# InteractEva: A Simulation-Based Evaluation Framework for Interactive AI Systems

**Yannis Katsis, Maeda F. Hanafi, Martín Santillán Cooper, Yunyao Li**
IBM Research AI
{yannis.katsis, maeda.hanafi, msantillancooper}@ibm.com, yunyaoli@us.ibm.com

## Abstract

Evaluating interactive AI (IAI) systems is a challenging task, as their output highly depends on the performed user actions. As a result, developers often depend on limited and mostly qualitative data derived from user testing to improve their systems. In this paper, we present *InteractEva*; a systematic evaluation framework for IAI systems. InteractEva employs (a) a user simulation backend to test the system against different use cases and user interactions at scale with (b) an interactive frontend allowing developers to perform important quantitative evaluation tasks, including acquiring a performance overview, performing error analysis, and conducting what-if studies. The framework has supported the evaluation and improvement of an industrial IAI text extraction system, results of which will be presented during our demonstration.

## Introduction

Classical AI systems are based on a two-step development workflow, where developers create an AI model based on labels provided by Subject Matter Experts (SMEs), which is then deployed and made available for SMEs to use. As a result, SMEs are not directly involved in the model building process and their feedback is only incorporated (if at all) after lengthy discussions with developers or other mediators (Amershi et al. 2014). To empower users and build better AI systems, the community has looked into building AI systems with humans-in-the-loop. A particularly popular approach has been *interactive ML/AI (IAI)* systems, which continuously interact with SMEs and incorporate their feedback to create ever-improving versions of the underlying AI models (Fails and Olsen Jr 2003; Amershi et al. 2014).

One challenge though with IAI systems is their *evaluation*. Since the resulting AI model depends on the performed user actions, *how can developers of such systems understand and track their performance accurately and efficiently?* A common technique is to drive evaluation from user testing. SMEs interact with the system and identify and report suboptimal cases, which are then replicated and debugged by developers. While user testing is very valuable, relying solely on it may lead to an ad-hoc whack-a-mole approach towards model improvement that is based only on *limited* evidence of mostly *qualitative* nature.

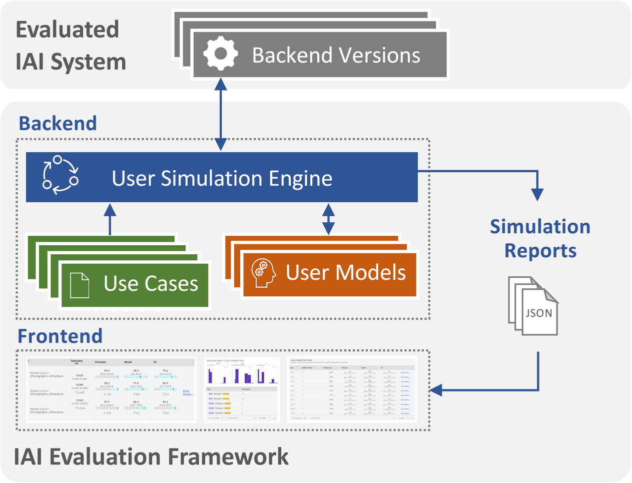

Figure 1: Evaluation framework architecture

To address this issue we present **InteractEva**; a novel evaluation framework for IAI systems tailored towards providing *data-driven, quantitative* guidance in the development of IAI systems. InteractEva leverages a *user simulation backend* to automate the evaluation process, while accounting for the plurality of potential user interactions. The quantitative simulation results can then be explored and analyzed at different granularities through an *interactive frontend* supporting several evaluation tasks, including (a) *acquiring a comprehensive overview of the AI model's performance*, (b) *conducting error analysis*, and (c) *performing flexible what-if studies*. The framework has been successfully used to support the development of Pattern Induction, a commercial IAI NLP extraction system, which will be showcased during our demonstration. A companion video can be found online [1].

**Related work.** Several works have looked into evaluating IAI systems (see (Boukhelifa, Bezerianos, and Lutton 2018; Sperrle et al. 2021) for surveys). These range from human-centered evaluations (focusing on user experience) to algorithm-centered evaluations (studying the robustness of the underlying algorithms). Our work falls under the latter category, but goes beyond prior work by leveraging simula-

[1] https://ibm.biz/BdfCXD

tions to build a systematic end-to-end evaluation framework for general IAI systems (which to the best of our knowledge is the first of its kind). User simulations have also been used to train/evaluate dialogue systems (Kreyssig et al. 2018; Zhang and Balog 2020). However, these works focus on creating simulators for the specific task that closely resemble real user behavior. In contrast, we focus on an end-to-end evaluation framework for general IAI systems (beyond dialogue systems) where different user simulators (referred to as *user models*) can be plugged in.

# IAI Evaluation Framework

Figure 1 shows InteractEva's architecture. The *backend* is responsible for running user simulations, whose results are stored in JSON-formatted simulation reports. These can then be loaded into the *frontend* to allow developers of the tested IAI system to perform various evaluation tasks.

## Backend

The backend is structured around a *user simulation engine* that can simulate large numbers of user interactions and evaluate the performance of the resulting models.

**Backend architecture.** The simulation engine interacts with three components designed to be flexible and capture the requirements of different IAI systems:

- *Use cases* – an extensible corpus of datasets and corresponding ground truth data for specific tasks, designed to reflect real-life use cases of the tested system.
- *User models* – interchangeable modules, each describing a particular class of user interactions. While work in dialogue systems looked into single statistical user models simulating an average user, we found that it is often beneficial to have multiple user models designed to test the effect of particular user behaviours (e.g., check how performance of the system changes when users perform action A vs B); thus enabling complex what-if analyses.
- *IAI system backend* – one or more versions of the backend of the tested IAI system and its associated API. This is used by the simulation engine to (a) programmatically submit user actions and (b) retrieve the predictions of the learned model with their explanations [2] (if available). Capturing model explanations is imperative for enabling error analysis as we will see next.

**Running the simulation.** For a given version of the IAI system and user model, the simulation engine iterates over all provided use cases. For each use case it queries the user model $k$ times to generate $k$ user simulations, referred to as *runs*. For each run, it executes the simulation, retrieves the predictions of the learned model by the IAI system and evaluates it against the ground truth using standard evaluation metrics (e.g., Precision/Recall/F1). The results of all runs over all use cases are stored in a JSON *evaluation report*, forming the foundation for a *quantitative* evaluation.

## Frontend

Evaluation reports can then be loaded into the frontend, where developers of the IAI system can analyze the simulation data and perform the following evaluation tasks:

**Overall performance analysis.** When invoked, the frontend shows evaluation results of the learned model for each use case. However, in contrast to classical evaluation systems that compute only *aggregate* results, InteractEva also shows the *min-max range* of the performance observed across all simulation runs. This helps developers identify not only cases where the system consistently underperforms, but also *long-tail edge cases*, which are especially important for improving AI models (Bornstein and Casado 2020).

**Error analysis.** Developers can subsequently drill down into specific use cases of interest to identify cases of suboptimal system performance together with their root causes. InteractEva enables error analysis by allowing developers to (a) *identify patterns across runs* (e.g., by computing the most common model explanations), as well as (b) *drill down into single runs* and step through them to inspect what the underlying model learns. This allows developers to debug runs directly through the evaluation framework and avoid the time-consuming task of manually creating and running test cases.

**What-if analysis.** Last but not least, developers can load several simulation reports to perform comparative analysis and test various hypotheses. They can compare performance across (a) backend versions (e.g., to verify whether their fixes worked and avoid unwanted regressions) or (b) user models (e.g., to compare the effect of different user actions).

Insights gained through these evaluation tasks can then be used to inform the development process (e.g., decide which part of the backend to improve based on the error analysis) and make data-driven decisions (e.g., decide which user actions to encourage based on the user model comparison).

# Demonstration

To demonstrate the effectiveness of InteractEva, we will showcase how it was used to evaluate and improve Pattern Induction; an industrial IAI system, currently available in Beta on IBM Watson® Discovery (IBM 2021). Pattern Induction is an IAI text extraction system that iteratively learns rule-based extractors by leveraging user-provided (a) examples of extractions and (b) boolean feedback to questions generated by an active learning component[3]. For the demo, we will pre-load into InteractEva's frontend, simulation reports generated during Pattern Induction's development. Based on them, we will guide the audience through a set of *real evaluation scenarios*, explaining how developers leveraged InteractEva to (a) systematically track system performance over time and (b) identify and resolve specific issues in Pattern Induction's backend; tasks that would be either impossible or substantially harder without InteractEva's support for simulation-driven quantitative evaluation.

[2] These can be native explanations of white-box models or explanations created through explainability techniques for black-box models (Xu et al. 2019; Danilevsky et al. 2020). In the IAI system used for the demonstration, explanations take the form of rules.

[3] For additional information please refer to the SEER system (Hanafi et al. 2017) on which Pattern Induction is based.

# References

Amershi, S.; Cakmak, M.; Knox, W. B.; and Kulesza, T. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, 35(4): 105–120.

Bornstein, M.; and Casado, M. 2020. How to improve AI economics by taming the long tail of data. https://venturebeat.com/2020/08/14/how-to-improve-ai-economics-by-taming-the-long-tail-of-data/. Accessed: 2021-09-16.

Boukhelifa, N.; Bezerianos, A.; and Lutton, E. 2018. Evaluation of interactive machine learning systems. In *Human and Machine Learning*, 341–360. Springer.

Danilevsky, M.; Qian, K.; Aharonov, R.; Katsis, Y.; Kawas, B.; and Sen, P. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 447–459. Suzhou, China: Association for Computational Linguistics.

Fails, J. A.; and Olsen Jr, D. R. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, 39–45.

Hanafi, M. F.; Abouzied, A.; Chiticariu, L.; and Li, Y. 2017. *SEER: Auto-Generating Information Extraction Rules from User-Specified Examples*, 6672–6682. New York, NY, USA: Association for Computing Machinery. ISBN 9781450346559.

IBM. 2021. IBM Watson Discovery. https://www.ibm.com/cloud/watson-discovery. Accessed: 2021-09-15.

Kreyssig, F.; Casanueva, I.; Budzianowski, P.; and Gašić, M. 2018. Neural User Simulation for Corpus-based Policy Optimisation of Spoken Dialogue Systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, 60–69. Melbourne, Australia: Association for Computational Linguistics.

Sperrle, F.; El-Assady, M.; Guo, G.; Borgo, R.; Chau, D. H.; Endert, A.; and Keim, D. 2021. A Survey of Human-Centered Evaluations in Human-Centered Machine Learning. *Computer Graphics Forum*, 40(3): 543–567.

Xu, F.; Uszkoreit, H.; Du, Y.; Fan, W.; Zhao, D.; and Zhu, J. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*, 563–574. Springer.

Zhang, S.; and Balog, K. 2020. *Evaluating Conversational Recommender Systems via User Simulation*, 1512–1520. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379984.