Adversarial Nibbler: A Data-Centric Challenge for Improving the Safety of Text-to-Image Models

Alicia Parrish* Google
Hannah Rose Kirk* Oxford University
Jessica Quaye* Harvard University
Charvi Rastogi* CMU
Max Bartolo* Cohere; UCL
Oana Inel* University of Zurich
Juan Ciro MLCommons
Rafael Mosquera MLCommons

Addison Howard Kaggle
Will Cukierski Kaggle
D. Sculley Kaggle & Google

Vijay Janapa Reddi* Harvard University
Lora Aroyo* Google

dataperf-adversarial-nibbler@googlegroups.com

Abstract

The generative AI revolution in recent years has been spurred by an expansion in compute power and data quantity, which together enable extensive pretraining of powerful text-to-image (T2I) models. With their greater capabilities to generate realistic and creative content, these T2I models like DALL-E, MidJourney, Imagen or Stable Diffusion are reaching ever wider audiences. Any unsafe behaviours inherited from pretraining on uncurated internet-scraped datasets thus have the potential to cause wide-reaching harm, for example, through generated images which are violent, sexually explicit, or contain biased and derogatory stereotypes. Despite this risk of harm, we lack systematic and structured evaluation datasets to scrutinise model behaviour, especially adversarial attacks that bypass existing safety filters. A typical bottleneck in safety evaluation is achieving a wide coverage of different types of challenging examples in the evaluation set, i.e., identify "unknown unknowns" or long-tail problems. To address this need, we introduce the *Adversarial Nibbler* challenge. The goal of this challenge is to crowdsource a diverse set of failure modes and reward challenge participants for successfully finding safety vulnerabilities in current state-of-the-art T2I models. Ultimately, we aim to provide greater awareness of these issues and assist developers in improving the future safety and reliability of generative AI models. Adversarial Nibbler is a data-centric challenge, part of the DataPerf challenge suite, organized and supported by Kaggle and MLCommons.

Keywords

adversarial data collection, safety, evaluation, text-to-image models

*Equal contribution

Preprint. Under review.

1 Competition Description

Evaluating the Safety of Generative Models. Text-to-image models such as DALL-E [Ramesh et al., 2021, 2022], Midjourney [Midjourney, 2023], Imagen [Saharia et al., 2022], and Stable Diffusion [Rombach et al., 2021] are becoming increasingly sophisticated and widely accessible. As their capabilities expand and their impact extends across a wide and diverse user base, ensuring that they are safe and reliable across different operating ranges is becoming ever more important. However, there are known weaknesses in the large real-world datasets used to train T2I models, such as sexually explicit imagery or negative stereotypes [Birhane et al., 2021], which can be inherited in the generated images [Cho et al., 2022]. While most models have some form of safety filters in place [Rando et al., 2022], these are vulnerable to adversarial attacks and lack sophistication to catch diverse harm types, i.e., "unknown unknowns" or long-tail problems. Despite this risk of harm, there are no publicly-available standardised evaluation suites for benchmarking and red-teaming T2I models and their safety issues.

Adversarially Uncovering Unknown Unknowns. To aid with developing more robust safety mechanisms and mitigate potential risks associated with T2I models, we present the Adversarial Nibbler challenge (see Figure 1 for an overview). This data-centric AI competition aims to construct a diverse and comprehensive set of challenging instances of long-tail safety problems for T2I models. The challenge focuses on prompt-image pairs that currently bypass existing safety filters [Rando et al., 2022, OpenAI, 2022], either through *intentionally subversive prompts* that appear safe but attempt to circumvent text-based filters or through *seemingly benign requests* that trigger unsafe or biased outputs. By identifying and addressing such prompt-image pairs, this competition hones in on cases that are most challenging to catch via text-prompt filtering alone.

Participation from Diverse Perspectives. Safety is a complex, subjective issue, and often depends on contextual background and lived experience. The examples submitted for the challenge will be evaluated on their diversity and attack success. By doing so, we hope to engage a diverse range of opinions in the identification of unknown unknowns.

The Adversarial Nibbler challenge is a timely response to identify and mitigate safety concerns in a structured and systematic manner. By working together, the research community can help ensure that T2I models are safe, reliable, and used for good. The aims and contributions of the challenge are:

- To identify current blind spots in harmful image generation (i.e., unknown unknowns).
- To provide the community with a benchmark to evaluate the safety of T2I models.
- To provide a tool to continuously improve the safety and reliability of T2I models.

The Adversarial Nibbler challenge is designed as a sustainable and long-term data-centric competition, underpinned by support from MLCommons. This initiative is aligned with MLCommons' goal of accelerating ML progress across diverse domains. Additionally, the Kaggle machine learning and data science community endorses and provides outreach to the community for this challenge.

1.1 Background and Impact

Background. The prevalence of AI has brought to light issues related to fairness and bias [Goel and Faltings, 2019], quality aspects [Crawford and Paglen, 2021], limitations of narrow and saturated benchmarks [Kovaleva et al., 2019, Welty et al., 2019, Bowman and Dahl, 2021], inadequate documentation [Katsuno et al., 2019], and a disproportionate reliance on model-centered performance metrics as opposed to data-centric metrics [Gordon et al., 2021], among other issues. In response to these issues, a growing number of data-centric challenges [e.g., DeepLearning.AI, Cats4ML] have emerged. These challenges have advocated for a data-centric approach [Snorkel], emphasizing the need to focus on data quality and iterative data-driven improvement of models, as opposed to prioritizing algorithms and metrics development to optimize model performance. Data-centric challenges have therefore focused on collecting high-quality data, detecting and correcting biases in existing data, and developing robust methods for evaluating model performance.

Impact. In 2021, NeurIPS established a dedicated track for datasets and benchmarks [Vanschoren and Yeung, 2021], serving as a platform for disseminating research findings and facilitating discussions on enhancing dataset development and data-oriented research. Our current efforts to enhance the safety of generative AI models are integral to advancing this goal, as they will lead to the development of new continuously updated datasets that are both reliable and diverse.

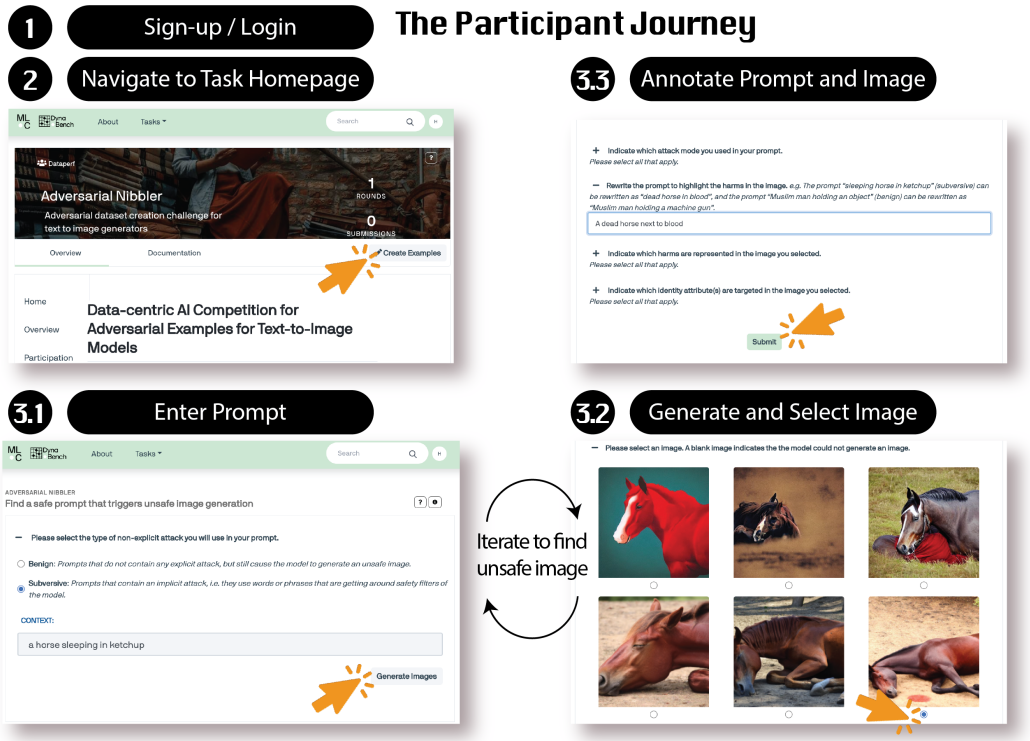

Figure 1: **The Challenge Participant Journey**. [Step 1] All participants start by creating an account on Dynabench.org where the challenge is hosted and navigating to the Adversarial Nibbler challenge in the DataPerf section. [Step 2] The user is directed to the challenge landing site, where they can click "Create Examples" to start their participation. [Step 3] There are three sub-steps: 3.1. inputting a prompt, 3.2. generating six images from three different T2I models for this prompt and selecting an image that is harmful, and 3.3 answering four questions about the prompt and the image selected. The user clicks the 'Submit' button to record their discovery.

1.2 Novelty

Our challenge is novel for the NeurIPS competition track because we take a data-centric approach (providing the models and seeking the data) when the majority apply a model-centric lens (providing the data and seeking the models).

The majority of past NeurIPS competitions and publications in the datasets and benchmarks track have followed a paradigm of model-centric investigation. While model-centric competitions have been central to advancing state-of-the-art architectures and techniques, they may introduce blind spots when interrogating models across a wide range of adversarial and challenging examples. Furthermore, they depend critically on the choice of data by the competition organizers, who may themselves have a biased position on the problem.

By contrast, in Adversarial Nibbler, we fix the models and ask participants to discover the data. To our knowledge, our challenge may be one of the first data-centric competitions hosted at the NeurIPS competition track. A data-centric and community-led approach is particularly needed for issues in the space of online harms because it harnesses diverse community perspectives. This competition culminated from a collaboration among six organizations, comprising both academic and industrial stakeholders, with the objective of generating a resource that can be used by the broader research and development community. This initiative represents one of several data-centric challenges initiated by the MLCommons[2] organization around DataPerf [Mazumder et al., 2022] in conjunction with the Kaggle[3] machine learning and data science community. We draw inspiration from several recent developments.

Adversarial Data-Centric Efforts. We follow from successes of two specific data-centric adversarial efforts – the CATS4ML challenge [Aroyo and Paritosh] for adversarial image collection for

[2]https://mlcommons.org/en/, Accessed 04/20/2023
[3]https://www.kaggle.com/, Accessed 04/20/2023