Supplemental Material For GenAI Arena

Dongfu Jiang* **Max Ku*** **Tianle Li***
Yuansheng Ni **Shizhuo Sun** **Rongqi Fan** **Wenhu Chen**
University of Waterloo
{dongfu.jiang, m3ku, t29li, wenhuchen}@uwaterloo.ca

1 Dataset Accessibility

- **URL to website/platform where the dataset/benchmark can be viewed and downloaded by the reviewers:**
  - Platform: https://hf.co/spaces/TIGER-Lab/GenAI-Arena
  - Dataset: https://huggingface.co/datasets/TIGER-Lab/GenAI-Bench
- **URL to Croissant metadata record documenting the dataset/benchmark available for viewing and downloading by the reviewers:**
  Croissant metadata can be downloaded on https://huggingface.co/datasets/TIGER-Lab/GenAI-Bench.
- **To reproduce our experimental results, see our HFSpace repository at:**
  https://huggingface.co/spaces/TIGER-Lab/GenAI-Arena/tree/main
- **The persistent dereferenceable identifier (DOI) of our dataset:**
  10.57967/hf/2499
- **License:**
  MIT License is applied.

2 Datasheets for Datasets

To illustrate the Dataset documentation and intended uses, we have completed the following fill-in of the Datasheets for Datasets:

2.1 Motivation

- **For what purpose was the dataset created?**
  To foster the research in aligning diffusion models further and analyze the user preferences.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
  TIGER Lab, University of Waterloo
- **Who funded the creation of the dataset?**
  University of Waterloo

2.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

The dataset consists of user votings on AI image generation, AI image edition, and AI video generation.

- **How many instances are there in total (of each type, if appropriate)?**
  - AI image generation: 1740
  - AI image edition: 919
  - AI video generation: 1070
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
  All possible instances.
- **What data does each instance consist of?**
  Each instance consists of 2 image/video outputs, the two model names, input prompt(s), and a user voting.
- **Is there a label or target associated with each instance?**
  Yes. Each row contains a label, which is the user voting.
- **Is any information missing from individual instances?**
  No.
- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**
  No. Each instance is treated independently.
- **Are there recommended data splits (e.g., training, development/validation, testing)?**
  No.
- **Are there any errors, sources of noise, or redundancies in the dataset?**
  Yes, there may be some noise in the dataset, because the voting collection process is public.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**
  Yes, it is self-contained.
- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**
  No.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
  We applied Llama Guard as an NSFW filter to ensure the dataset is appropriate for a wide range of audiences and protects users of the benchmark from exposure to potentially harmful or offensive content. However, the filter might have missed some NSFW elements in the filtering.

2.3 Collection Process

- **How was the data associated with each instance acquired?**
  The data are acquired from the logs of the GenAI-Arena platform.
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?**
  The data was collected through the GenAI-Arena platform, where its publicly available to let users to do the voting.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
  No, it is not a sample from a larger set.
- **Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?**

The general public is involved in the data collection process. GenAI-arena is a platform where visitors do their voting voluntarily.

- **Over what timeframe was the data collected?**
  In 4 months. The data collection period for our study spanned from February 2024 to June 2024.
- **Were any ethical review processes conducted (e.g., by an institutional review board)?**
  N/A

2.4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
  We applied Llama Guard as an NSFW filter to ensure the dataset is appropriate for a wide range of audiences and protects users of the benchmark from exposure to potentially harmful or offensive content.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**
  No.
- **Is the software that was used to preprocess/clean/label the data available?**
  Llama Guard is publicly available.

2.5 Uses

- **Has the dataset been used for any tasks already?**
  Yes, the dataset is used to evaluate the effectiveness of MLLMs as an evaluator.
- **Is there a repository that links to any or all papers or systems that use the dataset?**
  https://huggingface.co/datasets/TIGER-Lab/GenAI-Bench
- **What (other) tasks could the dataset be used for?**
  No.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
  No.
- **Are there tasks for which the dataset should not be used?**
  To fine-tune the Large Langauge model to improve the MLLM evaluation. This destroys the meaning of the benchmark.

2.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**
  Yes, its publicly available.
- **How will the dataset be distributed (e.g., tarball on the website, API, GitHub)?**
  On HuggingFace Dataset.
- **When will the dataset be distributed?**
  Its already available on https://huggingface.co/datasets/TIGER-Lab/GenAI-Bench
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
  The dataset will be distributed without any copyright restrictions and will be open source. It is permissible for anyone to use it for research, study, or other non-commercial purposes.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
  No.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
  No.

2.7 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**
  TIGER Lab, University of Waterloo
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
  {dongfu.jiang, m3ku, t29li, wenhuchen}@uwaterloo.ca
- **Is there an erratum?**
  No.
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**
  Yes, we will continuously update on the Hugging Face platform.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
  The dataset does not involve any personal information.
- **Will older versions of the dataset continue to be supported/hosted/maintained?**
  Dataset will be updated from time to time when more votes are collected.
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
  People can build on top of the dataset as long as they cite our work, according to MIT license.