2024 NIST Generative AI (GenAI)
Data Creation Specification for Generators
Text-to-Text (T2T)

Date: 2024-04-01

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.

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 understand cutting-edge technologies as well as sources for recommendations and guidance for responsible and safe use of digital content.

The pilot study of the 2024 GenAI evaluation will focus on the text modality in this document. In the pilot GenAI generator task, the objective of Text-to-Text Generators (T2T-G) is to automatically generate high-quality summaries given a statement of information needed ("topic") and a set of source documents to summarize. On the other side, the pilot Text-to-Text Discriminators (T2T-D) task is to detect if a target output summary was generated using a Generative AI system or a Human. The context of this evaluation assumes completely AI-generated content (ignoring cases where humans use AI tools to co-author content such as rephrasing, grammar correction, editing, etc.). Please see the discriminator evaluation plan for the T2T-D task for more details.

Participants are required to indicate if they are participating as a generator team, a discriminator team, or both. This document describes the task specification for "generator teams". Datasets (e.g., source articles) created by the NIST GenAI team will be available to generator participants as input for their data creation. Discriminator participants will run a detection system on generator data outputs using their own hardware platform and submit their detection system outputs to NIST for scoring and displaying results.

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 for 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 provide guidance to end users to help differentiate between the two types of content they may encounter on a daily basis. This pilot evaluation does not address the differentiation between "factual" and "fake", however, this remains a potential topic for research of interest for future challenge problems.

NIST GenAI is a series of Generative AI evaluations, and every evaluation will tackle a different task depending on the research interest of the AI community.

2.1 Text-to-Text Generators (T2T-G)

The T2T-G task for the generative AI models is: Given a topic and a set of about 25 relevant documents as input, create from the documents a brief, well-organized, fluent summary output which answers the need for information expressed in the **topic statement**. Participants should assume that the target audience of the summary is a supervisory information analyst who needs the summary in order to inform decision-making.

- All processing of documents and generation of summaries must be automatic.
- The summary can be no longer than 250 words (whitespace-delimited tokens).
- Summaries over the size limit will be truncated.
- No bonus will be given for creating a shorter summary.
- No specific formatting other than linear is allowed (e.g., plain text).

There will be about 45 topics in the test data for generator teams. This set of summaries from all generator teams will serve as the testing data for discriminator teams, who will work on detecting whether the written content is human-generated or AI-generated.

The summary output will be evaluated by determining how easy or difficult it is to discriminate AI-generated summaries from human-generated summaries, i.e., the goal of generators is to output a summary that is indistinguishable from human-generated summaries.

For more information and details about the task specifics for generator teams, please refer to Appendix A.

2.2 Protocol and Rules

The participants are NOT allowed to use the test dataset for purposes of training, modeling, or tuning their algorithms. All machine learning or statistical analysis algorithms must complete training, model selection, and tuning prior to running their system on 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.

Each trial consists of a topic and its corresponding source documents. 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 another task's data.

While participants may report their own results, they may not make advertising claims about their standing in the evaluation, regardless of rank, winning the evaluation, or claiming NIST endorsement of their system(s). The following language in the U.S. Code of Federal Regulations (15 C.F.R. § 200.113)14 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 may generate a report summarizing the system results with the anonymized team names. Participants may publish or otherwise disseminate these charts unaltered and with appropriate reference to their source.

3 Data Resources

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

4 Schedule

Date	Generators (G)
April 15, 2024	Data Specification available
May 1, 2024	Registration period open
June 3, 2024	NIST source article data available for Round-1
August 2, 2024	Round-1 data submission deadline
September 2, 2024	G-Scorer results for the Round-1 data available (Leaderboard)
September 3, 2024	NIST source article data available for Round-2
October 18, 2024	Round-2 data submission deadline
December 13, 2024	G-Scorer results for the Round-2 data available (Leaderboard)
January 2025	Close
February 2025	Results release for both G and D
March 2025	Pilot GenAI evaluation workshop

Appendix A Text-to-text Generator (T2T-G) Task and submission

A-1. Data Generation Instructions

NIST human assessors developed topics of interest. Each assessor created a topic and chose a set of 25 documents relevant to the topic. The testing dataset documents came from a corpus comprising multiple newswire articles. Topics and relevant documents will be distributed by NIST. Only GenAI generator participants who have completed and submitted all required data agreement forms will be allowed access. As the example below shows, each topic includes an id (num), title, and the required topic statement (narr). The "docs" tag indicates the source relevant documents to be used when generating the required summaries. Please check the published schedule (Section 4) for testing data release dates.

Example of topic:

```
<topic>
<num> topic_5445 </num>
<title> North Medical Center  </title>

<narr>
Describe the activities of John Smith and the North Medical Center.
</narr>

<docs>
article_2318
article_1721
article_1619
</docs>
```

A-2. Data Submission Guidelines

- Each team may submit up to 5 runs for a data generation package. Each run should include one summary per topic. We are aware of possible interactions between prompt generations and LLM outputs which we plan to address in the next round study.

- Each run should contain summaries for all topics; a run can not skip a topic or submit summaries for a subset of the topics.

- Each summary should be 250 words or less. Summaries of more than 250 words will be truncated.

- Summary content should be free from offensive text or inappropriate remarks. NIST has the right to exclude any summary or whole runs if the content proves to be inappropriate for the general public.

- Each run should include high-level metadata to characterize the generator system as requested by the below run format and DTD file. As explained in the DTD file, teams need to provide some required information/parameters, such as:

    - teamName: The name of the team as registered on the NIST GenAI website

    - trainingData: Name of training dataset or collection of different datasets or source data

- version: The version of their model
- priority: The priority of the submitted run (the lower number, the higher the priority). For any required manual review of submissions, NIST may need to limit effort to only the highest priority runs.
- trained: A boolean (T or F) to indicate if the run was the output of a trained system by the team specifically for this task (T) or the output of an already existing system that the team used to generate the outputs (F)
- desc: A high-level description of the system that generated this run
- link: A link to the model used to generate the run (e.g. GitHub, etc)
- topic: The topic id (the "num" field in the topic xml file)
- elapsedTime: The processing time of the model per topic to generate the summary after the topic and documents were given to it.

Example of a sample run:

```
<!DOCTYPE GeneratorResults SYSTEM "GeneratorResult.dtd">
<GeneratorResults teamName="participant_1">
  <GeneratorRunResult trainingData="OpenAI" version="1.0" priority="1" trained="T" desc="This run uses the top secret x-component" link="TBD">
    <GeneratorTopicResult topic="1" elapsedTime="5">
      this is a 250-word summary of topic 1
    </GeneratorTopicResult>
    <GeneratorTopicResult topic="2" elapsedTime="5">
      this is a 250-word summary of "topic 2"
    </GeneratorTopicResult>
    <!-- ... -->

    <GeneratorTopicResult topic="40" elapsedTime="5">
      this is a 250-word summary of topic 40
    </GeneratorTopicResult>
  </GeneratorRunResult>
</GeneratorResults>
```

A-3. Generator Data submission validation

- NIST will provide, prior to submission dates, a validator script to participants to validate their output XML file format as well as content specific to the task guidelines (e.g. topic ids, empty required attributes, etc). All generator teams should validate their runs before submitting them to NIST. Example of available DTD validators (via a shell script): xmllint --valid simple_sample.xml
- **Submission instructions:** according to the published schedule, the submission form will be open and available (via the GenAI website) for teams to submit their data outputs based on the specified

format. Please make sure to follow the schedule and submit on time, as extending the submission dates may not be possible.

- Upon submission, NIST will validate data outputs uploaded and report any errors to the submitter.

- Please take into consideration that submitting your data outputs indicates and assumes your agreement to the data transfer agreement and rules of behavior.

A-4. Metrics for Generator Data Outputs

Discriminator system detection scores will not be available until D-participants submit their results on G-participants' data submissions. Hence, after Round-1 for G-participants, we will be able to report only some key summary statistics of interest. However, after Round-2, we will be able to report performance measures for G-participant systems with respect to the task (described in Section 2.1) of generating text that is indistinguishable from human summaries. Please refer to the published schedule.

As stated in Section 2.1, our main interest is in evaluating the ability of humans and/or state-of-the-arts (SOTA) algorithms to discriminate between AI-generated summaries and human-generated summaries. More specifically, we want to assess the probability that a SOTA AI-generator can defeat a SOTA AI-detector and/or a human-detector.

Metrics used in the pilot study only evaluate the performance of humans plus LLMs together as a system. In a future study, we plan to investigate the interaction between human expertise in prompt generations and LLM outputs.

Simultaneously, we want to assess the probability that a SOTA discriminator system will identify a SOTA AI-generated output. This assessment can be done using data from this GenAI evaluation once we have all the discriminator scores available after conducting experiments on the discriminator systems using AI-generated and human-generated summaries. However, there are many metrics available in the literature that attempt to evaluate how good an AI-generated summary is. This evaluation is done by comparing summaries to the source documents. Such evaluation can also be done by comparing selected features from AI-generated summaries with the same features from human-generated summaries. The metrics listed below have been proposed in the literature for such assessments. We intend to use these metrics to investigate the possibility of building an automatic classifier for discriminating between AI-generated summaries and human-generated summaries. We will investigate the possibility of a fusion metric that is a good discriminator. Discrimination capability will be assessed using DET curves (or ROC curves) constructed from our empirical data.

Unlike most of the metrics available in the literature, the following two metrics – GLTR (Giant Language model Test Room) and RADAR (Robust AI-Text Detection via Adversarial Learning) – are designed to directly classify a document as AI-generated or human-generated. These will also be computed and reported after each stage of the GenAI evaluation.

A subset of the following automatic G-metrics (also called G-scorer) will be adopted to evaluate generator teams' data outputs:

1. **Syntactic Evaluation - automatic analysis of syntactic complexity** (https://www.benjamins.com/catalog/ijcl.15.4.02lu)

2. **BERTScore (Bidirectional Encoder Representations from Transformers Score)** - similarity score for each token in the candidate sentence with each token in the reference sentence using contextual embeddings (https://arxiv.org/pdf/1904.09675.pdf)

3. **BLEU (BiLingual Evaluation Understudy)** - A machine translation based n-gram overlap evaluation metric (https://aclanthology.org/P02-1040.pdf)

   This score is based on the idea of "modified word n-gram counts" when quantifying "precision" (number of words in a candidate sentence that appear in a reference sentence). The BLEU score is a geometric average of modified n-gram counts for n=1,2,...,N (N pre-assigned) penalized by what the authors call a "brevity penalty" (BP).

4. **METEOR (Metric for Evaluation of Translation with Explicit ORdering)** - A machine translation evaluation based on a generalized concept of unigram matching between the machine and human reference translations (https://aclanthology.org/W05-0909.pdf)

   This uses an alignment process between a reference document and a candidate document and then calculates a "penalized" F-score where the F-score is a weighted harmonic mean of precision and recall.

5. CIDEr (Consensus-based Image Description Evaluation)- An image description evaluation metric based on human consensus (https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.pdf)

   This is based on cosine similarity of n-grams of words between a candidate document and a reference document. The authors recommend an average score based on n=1, 2, 3, and 4. The score is a "consensus score" in the sense that it is an average over a number of reference documents. That is, each candidate document is compared with everyone of multiple reference documents, and an average is taken.

6. **CHRF (Character n-gram F-score)** - A metric based on Character n-gram F-score (https://www.statmt.org/wmt17/pdf/WMT70.pdf)

$$\mathrm{CHRF}\beta = (1+\beta^2)\frac{\mathrm{CHRP}\cdot\mathrm{CHRR}}{\beta^2\cdot\mathrm{CHRP}+\mathrm{CHRR}}$$

   β equal to 1 and 3 were considered by the authors.

   CHRP = Percentage of n-grams in the hypothesis which have a counterpart in the reference

   CHRR = percentage of character n-grams in the reference which are also present in the hypothesis.

7. Sentence Mover's Similarity - A metric based on word and sentence embeddings (https://aclanthology.org/P19-1264.pdf)

8. SummaQA - A metric based on Question Answering (https://aclanthology.org/D19-1320.pdf)

9. **SUPERT (SU**mmarization evaluation with **P**seudoreferences and b**ERT)** - A metric based on selected salient sentences from the source documents, using contextualized embeddings and soft token alignment techniques (https://aclanthology.org/2020.acl-main.124.pdf)

   An extension of the BERTscore . We measure the relevance of a summary in two steps: (i) identifying the salient information in the input documents to build a pseudo reference summary, and (ii) measuring the semantic overlap between the pseudo reference and the summary to be evaluated. The resulting evaluation method is called SUPERT.

10. BLANC(Bacronymic Language model Approach for summary quality estimatioN) - Measures the performance boost gained by a pre-trained language model with access to a document summary while carrying out its language understanding task on the document's text (https://arxiv.org/pdf/2002.09836.pdf)

11. **Misc. statistics (extractiveness, novel n-grams, repetition, length)** - https://aclanthology.org/N18-1065/

12. GLTR ((Giant Language model Test Room): A tool to support humans in detecting whether a text was generated by a model (http://gltr.io)

13. RADAR (Robust AI-Text Detection via Adversarial Learning): A robust AI-text detector via adversarial learning (https://radar.vizhub.ai)

A-5. Data Agreement

All generator teams submitting data generation outputs will be required to complete and sign a Data Transfer Agreement and a DUC (Document Understanding Conferences) Data Usage Agreement before uploading their data outputs.

Appendix B Generator Validator Script

G-Validator Script Usage

validate T2T-G system output

$ python validate.py -t summarization -s genai24_T2T-G_summarization_sysout.xml