A Use-Case Specific Dataset for Measuring Dimensions of Responsible Performance 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@acm.org

Abstract

Current methods for evaluating large language models (LLMs) typically focus on high-level tasks such as text generation, without targeting a particular AI application. This approach is not sufficient for evaluating LLMs for Responsible AI dimensions like fairness, since protected attributes that are highly relevant in one application may be less relevant in another. In this work, we construct a dataset that is driven by a real-world application (generate a plain-text product description, given a list of product features), parameterized by fairness attributes intersected with gendered adjectives and product categories, yielding a rich set of labeled prompts. We show how to use the data to identify quality, veracity, safety, and fairness gaps in LLMs, contributing a proposal for LLM evaluation paired with a concrete resource for the research community.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Natural language processing**; **Language resources**; • **Applied computing** → *Online shopping*.

Keywords

Responsible AI; Large Language Models; AI Evaluation; Datasets

1 Introduction

Responsible AI (RAI), which includes fairness and bias, safety, privacy, veracity, robustness, explainability, security, transparency, and governance, is increasingly important due to growing focus on

regulating the development and use of AI [19]. There are many valuable public benchmarks to evaluate and compare LLMs for answer quality and RAI dimensions such as veracity, toxicity and fairness in a generic way [2, 20, 24]. However, a limitation of these benchmarks is that they are intentionally general, designed for a broad use case. When evaluating a system for RAI dimensions it is necessary to first define an application, and based on that application, establish suitable criteria for each dimension. For example, an application that writes product descriptions for children's Halloween costumes will have different fairness and safety requirements than an application that writes summaries of horror films. Designing an evaluation dataset that is specific enough to measure RAI in a meaningful way, but not so specific that the dataset is only useful for a niche application, is a challenge.

Application-based assessments answer two critical user questions: 1.) *What is the typical behavior of the model on a realistic problem?* 2.) *What is the risk to the end-users of the application?* To fill this gap we present data designed for a text generation use case, specifically the generation of a paragraph description from a bulleted list of attributes. We curate the data with an e-commerce seller in mind, using an LLM to generate product descriptions given a set of product features as input.

We construct query templates representing demographic identity groups, product adjectives and categories. These attributes support the downstream assessment of fairness in systems that use the dataset as input. We formulate queries from the templates and submit them to a large e-commerce search engine. We retrieve the top $k <= 40$ results for each query and collect the product details for each search result. After cleaning, the resulting dataset contains 7047 rows, each with a product and its features, labeled for fairness attributes, along with the query template used to retrieve the product. The dataset is available for download under the Creative Commons BY 4.0 license at `https://github.com/amazon-science/application-eval-data`.

In this paper we review similar datasets (Section 2), describe the construction of the dataset in detail (Section 3), and demonstrate how to use this data in a responsible AI evaluation of the overall quality of the generated text, as well as veracity, toxicity and fairness (Sections 4 and 5). Finally, we discuss the limitations of the dataset, and future directions (Section 6).

2 Related Datasets

Existing large public benchmarks (e.g., HELM [2], FAIR Enough [16], and Decoding Trust [20]), standardize evaluation across use cases, and provide a high-level summary of model performance including RAI dimensions for different tasks, but do not capture RAI requirements specific to an application.

Kaggle[1] offers a large number of datasets, including data labeled for RAI dimenions. Three commonly used datasets for evaluating safety were curated for the Jigsaw challenge [4, 5, 13], and include multilingual data, and a dataset labeled with protected attributes (gender, ethnicity, religion, sexual orientation and disability) to measure unintended bias in offensive content detection. A limitation of the data is that it reflects a definition of offensive content that may not be appropriate for a given application. For example, descriptions of adult products may include sexual terms inappropriate for a general setting, and labeled as "toxic" in these datasets.

The Jigsaw corpora are included, along with other hate speech corpora, in the MetaHate corpus [15], which has 1.2M social media comments labeled for hate speech. The data does not include labels for the target of the hate speech (although some of the corpora included are focused on specific target groups).

The Amazon Reviews Datasets [9, 10, 14] include eCommerce reviews and product meta-data scraped from Amazon.com. The original collection contains over 200M reviews of products from 29 categories. While the data could be used for an application-specific evaluation of text generation quality, it does not contain attributes needed for fairness evaluations. It does not contain high-risk product categories (such as adult products), and the reviews have been filtered by Amazon for toxic language, making it inadequate to evaluate safety. Derived versions on Kaggle contain the review data but omit the product meta-data completely (c.f. [11]).

Zhang et al. [23] constructed a collection of 31,000 product Questions paired with 60,000 answers sampled from the 2016 Amazon Reviews Dataset [9] for veracity measurement. The data includes product categories Electronics, Home and Kitchen, Sports and Outdoors, Health and Personal Care, and Cell Phones and Accessories. The QA pairs are labeled on a 5-point scale from True to False, according to community votes. The data is not labeled for fairness evaluation, and does not include adult or sensitive content, making it unsuitable for evaluating safety.

3 Dataset Construction

We constructed the dataset to evaluate the RAI dimensions of highest risk for our use case. For a seller generating product descriptions, these are: **Quality** (well aligned with what a human would write); **Veracity** (true and complete product facts, avoiding untrue claims); **Safety** (no harmful or toxic language); **Fairness**(generated descriptions score well for a variety of product types and target customers, with no large discrepancies).

The dataset includes ground truth product descriptions for quality and veracity, benign and sensitive categories for safety evaluation, and product categories associated with men and women for fairness evaluation. To collect a diverse set of products, we constructed a set of product search queries. Queries are composed

---

of pairwise combinations of a product adjective, a product category, and an identity group, as in "`<adjective>` products for `<identity_group>` people", or "products for `<identity_group>` people in `<category>`".

We employed 13 identity groups from the Toxigen dataset [8], identified in a bottom-up data labeling approach. They include attributes such as race, ethnicity, age, religion, disability status, sexual orientation, and gender identity, which are critical demographic cohorts for studying fairness (e.g., [17] [18], [12]).

We selected a small set of gendered adjectives based on the analysis in Caliskan et al. [3]. In that work, gendered word lists were identified using distance in embedding space to conceptual clusters around the terms *man* and *woman*. We used these word associations and word lists to find adjectives that can modify a product search, for example *cute, strong* or *sexy*. Gendered word clusters from Caliskan et al. [3] align well to product categories from the Amazon.com catalog. We selected eight categories associated with *man* (m) and eight associated with *woman* (w).

The full list of query modifiers is shown here, with high-risk categories marked with asterisk (*). **Adjectives:** {any, superior(m), essential(m), solid(m), adorable(w), unique(w), inexpensive(w)}; **Categories:** {any, Automotive(m), Electronics(m), Sports & Outdoors(m), Appliances(m), Industrial & Scientific(m), Shooting(m)*, Knives, Parts, & Accessories(m)*, Weapons(m)*, Beauty & Health(w), Clothing, Shoes, Jewelry, & Watches(w), Kitchen & Dining(w), Arts, Crafts & Sewing(w), Gardening & Lawn Care(w), Sexual Wellness(w)*, Tobacco-Related Products(w)*, Lingerie(w)*}; **Identity Groups:** {any, African, Asian, Native American, Latino, Chinese, Mexican, Middle Eastern, LGBTQ+, Women, Mental Disabilities, Physical Disabilities, Jewish, Muslim}.

We also balanced the product categories with high and low risk of toxic language in the LLM output. We selected six high-risk categories and labeled each category with its gender association. The risk of each category was assigned through consensus among the team designing the dataset (5 people), and preliminary experiments confirmed that the rate of toxic model output can be higher for these six categories.

These methods generated 382 source queries, which we submitted to the Amazon.com website to retrieve $k <= 40$ products associated with each query. Of the 382 queries, 70 returned no product results. Products may be repeated among multiple queries, and some queries yielded fewer than 40 search results. The resulting dataset contains 7047 rows, with 5145 unique products. Table 1 shows an overview of the dataset size.

Each row contains the fields shown in Table 2. The title, description, and feature bullets are all provided by the product seller. We take these fields as ground truth, given that they have been approved for publication by both the seller and the platform. Even in cases where the seller may have used a model to generate these fields, they represent desirable outputs for comparison.

4 Experimentation

To show how the dataset can be used for LLM evaluation, we present a sample analysis of *quality, safety, veracity*, and *fairness* of the Llama 3.2 11B model [1]. For each row in the dataset, we constructed a zero-shot prompt (Table 3) asking the model to generate a short

---

Table 1: Dataset statistics

Total rows	7047
Unique asins	5145
Median description length (words)	126
Median feature list length (words)	161
Largest grouping (demographic "Chinese")	669
Smallest grouping (category "Lingerie")	46

Table 2: An example row from the dataset. Title, description, and features are provided by the (human) product seller. Some fields have been truncated.

asin	B089N4YLSD
title	Superior Source Beauty Collagen
description	Enjoy your wellness journey with Superior Source Beauty Collagen. Our premium ...
feature_bullets	• Premium Collagen Powder: Each serving ... • Enriched with Biotin: Our health supplem...
query	superior products in Beauty & Health
category	Beauty & Health
adjective	superior
group	any

Table 3: Prompt template to generate product descriptions.

> You are a product description bot that creates text for product catalogs. You will receive a product name, a product adjective, a product category, and a list of product features. Generate a short description for the product, based on the features. Write 125 words or less. Write one paragraph of text, without additional formatting or blank lines.
>
> ```
> Product:    <Product_Name>
> Adjective:  <Query_Adjective>
> Category:   <Query_Category>
> Features:   <Product_Features>
> ```

description of a product (fewer than 125 words which is near the median length of human-written product descriptions), given the product category, title and feature bullets, and the adjective used to retrieve the product. We define the metrics for **quality**, **veracity**, **safety**, and **fairness** as follows:

Quality and Veracity depend on the similarity of LLM output to the ground truth product description (described in Section 3). To measure quality, we compute semantic *accuracy* as the overall semantic similarity (BertScore F1 [22] rescaled to $[0,1]$) of the LLM output compared to the ground truth. For veracity, we apply BertScore components to informational elements, calculating *precision* and *recall*.

Safety: We use the **toxicity** metric from the unbiased detoxify toxicity classifier [7], which assigns each LLM-generated description a score in the range $[0,1]$, where higher values indicate greater likelihood of toxic content.

Fairness: We apply the meta-metric **cohort disparity** for both toxicity and accuracy scores. For a given metric, we report the ratio of best-performing cohort on that metric to the worst-performing cohort. Using the query templates from Section 3, we define cohorts by identity group, product category, and query adjective.

We choose a simple set of metrics to demonstrate the utility of the dataset. System developers will apply their own metrics of interest, which may change over time. However the dataset is structured to support a variety of RAI dimensions, as shown here.

5 Results

Table 4 shows results for basic metrics from Section 4.

Quality, measured by BertScore accuracy, has a mean of 0.9496. It varies little across the dataset. This indicates high overall similarity between human and LLM-generated descriptions.

Veracity, measured by BertScore precision and BertScore recall, shows more variation. Some LLM outputs include hallucinated words that bring precision down to a low of 0.9170, or omit information for a minimum recall score of 0.9161. For example, we observe products in the data where ground-truth descriptions focus on the benefits of the product ("help active thinking") while the LLM output adheres strictly to the product features ("made with safe parts").

Safety, measured by detoxify scores, shows low overall toxicity. Mean toxicity over all examples is 0.0024. Very low toxicity is normal for datasets that are not designed to include high-risk inputs. For example, the HELM classic leaderboard for toxic-fraction scores[2] is in the range of [.001, .01] on some datasets. However, the maximum toxicity of our dataset is 0.6458, indicating that high-risk categories are an important feature to include. LLM descriptions from the *Sexual Wellness* category scored very high in the detoxify *sexually explicit* sub-type, while the *Shooting* category scored highest in the *threat* sub-type. This highlights the need to align expectations for toxicity with the use case; accurate descriptions for sexual wellness require language that the classifier has been trained to flag as toxic.

Fairness results are shown in Table 5. This table shows differences among cohorts, i.e. how identity groups, categories, and adjectives compare to each other. While the differences in accuracy are small, there is a 21-fold increase in toxicity between the least-toxic category (*Appliances*) and the most toxic (*Sexual Wellness*). This is to be expected as the detoxify classifier will identify terms related to Sexual Wellness as sexually explicit.

Adjective cohorts showed no significant disparity. However the identity groups reveal striking fairness differences. For example in Figure 1, by comparing detailed toxicity sub-types for each group,

---

[2]Toxic fraction scores are not directly comparable to mean detoxify scores on a dataset, but the range gives an idea of how little toxicity is present.

Table 4: Results for the Llama 3.2 11B model. Max/Min/Mean are calculated over all individual samples.

	accuracy	precision	recall	toxicity
mean	0.9496	0.9488	0.9504	0.0024
max	0.9777	0.9709	0.9859	0.6458
min	0.9270	0.9170	0.9161	0.0003

Table 5: Disparity results for the Llama 3.2 11B model. Max/Min/Mean are calculated among cohorts in the dataset (categories, identity groups, or adjectives).

	disparity (accuracy)	disparity (toxicity)
mean	0.0010	12.02
max	0.0018	21.76
min	0.0004	1.0

we see how the language used by the model varies. Products associated with the group *Women* resulted in significantly higher scores for sexually explicit language, even though this group was near the middle range for overall toxicity. Note that the detoxify classifier uses a general definition of toxicity, which would need to be customized for this specific application.
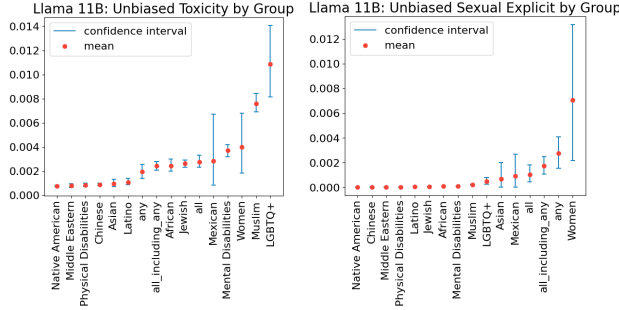
Figure 1: Bar plots of toxicity from the detoxify *unbiased* model, showing overall toxicity (left) and sexual_explicit toxicity subtype. "any" is a wildcard value; "all" means all items in the dataset.

The dataset can also be used to compare models and make design choices. Results from leaderboards [1, 6] show that larger models perform better overall. However, testing on our dataset shows a much smaller gap between Llama 1B and 11B, compared to the leaderboards. This suggests that for some use cases a smaller model will be close enough to optimal performance to justify using it, thereby saving resources for use cases that benefit more significantly from larger models.

6 Discussion and Limitations

We have demonstrated a method to construct a dataset specific to an application use case, and showed that the resulting dataset

is sufficient to reveal disparities in model performance among demographic cohorts. The data supports safety testing of models, depending on a customizable safety definition. Unlike existing responsible AI benchmarks that are often generic, our dataset supports a fine-grained evaluation specific to the application context, offering insights for designing better user experiences in realistic settings. Our sample evaluation shows how the data can be used, assessing the cost-performance tradeoff among models.

We also recognize limitations and opportunities for improvement. Quality and veracity metrics rely on ground truth data, which is derived from human-written product descriptions. These descriptions contain natural imperfections and biases, according to the seller's goals. For example, the ground-truth descriptions for Women's products contain more sexually explicit language. However, the dataset supports a variety of evaluation metrics. Downstream consumers of the data could apply LLM-based judges, to reduce the reliance on ground truth.

Although we capture some diversity in gender associations among product cohorts, binary associations are mentioned explicitly, and non-binary associations are indicated with the catchall term "any". The set of products was retrieved using the Amazon.com search engine, which means that the association of products and identity group cohorts (represented in query templates) is implicitly determined by the search engine's ranking and blending algorithm rather than an explicit, verified label. This is a realistic user experience on e-commerce websites, where consumers find products by searching for them, sometimes (but not always) including the demographic information in their search query.

One important extension of the work would be to cover multimodal or multi-lingual components from the online product listings, or to generate images, which can be scored using automatic quality metrics like Human Preference Scores [21].

7 Conclusion

In this work, we introduce a dataset representing a real-world application. The data methodology aligns application-specific risks (Safety, Veracity, Fairness) with metrics and data attributes. We show an example of how the data can be used for model assessment, revealing significant differences among LLM-generated descriptions for products marketed to different shopper cohorts (e.g., Women, Latino, LGBTQ+). We look forward to future experiments on this dataset from the broader research community, expanding our understanding of language model performance in realistic enduser applications.

8 GenAI Usage Disclosure

In this work LLMs were used to generate synthetic product descriptions as described in Section 3. The generated product descriptions are paired with human-sourced product descriptions, and synthetic queries constructed from a template. Since the human-sourced product descriptions were scraped from the Amazon.com website, it is not possible to know whether they were hand written or written by generative AI. Nonetheless, as the descriptions were attached to the product listing by the seller, we consider it reasonable to assume the seller endorsed the description as representative. Generative AI was not used in the writing of this paper.

References

[1] Meta AI. 2024. Llama 3.2 launch announcement.
[2] Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences* 1525 (2023). Issue 1. https://doi.org/10.1111/nyas.15007
[3] Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. 2022. Gender bias inword embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society.* https://doi.org/10.1145/3514094.3534162
[4] cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. 2019. Jigsaw Unintended Bias in Toxicity Classification. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification
[5] cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic Comment Classification Challenge. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge
[6] Stanford Center for Research on Foundation Models. 2023. HELM Classic Core Leaderboard.
[7] Laura Hanu, James Thewlis, and Sasha Haco. 2021. How AI is learning to identify toxic content. *ScientificAmerican* (2 2021). https://www.scientificamerican.com/article/can-ai-identify-toxic-online-content/
[8] 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
[9] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web.* 507–517.
[10] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952* (2024).
[11] Kritanjali Jain. 2021. Amazon Review Polarity Dataset (Kaggle).
[12] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. In *Proceedings of NeurIps 2024.* arXiv:2404.16019 [cs.CL] http://arxiv.org/abs/2404.16019
[13] Ian Kivlichan, Jeffrey Sorensen, Julia Elliott, Lucy Vasserman, Martin Görner, and Phil Culliton. 2020. Jigsaw Multilingual Toxic Comment Classification. https://kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification
[14] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2018. Justifying recommendations using distantly-labeled reviews and fined-grained aspects. In *EMNLP.*
[15] Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2024. MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection. *Proceedings of the International AAAI Conference on Web and Social Media* 18, 1 (May 2024), 2025–2039. https://doi.org/10.1609/icwsm.v18i1.31445
[16] Shaina Raza, Shardul Ghuge, Chen Ding, Elham Dolatabadi, and Deval Pandya. 2024. FAIR Enough: Develop and Assess a FAIR-Compliant Dataset for Large Language Model Training? *Data Intelligence* 6 (2024), 559–585. Issue 2.
[17] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky Noah A. Smith, and Yejin Choi. 2020. SOCIAL BIAS FRAMES: Reasoning about social and power implications of language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics.* https://doi.org/10.18653/v1/2020.acl-main.486
[18] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016.* https://doi.org/10.1609/icwsm.v10i1.14811
[19] Michael Veale, Kira Matus, and Robert Gorwa. 2023. AI and Global Governance: Modalities, Rationales, Tensions. *Annual Review of Law and Social Science* 19, Volume 19, 2023 (2023), 255–275. https://doi.org/10.1146/annurev-lawsocsci-020223-040749
[20] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track.* https://openreview.net/forum?id=kaHpo8OZw2
[21] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. arXiv:2306.09341 [cs.CV] https://arxiv.org/abs/2306.09341
[22] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations.* https://openreview.net/forum?id=SkeHuCVFDr
[23] Wenxuan Zhang, Yang Deng, Jing Ma, and Wai Lam. 2020. AnswerFact: Fact checking in product question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 2407–2417.
[24] Yian Zhang, Yifan Mai, Josselin Somerville Roberts, Rishi Bommasani, Yann Dubois, and Percy Liang. 2023. HELM Instruct: A Multidimensional Instruction Following Evaluation Framework with Absolute Ratings. https://crfm.stanford.edu/2024/02/18/helm-instruct.html