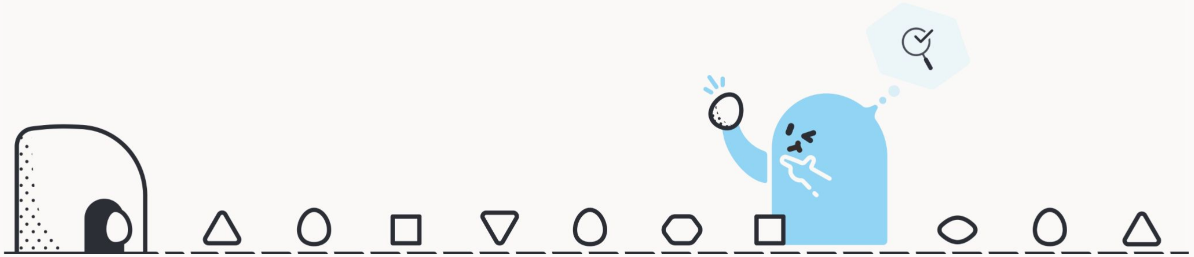
Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI



Avijit Ghosh, Zeerak Talat, Yacine Jernite, Irene Solaiman

&

Usman Gohar, Jennifer Mickel, Michelle Lin, Cedric Whitney, Lucie-Aimée Kaffee, Arjun Subramonian, Alberto Lusoli, Felix Friedrich

Agenda

9:15 - 9:30: Welcome and Introductions
9:30 - 10:30: Opening Panel – Reflections on the Landscape
10:30 - 11:30: Oral Session 1 – Provocations and Ethics in AI Evaluations (+breakouts)
11:30 - 12:30: Oral Session 2 – Multimodal and Cross-Cultural Evaluation Methods (+breakouts)
12:30 - 1:15: Lunch
1:15 - 2:30: Poster Session
2:30 - 3:00: Oral Session 3 – Systematic Approaches to AI Impact Assessment
3:00 - 3:30: Break
3:30 - 4:05: Breakouts
4:05 - 5:00: What's Next (+breakouts)
5:00 - 5:15: Closing

EvalEval

NEURAL INFORMATION PROCESSING SYSTEMS

Paper: Evaluating Social Impacts

Report
https://arxiv.org/abs/2306.05949

Why?

Help standardize how researchers and developers conduct broader impact assessments and how policymakers/regulators assess system risk

Eval Eval

NEURAL INFORMATION PROCESSING SYSTEMS

What's Going on Here

5. Toxicity and Bias Analysis

Alongside the benefits of scaling language models, it is crucial to analyse how scale impacts potentially harmful behaviour. Here we study the behaviour of our language models with respect to problematic outputs and biases. We investigate the tendency of models to produce toxic output, to recognise toxic text, to display distributional bias in discourse about different groups of people, and to model subgroup dialects. For each question we consider variation across model scale.

We choose evaluations and metrics which are commonly used in the field. However, various work has discussed the limitations of current metrics and evaluations (Blodgett et al., 2020, 2021; Sheng et al., 2019; Welbl et al., 2021; Xu et al., 2021a) and our analysis has uncovered further caveats, which we highlight in the following sections and Section 7.2. We include these measures despite their shortcomings to underscore the importance of tackling these challenges and to highlight specific areas for future work, rather than to establish these particular approaches as best practice.

10 Representational Bias Analysis

Pre-trained language models have been demonstrated to contain and amplify biases in underlying data (Sheng et al., 2021; Kurita et al., 2019; Dev et al., 2019). The importance of communicating the infrastructure of the model has also been emphasized (Mitchell et al., 2019). We provide a datasheet in Appendix D and a model card in Appendix E which detail the intended usage, datasets used, and more. In this section, we analyze PaLM for distributional biases related to social groups, and for toxicity in open-ended language generation. This analysis helps outline some of the potential risks of the model, although domain and task specific analysis is essential to truly calibrate, contextualize, and mitigate possible harms.

2.4 Harms of representation, allocation, and quality of service

Language models can amplify biases and perpetuate stereotypes.[40, 41, 42, 4
earlier GPT models and other common language models, both GPT-4-early
continue to reinforce social biases and worldviews.

The evaluation process we ran helped to generate additional qualitative evide
in various versions of the GPT-4 model. We found that the model has the poter
reproduce specific biases and worldviews, including harmful stereotypical and de
for certain marginalized groups. Model behaviors, such as inappropriate hedging

4 Bias & Toxicity Evaluations

To understand the potential harm of OPT-175B, we evaluate a series of benchmarks related to hate speech detection, stereotype awareness, and toxic content generation. While there may be shortcomings in these benchmarks (Blodgett et al., 2021; Jacobs and Wallach, 2021), these measurements provide a first step towards understanding the limitations of OPT-175B. We compare primarily against GPT-3 Davinci, as these benchmarks were not yet available to be included in Brown et al. (2020).

4.1 Hate Speech Detection

Using the ETHOS dataset provided in Mollas et al. (2020) and instrumented by Chiu and Alexander (2021), we measure the ability of OPT-175B to identify whether or not certain English statements are racist or sexist (or neither). In the zero-, one-,

6 Conclusions, Limitations and Societal Impact

Imagen showcases the effectiveness of frozen large pretrained language models as text encod
the text-to-image generation using diffusion models. Our observation that scaling the size o
language models have significantly more impact than scaling the U-Net size on overall perfor
encourages future research directions on exploring even bigger language models as text en
Furthermore, through Imagen we re-emphasize the importance of classifier-free guidance, a
introduce dynamic thresholding, which allows usage of much higher guidance weights tha
in previou
produces 1024 × 1024 sample
unprecede

5. Limitations & Societal Impact

Limitations While LDMs significantly reduce computational requirements compared to pixel-based approaches, their sequential sampling process is still slower than that of GANs. Moreover, the use of LDMs can be questionable when high precision is required: although the loss of image quality is very small in our $f = 4$ autoencoding models (see Fig. 1), their reconstruction capability can become a bottleneck for tasks that require fine-grained accuracy in pixel space. We assume that our superresolution models (Sec. 4.4) are already somewhat limited in this respect.

Societal Impact Generative models for media like imagery are a double-edged sword: On the one hand, they

9 Discussion and limitations

Perhaps the most noteworthy aspect of
safer dialog models with modest amour
However, our study and LaMDA still h

Collecting fine-tuning datasets brings t
time consuming, and complex process
longer contexts, and more metrics that
conversations. The complexity of capt
crowdworker rating quality against th

Panel:Reflections on the Eval Landscape

Abigail Jacobs

University of Michigan

Lee Wan Sie

IMDA Singapore

Su Lin Blodgett

Microsoft

Avijit Ghosh

Hugging Face
(Panel Moderator)

EvalEval

NEURAL INFORMATION PROCESSING SYSTEMS

Provocations & Ethics in AI Evaluation

10:30 - 10:55

Oral Session

- "Provocation: Who benefits from 'inclusion' in Generative AI?"
- "(Mis)use of nude images in machine learning research"
- "Evaluating Refusal"

Provocations & Ethics in AI Evaluation

Breakout

Discussion prompts: Please fill out during your breakout session

10:55 - 11:15

Report Back

11:15 - 11:30

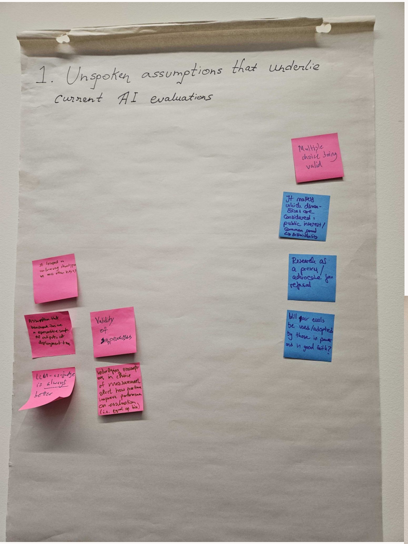
Discussion Prompts 10:55 - 11:15

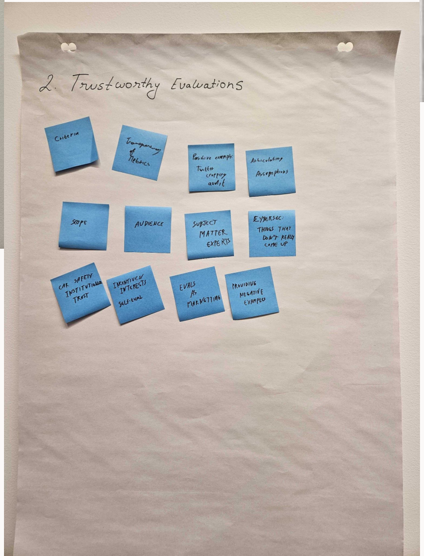
1. **Unspoken assumptions that underlie current AI evaluations**
   a. What are assumptions/choices in measurement that affect the results of AI evaluations?
   b. How do assumptions made in the development of evaluations affect the evaluation effectiveness and/or contribute to evaluation limitations?
2. **Trustworthy evaluations**
   a. What builds trust in an evaluation?
   b. What makes an evaluation not trustworthy?
3. **Human participation in evaluation**
   a. How can/should human feedback scale in sociotechnical evals?
   b. How should human participants be chosen?
   c. What are the costs of human participation? When is automated feedback appropriate?
4. **Conflicting values in evaluations**
   a. What are known conflicting values in existing evaluations?
   b. How should conflicting values be reckoned? What should be prioritized?
5. **Access for running evals**
   a. Who should be responsible for running evals (model developers, some third party, etc.)?
   b. What resources are needed per type of eval?
   c. How does this differ by type of broader impact and type of system/system component (modality, data vs. model)

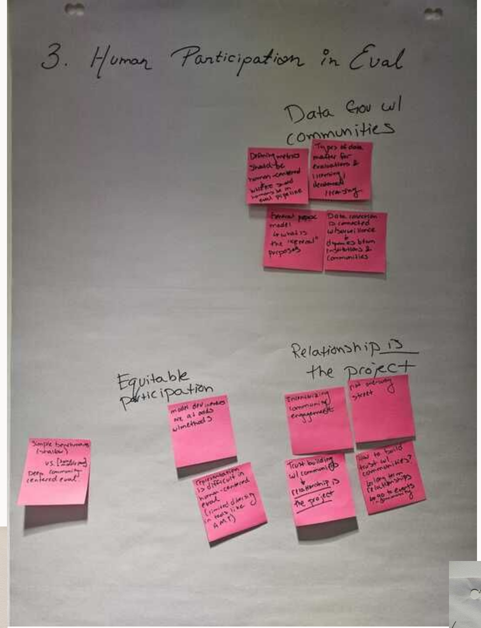
Eval Eval

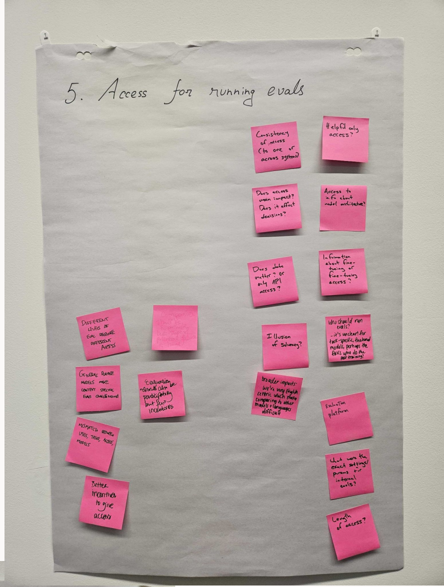
NEURAL INFORMATION PROCESSING SYSTEMS

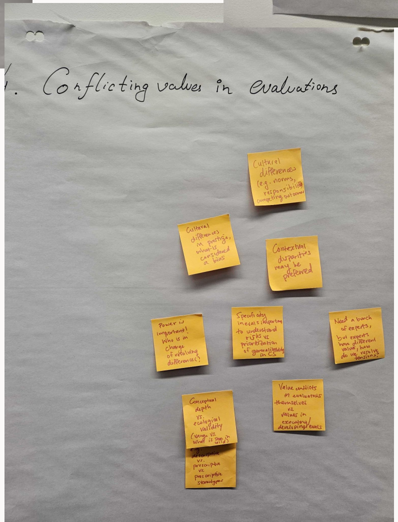
Results from Breakout 1











Multimodal & Cross-Cultural Evaluation

11:30-11:55

Oral Session

- "JMMMU: A Japanese Massive Multi-discipline Multimodal Understanding Benchmark"
- "Critical human-AI use scenarios and interaction modes for societal impact evaluations"
- "Cascaded to End-to-End: New Safety, Security, and Evaluation Questions for Audio Language Models"

NEURAL INFORMATION PROCESSING SYSTEMS

Multimodal & Cross-Cultural Evaluation

Breakout

Discussion prompts: Please fill out during your breakout session

11:55 - 12:15

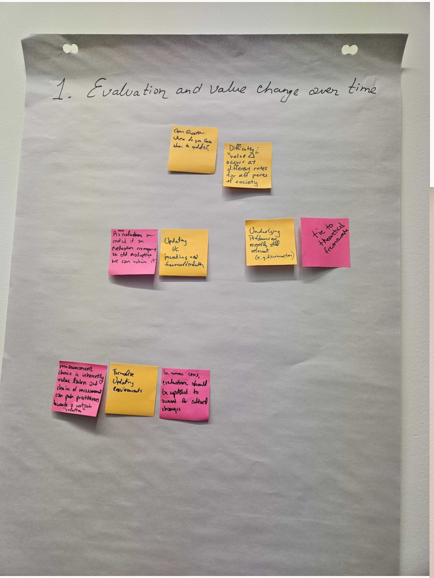
Report Back

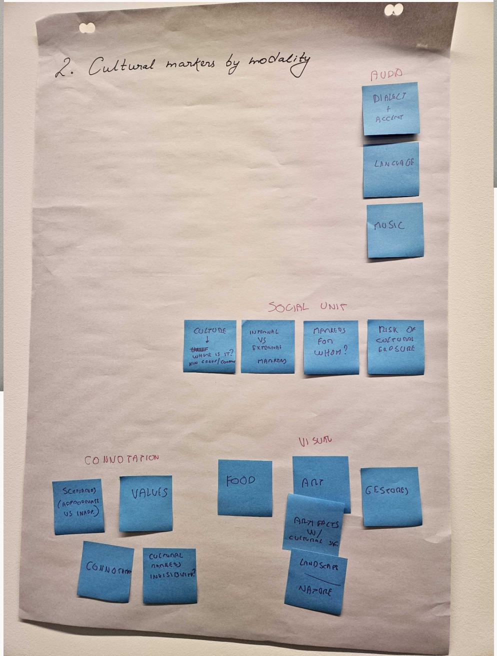
12:15 - 12:30

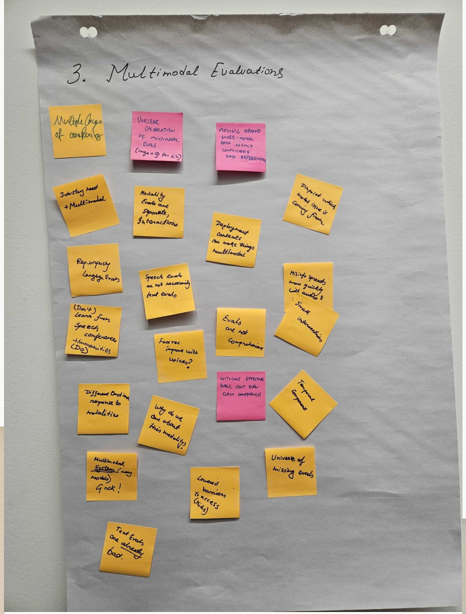
Discussion Prompts 11:55 - 12:15

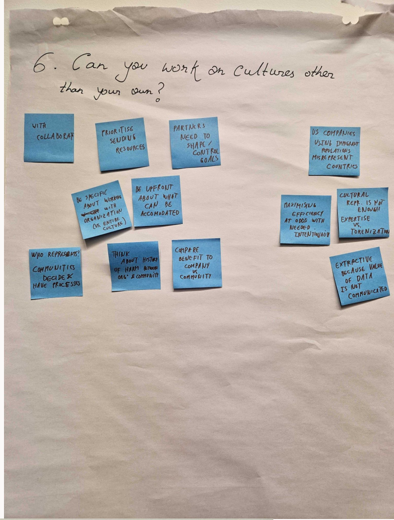
1. **Evaluation and value change over time**
   a. How do we ensure evaluations are relevant?
   b. Should evaluations be retired as values change and perspectives change?
2. **Cultural markers by modality**
   a. What is a cultural marker?
   b. What do cultural markers look like in language, image, audio, video?
3. **Multimodal Evaluations**
   a. How should the modality of an evaluation be prioritized?
   b. How do we improve gaps in building evaluations for underrepresented modalities?
4. **What is “cultural competence”?**
   a. What would it mean for a model to demonstrate knowledge of a culture and why/when would we want that?
5. **Scaling to other cultures**
   a. For low-resource regions, how should existing evaluations adapt to be inclusive, if at all?
   b. Should new evaluations be created per culture?
   c. How should evaluations be developed to ensure cultures are adequately represented?
   d. How should different groups within cultures be adequately represented within evaluations?
6. **Can you work on cultures other than your own?**
   a. Who represents a culture?
   b. If yes, how?
   c. If no, how to enable the work?

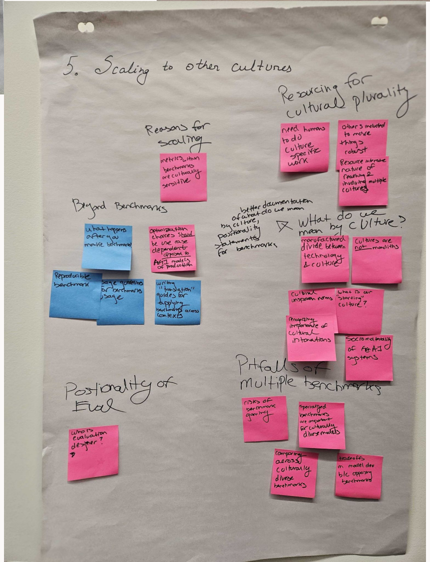
Results from Breakout 2











Lunch

12:30 - 1:15

Poster Session

1:15 - 2:30

Systematic Approaches to Impact Assessment

2:30 - 3:00

Oral Session

- "GenAI Evaluation Maturity Framework (GEMF)"
- "AIR-Bench 2024: Safety Evaluation Based on Risk Categories"
- "Evaluating Generative AI Systems is a Social Science Measurement Challenge"

Break (coffee in hall)

3:00 - 3:30

Systematic Approaches to Impact Assessment Breakout

Discussion prompts: Please fill out during your breakout session

3:30 - 3:50

Report Back

3:50 - 4:05

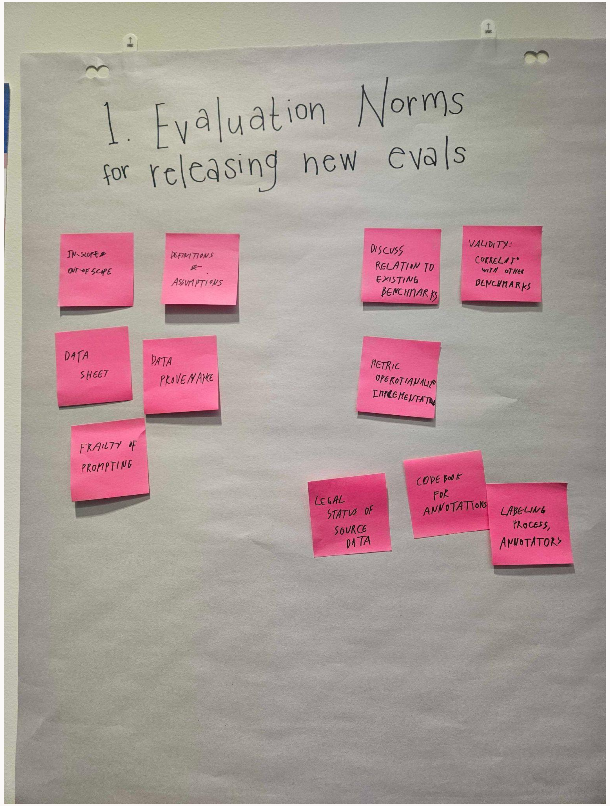
Discussion Prompts 3:30 - 3:50

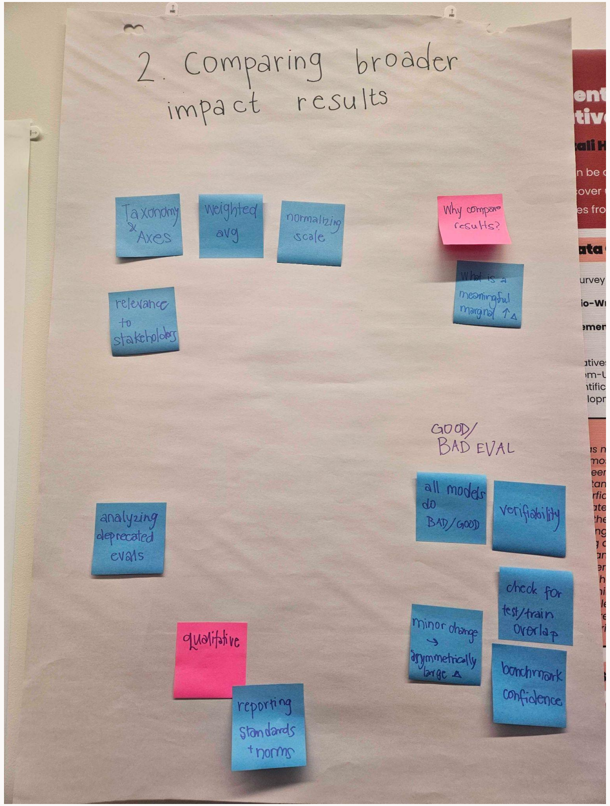
1. **Evaluation norms for releasing new evals**
   a. What should accompany a new evaluation release?
2. **Comparing broader impact results**
   a. How should evaluation results be compared or ranked?
3. **Metadata and Evaluation Selection**
   a. How can metadata (e.g., intended purpose, assumptions, limitations) in repositories aid evaluation selection?
4. **Evaluation Communication**
   a. How much information about evaluation results needs to be communicated?
   b. To whom should results be interpretable?
5. **Engagement with Social Sciences**
   a. What does successful social science engagement look like?
   b. Are there tools specifically designed for non-technical stakeholders to engage in the evaluation process? If not, how could such tools be developed?
   c. How should "borrowed" approaches be diversified if at all?
6. **System and Model Developer Responsibilities**
   a. What is needed from system/model developers?
   b. How should external evaluator access be systematically determined?
7. **Effective taxonomies**
   a. What makes a broader impact taxonomy useful?
   b. How do we avoid "death by a thousand taxonomies"?

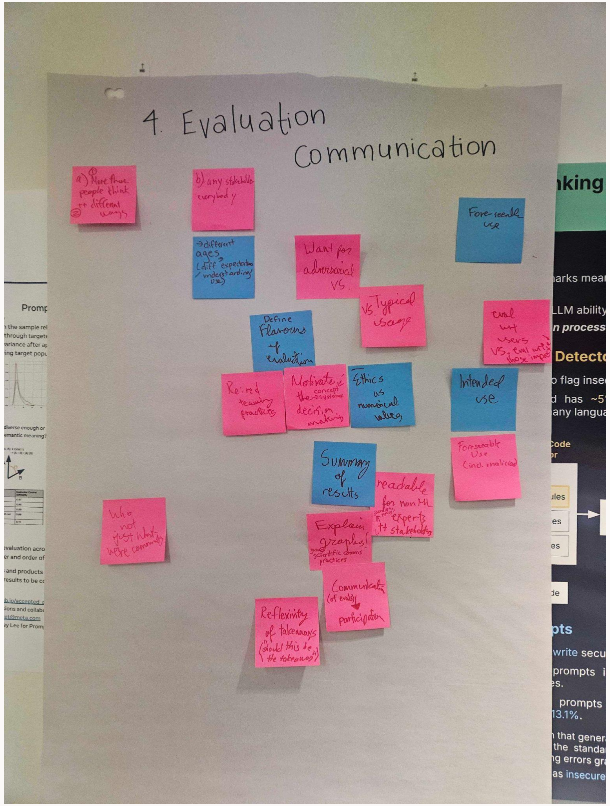
Eval Eval

NEURAL INFORMATION PROCESSING SYSTEMS

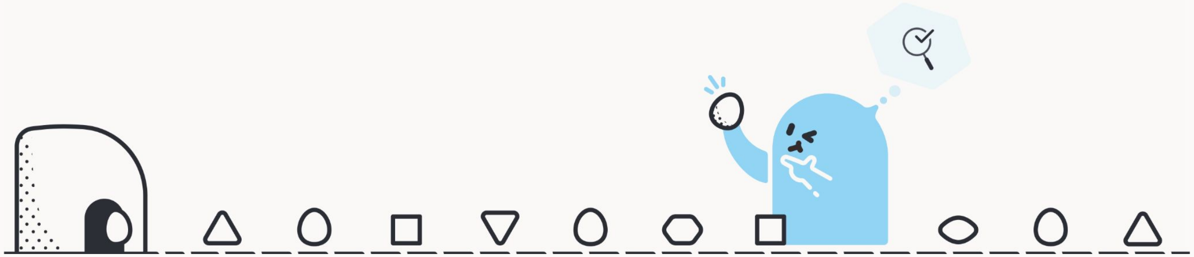
Results from Breakout 3







Next Steps…



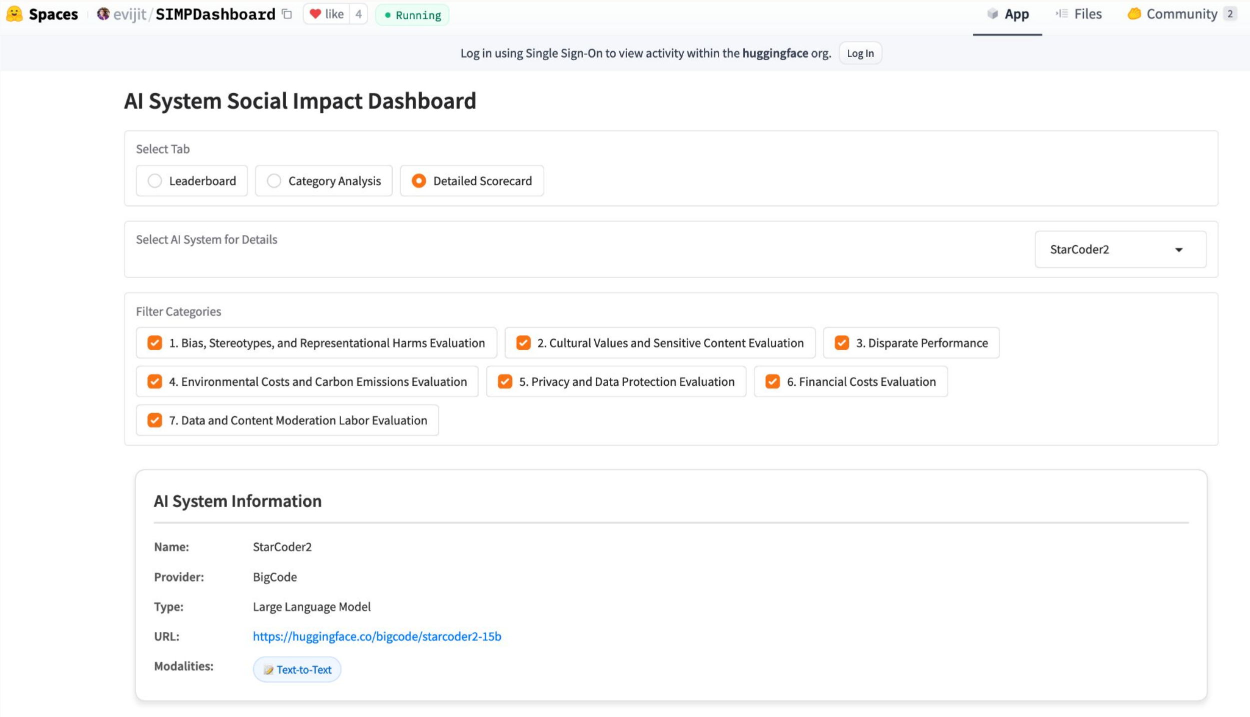
Springer Journal Publication

Authors: Opt-in to this Special Issue!

Eval Eval

NEURAL INFORMATION PROCESSING SYSTEMS

Social Impact Card Demo



Eval Eval

NEURAL INFORMATION PROCESSING SYSTEMS

What's Next: Coalition Working Groups

Thank you for sharing your thoughts, energy, and time with us!

If you'd like to continue working on these topics, fill out this form:

Eval Eval

NEURAL INFORMATION PROCESSING SYSTEMS

Broader Impact Evaluation Coalition

Breakout

4:20 - 4:50

Research Outputs

- **Eval documentation**
  - What should be documented when a new evaluation is created/released?
  - What is needed to document an evaluation (resources, access to information)?
- **Eval science and comparison**
  - What are essential criteria for good broader impact evaluations?
  - How should evaluations in a given broader impact category (e.g. bias) be chosen?
  - What are sufficient conditions for making an evaluation reproducible?
- **Broader Impacts card**
  - How can the Broader Impacts Card be most effective?
  - What should developers report and what should external evaluators report?

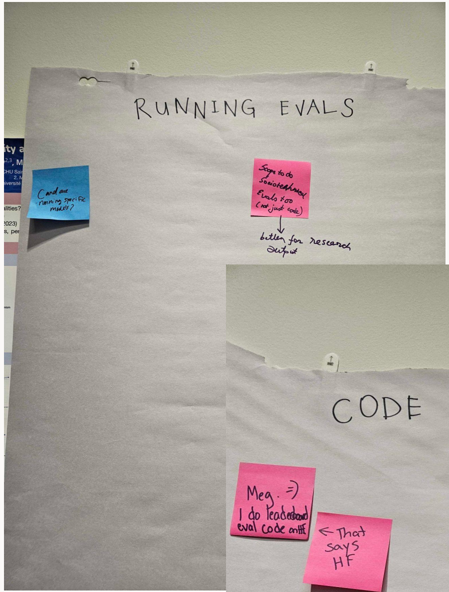
Infrastructure

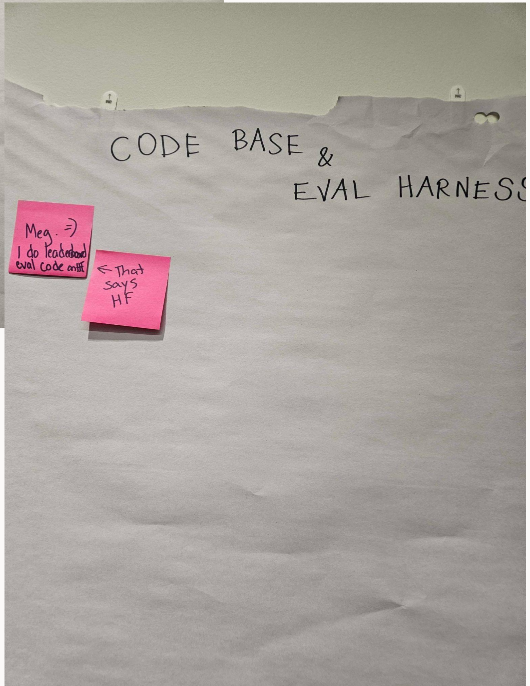
- **Eval harness**
  - What existing infrastructure has been useful?
  - What would most lower the barrier to run broader impact evals?
- **Running evals on chosen models**
  - What are high priority broader impact evals to run?

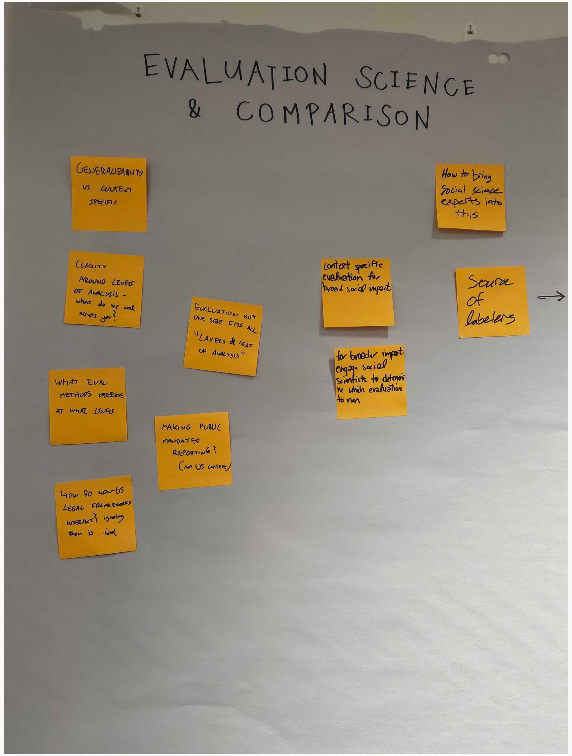
Eval Eval

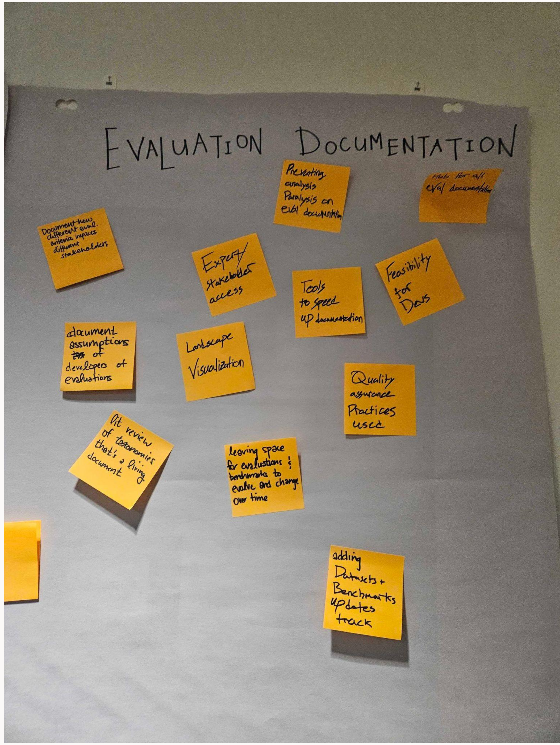
NEURAL INFORMATION PROCESSING SYSTEMS

Results from Coalition Breakout









Thanks !

Feedback Form

Coalition Form