manipulation technique definitions (i.e., 77.50%). During this process the raters also identified some missing data in the human-readable formatted probes which were relayed to the data team.

Evaluation 3.3 Construct Consistency and Inter-Rater Reliability Plan

Probe Data Sampling and Schedule

Following the process detailed above for Evaluation 3.2, in Evaluation 3.3 raters will also review approximately 25% of the sample of generated probes. Again, this construct consistency review only reviewed probes where human judgment was required in the development of the probes and that observers may or may not agree reflect the intended propaganda technique or intention of the probe. This was *not* for the purposes of quality control. Thus, Task 3.3.1: Deepfakes were not reviewed as these probes contained no propaganda technique or intention to be detected. Similarly, Task 3.3.2: Synthetic Audio and Task 3.3.3: Synthetic Media contain only generated media content with no specific propaganda techniques. For this reason, only Task 3.3.4 probes required construct consistency review. This process is detailed below.

This construct consistency review for Task 3.3.4 probes will examine assets contained within news articles to determine if the elements reflect the intended inconsistency (across intended assets in the article) and its characterization (appeal-to-fear or minimization). For example, do the assets of the news article demonstrate clear inconsistencies in the modalities they purport to? Do the manipulations used to produce the inconsistencies reflect the characterization intended as defined previously? Finally, note that non-targets (pristine) will not be reviewed as they have not been selected or manipulated to be either inconsistent or reflect a particular propaganda technique. For this construct consistency review, all raters will be provided with the definitions for Provided in section (pg. 37)

Table 23. Evaluation 3.3 Distribution of Probes *for Data Generation and Construct Consistency*

Tasks	Anticipated Journals*		Completion Status**	Targets to be Reviewed (25%)
	Targets	Non-Targets	Currently Submitted Journals*	
3.3.4.1 - Detect Inconsistencies between different article assets (Headline, Image and Body)				
3.3.4.1a - Detect Inconsistency: Headline vs. Image	120	120	134	34
3.3.4.1b - Detect Inconsistency: Headline vs. Body	120	120	134	34
3.3.4.1c - Detect Inconsistency: Image vs. Body	120	120	134	34
3.3.4.1d - Detect Inconsistency: Image/Headline vs. Body	120	120	134	34
3.3.4.2 - Characterize MMA Inconsistency: To Appeal To Fear (COVID)	240	240	268	67
3.3.4.3 - Characterize MMA Inconsistency: To Minimize (COVID)	240	240	268	67

Note. * = Each journal may have multiple probes built from it. ** as of 8/30/2022

The probe generation team reported that production across proportion of characterizations and manipulation techniques may be slightly skewed by characterization and manipulation technique over phases. To ensure that a representative sample of the produced probes was used, a stratified random sample of intended characterizations and manipulation techniques was pulled from the total probes produced by each phase to be coded.

Interrater Consistency Results

To be determined. Results of the inter-rater reliability analysis will be added to this document and the Evaluation Report as they progress.

References

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378-382.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons.

APPENDICES

Appendix A - Setting Static versus Dynamic Thresholds

Traditional Static Thresholds

The range of Log-Likelihood Ratio (LLR) values offer unique challenges—but also opportunities—for evaluation. In traditional evaluation of machine learning classification, a threshold must be set to make positive classifications from probes. The SemaFor evaluation team, however, will *not* be evaluating performers in this way.

In setting a static threshold, performers or the evaluation team must select a Log-Likelihood Ratio (LLR) value above a certain numerical threshold as a prediction of the positive class, and below, a prediction of the negative class. Metrics are then computed to score the success of the classifier based on that threshold's output. Despite the advantage of its simplicity, this approach does not give TA1 the best opportunity for analytics to perform, nor provide the most flexibility for TA2 performers to subsequently integrate these analytics.

First, a static threshold must be selected. This requires selecting either an arbitrary value or to implement an *objective function* that defines an optimal threshold. In the case of an arbitrary value, this may randomly favor the raw LLRs of some performers over others. In the case of a developing and implementing an objective function, this jeopardizes the mission by becoming a research task in itself. Accordingly, the SemaFor evaluation team will evaluate performers by using dynamic thresholds based on the LLRs of performer analytics.

Dynamic Thresholds

Rather than setting a static threshold on performer analytics' raw LLRs, the TA3 evaluation team will instead use dynamic thresholding. In the below example, the EER is used to establish a threshold for analytics. The blue ROC Curve has an EER where p(D) =.9, FAR =.1. This approach avoids the aforementioned challenges in setting thresholds, provides more information about the performance of the algorithm across contexts, and offers advantages to performers as well. Dynamic thresholding ensures that scores most accurately represent the best performance of the analytics, rather than a performer's ability to set a threshold. This also eliminates a penalty for a poorly-calibrated analytic. As mentioned, well-calibrated analytics will be critical for fusion, but not necessarily for evaluation.

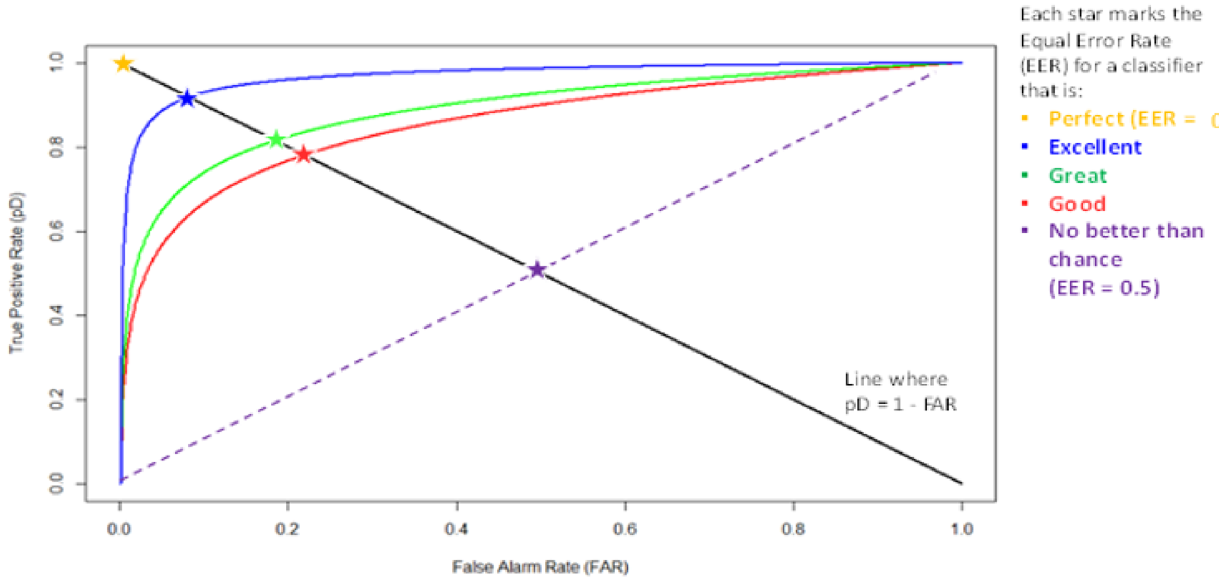


Figure 31. Example of use of the EER to establish a threshold for analytics.

Metrics are computed by sorting classifications by LLR per performer for each analytic and for each task. The evaluation team will dynamically set the LLR threshold at the point on the particular performer's ROC Curve where p(D) = 1 – FAR. Once this point is identified on the ROC Curve, that LLR value will be set as the performer's threshold to score individual probes. Thus, performers' particular LLR thresholds for a positive versus negative classification will differ. After the dynamic threshold has been set, all DAC outputs will be compared against ground truth. Single p(D) and FAR of a system-defined threshold will not be computed per analytic.

There are several advantages to this approach that are aligned with SemaFor's program-level goals. First, specific thresholds do not need to be set, but can still be oriented around program-wide goals (e.g., equal error rates). The EER provides a target to shoot for with a balanced tradeoff between an algorithm's correctness and completeness—that is, the ratio between the false acceptance rate and false rejection rate, over the available evaluation sample. A potential disadvantage of this approach is that this does not evaluate the quality of the LLRs. That is, the score is neither improved nor harmed for distance from threshold. However, this thresholding approach offers performers the best opportunity to demonstrate performance across the entire test sample.

We also recognize that while the EER seeks equal rates of errors (FAR = EER), the program goals are not equal, that is, FAR = .1, EER = .2. Therefore, we also propose to compute secondary metrics to provide additional insights where p(D) is determined at .1 FAR, which would be an operating point relevant to the program goals. This example is shown below, where an instance of a ROC Curve (blue) achieves the program goals p(D) = .8, FAR = .1 (blue dot), but where the EER identifies a different point p(D) = .8, FAR = .2 (red dot). This secondary metric will be used in the Leaderboard. Another metric under consideration is Area Under the

Curve (AUC) for FAR <= 0.1. This metric would indicate that the analytic generating the yellow-dotted curve is preferred over that generating the blue curve.

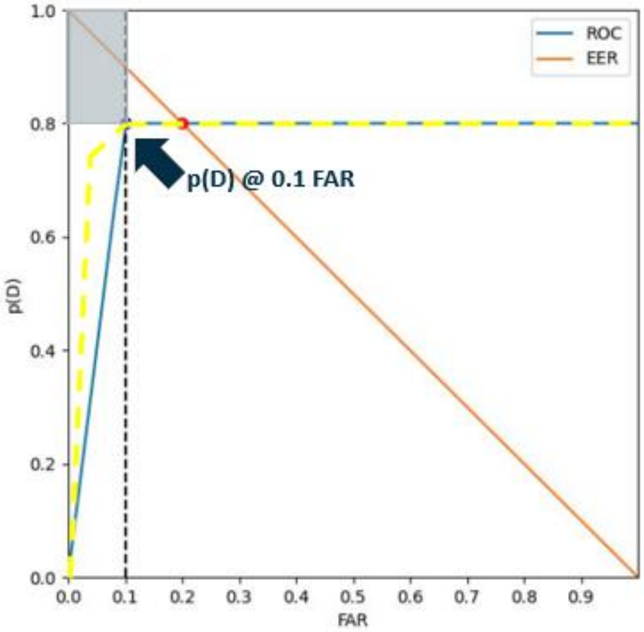


Figure 32. Example of an instance where a ROC curve (blue) achieves the program goals (blue dot), but where the EER identifies a different point (red dot).

In summary each analytic will be evaluated based on a dynamically set LLR threshold at the point on the particular performer’s ROC Curve where p(D) = 1 – FAR, as well as a secondary metric where p(D) is determined at .1 FAR.