2024 NIST GenAI (Pilot)
Evaluation Plan for Discriminators
Text-to-Text (T2T)

Released: 2024-04-01

Updated: 2024-08-30

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

2024-05-02 Updated Protocol and Rules

2024-06-17 Updated the GeneratorID and DiscriminatorID definition

2024-07-08 Updated the system output filename to be included the cutoff (threshold)

2024-08-30 Updated 6.2 Area Under the ROC Curve (AUC) description and removed Schedule Section.

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 the digital age, including issues related to mis/disinformation and trustworthy information in digital content. With the advancements in generative AI technology, it is becoming increasingly difficult to distinguish AI-generated from human-generated, which can potentially cause an information crisis.

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 (e.g., allies/adversaries) and AI content discriminators (e.g., detectors /defenders). 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 should benefit future research directions and approaches to understanding cutting-edge technologies as well as sources for recommendations and guidance for responsible and safe 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).

In this 2024 GenAI pilot study, the Text-to-Text Discriminators (T2T-D) task is a detection evaluation to measure how well systems can automatically detect AI-generated content vs human-generated content in multiple modalities, such as text, audio, image, and video. In this document, the task will focus on the text modality only.

The pilot GenAI evaluations provide data, including dry-run sets and test sets, created by both G-participants and the NIST GenAI team. This allows D-participants to validate/develop and run a system on their own hardware platform. Discriminator participants can then submit their system outputs to a web-based leaderboard, where scores and results are 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 approves the data. However, NIST will provide pilot data generated by the NIST GenAI team, aiming for D-participants to start the development of their systems. NIST reports performance measures for D-participant system outputs, displayed through a leaderboard, using either NIST pilot data or the evolved G-participants data. 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 understand system behavior in detecting AI-generated vs human-generated content. This includes characteristics of undetectable AI-generated content, how human content differs from AI content, and how the conclusions of the task can potentially provide guidance to end users to help differentiate between the two types of content they may encounter in their day-to-day lives.

2.1 Detection task

The pilot GenAI task for D-participants is text-to-text discriminator (T2T-D) detection, a detection task focused on determining if a target output was generated by Generative AI or humans. The T2T-D detection task will detect if a target text summary was generated using large language models (LLMs) such as ChatGPT.

For each T2T-D trial consisting of a single summary, the T2T-D detection system must render a confidence score (any real number), with higher numbers indicating the target text summary is more likely to have been generated using LLM-based models. The primary metric for measuring detection performance will be the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) (see Section 6).

2.2 Protocol and Rules

The evaluation participants agree not to probe the test media (text, audio, image, or video) 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 are NOT allowed to use publicly available NIST data, however, they may use other 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.

Each participant is allowed to submit system output for evaluation only once per 24-hour period.

All trials must be processed independently of each other within a given task and across all tasks, meaning content extracted from the data must not affect the processing of other data.

While participants may report their own results, participants may not make advertising claims about their standing in the evaluation, 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 with the team names anonymized for conditions of interest. Participants may publish or otherwise disseminate these charts unaltered and with appropriate reference to their source.

3 Data Resources

A text summary dataset generated by the NIST GenAI team, named "GenAI Pilot Data", will be used to compile the dry-run (validation purpose) and/or test sets for D-participants. 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, one at a time; the GenAI team will facilitate the evolution of the G-participants data creation process depending on a given task.

The text summary data, created by either humans or LLMs, given a topic and a set of about 25 relevant documents, will serve as the testing data for D-participants who will work on detecting whether the written content is human-generated or AI-generated.

NIST will make all necessary data resources available to D-participants. Each team will receive access to data resources upon completing all needed data agreement forms and based on the published schedule of each task data release date. Please refer to the schedule for T2T data release dates.

GenAI Pilot Data will be organized in a directory structure as follows:

README.md	A helpful documentation file in markdown format
genai24_T2T-D_detection_ref.csv	The reference file with ground truth for the text-to-text discriminator (T2T-D) detection task
validation.json	The recommended split of file_id into validation subsets
annotations/	A subdirectory with metadata (format should be decided later)
text/	A subdirectory containing text summary trials organized by <modality_id>/<task_id>/<dataset_id>/<topic_id>/<file_id>.txt

The following format constitutes the reference file for the T2T-D detection task.

genai24_T2T-D_detection_ref.csv	
DatasetID	(string) The ID of the dataset release (e.g., GenAI24-PL-set1)
TaskID	(string) The globally unique ID of tasks. Tasks could be summarization, generation, translation, question-answering (e.g., Detection)
TopicID	(string) The globally unique ID of the topic (e.g., D0701)
FileID	(string) The globally unique ID of the text summary trials (e.g., xxx_000011.txt)
GeneratorID	(string) The site name of generator participants (e.g., G_NIST_site)
IsTarget	AI-generated ('Y') and Human-generated ('N')

Example of the CSV file with delimiter "|".

DatasetID	TaskID	TopicID	FileID	GeneratorID	IsTarget
GenAI24-PL-set1	detection	topic_0000	file_0001.txt	G_NIST_site	Y

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
	genai24_T2T-D_dryrun_index.csv	The system input file (dry-run set) for the text-to-text discriminator (T2T-D) detection task
	files/	A flat subdirectory containing trials organized by <file_id>.*

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

DatasetID	TaskID	FileID
GenAI24-dryrunset	detection	file_0001.txt

Example of files

- Filename: file_0001.txt
- Contents: Plain text without new line delimiter or new paragraph delimiter
  (e.g. "Morris Dees and the Southern Poverty Law Center (SPLC) played a central role in a series of articles detailing their legal actions against the Aryan Nations", …)

3.2 Test Set

The GenAI test set is delivered as a single gzipped tarball for the T2T-D detection task, which is unpacked to the following contents and subdirectory structure:

	README.md	A helpful documentation file in markdown format
	genai24_T2T-D_detection_index.csv	The system input file for the text-to-text discriminator (T2T-D) detection task
	files/	A flat subdirectory containing trials organized by <file_id>.*

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

DatasetID	TaskID	FileID
GenAI24-PL-set1	detection	file_0001.txt

Example of files

- Filename: file_0001.txt
- Contents: Plain text without new line delimiter or new paragraph delimiter
  (e.g. "Morris Dees and the Southern Poverty Law Center (SPLC) played a central role in a series of articles detailing their legal actions against the Aryan Nations", …)

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 (texts or images) as input(s), systems perform detection tasks.

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

genai24_T2T-D_detection_index.csv	
DatasetID	(string) The ID of the dataset release (e.g., GenAI24-PL-set1)
TaskID	(string) The globally unique ID of tasks. Tasks could be summarization, generation, translation, question-answering (e.g., detection)
FileID	(string) The globally unique ID of the text summary trials (e.g., file_0001.txt)

Example of the CSV file with delimiter "|".

DatasetID	TaskID	FileID
GenAI24-PL-set1	detection	file_0001.txt

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 with 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., `genai24_t2t_d_sys_model-01.csv`). The system output CSV file for the T2T-D detection task must follow the format below:

genai24_t2t_d_sys_model-01.csv	
DatasetID	(string) The ID of the dataset release, e.g., GenAI24-PL-set1
TaskID	(string) The ID of the summary files, 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_GPT4.0
FileID	(string) The globally unique ID of the text summary trials (e.g., file_0001.txt)
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	ConfidenceScore
GenAI24-PL-set	detection	D-NIST-site	MySys_GPT4.0	file_0001.txt	0.7

5.2 Validation/Submission

5.2.1 Validation

The file_id column in the system output [submission-file-name].csv must be consistent with the file_id 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 on the webpage (https://ai-challenges.nist.gov/t2t). 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

NIST GenAI team may 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.

6.1 Receiver Operating Characteristic (ROC)

The receiver operating characteristic (ROC) curve is a graphical performance analysis tool. Macmillan and Creelman[1] provide 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 human-generated),
FP stands for False Positive (those incorrectly detected as AI-generated), and
TN stands for True Negative (those correctly detected as human-generated).

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

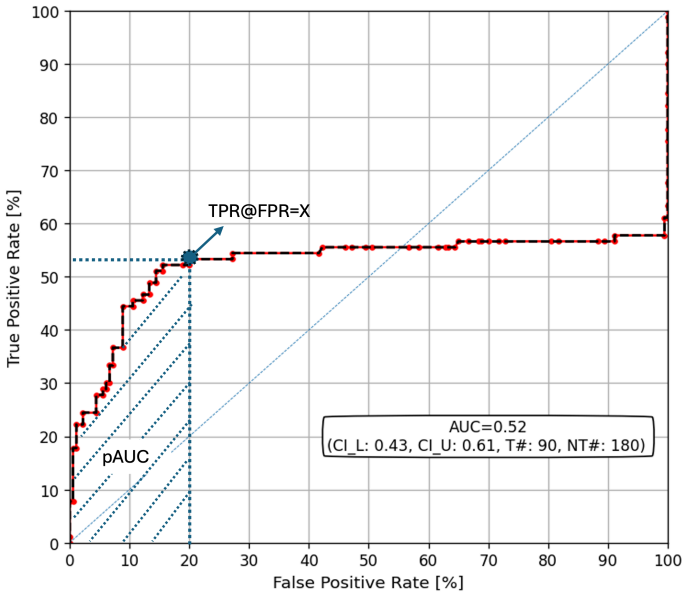

Figure 1: ROC and AUC

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. The AUC value of a system output has a value between 0 and 1.0. 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.

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

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 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.

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 at 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.

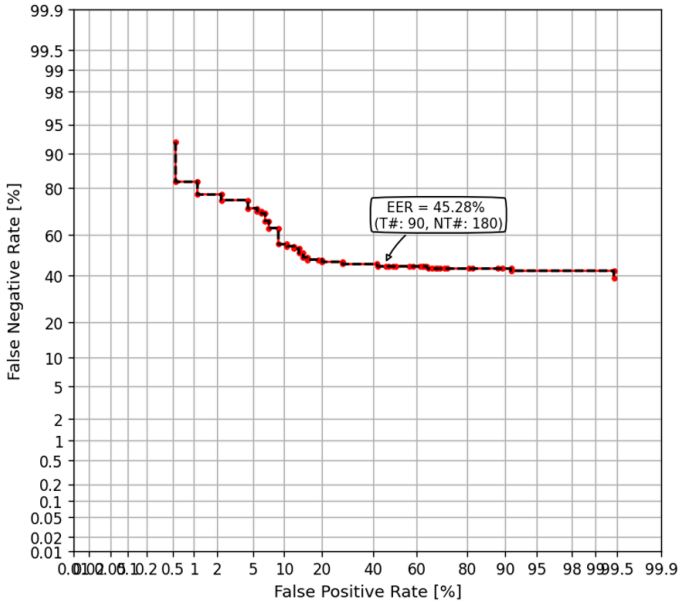

Figure 2: DET and EER

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.

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

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

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

Appendix A DISCRIMINATOR VALIDATOR AND SCORER USAGE

D-Validator Script Usage

validate T2T-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/genai24_T2T-D_detection_ref.csv \
-x /path/to/genai24_T2T-D_detection_index.csv \
-s /path/to/genai24_T2T-D_detection_sysout.csv \
--plotType [det, roc]