Social Impact Scorecards

Avijit Ghosh, Hugging Face 🤗

Hugging Face is working on an initiative that could contribute to our collective work on evaluation frameworks. This proposal draws from our large collaborative paper "Evaluating the Social Impact of Generative AI Systems in Systems and Society", which provides a comprehensive framework for evaluating generative AI systems across modalities.

The paper and this proposal connect with two upcoming initiatives that could strengthen our VRT efforts:

1. NeurIPS 2024 Workshop: "Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI" will bring together experts to challenge the existing eval landscape and to develop guidance for social impact evaluation standardization. This workshop will help refine evaluation approaches and methodologies that could inform VRT implementation.

2. Social Impact Measurement Consortium: A growing collaboration of researchers and practitioners working to establish best practices for measuring and documenting AI system impacts. The consortium aims to create shared resources and standardized approaches that complement voluntary reporting efforts. As a first step, we plan to ask consortium members to fill out our social impact scorecard for models they have developed, and launch a website with a leaderboard and other analysis of score cards, following the success of Hugging Face's Open LLM Leaderboard initiative.

The Social Impact Scorecard teaser presented below could hopefully involve AISIC's VRT design. We aim to design a structured template that incorporates lessons learned from both research and practice, and most importantly, is unambiguous and easy for practitioners to fill out.

For more information, or if you would like to be involved, please connect at avijit@huggingface.co !

Social Impact Scorecard Sample

Model Information

- **Model Name: _____**
- **Model Provider: _____**
- **Modality (select all that apply):**
  [ ] Text-to-Text
  [ ] Text-to-Image
  [ ] Image-to-Text
  [ ] Image-to-Image
  [ ] Audio
  [ ] Video
  [ ] Multimodal (specify): _____

Instructions

For each question, select Yes, No, or Not Applicable.
Indicate the source of information as:

- 1P: 1st Party/Developer
- 3P: 3rd Party
- Both: Both 1st Party and 3rd Party

Category: *Disparate Performance*

1. Subgroup Performance Analysis

Explainer: Has the system been evaluated for disparate performance across different subpopulations?

Please select all evaluation types that are applicable to your model:

- ☐ Non-aggregated (disaggregated) evaluation results across subpopulations, including feature importance and consistency analysis
- ☐ Metrics such as subgroup accuracy, calibration, AUC, recall, precision, min-max ratios
- ☐ Worst-case subgroup performance analysis, including performance on rare or underrepresented cases

- ☐ Expected effort to improve model decisions from unfavorable to favorable
- ☐ Coverage metrics to ensure wide representation of subgroups and identify dataset skew
- ☐ Intersectional analysis examining performance across combinations of subgroup characteristics
- ☐ Critical examination of how "performance" itself might be conceptualized differently across groups

If any are selected:

1. **Yes: Source is ___**
2. **No: These applicable evaluations have not been performed**

If none are selected:

3. **Not applicable: None of these evaluations are applicable to this model**

2. Language and Accent Performance

Explainer: Has the system been assessed for performance across different languages and dialects?

Please select all evaluation types that are applicable to your model:

- ☐ Cross-lingual prompting on standard benchmarks
- ☐ Examination of performance across dialects
- ☐ Analysis of hallucination disparity across languages
- ☐ Multilingual knowledge retrieval evaluations
- ☐ Comparison of performance to the highest-performing language or accent
- ☐ Assessment of low-resource language performance

If any are selected:

1. **Yes: Source is ___**
2. **No: These applicable evaluations have not been performed**

If none are selected:

3. **Not applicable: None of these evaluations are applicable to this model**

3. Generation Quality Assessment

Explainer: Has the system's generation quality been evaluated across different concepts, categories, and cultural representations?

Please select all evaluation types that are applicable to your model:

- ☐ Analysis of generation quality and realism across different content types and categories
- ☐ Assessment of cultural representation and stereotyping in generated content

- ☐ Evaluation of generation consistency and quality across demographic groups, including underrepresented populations
- ☐ Analysis of systematic patterns or biases in generated content
- ☐ Assessment of mitigation efforts' impact on generation quality across groups

If any are selected:

1. **Yes: Source is ___**
2. **No: These applicable evaluations have not been performed**

If none are selected:

3. **Not applicable: None of these evaluations are applicable to this model**

4. Dataset Disparities Evaluation

Explainer: Has the system been evaluated for disparities stemming from dataset issues?

Please select all evaluation types that are applicable to your model:

- ☐ Analysis of geographic biases in data collection
- ☐ Examination of disparate digitization of content globally
- ☐ Assessment of varying levels of internet access for digitizing content
- ☐ Evaluation of content filter impacts on data representation

If any are selected:

1. **Yes: Source is ___**
2. **No: These applicable evaluations have not been performed**

If none are selected:

3. **Not applicable: None of these evaluations are applicable to this model**

5. Evaluation of Systemic Issues

Explainer: Has the evaluation considered systemic issues that may lead to disparate performance?

Please select all evaluation types that are applicable to your model:

- ☐ Assessment of systemic barriers in dataset collection methods
- ☐ Examination of infrastructure biases favoring certain languages or accents
- ☐ Consideration of positive feedback loops in model-generated or synthetic data
- ☐ Assessment of whether interventions to address disparities have introduced new biases
- ☐ Evaluation of systemic barriers affecting different groups' representation in the data

If any are selected:

1. **Yes: Source is ___**
2. **No: These applicable evaluations have not been performed**

If none are selected:

3. **Not applicable: None of these evaluations are applicable to this model**

6. Mitigation and Future Considerations

Explainer: Has the evaluation considered challenges and mitigations related to model behavior on edge cases and evolving performance needs?

Please select all evaluation types that are applicable to your model:

- ☐ Analysis of model behavior on rare or outlier cases
- ☐ Evaluation of trade-offs between memorization and generalization for uncommon patterns
- ☐ Assessment of feature predictiveness across different contexts and frequencies
- ☐ Critical examination of how performance metrics may need to adapt for different use cases
- ☐ Testing of model adaptation strategies for previously unseen patterns or groups

If any are selected:

1. **Yes: Source is ___**
2. **No: These applicable evaluations have not been performed**

If none are selected:

3. **Not applicable: None of these evaluations are applicable to this model**