SPHERE: An Evaluation Card for Human-AI Systems

sphere-eval.github.io/

Qianou Ma[1*] (qianouma@cmu.edu), **Dora Zhao**[2*] (dorothyz@stanford.edu),
Xinran Zhao[1], **Chenglei Si**[2], **Chenyang Yang**[1], **Ryan Louie**[2],
Ehud Reiter[3], **Diyi Yang**[2+], **Tongshuang Wu**[1+]
(*Equal contribution, +Equal contribution)

[1]Carnegie Mellon University,
[2]Stanford University,
[3]University of Aberdeen

Introducing a **comprehensive template** for designing and documenting **human-AI system evaluations**

Motivation

- Mounting concern over how to evaluate generative systems as existing methods fall short
- Human-AI systems introduce additional complexities as we must consider not only model performance but also the impact on users
- We introduce **SPHERE cards**, which provide a holistic overview on how to design and document the evaluations of human-AI systems

SPHERE (**S**ubject-**P**rocess-**H**andler-**E**lapsed-**R**obustness **E**valuation) Cards

What is being evaluated?

What components and system design goals are evaluated?

When is the evaluation conducted?

What is the time scale over which the evaluation is conducted?

How is the evaluation conducted?

What is the scope of evaluation? What methods are used?

How is the evaluation validated?

How do we ensure the validation of the evaluation?

Who participates in the evaluation?

What are the automated or human evaluators used?

Example SPHERE Card: AngleKindling
(Petridis et al. 2023)

What is being evaluated?

The *system's effectiveness* (# of pursuable angles), *efficiency* (mental demand), and *satisfaction* (perceived helpfulness)

How is the evaluation conducted?

Extrinsic within-subject study with a *quantitative* post-task survey and *qualitative* interviews

Who is participating in the evaluation?

12 professional journalists *(domain experts, intended users)*

When is the evaluation conducted

Short-term sessions of up to 60 minutes

How is the evaluation validated?

Counterbalancing of tool order to reduce learning effects for *validity*

Recommendations

1. **Evaluations should reflect real-world use**
   Test systems in the real-world with evaluators from relevant stakeholder groups.
2. **Cross-verify results across methods**
   Triangulate results across from different data sources or methods and ground methods in existing practices
3. **Rigorously evaluate the evaluation**
   Expand practices for measuring reliability / validity and document practices for replication

Definition: A ***human-AI system*** harnesses AI capabilities that are exposed to end users through an interface. These systems consist of an AI model, a user interface, and logic that converts user entry into input for the model (Lee et al., 2023; Amershi et al., 2019).

ACL 2025 VIENNA JULY 27 - AUGUST 1