# Rethinking Model Evaluation as Narrowing the Socio-Technical Gap

Q. Vera Liao  Ziang Xiao
Microsoft Research, Montreal
Johns Hopkins University

Q. Vera Liao (veraliao@microsoft.com)
Microsoft Research
Johns Hopkins Whiting School of Engineering

## Challenges in Model Eval

- **Diverse capability:** Large generative models and diverse capabilities
- **Homogenization**: Using limited models for a wide range of applications.
- **The socio-technical gap**: The divide between what technology can do and what people need.

## Goals for Rethinking Model Eval

**Goal 1:** Studying people's needs, values, and activities in downstream use cases of models, and distilling principles and representations that can guide the development and evaluation of ML technologies.

**Goal 2:** Developing evaluation methods that can provide valid assessments for whether and how much human needs in downstream use cases can be satisfied.

## Bridging the Socio-technical Gap

**Context Realism:** Realistic proxy for how the technology will be used in a downstream use case

**Human Requirement Realism:** Realistic proxy for what requirements people involved in the use case have for technology

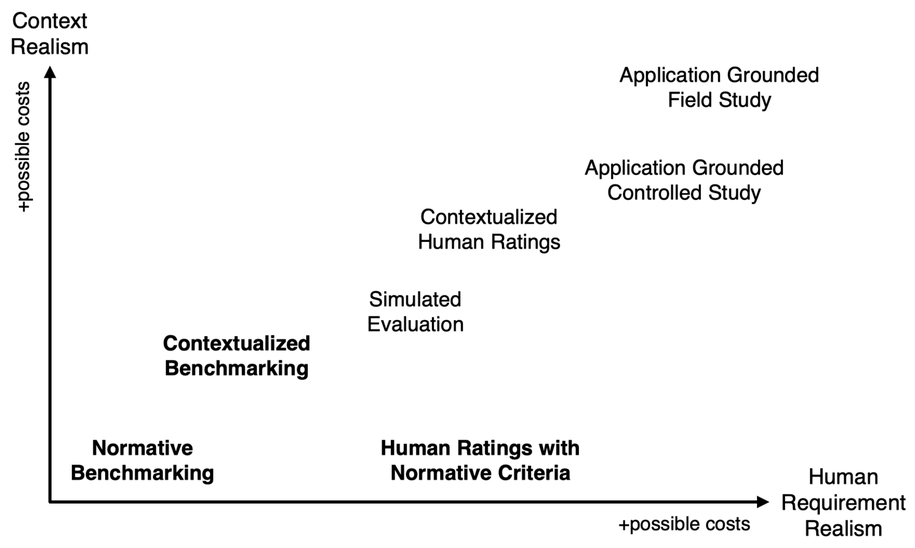


Fig. 1. Mapping of HCI and NLG **(in bold)** evaluation methods on the two dimensions of realism

## Opportunities for LLM Evaluation

- Contextualized human rating protocols.
- Use case grounded simulated evaluation.
- Application-grounded evaluation (controlled and field studies).

## Recommendations and Open Questions

- **Develop evaluation metrics for human-desirable constructs:** The community should focus on identifying human-desirable constructs of model outputs by studying people's needs and values in downstream use cases.
- **Formalization and validation "from the outside in":** Methods down the realism spectrum should be informed and validated by the upstream methods.
- **What makes a useful representation of "downstream use cases" for LLMs?:** Considering the discriminative power (e.g., are the socio-requirements in different use cases sufficiently different) and the appropriate or practical level of abstraction (e.g., practical for comprehensive benchmarking across all use cases).
- **How should "lowering LLM evaluation costs" be defined and justified?:** Unpacking types of costs, including model evaluation specific ones such as computing (including environmental impact), and further articulating "costs" and "benefits" of different evaluation methods.

## Lessons from Explainable AI(XAI) and HCI

### XAI:

- Human-centered evaluation
- Interpretability in Evaluation
- Study of socio-requirements in downstream use cases
- Multi-metric evaluation

### HCI:

- Embrace diverse evaluation methods
- Develop evaluation metrics for human-desirable constructs
- Formalization and validation "from the outside in
- Balancing costs and realism