A Datasheet Describing a Use-Case Specific Dataset for Measuring Fairness, Toxicity and Veracity in LLM-generated Text

Alicia Sagae
Amazon AWS AI/ML
Seattle, Washington, USA
aksagae@amazon.com

Chia-Jung Lee
Amazon AWS AI/ML
Seattle, Washington, USA
cjlee@amazon.com

Sandeep Avula
Amazon AWS AI/ML
Seattle, Washington, USA
sandeavu@amazon.com

Brandon Dang
Amazon AWS AI/ML
Seattle, Washington, USA
dangbran@amazon.com

Vanessa Murdock
Amazon AWS AI/ML
Seattle, Washington, USA
vmurdock@amazon.com

This datasheet template is taken from the proposal by Gebru et al. [1]. For readability, we paraphrased some of the questions (although the responses are to the full original question), and omitted questions that are not applicable to the dataset. Some questions are covered in more depth in the accompanying paper "A Use-Case Specific Dataset for Measuring Fairness, Toxicity and Veracity in LLM-generated Text" by Sagae et al.

1 Motivation

1.1 For what purpose was the dataset created?

The dataset was created to evaluate generative AI systems on fairness, toxicity and veracity. To define each of these dimensions, it is necessary to first define a narrow use case. For example, a system that generates ad copy for children's Halloween costumes has a different standard for safety than a system that generates summaries of true crime novels for an adult audience.

1.2 Who created this dataset?

The dataset was created by a centralized Responsible AI team in Amazon AWS. The dataset was created as a public dataset, not in use in any production system in Amazon or AWS.

1.3 Who funded the creation of the dataset?

The dataset was funded by Amazon AWS in the sense that it was created by employees of Amazon AWS as part of their regular work, using compute infrastructure and other resources provided by Amazon AWS.

1.4 Other Comments

The data was sourced from publicly available information: product queries to `amazon.com` along with product details available on `amazon.com`. Details include product features, product description, and price. No seller or buyer information was used in the data sourcing process.

2 Composition

2.1 What do the instances that comprise the dataset represent?

Each instance represents a product in the `amazon.com` catalog at time of writing for the paper (June 2025). Each product was retrieved in response to a query, and the dataset also stores the query for each product. The query represents a target market for the product. A given product may have more than one target market, and appear once in the dataset for each query that retrieved it.

2.2 How many instances are there in total (and of each type)?

The total size is 7584 rows. This includes 5528 unique products (ASINs).

2.3 Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset contains a sample of products from the `amazon.com` catalog, based on submitting templated queries for product retrieval. We retrieved up to 40 products per query. Queries were designed to collect a diverse set of products that support comparison among product cohorts. Cohorts include identity groups, product categories, and product adjectives.

2.4 What data does each instance consist of?

Each item in the dataset corresponds to an item for sale, and includes an ASIN (product identifier), vendor-created title, vendor-created product feature list, vendor-created product description in English, as well as the query used to retrieve the product.

Queries include pairwise combinations among a product category (e.g., Beauty & Health), an identity group (e.g.,Women), and a product adjective (e.g., cute). Each product is annotated with the terms of the query used to retrieve it. If a term was not used, the product is annotated with the value “any”.

2.5 Is there a label or target associated with each instance?

Each item in the dataset is labeled with the identity group, product category, and adjective used to retrieve the product. A product can appear multiple times, associated with different queries.

2.6 Is any information missing from individual instances?

We downsampled the retrieved products to keep only products with a non-empty value in the vendor-provided description field and in the vendor-provided feature list field.

2.7 Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

All products that were retrieved for a particular cohort (e.g., Women) can be identified using the query annotations described above. All instances that contain the same product will have the same value in the `asin` field.

2.8 Are there recommended data splits (e.g., training, development/validation, testing)?

The dataset is intended specifically for evaluation purposes and should not be split.

2.9 Are there any errors, sources of noise, or redundancies in the dataset?

Products may appear more than once. Some vendor-provided descriptions contain formatting characters, such as asterisks for bullet points. Some vendor-provided descriptions contain a limited amount of HTML noise due to scraping artifacts.

2.10 Is the dataset self-contained, or does it link to or otherwise rely on external resources?

The dataset is self-contained.

2.11 Does the dataset contain data that might be considered confidential?

The dataset does not contain confidential information.

2.12 Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

The dataset includes product categories that were intended to test toxicity handling in large language models. These categories include Shooting; Knives, Parts, & Accessories; Weapons; Sexual Wellness; Tobacco-Related Products; and Lingerie.

2.13 Does the dataset relate to people?

The dataset associates products with target markets, identified by identity group names.

2.14 Does the dataset identify any subpopulations (e.g., by age, gender)?

The dataset includes 13 identity groups from the Toxigen dataset [2], identified in a bottom-up data labeling approach. They include African, Asian, Native American, Latino, Chinese, Mexican, Middle Eastern, LGBTQ+, Women, Mental Disabilities, Physical Disabilities, Jewish, Muslim, and "any"(wildcard).

2.15 Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

No.

2.16 Does the dataset contain data that might be considered sensitive in any way?

The data associates each product with an identity group that constitutes a target market for that product. This association was automatically derived from the `amazon.com` product retrieval engine.

3 Collection Process

3.1 How was the data associated with each instance acquired?

The data is directly observable from the `amazon.com` website.

3.2 What mechanisms or procedures were used to collect the data?

We used the `amazon.com` retrieval engine to identify products and to scrape the publicly available details for each product.

3.3 If the dataset is a sample from a larger set, what was the sampling strategy?

The sample is based on taking the first 40 products returned by the retrieval engine in response to a query, or fewer products if the engine returns fewer.

3.4 Who was involved in the data collection process?

Data collection was executed automatically by software written by full-time employees.

3.5 Over what time frame was the data collected? Does this time frame match the creation time frame of the data associated with the instances?

The data was created between March-June 2025.

3.6 Were any ethical review processes conducted (e.g., by an institutional review board)?

No.

3.7 Does the dataset relate to people?

The dataset does not involve human annotation or human subjects.

4 Preprocessing/cleaning/labeling

4.1 Was any preprocessing/cleaning/labeling of the data done?

The product pages were cleaned of HTML using the BeautifulSoup HTML parser, and products with empty vendor-provided descriptions or feature lists were removed.

4.2 Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?

This data is available internally to the research team.

4.3 Is the software used to preprocess/clean/label the instances available?

BeautifulSoup is an open source tool[1].

5 Uses

5.1 Has the dataset been used for any tasks already?

The dataset was used to evaluate two large language models in the accompanying paper.

5.2 Is there a repository that links to any or all papers or systems that use the dataset?

The paper will be made available in the GitHub repo after publication.

5.3 What (other) tasks could the dataset be used for?

The dataset could be used for a variety of tasks related to content generation, in the domain of product description and advertising. This domain supports a variety of specific use cases.

5.4 Are there tasks for which the dataset should not be used?

The dataset should not be used to reverse-engineer identity group labels for amazon products. The dataset should not be used to infer characteristics of an identity group from the products associated with it. Uses should be restricted to using the data as input (prompts) or ground-truth for evaluation of generative models. These models may include text, image-text, or other multimodal systems.

6 Distribution

6.1 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.

6.2 How will the dataset will be distributed ?

The dataset is available for download[2] under the Creative Commons BY 4.0 license.

6.3 When will the dataset be distributed?

The dataset was published to GitHub in May 2025.

6.4 Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The dataset is available for download under the Creative Commons BY 4.0 license.

6.5 Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

6.6 Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No.

7 Maintenance

7.1 Who will be supporting/hosting/maintaining the dataset?

Amazon AWS.

7.2 How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Comments can be submitted to the GitHub project. Authors will monitor the repository for comments.

7.3 Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Updates will be made as needed to the GitHub repository.

[1] https://www.crummy.com/software/BeautifulSoup/

[2] https://github.com/amazon- science/application-eval-data

7.4 If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

Pull requests to the GitHub repository will be monitored by the authors.

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. Issue 12. https://doi.org/10.1145/3458723

[2] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. TOXIGEN: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Vol. 1. https://doi.org/10.18653/v1/2022.acl-long.234