H Dataset-related supplementary material

H.1 Licenses

All code and data are released under the MIT license.

H.2 Statement of responsibility

The authors confirm that that they bear all responsibility in case of violation of rights and confirm that the data is released under MIT license.

H.3 Croissant metadata

The Croissant (Akhtar et al., 2024) metadata for the dataset can be found at the following url: https://huggingface.co/api/datasets/JailbreakBench/JBB-Behaviors/croissant.

I Data card

We report information about the dataset following the guidelines of Pushkarna et al. (2022).

I.1 Summary

- Dataset name: JBB-Behaviors
- Dataset link: https://huggingface.co/datasets/JailbreakBench/JBB-Behaviors/
- Datacard author: Edoardo Debenedetti, ETH Zurich

I.2 Authorship

I.2.1 Publishers

- Publishing organizations: University of Pennsylvania, ETH Zurich, EPFL, Sony AI
- Industry types: Academic - Tech, Corporate - Tech
- Contact details:
  - Publishing POC: Edoardo Debenedetti
  - Affiliation: ETH Zurich
  - Contact: edoardo.debenedetti@inf.ethz.ch

I.2.2 Dataset Owners

- Contact details:
  - Dataset Owner: Edoardo Debenedetti
  - Affiliation: ETH Zurich
  - Contact: edoardo.debenedetti@inf.ethz.ch
- Authors:
  - Patrick Chao, University of Pennsylvania
  - Edoardo Debenedetti, ETH Zurich
  - Alexander Robey, University of Pennsylvania
  - Maksym Andriushchenko, EPFL
  - Francesco Croce, EPFL
  - Vikash Sehwag, Sony AI
  - Edgar Dobriban, University of Pennsylvania
  - Nicolas Flammarion, EPFL
  - George J. Pappas, University of Pennsylvania

- Florian Tramèr, ETH Zurich
- Hamed Hassani, University of Pennsylvania
- Eric Wong, University of Pennsylvania

I.2.3 Funding Sources

No institution provided explicit funding for the creation of this benchmark. However, Patrick Chao and Edgar Dobriban are supported in part by the ARO, the NSF, and the Sloan Foundation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Army Research Office (ARO), or the Department of Defense, or the United States Government. Maksym Andriushchenko is supported by the Google Fellowship and Open Phil AI Fellowship. Edoardo Debenedetti is supported by armasuisse Science and Technology. Alexander Robey, Hamed Hassani, and George J. Pappas are supported by the NSF Institute for CORE Emerging Methods in Data Science (EnCORE). Alexander Robey is also supported by an ASSET Amazon AWS Trustworthy AI Fellowship. Eric Wong is supported in part by Amazon Research Award "Adversarial Manipulation of Prompting Interfaces."

I.3 Dataset overview

- Data subjects: Others (Behaviors that a human might want to elicit in a language model)
- Dataset snapshot:
  - Total samples: 200 (100 harmful behaviors and 100 benign behaviors)
- Content description: The datasets comprise of a set of prompts that aim at eliciting specific behaviors from language models.

I.3.1 Sensitivity of data

- Sensitivity types: Others (data that could be disturbing for some readers)
- Fields with sensitive data:
  - Intentionally Collected Sensitive Data: None
  - Unintentionally Collected Sensitive Data: None
- Risk types: Indirect risk
- Security handling: We add a disclaimer in the dataset README file.

I.3.2 Dataset version and maintenance

- Maintenance status: Actively Maintained
- Version details:
  - Current version: v1.0
  - Last updated: 06/2024
  - Release date: 06/2024
- Maintenance plan:
  - Versioning: We will establish minor updates to the dataset, in case we realize there are some errors and/or inconsistencies.
  - Updates: We are not planning to release major updates.
- Next planned updates: We don't have a timeline yet.
- Expected changes: N/A

I.4 Example of data points

- Primary data modality: Text Data
- Sampling of data points:
  - Demo Link: https://huggingface.co/datasets/JailbreakBench/JBB-Behaviors/viewer/behaviors/harmful

  - Typical Data Point Link: https://huggingface.co/datasets/JailbreakBench/JBB-Behaviors/viewer/behaviors/harmful?row=0
- Data fields:
  - **Behavior**: A unique identifier describing a distinct misuse behavior
  - **Goal**: A query requesting an objectionable behavior
  - **Target**: An affirmative response to the goal string
  - **Category**: A broader category of misuse from OpenAI's usage policies[3]
  - **Source**: The source from which the behavior was sourced, i.e., Original, Trojan Detection Challenge 2023 Red Teaming Track/HarmBench (Mazeika et al., 2023, 2024), or AdvBench (Zou et al., 2023)

I.5 Motivations and intentions

I.5.1 Motivations

- Purpose: Research
- Domains of application: Machine Learning, Large Language Models, AI Safety
- Motivating factors: Studying the robustness of LLMs and their defenses against jailbreak attacks, studying the effectiveness of jailbreak attacks.

I.5.2 Intended use

- Dataset use: Safe for research use
- Suitable use cases: Testing robustness of LLMs and their defenses against jailbreak attacks, testing the effectiveness of jailbreak attacks.
- Unsuitable use cases: Using this benchmark to evaluate the robustness of LLMs and defenses by using only the existing attacks (especially, only against the existing precomputed jailbreak prompts), without employing an adaptive attack with a thorough security evaluation.
- Citation guidelines: To be decided upon acceptance.

I.6 Access, retention, & wipeout

I.6.1 Access

- Access type: External – Open Access
- Documentation link: https://github.com/JailbreakBench/jailbreakbench/?tab=readme-ov-file#accessing-the-jbb-behaviors-datasets
- Pre-requisites: None
- Policy links: None
- Access Control Lists: None

I.7 Provenance

I.7.1 Collection

- Methods used:
  - Artificially Generated
  - Authors creativity
- Methodology detail:
  - Source: Authors, Zou et al. (2023); Mazeika et al. (2023, 2024)
  - Is this source considered sensitive or high-risk? No

[3] https://openai.com/policies/usage-policies

- Dates of Collection: 11/2023 – 05/2024
- Primary modality of collection data: Text Data
- Update Frequency for collected data: Static
- Additional Links for this collection:
  - Zou et al. (2023): https://github.com/llm-attacks/llm-attacks/blob/0f505d82e25c15a83b6954db28191b69927a255d/data/advbench/harmful_behaviors.csv
  - Mazeika et al. (2023, 2024): https://github.com/centerforaisafety/tdc2023-starter-kit/tree/main/red_teaming
- Source descriptions: As described in Figure 3, some of the behaviors are sourced from Zou et al. (2023); Mazeika et al. (2023, 2024). Such behaviors are clearly marked as derived from those works also in the dataset itself. The behaviors are curated so that they are unique and–once the original behaviors are added–they are uniformly distributed across the categories of misuse from OpenAI usage policies. The behavior marked as "Original" in the dataset were created by the authors. Some of the behaviors were created with the assistance of LLMs to get inspirations on the types of behaviors, but without taking the LLM outputs verbatim.
- Collection cadence: Static.
- Data processing: We ensure that the behaviors are unique and uniformly distributes across the categories of misuse from OpenAI usage policies.

I.8 Human and Other Sensitive Attributes

There are no human or other sensitive attributes.

I.9 Extended use

I.9.1 Use with Other Data

- Safety level: Safe to use with other data
- Known safe/unsafe datasets or data types: N/A

I.9.2 Forking and sampling

- Safety level: Safe to fork. Sampling not recommended as the dataset is not particularly large in the first place.
- Acceptable sampling methods: N/A

I.9.3 Use in AI and ML systems

- Dataset use: Validation
- Usage guidelines: The benchmark can be used to assess the robustness of models and defenses, as well as the effectiveness of attacks.
- Known correlations: N/A

I.10 Transformations

I.10.1 Synopsis

- Transformations applied: No transformations were applied to the data.
- Fields transformed: N/A.
- Libraires and methods used: Manual changes.

I.11 Known applications and benchmarks

- ML Applications: large language models

- Evaluation results and processes: We show the evaluation results and methodology in the main paper, in Section 4.