Identifying human-AI use scenarios and interaction modes for societal impact evaluations

Lujain Ibrahim
University of Oxford
lujain.ibrahim@oii.ox.ac.uk

Saffron Huang
Collective Intelligence Project

Lama Ahmad
OpenAI

Markus Anderljung
Centre for Governance of AI

Abstract

Most real-world AI applications involve human-AI interaction, yet current evaluations, such as common benchmarks, do not. These evaluations typically assess the safety of models in isolation, thereby falling short of capturing the complexity of human-model interactions. While there are challenges in generalizing findings from human interaction evaluations at the individual-level to broader societal effects, such evaluations are nonetheless crucial for societal impact evaluation. They offer valuable insights into how AI systems affect individual users, which can inform interventions with significant societal implications. For instance, understanding how individuals engage with non-factual model outputs can guide effective labeling strategies for AI-generated content. This not only helps individuals recognize synthetic media but also addresses broader concerns about misinformation and trust. As human interaction evaluations become increasingly important, in this paper, we outline the *evaluation scenarios* and the human-model *interaction modes* the field needs to evaluate to better understand the societal impact of generative models.

1 Human interaction evaluation parameters

To effectively evaluate the risks and harms associated with human-model interactions, it is essential to identify two key parameters: (1) the harmful *use scenario*, which specifies the context in which the model is (mis)used, and (2) the *interaction mode*, which describes the nature of humans' interactions with the model. Identifying these two parameters enables researchers and practitioners to systematically design safety evaluations by mapping potential risks to specific interaction contexts, such as detecting overreliance in collaborative tasks or emotional attachment in conversational interactions, and selecting appropriate metrics to measure these risks.

1.1 Possible harmful use scenarios

In human-computer interaction (HCI) research, user goals or objectives have been shown to shape how users engage with systems and thus influence the outcomes of these engagements [Subramonyam et al., 2024]. Thus, here, we group harmful use scenarios according to user objectives in the interaction. In each scenario, we also consider the affected parties of any harm caused by use. We propose in Table 1 four scenarios which we believe address some of the most salient failure modes of current concern [Mitchell, 2024]. Additionally, these scenarios may be grounded in a specific use domain (e.g., medicine, education) to target domain-specific considerations.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Scenario	Misuse / adversarial testing	Unintended harm: personal impact	Unintended harm: external impact
Objective	User intentionally uses model to inflict harm on another person, group of people, or system	User uses model, gets harmed in the process	User uses model, unintentionally harms another person, group of people, or system
Affected parties	External subjects	User	External subjects
Example(s)	Influence operations, cybersecurity attacks, hate speech	Exposure to harmful stereotypes in model output	Decision-maker trusts inaccurate model judgment hurting decision-subject

Table 1: Three primary harmful use scenarios and examples of each

1.2 Common human-model interaction modes

For each use scenario, there exists multiple possible *interaction modes*, visualized in Figure 1. Interaction modes define the nature of the human-model relationship in completing certain tasks towards the objective [Gao et al., 2024, Händler, 2023]. These tasks may be goal-oriented tasks focused on specific outcomes (e.g., summarization), or open-ended tasks which are exploratory and without a clear endpoint (e.g., social dialogue).

Based on observed use cases and studies on real-world usage data, we taxonimize five main modes of prototypical human-model interactions that human-interaction evaluations can target [Ouyang et al., 2023, Zhao et al., 2024]:

- **Collaboration**: human and model work in tandem towards completing joint goal-oriented tasks (e.g., human and model write a report together, where the model generates text and the human iteratively edits and refines it).
- **Direction**: human instructs the model to complete specific goal-oriented tasks (e.g., human gives model a set of instructions to generate a marketing campaign).
- **Assistance**: model provides support to human in completing specific goal-oriented tasks (e.g., human makes a decision with model input and assistance).
- **Cooperation**: human and model undertake separate but complementary goal-oriented tasks. Unlike collaboration, where involvement is mutually integrated, cooperation involves distinct contributions towards the same goal but without shared execution (e.g., human and model code different sections of the same computer program).
- **Exposure**: human observes or is exposed to a single or discrete set of pre-produced model output (e.g., human reads a model-generated message).
- **Exploration**: human engages in open-ended tasks with model (e.g., human and model engage in open-ended dialogue and discussion).

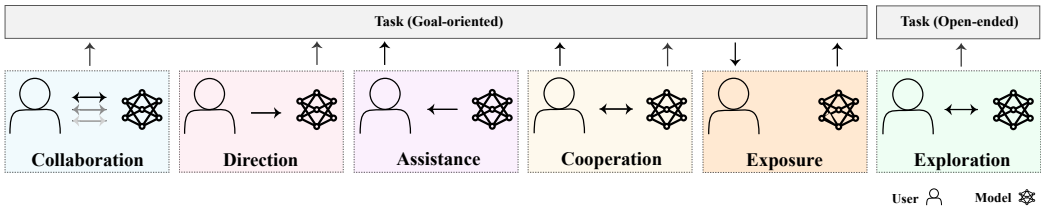

Figure 1: Taxonomy of human-LLM interaction modes. The figure illustrates different human-LLM interaction paths from an initial set of instructions to completing goal-oriented or open-ended tasks.

References

J. Gao, S. A. Gebreegziabher, K. T. W. Choo, T. J. J. Li, S. T. Perrault, and T. W. Malone. A taxonomy for human-llm interaction modes: An initial exploration. arXiv preprint arXiv:2404.00405, 2024.

Thorsten Händler. A taxonomy for autonomous llm-powered multi-agent architectures. 10 2023. doi: 10.5220/0012239100003598.

Margaret Mitchell. Ethical ai isn't to blame for google's gemini debacle, Feb 2024. URL `https://time.com/6836153/ethical-ai-google-gemini-debacle/`.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. The shifted and the overlooked: a task-oriented investigation of user-gpt interactions. *arXiv preprint arXiv:2310.12418*, 2023.

Hariharan Subramonyam, Roy Pea, Christopher Lawrence Pondoc, Maneesh Agrawala, and Colleen Seifert. Bridging the gulf of envisioning: Cognitive design challenges in llm interfaces, 2024.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild, 2024.