2025 NIST GenAI (Pilot)
Evaluation Plan for Image Discriminators

Released: 2025-03-13

Updated: 2025-03-13

GenAI Eval Team

National Institute of Standards and Technology
100 Bureau Dr, Gaithersburg, MD 20899

Contact: genai-poc@nist.gov

Disclaimers

Certain commercial equipment, instruments, software, or materials are identified in this document to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor necessarily the best available for the purpose. The descriptions and views contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NIST or the U.S. Government.

Updates

2025-03-13 Initial Release

Table of Contents

1 Introduction

In recent years, digital content from generative Artificial Intelligence (AI), including deepfakes, has had unprecedented growth and proliferation across various modalities, including image, video, audio, and text. This surge in generative AI presents both opportunities and challenges. The technologies have facilitated creative expression, enabling artists, designers, and writers to generate visually stunning content as well as fast professional written content. On the other hand, it has raised concerns regarding the authenticity and integrity of media in digital content. With the advancements in generative AI technology, it is becoming increasingly difficult to distinguish AI-generated content from human-generated content. Our goal is to advance measurement science and provide meaningful assessment for tools that can help differentiate between the two.

In this NIST Generative AI (GenAI) program, we invite and encourage participating teams from academia, industry, and other research labs to support research in Generative AI. GenAI is an evaluation series that provides a platform for testing and evaluation to measure the performance of AI content generators and AI content discriminators (e.g., detectors). The platform is planned to support multiple modalities and technologies enabled by both sides of the generative spectrum, "generators" and "discriminators."

Generator (G) teams will be tested on their system's ability to generate content that is indistinguishable from human-generated content. For the pilot study, the evaluation will help determine strengths and weaknesses in their approaches, including insights about how and when humans and/or AI can detect AI-generated content.

Discriminator (D) teams will be tested on their system's ability to differentiate between AI-generated content and human-generated content. Lessons learned from both sides of teams are expected to help develop future research directions and approaches for understanding cutting-edge technologies as well as sources for recommendations and guidance for responsible use of digital content.

Participants are required to select if they are participating as a generator team, a discriminator team, or both. This document describes the evaluation plan for the "discriminator team." It covers task definitions, task conditions, file formats for system inputs and outputs, evaluation metrics, and protocols for participating in GenAI evaluations (see details at https://ai-challenges.nist.gov/genai).

The goal of the 2025 GenAI pilot study is to measure how well discriminator systems can automatically detect AI-generated content in multiple modalities, such as text, audio, image, and video. In this challenge, the task will focus on the image modality only.

The pilot GenAI evaluations will provide testing datasets, created by G-participants and supplemented by the NIST GenAI team. D-participants can then submit their system outputs to a web-based leaderboard, where scores and results will be displayed.

The data from G-participants will only be accessible to D-participants once the G-participants submit their data packages to NIST and the NIST GenAI team reviews and approves the data (e.g., identifying inappropriate content). NIST reports performance measures for D-participant system outputs, displayed through a leaderboard. Please refer to the published schedule for details.

Data resources as well as GenAI Scorer and Format Validator scripts are available for download at resources.

Any questions or comments concerning the GenAI evaluation series should be sent to genai-poc@nist.gov.

2 Tasks

The primary goal of the pilot GenAI evaluations is to measure system behavior in detecting whether or not the images in the test set are AI-generated.

2.1 Detection Task

D-participants will be given a collection of images, some of which are generated using AI image generators, while others are either photographs or frames from video or movie clips, i.e., not generated by AI.

D-participants task is to attempt to classify the images as AI-generated or not by providing a confidence score (any real number between 0 to 1), with higher numbers indicating the target image is more likely to have been generated using AI.

2.2 Protocol and Rules

The evaluation participants agree not to probe the test media via manual/human means, such as looking at the media or annotating media to produce the authorship information prior to, during, and after the evaluation. The participants are NOT allowed to use the test dataset for purposes of training, modeling, or tuning their algorithms. The participants may use any publicly available data that complies with applicable laws and regulations to train their models. All machine learning or statistical analysis algorithms must complete training, model selection, and tuning prior to running the GenAI test data; learning/adaptation during processing is not permissible. The participants are not allowed to use the generators' topics files by any means to extract information that can assist their systems in predicting the testing image source.

All images must be processed independently of each other within a given round of evaluation and across all evaluation rounds, meaning content extracted from the data must not affect the processing of data from another rounds' data.

While participants may discuss their own results, participants may not make advertising claims about their standing in the evaluation including results of other teams, regardless of rank, winning the evaluation, or claim NIST endorsement of their system(s). The following language in the U.S. Code of Federal Regulations (15 C.F.R. § 200.113 (d)) shall be respected: NIST does not approve, recommend, or endorse any proprietary product or proprietary material. No reference shall be made to NIST or to reports or results furnished by NIST in any advertising or sales promotion which would indicate or imply that NIST approves, recommends, or endorses any proprietary product or proprietary material or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of NIST test reports or results.

At the conclusion of the evaluation, NIST will generate a report summarizing the system results as tested under specific conditions of interest with the team names anonymized. Participants may publish or otherwise disseminate these charts unaltered and with appropriate reference to the original NIST report results.

3 Data Resources

Once the G-participants submit their data to NIST and gain subsequent approval of the data by the NIST GenAI team, the D-participants will have access to the G-participants' data packages.

NIST will make all necessary data resources available to D-participants. Each team will receive access to data resources upon completing the required data agreement forms and based on the published release dates for each task. Please refer to the schedule for data release dates.

3.1 Dry-run Set

For the purpose of validating system output format, the GenAI dry-run set is delivered as a single gzipped tarball for each task, which is unpacked to the following contents and subdirectory structure:

README.md	A helpful documentation file in markdown format
genai25_Image-D_dryrun_index.csv	The system input file (dry-run set) for the image discriminator (Image-D) detection task
files/	A flat subdirectory containing trials (image files) organized by <file_id>.*

Example of the index CSV file with delimiter "|".

DatasetID	TaskID	FileID
GenAI25-Image-dryrunset	detection	file_0001.webP

Example of files

- file_0001.webP
- file_0002.webP

3.2 Test Set

The GenAI test set is delivered as a single gzipped tarball for the detection task, which is unpacked to the following contents and subdirectory structure:

README.md	A helpful documentation file in markdown format
genai25_Image-D_detection_index.csv	The system input file for the image discriminator (Image-D) detection task
files/	A flat subdirectory containing trials (image files) organized by <file_id>.*

Example of the index CSV file with delimiter "|".

DatasetID	TaskID	FileID
GenAI25-G1-Image-set1	detection	file_0001.webP

Example of contents of files/ : file_0001.webP , file_0002.webP, …etc

4 System Input

For a given task, a system's input is the task index file called <modality_id>_<dataset_id>_<task_id>_index.csv . Given an index file, each row specifies a test trial. Taking the corresponding media (e.g. texts or images) as input(s), systems perform detection tasks.

The following format constitutes the index file for the D-participant system input:

genai25_Image-D_detection_index.csv	
DatasetID	(string) The ID of the dataset release (e.g., GenAI25-G1-Image-set1)
TaskID	(string) The globally unique ID of tasks.
FileID	(string) The globally unique ID of the trials (e.g., file_0001.webP)

Example of the CSV file with delimiter "|".

DatasetID | TaskID | FileID |

GenAI25-G1-Image-set1 | detection | file_0001.webp |

5 System Output

5.1 System Output File

The system output file must be a CSV file with the separator "|". **Please include the optimal cutoff (threshold) for the confidence score for binary classification in your file name. For example, "cutoff-50" means the threshold is 0.5.** The filename for the output file must be a user-defined string that identifies the submission with **no spaces or special characters** besides '_-.' (e.g., `genai25_image_d_sys_model-01_cutoff-50.csv`). The system output CSV file for the Image-D detection task must follow the format below:

genai25_image_d_sys_model-01_cutoff-XX.csv	
DatasetID	(string) The ID of the dataset release, e.g., GenAI25-G1-Image-set1
TaskID	(string) The unique ID of the task, e.g., detection
DiscriminatorID	(string) The site name of Discriminator (D) participants, e.g. D-NIST_site
ModelVersion	(string) The system model version on D-participant submission e.g., MySystem_DallE
FileID	(string) The globally unique ID of the trials (e.g., file_0001.webp)
ConfidenceScore	(float) in the range [0,1], the larger, the more confidence that the output is AI generated

Example of the CSV file with delimiter "|".

DatasetID | TaskID | DiscriminatorID | ModelVersion | FileID | ConfScore

GenAI25-G1-Image-set | detection | D-NIST-site | MySys_DallE | file_0001.webp | 0.7

5.2 Validation/Submission

5.2.1 Validation

The FileID column in the system output [submission-file-name].csv must be consistent with the FileID in the <modality_id>_<dataset_id>_<task_id>_index.csv file. The row order may change, but the number of the files and file names from the system output must match the index file.

To validate your system output locally, D-Participants may use the command-line command as shown in Appendix A.

5.2.2 Submission

System output submission to NIST for subsequent scoring must be made through the web platform using the submission instructions described before (System output). To prepare your submission, you will first make **.tar.gz** (or **.tgz**) file of your system output CSV file via the UNIX command 'tar zcvf [submission_name].tgz [submission_file_name].csv' and then upload the system output tar file under a new or existing 'System' label. This system label is a longitudinal tracking mechanism that allows you to track improvements to your specific technology over time.

Please submit your files in time for NIST to deal with any transmission errors that might occur well before the due date. Note that submissions received after the stated due dates for any reason will be marked late and may not be scored. Please refer to the published schedule for details.

5.2.3 System Descriptions

The NIST GenAI team will request a system description for the top-performing systems in their submissions. Documenting a system is vital to interpreting evaluation results. Please make sure you document this information while developing your system and submitting your results. A system description should include, but is not limited to, the following information:

Section 1. Submission Identifier(s)

List the submission IDs for which system outputs were submitted; the GenAI team can help identify Submission IDs as needed.

Section 2. System Description

A brief technical description of your system and the system model used.

Section 3. System Hardware Description and Runtime Computation

Describe the computing hardware setup(s) and report the number of CPU and GPU cores. A hardware setup is the aggregate of all computational components used.

Section 4. Training Data and Knowledge Sources

List the resources used for system development and runtime knowledge sources, if any.

Section 5. References

List pertinent references, if any.

6 Performance Metrics

This section describes the metrics that will be used for measuring system performance.

The discriminator performance will be summarized using Area Under the Curve (AUC), TPR@FPR, EER, and two Brier scores – A Brier score for system performance on human generated images, denoted by BrierN ('N' for nontarget) and a Brier score for system performance on AI generated images, denoted by BrierT ('T' for target).

AUC values close to 1 indicate good performance for discriminators but poor performance for generators. On the other hand, AUC scores less than or equal to 0.5 indicate poor performance for discriminators but good performance for generators; when the AUC value is 0.5, this suggests that the discriminator is no better than a random guesser. AUC values between 0.5 and 0 suggest that the discriminator is assigning lower confidence scores for more AI-generated content than for human-generated content.

BrierT scores close to zero indicate good performance for discriminates but poor performance for generators. BrierN scores close to 0 indicate good performance for discriminators. Since BrierN scores are Brier scores evaluated only using human-generated content, they do not carry any direct information regarding the ability of the generators to deceive the discriminators. A desirable outcome for discriminator systems is to have both Brier scores close to zero and for generator systems, it is to have BrierT scores close to one.

6.1 Receiver Operating Characteristic (ROC)

The receiver operating characteristic (ROC) curve is a graphical performance analysis tool. John A. Swet[1] provides detailed information about ROC curves for system evaluation. Here is a brief description of the curve. In what follows,

TP stands for True Positive (those correctly detected as AI-generated),
FN stands for False Negative (those incorrectly detected as non-AI (authentic)),
FP stands for False Positive (those incorrectly detected as AI-generated), and
TN stands for True Negative (those correctly detected as non-AI (authentic)).

The vertical axis is the True Positive Rate (TPR), where $TPR = TP/(TP + FN)$, the horizontal axis is the False Positive Rate (FPR), where $FPR = FP/(TN + FP)$, which is also known as the False Acceptance Rate or False Alarm Rate. Figure 1 illustrates the ROC curve example as the red curve.

[1]John A. Swets, Signal Detection Theory and ROC Analysis in Psychology and Diagnostics, Psychology Press, 2014 (https://doi.org/10.4324/9781315806167)

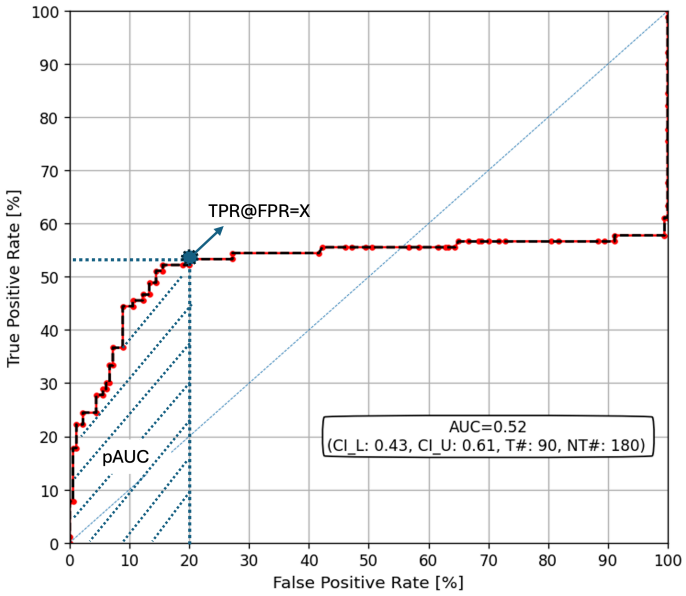

Figure 1: ROC and AUC

6.2 Area Under the ROC Curve (AUC)

The area under the ROC curve (AUC) is a score metric for the detection system. The AUC score quantifies the overall ability of a system to discriminate between two classes (in our case, the two classes as AI-generated content and human-generated content). The AUC score system has a value between 0 and 1. A system no better at identifying true positives than random guessing has an AUC of 0.5. A perfect system (no false positives or negatives) has an AUC of 1.0. A system that is worse than random guessing will have an AUC value less than 0.5.

Partial AUC (pAUC) is AUC at a specified False Positive Rate (FPR), shown as the shaded blue region under the ROC curve in Figure 1.

6.3 True Positive Rate (TPR) at False Positive Rate (FPR)

Another score metric used for the detection system is the True Positive Rate (TPR) rate at a specified False Positive Rate (FPR), namely TPR@FPR=x. It is illustrated as the blue point in Figure 1. In this task, we plan to use a True Positive Rate at a False Positive Rate equal to 0.1.

6.4 Detection Error Tradeoff (DET) and Equal Error Rate (EER)

The Detection Error Tradeoff (DET) curve is used as one of the graphical performance analysis tools. The horizontal axis is the False Positive Rate (FPR) and the vertical axis is the False Negative Rate (FNR). Martin et al[2] provide detailed information about DET curves for detection system evaluation. Equal Error Rate (EER) is the point at which the False Positive Rate (FPR) and False Negative Rate (FNR) are equal. Figure 2 illustrates a DET curve and EER.

[2] Martin, A., Doddington, G., Kamm, T., Orlowski, M., Przybocki, M., "The DET Curve in Assessment of Detection Task Performance," Eurospeech 1997, pp 1895-1898.

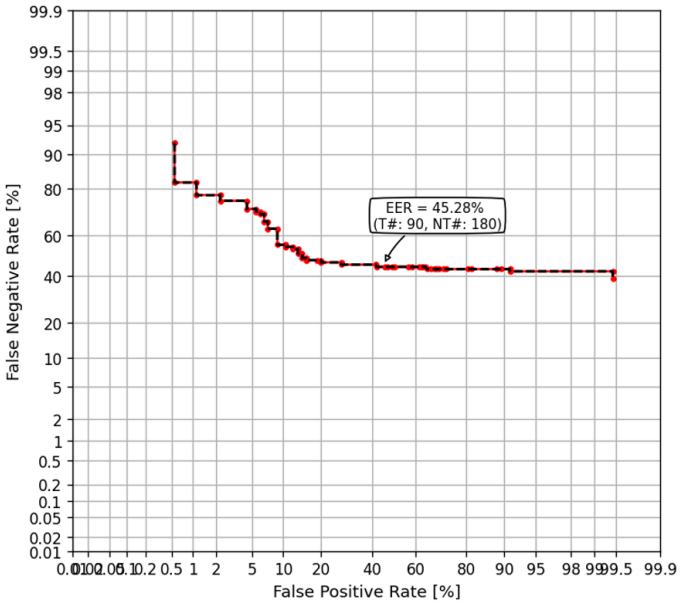

Figure 2: DET and EER

6.5 Brier Score

The Brier Score (BS)[3] is more like a cost function that measures how far your predictions are from the true values. It is usually used to calibrate the probabilities of the models and measures the mean square error between the predicted probability assigned to the possible outcomes for an event i and the actual outcome o_i.

$$BS = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2$$

where p_i is the prediction probability of occurrence of the event and o_i is equal to 1 if the event occurred and 0 if not.

In our experience we have found that it is informative to report two separate Brier scores, one that we call BrierT ('T' for target) and another we call BrierN ('N' for nontarget). These are defined as follows.

$$BrierN = \frac{1}{n0}\sum_{i=1}^{n0} p_i^{\ 2}$$

$$BrierT = \frac{1}{n1}\sum_{i=1}^{n1}(p_i - 1)^2$$

where n0 = number of human generated images and n1 = number of AI-generated images.

[3] Brier GW "Verification of forecasts expressed in terms of probability" Mon Weather Rev 1950.

APPENDIX A DISCRIMINATOR VALIDATOR AND SCORER USAGE

D-Validator Script Usage

validate Image-D system output

$ python validate_discriminator_sysout.py -x /path/to/index.csv -s /path/to/sysout.csv

D-Scorer Script Usage

run DetectionScorer with system output and reference files.

$ python DetectionScorer.py -t detection \
-r /path/to/genai25_Image-D_detection_ref.csv \
-x /path/to/genai25_Image-D_detection_index.csv \
-s /path/to/genai25_Image-D_detection_sysout.csv \
--plotType [det, roc]