Theory and Toolkits for User Simulation in the Era of Generative AI: User Modeling, Synthetic Data Generation, and System Evaluation

Krisztian Balog, Nolwenn Bernard, Saber Zerhoudi & ChengXiang Zhai

Tutorial at the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25) July, 2025

Presenters

Krisztian Balog
University of Stavanger
Norway

Nolwenn Bernard
TH Köln
Germany

Saber Zerhoudi
University of Passau
Germany

ChengXiang Zhai
University of Illinois at Urbana-Champaign
USA

Welcome and Introduction

Objectives

This tutorial aims to equip participants with a solid understanding of the goals, underlying principles, and diverse applications of user simulation within interactive AI, spanning system evaluation, model training, and user modeling.

- Overview of key simulation methodologies
- Discussion of practical resources available

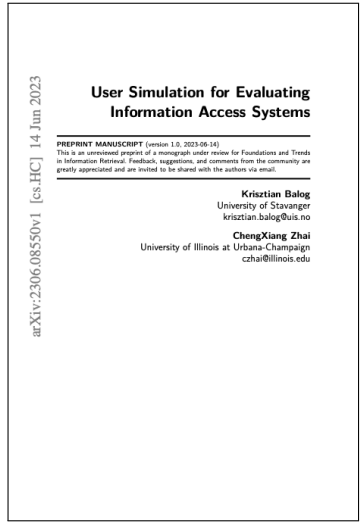
Editions

- Previous editions
  - CIKM'23 and SIGIR-AP'23
  - AAAI '24
  - WWW '24
- This edition
  - Broader perspective on the various roles of user simulation (evaluation, model training, user modeling)
  - Coverage of more recent work, in particular techniques leveraging LLMs
  - Coverage of practical resources

Overview

Resources

- FnTIR book: `https://arxiv.org/abs/2306.08550`
- Website: `https://usersim.ai`
  - Tutorials and slides
  - Annotated bibliography
  - List of toolkits
- Mailing list: usersim@googlegroups.com
- Slack channel: ACM SIGIR / #usersim



Background, Motivation, and Context

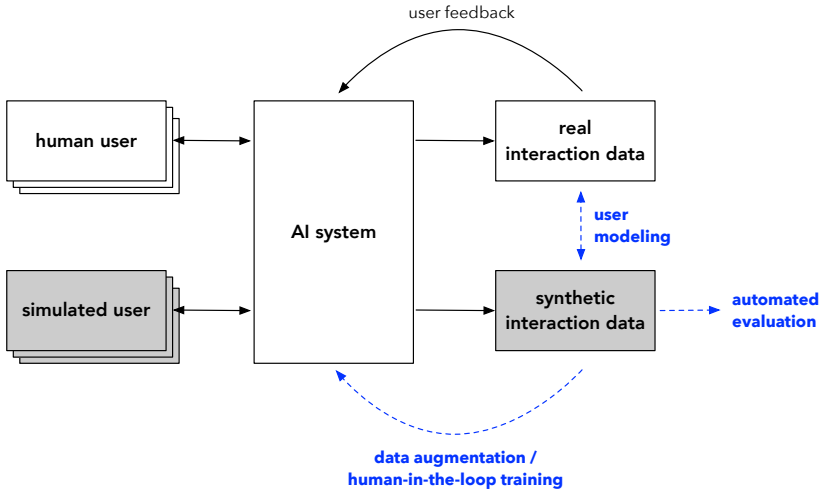
Generative AI: An Information Retrieval Perspective

- Generative AI for Information Retrieval
  - Search engines and recommender systems used to be the major real-world AI application systems
  - Generative AI enabled more intelligent search engines and recommender systems, evolving into intelligent agents that can support conversational information access
- Information Retrieval for Generative AI
  - IR is essential to support Generative AI (e.g., Retrieval-Augmented Generation)
  - Intelligent agents would need to learn how to effectively exploit a search engine (tool learning), and perhaps also create their own search engines (tool creation)
- Many challenges in IR research are also relevant to Generative AI (e.g., user modeling, interaction algorithms, evaluation, personalization)
- User simulation is key to addressing many of those challenges!

Intelligent Interactive Systems

- An intelligent interactive system (IIS) interactively supports a user to finish a task
- User and system take turns to make "moves" in a collaborative "board game" with the objective of helping a user finish the task with minimum overall effort (including cognitive effort)
- System needs to have a model of the user in order to optimize its collaboration with the user in a personalized manner
- Information Access Systems as a special case: search engines, recommender systems, and conversational assistants
- General challenges (user simulation can address all):
  - How to evaluate an IIS with reproducible experiments?
  - How to adapt its interaction with each individual user (model a user)?
  - How to obtain interaction data to train an intelligent interaction algorithm?

User Simulation and Applications



User Simulation

- Formally/computationally/mathematically define a user in the context of finishing a task using an interactive system, including particularly specifying how the user would behave in each interaction context/scenario
- Configuration variables for user simulation:
  1. Task (T): a user's behaviour varies according the task
  2. System (S): a user's behaviour depends on the system (functions) that the user interacts with
  3. User information (U): different users may behave differently when finishing the same task using the same system
- As a computation problem: Given T, S, U, create an agent to simulate every action that user U may take when finishing T by using system S

Partial vs. Complete User Simulation

- Simulation of an action of a user: Given an interaction context (system environment), predict what action a user would take (e.g., given a snippet in a list of search results, predict whether a user would click on it)
- Simulation of a sequence of actions of a user: Given an interaction context, predict the whole sequence of multiple actions that a user would take (need to consider dependency between actions)
- Simulation of a user's interactions in a whole session of finishing a task (there may be multiple sequences of interactions)
- Simulation of a user's general preferences and behaviour across tasks

Applications of User Simulation

- User simulation has many uses, including
  - Performing **large-scale automatic evaluation** of interactive systems (i.e., without the involvement of real users)
  - Gaining **insight into user behaviour** to inform the design of systems and evaluation measures
  - **Analyzing system performance** under various conditions and user behaviours (answering what-if questions, such as “What is the influence of X on Y?”)
  - **Generating synthetic data** with the purpose of training machine learning models, especially reinforcement learning
- simulation cannot be ”perfect” since real users are not well-defined; it only needs to be good enough for the attended application

Requirements and Desiderata

- Requirements
  - **Validity**: The behaviour of simulated users must align with empirical observations of real user behaviour in similar contexts
  - **Variation**: Simulated users should exhibit a range of behaviours, reflecting the diversity of real users
  - **Adaptability**: Simulated users should be able to learn from their interactions with the system, update their expectations, and adjust their behaviour accordingly
  - ...
- Desirable properties:
  - **Interpretability**: simulated behaviour can be understood and adjusted through controllable parameters
  - **Cognitive plausibility**: simulated users should behave in a way that is consistent with human cognition and decision-making processes
  - ...
- *These requirements and desiderata are articulated in qualitative terms; quantitative measures are an open area of research*

Requirements Differ based on Purpose of Simulation

- Requirements differ for simulators depending on intented use (training vs. evaluation)
- **Training objective**: Simulated users should act the same way as real users would act in a given situation (similarity of *policies*)
- **Evaluation objective**: Be predictive of system performance with real users (similarity of *evaluation outcomes*)
- It has been shown that optimizing for one objective (training) does not necessarily imply improvement on another objective (evaluation) (Bernard and Balog, 2024)

Simulation Approaches

- Two broad approaches:
  - *Model-based*: can be rule-based (based on knowledge about how users behave) or interpretable probablistic models (parameters set heuristically or estimated based on observed user data)
  - *Data-driven*: maximize accuracy of fitting any observed real user data, without necessarily imposing interpretability (supervised ML)
- Accurate simulation of observable behaviour may require simulation of latent behaviour (e.g., cognitive state of a user), which makes simulation more interpretable (via interpretable generative models)
- Interpretability is desirable to enable the testing of verifiable hypotheses about users and ensure that evaluation results are meaningful
  - Varying the parameters corresponds to the simulation of different kind of users

User Simulation as an Interdisciplinary Research Topic

- Information retrieval
  - Interactive IR
  - Recommender systems
  - Conversational search and recommendation
- Dialogue systems
- User modeling
- Broadly relevant to many fields: AI, Human-Computer Interaction, Psychology, etc. (see (Balog and Zhai, 2024) for a more complete discussion)

Some Background: IR & RecSys

Both search and recommendation address the problem of providing users with items that are estimated to be relevant to the user's information need, preferences, and/or context, often presented as a ranked list

- Early simulation work in IR
  - Synthetic queries and documents to analyze the effect of changes in query characteristics on the number of documents retrieved (Cooper, 1973)
  - Effectiveness of relevance feedback (Spärck Jones, 1979; Harman, 1992)
- "Second wave" with Interactive IR in the 2000s
  - Relevance feedback (Leuski, 2000; Keskustalo et al., 2008)
  - Query generation (Azzopardi and de Rijke, 2006; Baskaya et al., 2012)
  - Scanning/examination/stopping behaviour (Turpin et al., 2009; Baskaya et al., 2013; Maxwell et al., 2015)

Some Background: Interactive IR

While IR tends to have a strong system focus, interactive information retrieval (IIR) focuses more on users and how they interact with the retrieval system

- Early studies pointing out user effort as an important factor (Cleverdon and Kean, 1968; Salton, 1970)
- Early IIR measures can be categorized around relevance, efficiency, utility, user satisfaction, and success (Su, 1992)
- Important research finding: discrepancy between interactive and non-interactive evaluation results
  - No significant relationship between the effectiveness of a search engine, measured by Mean Average Precision, and real user success in a precision-oriented task (Turpin and Scholer, 2006)
  - Users can adapt their behaviour and can be just as successful with a degraded search system than with a standard one (Smith and Kantor, 2008)

Some Background: Dialogue Systems

The goal of task-based dialogue systems is to help the user accomplish some task, such as make a restaurant reservation or buy a product

- Important idea: modeling human-computer dialogue formally as a Markov Decision Process (MDP) (Levin et al., 2000; Young, 1999)
- Simulation has become the predominant form of dialogue policy learning (Schatzmann et al., 2006; Young et al., 2010)
- Using simulation for evaluation is much less studied

Some Background: User Modeling

User simulation can be regarded as developing a complete and operational user model

- Descriptive vs. formal models
  - *Descriptive models* can provide reasoning and (post-hoc) explanation behind user behaviour
  - *Formal models* are expressed mathematically and have predictive power about why users behave in a certain way

Summary

- User simulation is increasingly important, not just for IR, but also for AI in general, an essential step toward artificial general intelligence
  - It's required for reproducible evaluation of any interactive AI system
  - It enables rigorous and full user modeling and counterfactual analysis
  - It's required for personalizing user interaction in an adaptive and optimal way
  - It enables generation of synthetic interaction data for training machine learning algorithms
- Most past work was on simulation of users in information access, especially simulation of search users, but has been attracting broad attention in related fields
- New opportunities and challenges
  - Major opportunity: A new era of user simulation due to availability of powerful LLMs; interdisciplinary research collaborations; ...
  - Many challenges: Validation of user simulation; Lack of research infrastructure and interdisciplinary community support; ...

Foundations of User Simulation: Behavior Models, Formalisms, and Metrics

Foundations of User Simulation: Behavior Models, Formalisms, and Metrics

- Cognitive Models
- Process Models
- Strategic Models
- Choice and Decision Making in Recommender Systems
- Mathematical Framework

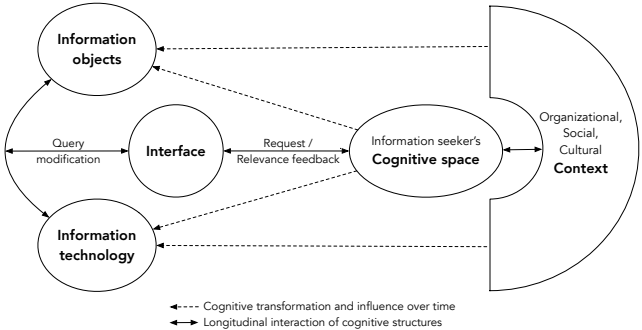
Cognitive Models

Focus on the **cognitive processes** underlying the information-seeking activity (individual's internal representation of a problem situation).

- Belkin's Anomalous State of Knowledge (ASK) hypothesis
  - *"An information need arises from a recognized anomaly in the user's state of knowledge concerning some topic or situation and that, in general, the user is unable to specify precisely what is needed to resolve that anomaly"* (Belkin et al., 1982)
  - Proposes a specific reason as to why people engage in an information-seeking behaviour
  - Assumes the presence of a human intermediary and proposes the ASK to be resolved via co-operative *dialogue* between the user and the intermediary

Cognitive Models

- Information seeking and retrieval (IS&R) research framework (Ingwersen and Järvelin, 2005)
  - Detailed description of essential processes from both the user and system perspectives
  - Emphasizes the *interaction* between the information seeker(s) and the environment surrounding that individual
  - Remains at a very high level of conceptualization



Foundations of User Simulation: Behavior Models, Formalisms, and Metrics

- Cognitive Models
- Process Models
- Strategic Models
- Choice and Decision Making in Recommender Systems
- Mathematical Framework

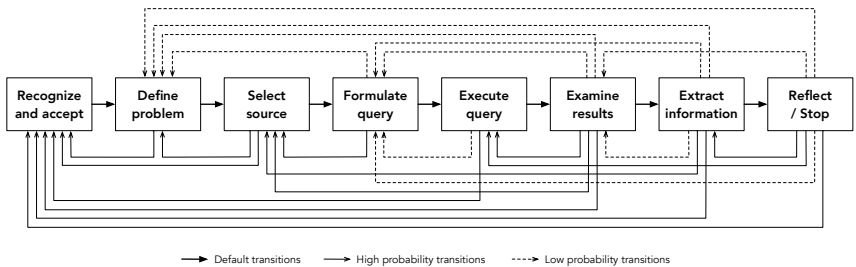
Process Models

Represent the **different stages and activities** during the search process.

- Kuhlthau (1991) identifies six stages:
  1. *Initiation*, recognizing a need for information
  2. *Selection* of the general topic and approach that is expected to yield the best outcome
  3. *Exploration* of the general topic in order to further personal understanding
  4. *Formulation*, where a focused perspective on the topic emerges
  5. *Collection* of the information related to the focused topic
  6. *Presentation*, which completes the search and prepares the results to be presented or used.
- These stages characterize complex information needs and are not necessarily representative for more light-weight tasks

Process Models

- Marchionini (1995) decompose information-seeking into eight sub-processes
  - Sub-processes do not necessarily follow each other in a sequential order, but may develop in parallel and at different rates
  - Sub-processes are further categorized into three classes: (1) understanding, (2) planning and execution, and (3) evaluation and use
    - (1) is mainly a mental activity,(2) and (3) are both mental and behavioural activities



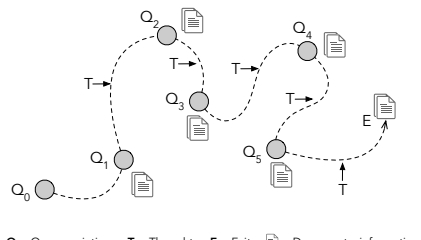
Foundations of User Simulation: Behavior Models, Formalisms, and Metrics

- Cognitive Models
- Process Models
- Strategic Models
- Choice and Decision Making in Recommender Systems
- Mathematical Framework

Strategic Models

Describe **tactics** (high level search strategies) that users employ when searching for information, using analogies from the physical world.

- *Berry-picking model* (Bates, 1989)
  - Considers information seeking analogous to foragers looking for food
  - It assumes that searchers' needs are not satisfied by a single set of retrieved results, scattered like berries on bushes
  - As searchers encounter new pieces of information along the way, those might give them new ideas and directions to follow
  - The model is supported by observational studies (O'Day and Jeffries, 1993; Borgman, 1996)



Strategic Models

- *Information foraging theory* (Pirolli and Card, 1999)
  - Applies ideas from optimal foraging theory ⇒ the searcher maximizes the rate of gaining valuable information over time
    - Optimal foraging theory explains how animals maximize their fitness while they search for food (i.e., gain the most energy for the lowest cost)
  - *Patch* is an area where food can be acquired ⇒ SERP
    - Foragers need to decide how long they want to stay in a patch before moving to the next patch ⇒ examine SERP vs. issue a new query
  - *Scents* indicate to animals their chances of finding prey ⇒ *information scent* are cues presented to on web pages or SERPs
    - When information scent starts to decrease, searchers transition to other information sources

Foundations of User Simulation: Behavior Models, Formalisms, and Metrics

- Cognitive Models
- Process Models
- Strategic Models
- Choice and Decision Making in Recommender Systems
- Mathematical Framework

Choice and Decision Making in Recommender Systems

The ASPECT model (Jameson et al., 2014) distinguishes six human *choice patterns*.

- *Attribute-based choice*: options can be described in terms of attributes, some of which are considered more important than others
- *Consequence-based choice*: consider the consequences of choosing a particular option
- *Experience-based choice*: the person has past experience either with the given choice situation or with particular options
- *Socially-based choice*: people often let their decisions influenced by the choices or advice of others
- *Policy-based choice*: choices can be made according to a specific policy (more common in an organizational setting)
- *Trial-and-error based choice*: a person may opt to randomly select an option to assess it (esp. when none of the above patterns leads to a clear decision)

Foundations of User Simulation: Behavior Models, Formalisms, and Metrics

- Cognitive Models
- Process Models
- Strategic Models
- Choice and Decision Making in Recommender Systems
- Mathematical Framework

Mathematical Framework

Markov decision process (MDP)

- Formally be described by a finite state space $\mathcal{S}$, a finite action set $\mathcal{A}$, a set of transition probabilities P, and a reward function R
- At a given point in time, the agent is in state $s \in \mathcal{S}$, and by executing action $a \in \mathcal{A}$, they transition into a new state s' according to the transition probability $P(s'|s, a)$ and receive reward $R(a, s)$
- The Markov property ensures that this transition depends only on the current state and action (which simplifies modeling and reduces computational complexity)

Example

Routing problems, such as the traveling salesman problem.

- Salesman = agent
- Routes available = the actions that the agent can take while in the current state
- Rewards = the costs of taking specific routes
- Goal = the optimal policy that lowers the overall cost for the entire duration of the trip

Using MDPs for User Simulation

- *State*: needs to encompass the high-level state in the information-seeking process, and the user's mental/cognitive state (goal, intent, preferences, emotional states, etc.); states are not fully observed, leading to Partially Observable MDP (POMDP)
- *Actions*: explicit and implicit actions the user might take
- *State transitions*: easier to model when we consider only explicit states and explicit actions; more complicated with implicit state variables and incompletely specified interaction environment
- *Reward (and Cost)*: models a user's objective of information seeking and the effort a user must make in order to achieve the goal
- *Policy*: determines how to choose an action in each state
  - Can be simple but interpretable models or machine-learned non-interpretable predictive models of user behaviour

Use of MDPs in RL vs. in User Simulation

Reinforcement learning

- The main focus revolves around finding an optimal policy (that maximizes the expected cumulative reward over time)
- Designing effective reward functions is crucial
- Transition probabilities are often observed from an external environment

User simulation

- Policy is based on an explicit model of user behaviour; does not need to be optimal, but needs to be controllable by the system designer
- The reward function can be used to encapsulate the costs and rewards based on observed data (from logs or user studies)
- Transition probabilities are also modeled explicitly based on some model of user behaviour

Duality of User Simulation Agent and AI Agent

- Interaction of an AI agent and an human agent (Yang and Zhai, 2025)

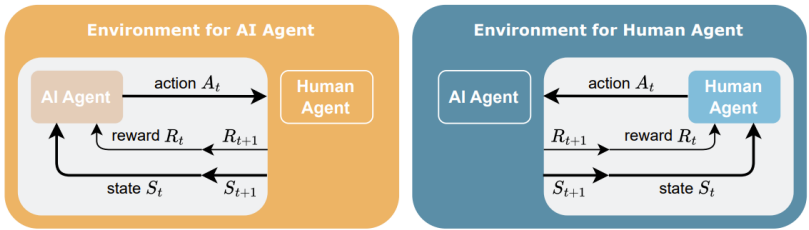


Figure 2: The agent-environment interaction from *AI/human-agent-centric* perspective.

- Implementation of Theory of Mind (Premack and Woodruff, 1978) in AI agent is needed to achieve artificial general intelligence (AGI)
- User simulation is thus an essential step toward AGI

User Simulation for Evaluation: General Idea

- A collection of user simulators are constructed to approximate real users
- A collection of task simulators are constructed to approximate real tasks
- Both user simulators and task simulators can be parameterized to enable modeling of variation in users and tasks
- Evaluation of a system
  - Have a simulated user perform a simulated task by using (interacting with) the system
  - Compute various measures based on the entire interaction history of the whole "task session"
- The method can be used to evaluate any interactive AI system

A General Formal Framework for Simulation-based Evaluation (Zhang et al., 2017)

- Let S be a system, U be a user, and I be the whole process of the interaction of U and S to finish task T
- Measure the system's performance based on I. From a user's perspective, we can measure the performance in two dimensions:
  - Interaction Reward, $R(I, T, U, S)$: the total reward the user has received via the interaction
  - Interaction Cost, $C(I, T, U, S)$: the total cost of the interaction
- In general, the more interaction actions the user makes, the more reward the user can potentially receive and the more cost the user would have to bear (since the user needs to make more effort)
- If one single measure is needed, the reward and cost can be combined, which can be in many different forms

Classic IR Simulator

- Task: find (all) relevant documents
- Interface card: document (snippet)
- User simulator
  - User actions: click, skip (and read next), or stop
  - User always clicks a relevant document when encountering one
  - User always skips a non-relevant document when encountering one
  - User will stop when the effort/cost reaches a budget (or when the user finds the first relevant document in the case of Mean Reciprocal Rank)
- Lap reward: 1 (relevant doc); 0 (non-relevant doc) $\Rightarrow$ Cumulative reward: # relevant docs
- Lap cost: 1 (for scanning each doc/snippet) $\Rightarrow$ Cumulative cost: # docs scanned by the simulated user
- User state: cumulative reward and cost

Classic IR Metrics

- Precision: $R(I, T, U, S)/C(I, T, U, S)$
- Recall: $R(I, T, U, S)/N$, $N =$ maximal possible reward
- Remarks
  - Assumes user stops when the list is exhausted
  - Precision@K and Recall@K: K = cost budget
  - Precision emphasizes more on cost
  - Recall emphasizes more on task completion

Average Precision

- Variable-recall simulator
  - Classical IR simulator with the task of finding N' relevant documents ($N' \in [1..N]$)
  - Stops and only stops when the task is finished
- Average Precision (AP)
  - Average $R(I,T,U,S)/C(I,T,U,S)$ across N variable-recall simulators with N' ranging from 1 to N respectively
  - AP@K: K $=$ cost budget

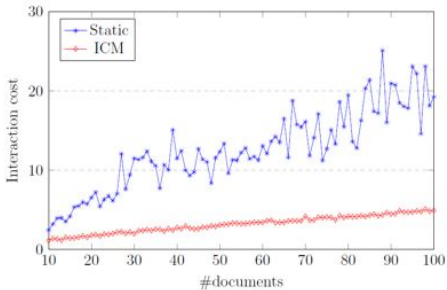
Application of Framework: Evaluating of Tag-based Search Interfaces (Zhang et al., 2017)

- Examples of an interactive search interface beyond ranking
  - Traditional interface: static layout
    - Medium screen: tag list alongside document list
    - Small screen: only tag list or document list at a time, and user needs to click "switch" to switch between the two lists
  - Interface Card Model (ICM) interface: dynamic layout (Zhang and Zhai, 2015)
  - Evaluation based on simulators
    - Task: find target document(s)
    - Simulator never stops until task is completed
    - Metrics: interaction cost

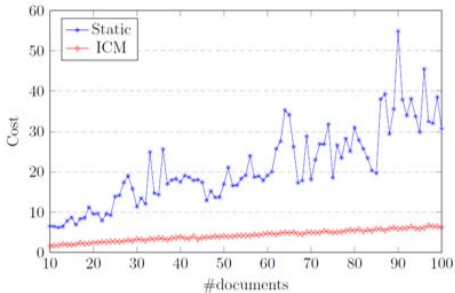
Results of Simulation-based Evaluation (Zhang et al., 2017)

Interface Card Model has consistently lower interaction cost than the static interface

Medium Screen



Small Screen



Validation from Real User Experiment

- Real user experiment (Zhang et al., 2017)
  - ICM is more efficient than static interface
  - The difference is higher on small screen than on medium screen
  - These results are consistent with results of simulation-based evaluation
- Insights about real user behavior
  - Users can well utilize the tag list on the medium screen, but cannot make full use of the tag list on the small screen

Screen size	Sample size	Workers' average
Small	42	$\hat{\tau}_1 = 0.845$, $\hat{\tau}_2 = 0.370$
Medium	38	$\hat{\tau} = 0.211$

Table 6.2: Real user action averages

Summary

- A user simulation agent can be modeled generally with a Partially Observable Markov Decision Process (POMDP)
- A realistic user simulator should be designed based on knowledge about the behaviour of users, especially cognitive models that users are known to follow during information seeking
- User simulation can be used as a general methodology for evaluating any interactive AI system with reproducible experiments
  - It covers traditional measures as special cases, thus naturally generalizes traditional measures
  - It enables definition of new meaningful measures
  - It enables evaluation of sophisticated interactive interfaces, which would otherwise be impossible with the current evaluation methods.

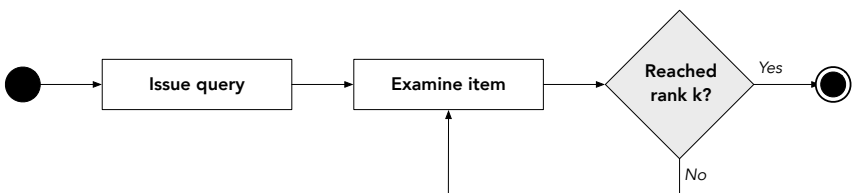
Simulating Interactions

Simulating Interactions

- Interactions with Search Engines
- Interactions with Conversational Assistants

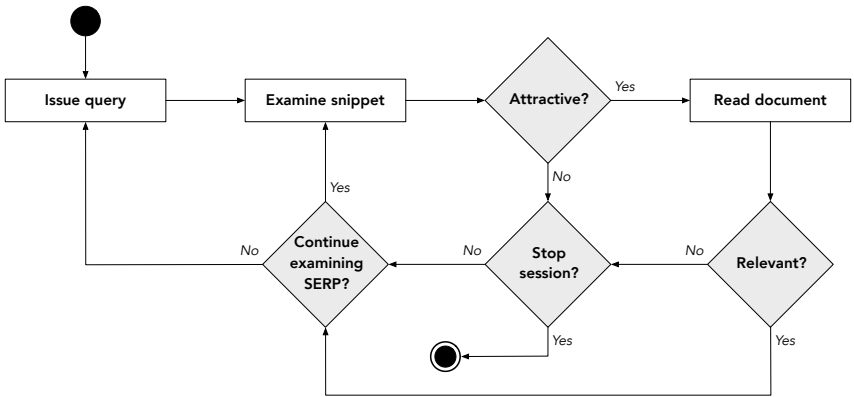
Workflow Models

- Simulation relies on simplified models (of workflows and user behaviour), which allows for "unnecessary complications" to be abstracted away
- The main research challenge is determining what elements of human behaviour to capture in these abstractions, while keeping the models as simple as possible



Naive searcher model, corresponding to highly abstracted user

Search Workflows



Searcher model by Baskaya et al. (2013)

Search Workflows

Complex Searcher Model, proposed by Maxwell et al. (2015) and then further updated in (Maxwell and Azzopardi, 2018)

(A) Topic examination

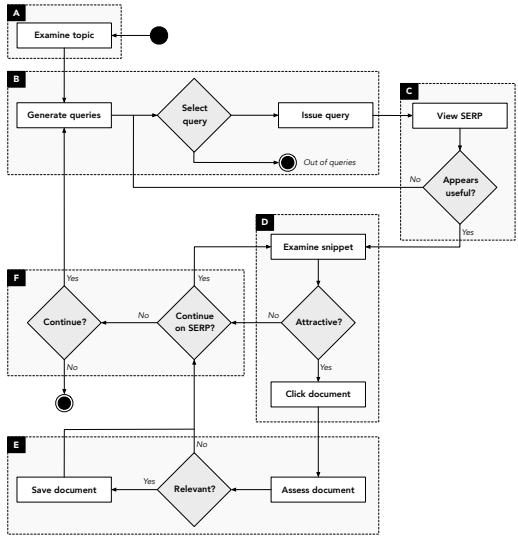
(B) Querying

(C) SERP examination

(D) Result summary examination

(E) Document examination

(F) Deciding to stop



Simulating Queries

Known item search: the searcher aims to re-find a previously encountered item but struggles to formalize the query

- Keyword queries (Azzopardi et al., 2007)
  - Pick a document from a collection and generate a query that is likely to retrieve that document by sampling terms from the document's (unigram) language model
- Tip-of-the-Tongue query generation (TREC 2024 TOT track) (He et al., 2025)
  - Queries are generated by prompting an LLM

> Let's do a role play. You are now a person who watched a movie {ToTObject} a long time ago and forgot the movie's name. You are trying to recall the name by posting a verbose post in an online forum like Reddit describing the movie. Generate a post of length of about 200 words about the movie ToTObject. Your post must describe a vague memory of a movie without mentioning its exact name. People in the forum must have a hard time figuring out which movie you are looking for. The answer should be hard to find in search engines, so do not write too obvious search terms. I will provide you a basic information about the movie, and you have to follow the guidelines to generate a post.
>
> Information about {ToTObject}: {WikipediaSummary}
>
> Guidelines:
> MUST FOLLOW: [...]
> COULD FOLLOW: [...]

Simulating Queries

Evolution of queries during a search session

- Query reformulations (Baskaya et al., 2012)
  - Assumes that a fixed set of terms available for each topic, from which queries may be constructed (e.g., from TREC topic definitions)
  - Five prototypical strategies based on term level changes (adding a new term, replacing a term, etc.)
- Queries within a search session (Carterette et al., 2015)
  - Takes a topic description and previous queries as input
  - Language model from which query terms are sampled is continuously updated based on the results the user has seen for previous queries in the session

Simulating Queries

LLM-based query generation

- Doc2Query (Nogueira et al., 2019): generates a set of quesions that a document may answer to enrich documents during indexing
  - Also effective for simulating interactive search sessions (Engelmann et al., 2023)
  - Can be extended to factor in the user's changing knowledge state (terms from seen documents) (Engelmann et al., 2024)
- Prompting an LLM to generate query variants
  - Using TREC-like topic descriptions as input (Engelmann et al., 2024)

```
Please generate one-hundred keyword queries about <title>.
<description> <narrative>
```

  - In-context learning (Alaofi et al., 2023)

```
<Task Description>
<Input Backstory> ⇒ <Output Query Variants>
<Input Backstory> ⇒
```

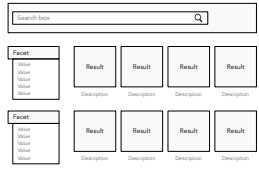
Simulating Scanning Behaviour

- Concerned with how the user processes the list of results presented to them in response to their search query
- Commonly, **sequential browsing** is assumed
- *Cascade model* (Craswell et al., 2008)
  - The user examines each result and decides whether the snippet is deemed relevant enough to warrant a click
  - Snippets below a clicked result are not examined (i.e., the user would stop after having found a relevant result)
- *User browsing model* (Dupret and Piwowarski, 2008)
  - At each rank position, the user first decides whether to look at the snippet or not ("attractive" or not)
  - Then, resume the scan of the result list from the next rank position (whether the result gets clicked or not)

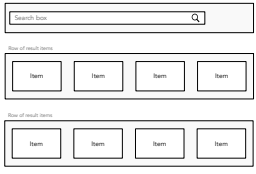
Complex Presentation Layouts

Current approaches rarely consider modern SERPs and alternative presentation layouts, where the top-down traversal assumption is challenged

(a) A traditional "ten blue links" layout.



(b) A product search layout.



(c) A video recommendation layout.

(d) An advertisement layout.

Simulating Clicks

- Mimic a user's decision on whether to click on a search result (to view it in detail) after being exposed to a result (snippet)
- Often integrated with the modeling of scanning behaviour
- Many tradeoffs to be made, especially interpretability vs. prediction accuracy
  - *Position-based simulation*: clicking probability only depends on the rank positions:
    - $P(Click = 1|Rank = i, R_1, R_2, ..., R_k) \approx P(Click = 1|Rank = i)$
    - Naive but generally applicable to any simulation scenario
  - *Content-based simulation*: snippet content is used to model the probability of clicking
    - Intuitively more accurate, but learned models are prone to overfitting and may lose interpretability
- Some click models may not be realistic for simulation purposes
  - Click decision is generally made based on the information shown in the result snippet of a result without having access to the whole document
  - User's prior background knowledge about the query topic is also relevant
    - For example, an expert user may be able to recognize a relevant document based on just a short snippet, where a novice user might not

Simulating Document Processing

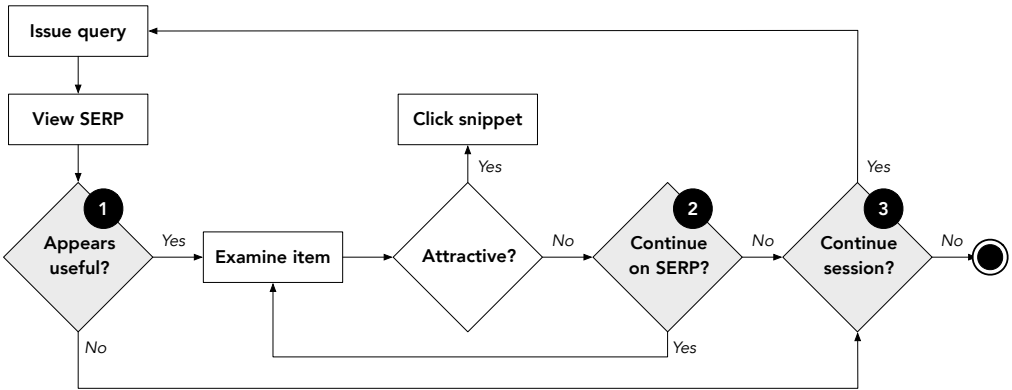
- Processing (i.e., reading and understanding) a document requires an **effort** from the user and yields some **utility** to them (enabling the user to acquire new information, thus changing cognitive state)
- *Dwell time* is often used as a proxy for effort
  - Time (in seconds) needed to process a document of length l, measured in words (Smucker and Clarke, 2012)

    $$T_D(l) = al + b$$

    User is reading at a rate of a seconds per word, and then uses a constant amount of b seconds to make an assessment about the document's relevance
- *Relevance* is used as a proxy for utility
  - Commonly, leveraging ground truth relevance assessments in existing test collections
  - Alternatively, predict whether the user would find the document relevant
    - Represent the user's knowledge state as a language model that evolves based on the documents encountered (Maxwell and Azzopardi, 2016a)
  - Note that utility is meant to be a broader concept than topical relevance!
    - Includes quality, novelty, importance, credibility, etc.
    - Encompasses everything that the user values, e.g., a witty or engaging writing style

Simulating Stopping Behaviour

Users can decide to stop the search process at various points



Excerpt from the updated Complex Searcher Model (Maxwell and Azzopardi, 2018), highlighting various stopping decision points: (1) SERP-level stopping, (2) query-level stopping, and (3) session-level stopping

Simulating Stopping Behaviour

- Several user studies (interviews) to understand *why* people decide to stop
- Users do not apply predetermined criteria, but rather base stopping decisions on the feeling of "good enough"
  - Factors include time constraints, diminishing returns of further information seeking, and increasing redundancy of information encountered
- Different heuristic rules to quantitatively characterize the sense of "good enough," for example,
  - *Satisfaction*: encountering a predefined number of relevant snippets
  - *Searcher frustration*: observing a certain number of non-relevant snippets
  - *Satisfaction or frustration*: stopping as soon as one of the two conditions is met
  - *Time-based*: total amount of time spent on the SERP or time elapsed after the last relevant document found

Simulating Interactions

- Interactions with Search Engines
- Interactions with Conversational Assistants

Conversational AI

- High-level categorization of systems
  - *Goal-driven* (a.k.a. *task-oriented*): aiming to assist users to complete some specific task $\Leftarrow$ our focus
    - **Conversational information access: tasks with an underlying information need, which can be satisfied through a conversation**
    - Includes the tasks of search, recommendation, and question answering (boundaries often blurred)
  - *Non-goal-driven* (a.k.a. *chatbots*): aiming to carry on an extended conversation ("chit-chat"), usually with the purpose on entertainment

Challenges

Traditional search and recommender systems	**Conversational information access**
Limited set of user actions allowed by the system's UI	User intents need to be inferred from free text
Interactions are either driven by the user (search) or by the system (recommendation)	*Mixed initiative*: the user and system both actively participate in addressing the user's information need
Results are restricted to a ranked list of items	Results can be text of arbitrary length (incl. semi-structured elements and questions posed to the user)

$\Rightarrow$ More advanced natural language understanding capabilities are required

Conceptualization of Conversational Information Access

- Dialogue is a sequence of *turns*
- Each turn is a natural language *utterance* from either the user or the system
- *Dialogue act* represents the function or high-level intention of an utterance
  - Typically represented as tuples: *intent* and (optionally) slot-value pairs (e.g., `AFFIRM` or `INFORM(a=x,b=y,...)`)
  - The set of dialogue acts needs to be designed specific to the objectives of the dialogue application (various taxonomies exist)

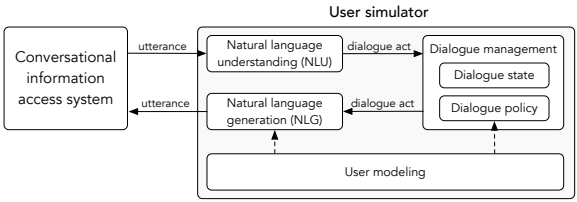
Conceptualization: Dialogue Structure

Dialogue structure: A characterization of dialogues in terms of overall organization, sequencing, and components.

- Three stages in e-commerce conversational search (Zhang et al., 2018)
  - Initiation, conversation, and display
- Mixed-initiative conversational search (Aliannejadi et al., 2021)
  - Querying, feedback, and browsing
- Transition patterns in information-seeking conversations (Qu et al., 2018)
  - START $\Rightarrow$ original question ($\Rightarrow$ potential answer $\Rightarrow$ further details)x3 $\Rightarrow$ potential answer $\Rightarrow$ positive feedback $\Rightarrow$ END
- Context-driven recommendation in the restaurant domain (Lyu et al., 2021)
  - (1) Preference elicitation and refinement in the first stage, (2) inquiry and critiquing in subsequent stages, (3) additional comparisons

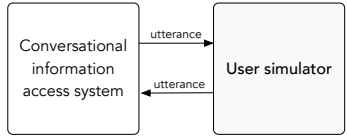
Simulator Architectures

Modular systems



- Model user responses semantically on the level of dialogue acts, then generate the corresponding natural language utterances

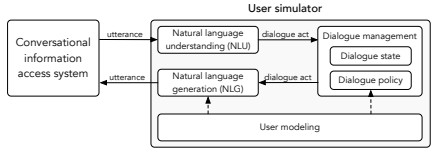
End-to-end systems



- Operate on the utterance level (generate textual responses directly)
- Might yield more fluent dialogues, but do not allow for interpretable user behaviour

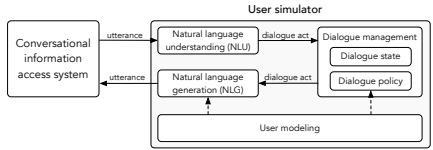
Modular Systems

- *Natural language understanding (NLU)*: converting the (raw) system utterance into an internal semantic representation (dialogue act)
  - *Intent detection* is naturally approached as a classification task
  - *Slot filling* is a sequence labelling problem
- *Dialogue management*: maintaining the dialogue state and determining the next user action
  - The *dialogue state* is based on the notion of a *semantic frame*: collection of slots that together specify what the system needs to know to complete a given task
  - The *dialogue policy* determines how the user should respond



Modular Systems

- *Natural language generation (NLG)*: turning the generated response from a structured representation (dialogue act) into natural language
  - Template-based, retrieval-based, text generation, and hybrid methods
- *User modeling*: capturing the characteristics of individuals that would influence how they interact with the system
  - Information about the user's goal, knowledge, preferences, personal characteristics (e.g., patience), and beliefs about the system



Simulation of Conversational Interactions

- **Task-oriented dialogue** $\Leftarrow$
- Conversational recommendation
- Conversational search

User Dialogue Policy

- Here: task-oriented dialogue in a restricted "slot-filling" sense
  - A *domain ontology* describes the specific intents, slots, and entities that can be talked about
  - The user can specifying their constraints in terms of *informable slots* and requesting information on *requestable slots*
  - Appropriate for modeling user goals in some scenarios (e.g., item recommendation), while others (e.g., exploratory search) are open research problems
- Dialogue is represented as a sequence dialogue acts by the system (a_i^s) and the user (a_i^u) as they take turns: $a_0^s \rightarrow a_0^u \rightarrow a_1^s \rightarrow a_1^u \rightarrow \cdots \rightarrow a_{t-1}^s \rightarrow a_t^s$
- The policy π determines what action a_{t+1}^u the user should take next, given the dialogue history

Simulation of Conversational Interactions

- **Task-oriented dialogue**
  - Statistical user models ⇐
  - Sequence-to-sequence models
  - Large language models (prompting)
- Conversational recommendation
- Conversational search

Statistical User Models: N-grams Models (Eckert et al., 1997)

- Next response based on the dialogue history (resembling the estimation of language models):
$$\pi(s_t) = P(a^u_{t+1}|a^s_t, a^u_t, a^s_{t-1}, a^u_{t-1}, \ldots, a^u_0, a^s_0)$$
- Strong simplifying assumption to condition the next user action exclusively on the preceding system action:
$$\pi(s_t) = P(a^u_{t+1}|a^s_t)$$
- Conditional probabilities estimated from an annotated dialogue corpus
- No information about the user's goal, no constraints on the simulated user behaviour $\Rightarrow$ fails to produce realistic dialogues
  - Placing constraints on the dialogue flow yields somewhat more realistic dialogues (Levin et al., 2000), but the consistency between user responses across the dialogue is still not guaranteed

Statistical User Models: Goal-directed User Model with Memory (Pietquin, 2004)

- Explicit representation of the user goal as a sequence of slot-value pairs with priority: $G = \langle (slot_1, value_1, prior_1), \ldots, (slot_n, value_n, prior_n) \rangle$
  - When the user is prompted for the relaxation of some attribute, slot-value pairs with a higher priority are less likely to be relaxed
- Dialogue history at time t is represented as a vector $h_t = \langle c_1, \ldots, c_n \rangle$
  - c_i is the count of the occurrences a value is provided for the corresponding $slot_i$
  - Enables the simulator to disclose new information to the system if mixed initiative is supported
- Allows for automatic evaluation in terms of full or partial task completion (given how goals are represented)

Statistical User Models: Agenda-based Simulator (Schatzmann et al., 2007)

- Factors the user state into an agenda and a goal $s_t = (A_t, G_t)$
- Agenda A_t is a stack-like structure, representing the pending intentions of the user
- Goal is a tuple $G_t = (C_t, R_t)$, where
  - C_t is a set of domain-specific constraints the user wants to impose on the dialogue
  - R_t specify requests, i.e., slots whose values are initially unknown to the user and will need to be filled out during the conversation
- For example (restaurant recommendation): looking for the name, address, and phone number of a centrally located bar serving beer:

$$C_0 = \begin{bmatrix} \text{type} & = & \text{bar} \\ \text{drinks} & = & \text{beer} \\ \text{area} & = & \text{central} \end{bmatrix} \qquad R_0 = \begin{bmatrix} \text{name} & = & \\ \text{addr} & = & \\ \text{phone} & = & \end{bmatrix}$$

Statistical User Models: Agenda-based Simulator (Schatzmann et al., 2007)

- Agenda initialization
  - All goal constraints set to `INFORM` acts and all goal requests set to `REQUEST` acts
  - `BYE` added at the bottom of the agenda to close the dialogue

$$A_0 = \begin{bmatrix} \text{INFORM}(\text{type} = \text{bar}) \\ \text{INFORM}(\text{drinks} = \text{beer}) \\ \text{INFORM}(\text{area} = \text{central}) \\ \text{REQUEST}(\text{name}) \\ \text{REQUEST}(\text{addr}) \\ \text{REQUEST}(\text{phone}) \\ \text{BYE} \end{bmatrix}$$

- As the conversation progresses, the agenda and goal are dynamically updated
  - Next user action simplifies to popping items from the top of the agenda
  - Agenda updates are push operations, where dialogue acts get added on top of the agenda

Simulation of Conversational Interactions

- **Task-oriented dialogue**
  - Statistical user models
  - Sequence-to-sequence models $\Leftarrow$
  - Large language models (prompting)
- Conversational recommendation
- Conversational search

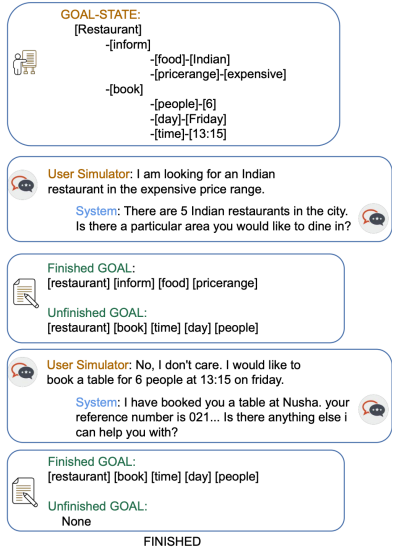
Sequence-to-sequence Models

User simulator is learned fully data-driven from dialogue corpora

- Variations on input/output
  - Utterance→Utterance (Crook and Marin, 2017)
    - No explicit goal modeling
  - Dialogue act→Dialogue act (Gür et al., 2018)
  - Feature vector→Dialogue act (El Asri et al., 2016; Lin et al., 2021)
    - Features encode dialogue history and goal state
  - Feature vector→Utterance (Kreyssig et al., 2018)
  - Context→Dialogue act + utterance (Lin et al., 2022)

Goal-State Guided Simulator (Cheng et al., 2022)

- Track goal states based on user and system actions
  - Add current goal states at the front of user simulator inputs
  - Predict the user actions and then obtain finished goals by combining both user actions and dialogue system actions
  - The user simulator will generate user utterances based on these unfinished goals at next turn
- Simply concatenate goal states with utterances and approach it as a text generation task



Metaphorical User Simulator (Sun et al., 2023)

- Introduces a *metaphor* module that is responsible for retrieving similar dialogue records from a corpus
  - Two-stage retrieval process: (1) TF-IDF-based candidate generation and (2) a learnable ranker that considers utterances, user preferences, and dialogue state
- The policy predicts the next user action and user satisfaction jointly based on the metaphor and the dialogue context

$\mathcal{P}$	{domain=hotel; parking=yes; price=cheap;}
$\mathcal{U}$	{**user**: I am looking for a place . . . . **system**: Okay, do you have. . . . }
$\mathcal{S}$	{**user**: inform type=hotel, price=cheap. **system**: request area. }
a_{t+1}	{**user**: <neutral>; inform parking=yes. }
u_{t+1}	{**user**: No, I just need to have parking . }

Other Research Threads in a TOD Context

- Estimating user satisfaction (Sun et al., 2021; Hu et al., 2023)
  - Enables more human-like simulation (conditioning user response based on their satisfaction)
  - Incorporate signal into TOD training to enhance the quality of system responses
- Joint training of the dialogue system and the user simulator (Tseng et al., 2021)
  - Improving both the conversational agent and the user simulator via self-play
- Multi-task learning (Kim and Lipani, 2022)
  - Predicting user satisfaction, next action and utterances at the same time in a multi-task learning setting

Simulation of Conversational Interactions

- **Task-oriented dialogue**
  - Statistical user models
  - Sequence-to-sequence models
  - Large language models (prompting) $\Leftarrow$
- Conversational recommendation
- Conversational search

Large Language Models

Generate user-side utterances in the conversation by prompting an LLM

- Using in-context learning (Terragni et al., 2023; Davidson et al., 2023; Li et al., 2022)
  - Manually or randomly selected example conversations
  - Retrieving more relevant examples (based on slots in the target user goal) from a corpus
- Self-verification of responses (Li et al., 2022; Luo et al., 2024b)
- Simulators can play both the user and system roles in a conversation Data Augmentation for Conversational AI (Soudani et al., 2023)

Simulation of Conversational Interactions

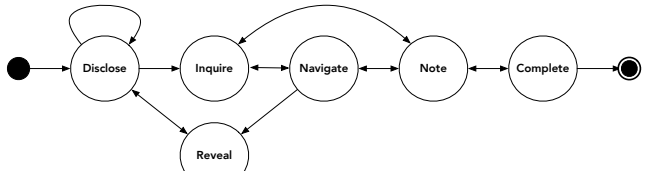
- Task-oriented dialogue
- **Conversational recommendation** $\Leftarrow$
- Conversational search

User Simulation for Conversational Recommendation

- Goals: elicit user preferences using natural language interactions, point users to potential items of interest, and process feedback by users on the made suggestions
- Conversational recommendation can naturally be framed in the classical sense of task-oriented dialogue systems:
  - Find items that satisfy the set of constraints expressed by the user, which can be represented in terms of slot-value pairs: $C = \langle (slot_1, value_1), \ldots, (slot_n, value_n) \rangle$

Agenda-based Simulation (Zhang and Balog, 2020)

- Agenda-based dialogue policy, guided by an *interaction model*
  - Interaction model specifies the set of user actions and expected system response for each user action
  - The latter allows the simulator to determine whether the system responds to the user with an appropriate action (i.e., "understood" the user)
- Users are characterized by their *preference model*, which is a a knowledge structure with $(slot, value, pref)$ triples
  - Grounded in actual user preferences, by randomly subsampling item ratings from a dataset of historical user-item interactions
  - To ensure the consistency of preferences, a *personal knowledge graph* is used



Large Language Models

- Zero-shot prompting an LLM (Wang et al., 2023)
- Integrating dialogue generation with existing recommendation models (Wang et al., 2023)
- Controlling the behaviour of the simulator using a single prompt is challenging (Liang et al., 2024; Zhu et al., 2024)

Limitations

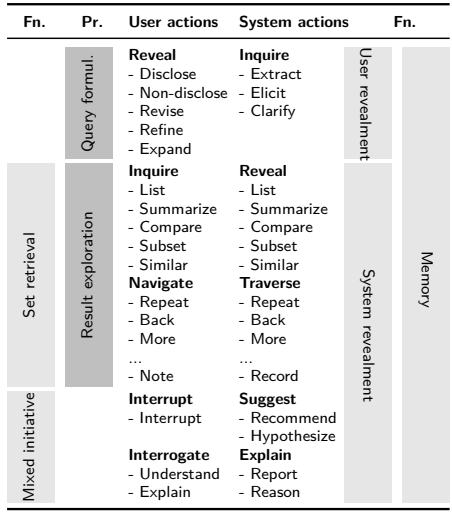
- Simulators differ from real users; this can be mitigated to some extent by prompting strategies and model choice (Yoon et al., 2024)
  - Simulators mention less diverse items compared to real users $\Rightarrow$ Prompting with interaction history enhances item diversity
  - Simulators may poorly represent real user preferences $\Rightarrow$ Adding varying levels of 'pickiness' improves preference alignment
  - Simulators express preferences differently from real users
    - Real users often express opinions in subtler ways
    - Predictable which aspects simulators will mention
    - Biased towards positive sentiment
  - Simulators struggle to generate a diverse pool of personalized requests
  - Simulators may not capture subtle nuances in requests, and thus reject relevant recommendations

Simulation of Conversational Interactions

- Task-oriented dialogue
- Conversational recommendation
- Conversational search $\Leftarrow$

User Simulation for Conversational Search

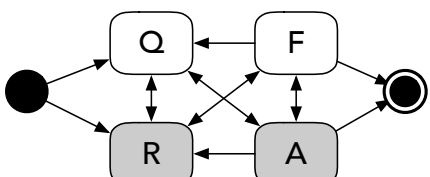
- Taxonomy of user and system actions by Azzopardi et al. (2018)
  - Fn: conversational functionality according to (Radlinski and Craswell, 2017)
  - Pr: search process in (Trippas et al., 2018)
- A task-oriented approach is challenging as the user's information need is often not well-defined and it's difficult to capture progress towards task completion



Conceptualization

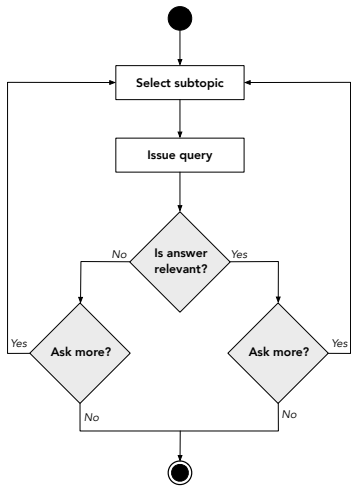
QRFA (Vakulenko et al., 2019): generic model of conversational information seeking processes.

- Four basic classes: 2 for user and 2 for system (proactive and reactive)
  - User: **Query** and **Feedback** $\Leftarrow$ Actions to simulate
  - System: Request and Answer



Simulating User Questions (Lipani et al., 2021)

- It is assumed that the user's goal is to learn about a set of subtopics by interacting with the system
- Both user queries and system responses are represented as *subtopics*
- At each dialogue turn the user asks about a particular subtopic
- Based on the relevance of the system's response, the user will ask further questions (about the same subtopic or a different one) or stop querying



Simulating User Questions (Lipani et al., 2021)

- The user dialogue policy is based on the notion of persistence in querying the system, depending on the relevance of the answer to the previous query
- Start with a query in turn 1
- For any subsequent turn t
  - Leave with probability $P(L_t = l|Q_t = q, R_t = r)$ if the system response was relevant
  - Leave with probability $P(L_t = l|Q_t = q, R_t = \bar{r})$ if the result was not relevant
  - Both probabilities are estimated from user logs
- Overall, the following data components are required:
  - A sample of information needs (i.e., topics)
  - For each topic, a pre-defined set of subtopics
  - Subtopic-level relevance judgments
  - A dialogue dataset with subtopic annotations for the estimation of state transition probabilities

Simulating Answers to Clarifying Questions (Salle et al., 2021)

- Simulating how a user would respond to clarifying questions that are in the form: "Are you looking for *[facet]*?"
- *User intent model*: represents the user's information need and estimates whether the clarifying question matches the user's intent
  - Implemented by fine-tuning a BERT model for binary classification
- *Persona model*: specifies personal user characteristics
  - *Cooperativeness* ($\in [0, 1]$): the user's willingness to help the system by giving an informative answer (e.g., "No, I'm looking for *[intent]*") vs. simply "Yes" or "No")
  - *Patience*: maximum effort (number of turns) the user is willing to spend interacting with the system

Simulating Answers to Clarifying Questions (Sekulić et al., 2022)

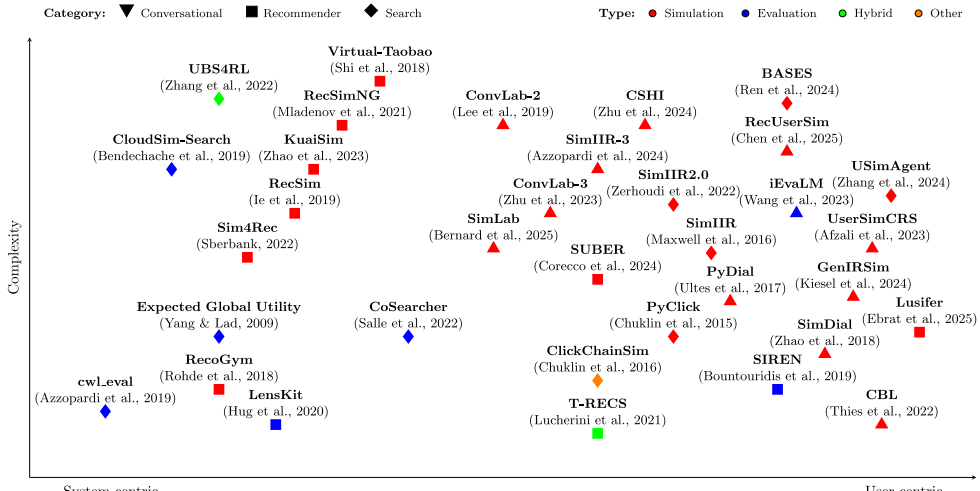
- Fine-tuning a transformer-based large language model (LLM) for the task of answering clarifying questions
- DoubleHead GPT-2 with language modeling and classification losses
- Training input part 1 is given as the sequence `in[SEP]q[SEP]cq[bos]a[eos]`
  - in: textual description of the user's information need
  - q: user's query
  - cq: clarifying question asked by the system
  - a: answer given by the user
  - `[bos]` and `[eos]` are special tokens indicating the beginning and end of a sequence
  - `[SEP]` is a separation token
- Training input part 2: distractor answer and a binary label indicating which of the answers is preferable
  - Distractor answers are sampled from the training dataset heuristically

User Simulation for Conversational Search (Owoicho et al., 2023)

- Generating a variety of utterances by few-shot prompting a ChatGPT model:
  - Queries to seek information
  - Answers to clarifying questions
  - Feedback to system responses
- Note: LLM-based approaches generate answers that are fluent and natural-sounding, they work much like black boxes
  - The behaviour of the simulated user can be controlled only indirectly and only to a certain extent via training examples

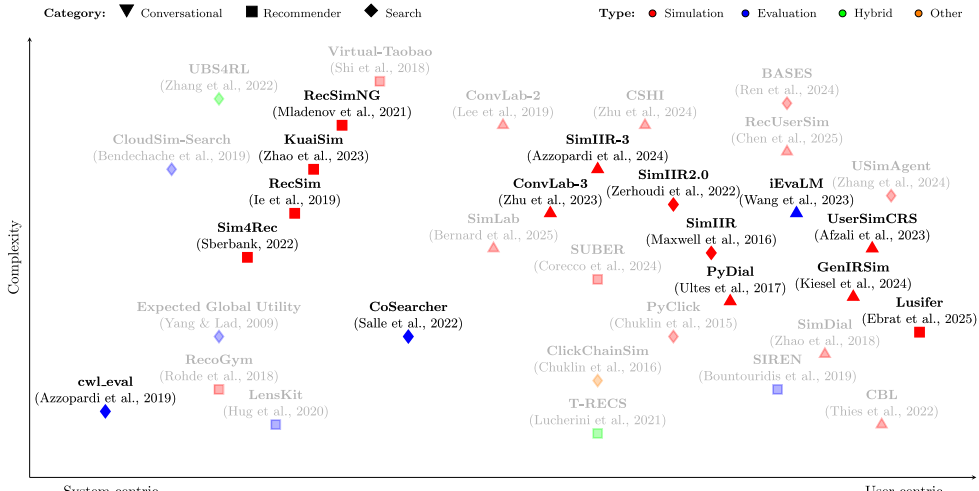
Simulation Toolkits and Frameworks

Landscape of Simulation Frameworks



Simulation frameworks and tollkits plotted by *system- vs. user-centric* orientation and *complexity*.

Landscape of Simulation Frameworks



Simulation frameworks and tollkits plotted by *system- vs. user-centric* orientation and *complexity*.

Simulation Toolkits and Frameworks

- Simulating Search Interactions ⇐
- Simulating Recommendation Interactions
- Simulating Conversational Systems

Simulating Search Interactions (1/4)

- **cwl_eval** (Azzopardi et al., 2019)
  - *Architecture*: An evaluation framework built on the Cost–Utility–Length model.
  - *Scope*: Unifies a wide range of IR metrics (from Average Precision and nDCG to Rank-Biased Precision) by modeling the utility (gain) and cost of examining search results. Its primary task is to generate measurements about predicted user interactions with a static ranked list.
  - *Ease of Use*: Python package (`cwl-eval`)—easy to add to any pipeline.
  - *Extensibility*: Specify one function, derive many related measurements (e.g., Expected Utility, Cost, Depth).

Simulating Search Interactions (2/4)

- **SimIIR** (Maxwell and Azzopardi, 2016b)
  - *Architecture*: An open-source toolkit implementing the *Complex Searcher Model (CSM)*. Simulations are configured via an XML file specifying four components: output logging, topics, user behavioural models, and a search interface (e.g. a Whoosh index).
  - *Scope*: Enables interactive IR experiments by modelling the full search loop—querying, snippet examination, relevance assessment, and stopping—so researchers can study user–system interplay.
  - *Ease of Use & Extensibility*: XML-driven and highly modular; swapping in new behavioural modules is straightforward.
  - *Dependencies*: External tools such as `ifind` and `trec_eval`; a search backend (e.g. `Whoosh`); and TREC-style topics and relevance judgement files.

Simulating Search Interactions (3/4)

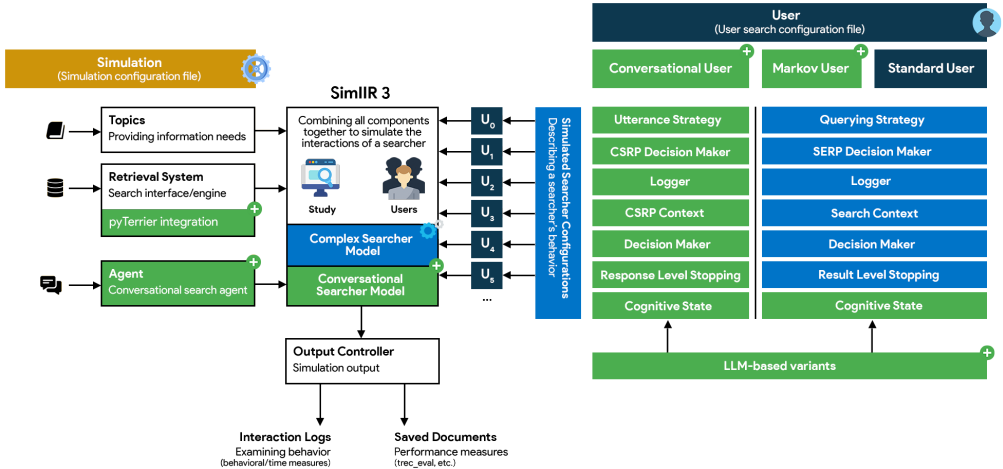
- **SimIIR 2.0** (Zerhoudi et al., 2022)
  - *Architecture*: Extends the original SimIIR by swapping static modules for dynamic, data-driven components trained on real user logs. Adds a dynamic query generator and Markov decision models.
  - *Scope*: Models richer search sessions via multiple user types (e.g. "exploratory" vs. "lookup") with learned behaviours; the generator can formulate new queries from terms seen during the session.
  - *Ease of Use & Extensibility*: Keeps SimIIR's modular design, updated to modern Python; serves as a platform for implementing and comparing query or user-behaviour simulators.
  - *Dependencies*: Needs access to interaction logs to train the Markov models that drive user decisions.

Simulating Search Interactions (4/4)

- **SimIIR 3.0** (Azzopardi et al., 2024)
  - *Architecture*: Updated for generative & conversational IR; Markov user models with cognitive states; LLM wrapper (e.g. LangChain).
  - *Scope*: Simulates conversational search and RAG workflows; LLM components handle query, relevance, stopping.
  - *Ease of Use & Extensibility*: Community-led; integrates PyTerrier; same loop, more powerful parts.
  - *Dependencies*: Modern LLM libraries + PyTerrier.

Simulating Search Interactions (4/4)

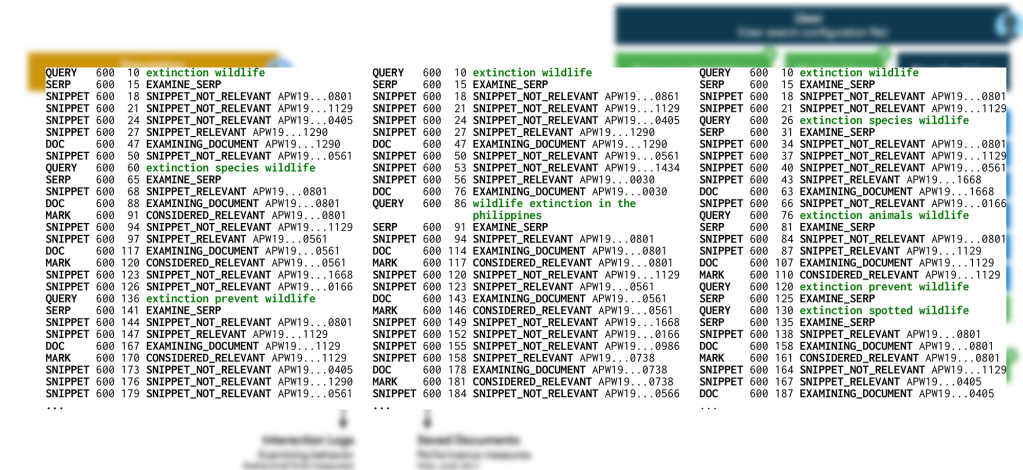
- **SimIIR 3.0** (Azzopardi et al., 2024)



SimIIR 3.0 Framework by (Azzopardi et al., 2024).

Simulating Search Interactions (4/4)

- **SimIIR 3.0** (Azzopardi et al., 2024)



SimIIR 3.0 simulated search logs example.

Simulation Toolkits and Frameworks

- Simulating Search Interactions
- Simulating Recommendation Interactions ⇐
- Simulating Conversational Systems

Simulating Recommendation Interactions (1/5)

- **RecSim** (Ie et al., 2019)
  - *Architecture*: Python/TensorFlow simulator–agent platform with three core components: user model, document model, and user-choice model.
  - *Scope*: Enables RL research in recommender settings by creating synthetic user environments.
  - *Ease of Use & Extensibility*: Highly extensible (custom user dynamics, choice models, agents). Jupyter tutorials help navigate its abstract component design.
  - *Dependencies*: TensorFlow and, for some agents, the Dopamine framework.

Simulating Recommendation Interactions (2/5)

- **RecSim NG** (Mladenov et al., 2021)
  - *Architecture*: Probabilistic platform (Edward2 + TensorFlow) modelling the environment as composable dynamic Bayesian networks (DBNs) for principled uncertainty.
  - *Scope*: From single user–agent loops to full ecosystems (users, creators, advertisers), with data-driven parameter learning.
  - *Ease of Use & Extensibility*: Greater power but higher complexity than RecSim; DBN structure affords deep extensibility for causal modelling.
  - *Dependencies*: TensorFlow, Edward2.

Simulating Recommendation Interactions (3/5)

- **Lusifer** (Ebrat et al., 2024)
  - *Architecture*: An LLM-driven simulation environment. The LLM maintains and updates a user profile at every interaction step, explaining preference shifts in natural language.
  - *Scope*: Generates dynamic user feedback—ideal for RL recommender training; captures concept drift and cold-start items.
  - *Ease of Use*: Needs access to a strong LLM and good prompt engineering; behavior is controlled via prompts.
  - *Extensibility*: Domain-agnostic—works wherever rich item metadata is available.
  - *Dependencies*: Powerful LLM (GPT-3+ class), item-content dataset, and some starter user-history data.

Simulating Recommendation Interactions (4/5)

- **Sim4Rec** (Volodkevich et al., 2025)
  - *Architecture*: A Python framework built on PySpark for large-scale simulation.
  - *Scope*: Models iterative learning—each new batch of simulated user logs refits the recommender to mimic online training cycles.
  - *Ease of Use*: Best suited to users already comfortable with the Spark ecosystem.
  - *Dependencies*: PySpark.

Simulating Recommendation Interactions (5/5)

- **KuaiSim** (Zhao et al., 2023)
  - *Architecture*: Comprehensive RL environment with a user model that yields multi-level feedback (clicks, likes, follows) across sessions; integrated evaluation protocols and baseline agents.
  - *Scope*: Targets RL recommender research; leverages the large-scale, multi-behaviour KuaiRand dataset. Supports list-wise, whole-session, and cross-session retention tasks.
  - *Ease of Use*: Open-sourced on GitHub with example agents; familiar to anyone who has used frameworks like RecSim.
  - *Extensibility*: Modular—plug in new algorithms or tweak user-behaviour parameters such as novelty preference.
  - *Dependencies*: Python, standard DL/RL libraries; data coupling with KuaiRand.

Simulation Toolkits and Frameworks

- Simulating Search Interactions
- Simulating Recommendation Interactions
- Simulating Conversational Systems $\Leftarrow$

Simulating Conversational Systems (1/6)

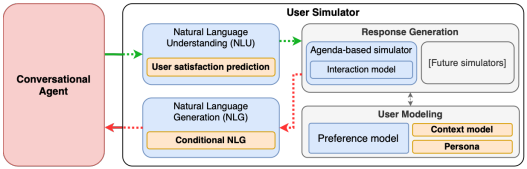
- **ConvLab-3** (Zhu et al., 2022)
  - *Architecture*: A flexible toolkit whose cornerstone is a Unified Data Format bridging diverse dialogue datasets and models; modular NLU, DST, Policy, NLG, plus a robust RL framework.
  - *Scope*: Targets task-oriented dialogue, offering data-driven user simulators (TUS, GenTUS) for RL training and interactive evaluation across datasets such as MultiWOZ and Schema-Guided.
  - *Ease of Use & Extensibility*: Extensive docs and tutorials; unified format lets newcomers add new datasets or models with minimal effort, while experts can swap components freely.
  - *Dependencies*: `PyTorch`, `transformers`.

Simulating Conversational Systems (2/6)

- **PyDial** (Ultes et al., 2017)
  - *Architecture*: Open-source toolkit with modular pipeline—Semantic Parser, Belief Tracker, Policy, and NLG. Multi-domain support via a Topic Tracker delegating input to domain-specific pipelines.
  - *Scope*: Multi-domain statistical task-oriented dialogue. Includes an agenda-based user simulator for training RL policies without expensive human trials.
  - *Ease of Use & Extensibility*: Domain-independent; one config file controls the whole system. Each module has a "manager" concept that makes swapping components easy.
  - *Dependencies*: Pure-Python toolkit predating large transformers—runs without heavyweight DL frameworks.

Simulating Conversational Systems (3/6)

- **UserSimCRS** (Afzali et al., 2023)
  - *Architecture*: Two Python libs—`DialogueKit` (generic dialogue) and `UserSimCRS` (agenda-based + LLM user simulation for persona, context, satisfaction).
  - *Scope*: Evaluation of black-box Conversational Recommender Systems (CRS) with realistic, low-data user simulations.
  - *Ease of Use & Extensibility*: Ships with baselines and the MovieBot case study. Modular design (built on `DialogueKit`) eases future model swaps or feature adds.



UserSimCRS architecture by (Afzali et al., 2023)

Simulating Conversational Systems (4/6)

- **CoSearcher** (Salle et al., 2021)
  - *Architecture*: Simulation framework for conversational search refinement and clarification, using a stochastic, parameterized user simulator.
  - *Scope*: Focuses on query-clarification dialogs; exposes **cooperativeness** and **patience** parameters to study their effect on success.
  - *Ease of Use & Extensibility*: Enables large-scale experiments impossible with real users; code available on GitHub.
  - *Dependencies*: Standard Python libraries.

Simulating Conversational Systems (5/6)

- **iEvaLM** (Wang et al., 2023)
  - *Architecture*: Interactive evaluation framework for CRS that leverages an LLM as a user simulator—allowing realistic tests of other systems, including ChatGPT-style models.
  - *Scope*: Rethinks CRS evaluation in the LLM era; the simulator adopts a persona grounded in true user preferences and can handle attribute Q&A as well as free-form chit-chat.
  - *Ease of Use & Extensibility*: Highly configurable; researchers can script systematic, dynamic evaluations with minimal effort.
  - *Dependencies*: Requires API access to the LLMs used for simulation and system evaluation.

Simulating Conversational Systems (6/6)

- **GenIRSim** (Kiesel et al., 2024)
  - *Architecture*: LLM-based platform with metrics, baselines, and configurable user simulators.
  - *Scope*: Tailored to Generative IR evaluation—especially Touché debate tasks; scores quality, relevance, clarity.
  - *Ease of Use*: Built for shared-task participants; a public demo is available.
  - *Extensibility & Dependencies*: Simulators support varied argument strategies; relies on large-language models throughout.

Simulator Quality: Validation Principles and Methods

Validating Simulators

- Motto: *"Simulation does not need to be perfect in order to be useful"*
- **How can we evaluate if a simulator imitates the behaviour of real users *sufficiently well*?**

Requirements and Desiderata

- Requirements
  - **Validity**: The behaviour of simulated users must align with empirical observations of real user behaviour in similar contexts
  - **Variation**: Simulated users should exhibit a range of behaviours, reflecting the diversity of real users
  - **Adaptability**: Simulated users should be able to learn from their interactions with the system, update their expectations, and adjust their behaviour accordingly
  - ...
- Desirable properties:
  - **Interpretability**: simulated behaviour can be understood and adjusted through controllable parameters
  - **Cognitive plausibility**: simulated users should behave in a way that is consistent with human cognition and decision-making processes
  - ...

Requirements Differ based on Purpose of Simulation

- Requirements differ for simulators depending on intented use (training vs. evaluation)
- **Training objective**: Simulated users should act the same way as real users would act in a given situation (similarity of *policies*)
- **Evaluation objective**: Be predictive of system performance with real users (similarity of *evaluation outcomes*)
- It has been shown that optimizing for one objective (training) does not necessarily imply improvement on another objective (evaluation) (Bernard and Balog, 2024)

Validation Approaches

- Validation requires a **comparison against human users**
- Would a simulated user produce data that matches the characteristics of real user data?
  - Specific characteristics (e.g., dialogue length, distribution of dialogue acts)
  - (In)distinguishability of data produced by simulated vs. real users
- Would a simulated user lead to similar performance measures to what is obtained from real users?
  - Weak requirement: reproduce the same *relative* ranking of systems
  - Stronger requirement: closely approximate the *absolute* measures performance

Comparing Specific Characteristics

- Search
  - Query characteristics (e.g., length, terms), query reformulation patters
  - Click patterns (e.g., clicking on relevant vs. non-relevant results)
- Conversational agents
  - *Utterance-level*: commonly, human raters evaluate the generated responses along different dimensions (e.g., naturalness, usefulness, grammar)
  - *Dialogue-level*:
    - *High-level dialogue features*: avg. dialogue length, ratio of user vs. system actions, etc.
    - *Dialogue style*: distribution of dialogue acts, user cooperativeness (proportion of slot values provided when requested), etc.
    - *Dialogue efficiency*: success (or task completion) rate, reward, completion time, etc.

(In)distinguishability of Data

- Human evaluation protocol for conversations (Zhang and Balog, 2020)
  - Assessors are given transcripts of two conversations, in random order
  - They have to guess which of the two is the generated by a human
  - The more often the simulated response 'fools' the human assessor, the more realistic it is
    - However, human assessors can be tricked

Realism of LLM-based Simulations

- Comparing dialogues produced by real humans vs. simulated users
- Lack of the natural variation found in human interactions (Terragni et al., 2023; Davidson et al., 2023; Wang et al., 2024; Yoon et al., 2024)
  - E.g., mention less diverse items
- Poor representation of human preferences or expression of preferences (Yoon et al., 2024)
  - E.g., mention the same aspects
- Generation of overly "perfect" responses, potentially leading to the simulation of unrealistic "superusers" (Wang et al., 2023, 2024)
  - E.g., already full knowledge of the item that's being recommended

Performance Prediction

- How well can a simulator predict the performance of a system with real users?
- Commonly: System ranking correlation
  - However, there is often a limited number of systems to compare

Method	Reward	Success Rate
Real users	**A** (8.88) > **B** (7.56) > **C** (6.04)	**B** (0.864) > **A** (0.833) > **C** (0.727)
QRFA-Single	**A** (8.04) > **B** (7.41) > **C** (6.30)	**B** (0.836) > **A** (0.774) > **C** (0.718)
CIR6-Single	**A** (8.64) > **B** (8.28) > **C** (6.01)	**B** (0.822) > **A** (0.807) > **C** (0.712)
CIR6-PKG	**A** (11.12) > **B** (10.65) > **C** (9.31)	**A** (0.870) > **B** (0.847) > **C** (0.784)

Performance of conversational agents using real vs. simulated users in (Zhang and Balog, 2020)

Testing Simulators

- Testing the **reliability** of simulators: Does the user simulator behave as expected for it intended use? (testing whether it matches human expectation; not the same as validation)
- **Tester-based framework** (Labhishetty and Zhai, 2021, 2022)
  - Tester: System A is expected to perform better than system B under a certain condition (e.g., for a certain kind of queries)
  - Simulator passes the test if the expected behavior is observed
  - Reliability of a user simulator and reliability of a Tester can be estimated jointly

Application of Tester-based Framework in TOD (Sun et al., 2023)

- Context tester: Having more context for training should improve recommendation performance
- Recommender tester: Different retrieval capabilities (removing keywords in the query)
- Domain tester: Controlling the amount of within-domain training data

Resources for Validation: Benchmarks and Protocols

Resources for Validation

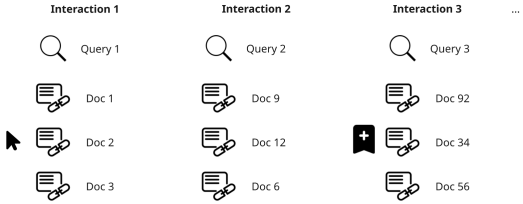
- Search interactions $\Leftarrow$
- Recommendation interactions
- Conversational interactions

Validation of Search Interaction Simulators (1/4)

- **UQV100** (Bailey et al., 2016)
  - *Data*: Query variants for 100 topics from the TREC 2013 and 2014 Web Tracks, collected from crowdworkers and relevance judgments.
  - *Potential for validation*: Comparison of query characteristics and system performances between simulated and human queries.
  - *Example of validation*: Query Variant Simulator (Breuer et al., 2022).
- Examples of query variants for the backstory: You have heard quite a lot about cheap computing as being the way of the future, including one recent model called a Raspberry Pi. You start thinking about buying one, and wonder how much they cost.
  - *QV1*: "amazon raspberry pi"
  - *QV2*: "best deal raspberry pi computer"
  - *QV3*: "buy Raspberry Pi"

Validation of Search Interaction Simulators (2/4)

- **TREC Session Track (2011-2014)** (Carterette et al., 2016)
  - *Data*: Search sessions including queries, documents retrieved and saved, clicks, and relevance judgments.
  - Released four different datasets with variants of properties, such as different user population and different topics.
  - *Potential for validation*: Comparison of query characteristics, session characteristics, and system performances between simulated and human sessions.
  - *Examples of validation*: Search session simulator (Hagen et al., 2016) and query simulation in search session (Günther and Hagen, 2021).



Validation of Search Interaction Simulators (3/4)

- **Archive Query Log** (Reimer et al., 2023)
  - *Data*: Query logs spanning 25 years across 550 search providers, including 356 million queries.
  - *Potential for validation*: Comparison of query characteristics between simulated and real users.

Service	Query	Snippet Title	Snippet Rank
Google	g1 sim free	HP EliteBook 840 G1 plus FREE Mobile Broadband 3-in-1 SIM	6
StackOverflow	node.js+slack	How to pass SlackEvents data to client side?	21
Google	"Georgia Lancaster"	Farms Compete With Suburbs.	2

Table: Partial examples of queries from the Archive Query Log.

⇒ The majority of large-scale query logs are proprietary. Exceptions include the **AOL Query Log** (Pass et al., 2006) and **MSN dataset** (Zhang and Moffat, 2006) but they are not available anymore.

Validation of Search Interaction Simulators (4/4)

- **ORCAS** (Craswell et al., 2020)
  - *Data*: Click dataset with ∼10 million queries connected to TREC Deep Learning documents, including 18 million clicked query-document pairs.
  - *Potential for validation*: Comparison of query characteristics and clicking behavior between simulated and real users.

Query ID	Query	Document ID	URL
10103699	why is the sky blue	D1968574	`http://www.sciencemadesimple.com/sky_blue.html`
9265503	github	D1265400	`https://desktop.github.com/`
9265503	github	D3438005	`https://github.com/`
5584651	"climate change"	D55701	`https://climate.nasa.gov/`
5584651	"climate change"	D568385	`https://climate.nasa.gov/evidence/`

Table: Examples of clicked query-document pairs from ORCAS.

Resources for Validation

- Search interactions
- Recommendation interactions $\Leftarrow$
- Conversational interactions

Validation of Recommendation Interaction Simulators (1/3)

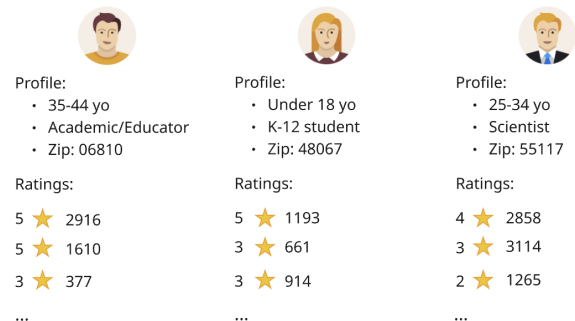
- **MIND** (Wu et al., 2020)
  - *Data*: User click logs for news recommendation from Microsoft News over 6 weeks.
  - *Potential for validation*: Comparison of clicking behavior between simulated and real users.
  - *Example of validation*: MINDSim (Luo et al., 2022).

Column	Content
Impression ID	91
User ID	U397059
Time	11/15/2019 10:22:32 AM
History	N106403 N71977 N97080 N102132 N97212 N121652
Impressions	N129416-0 N26703-1 N120089-1 N53018-0 N89764-0 N91737-0 N29160-0

Example of an entry in MIND dataset

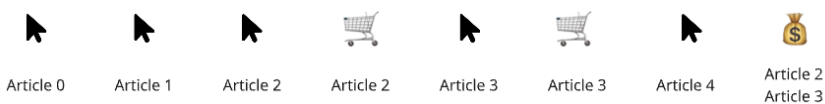
Validation of Recommendation Interaction Simulators (2/3)

- **MovieLens 1M Dataset** (Harper and Konstan, 2015)
  - *Data*: ∼1 million ratings from ∼6,000 users on ∼3,900 movies. It includes ratings, movie metadata, and user demographics.
  - *Potential for validation*: Comparison between simulated and real users' ratings overall and across different user populations.
  - *Example of validation*: User behavior simulation with LLMs (Wang et al., 2025).



Validation of Recommendation Interaction Simulators (3/3)

- **OTTO Recommender Systems Dataset** (Normann et al., 2023)
  - *Data*: ∼12 million session logs from an e-commerce platform, including clicks, add-to-cart, and order actions.
  - *Potential for validation*: Comparison of session characteristics between simulated and real users.



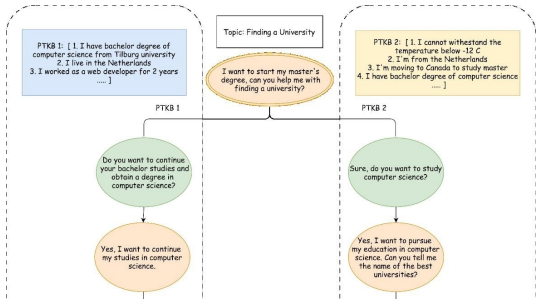
Example of a session from the OTTO Recommender Systems Dataset.

Resources for Validation

- Search interactions
- Recommendation interactions
- Conversational interactions $\Leftarrow$

Validation of Conversational Simulators (1/4)

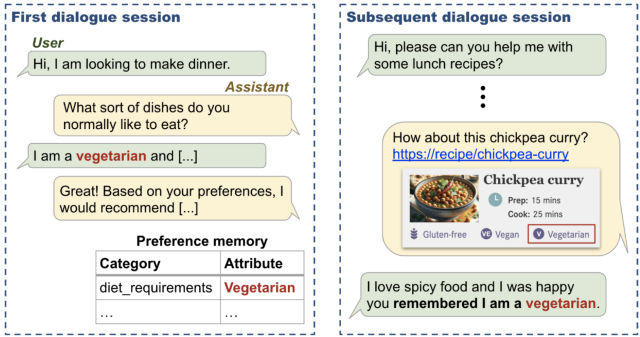
- **TREC Interactive Knowledge Assistance Track (2023-present)** (Aliannejadi et al., 2024a,b)
  - Successor of the TREC Conversational Search Track (Dalton et al., 2020) with emphasis on personalization.
  - *Data*: Personalized conversations for a given set of topics with response assessments based on user profiles and preferences.
  - *Potential for validation*: Comparison between simulated and human conversations characteristics. Comparison of system performances with simulated and human conversations.



Partial example of personalized conversations on a given topic.

Validation of Conversational Simulators (2/4)

- **LAPS** (Joko et al., 2024)
  - *Data*: ∼1,400 multi-session and multi-turn conversations annotated with user preferences. Conversations are created by crowdworkers assisted by a large language model.
  - *Potential for validation*: Comparison between simulated and human conversation characteristics.



Snippet from a multi-session dialogue in the recipe domain.

Validation of Conversational Simulators (3/4)

- **BIDD-1k** (Trippas et al., 2024)
  - *Data*: 1,000 anonymized prompts (some in the same session) from Google Bard[1] interactions with crowdworkers.
  - *Potential for validation*: Comparison between human and simulated prompt characteristics.

Session ID	Prompt
4350	use less verbosity
4350	2) Now, please imagine what it would be like to have network connections with people at the very bottom of the ladder. Compared with having those lower-class connections, what is something unique about your own network, i.e., having network connections with individuals who are relatively HIGH in social class?
4500	what are cucumbers?
4388	Can you tell me some physical differences between a songbird and a woodpecker?

Table: Examples of prompts from BIDD-1k.

[1] Predecessor of Google Gemini

Validation of Conversational Simulators (4/4)

- **Chatbot Arena** (Chiang et al., 2024)
  - *Data*: Prompts and human pairwise preferences (votes) for different LLMs from a crowdsourced comparison platform. Multiple datasets are available at `https://huggingface.co/lmarena-ai`. A leaderboard is maintained for LLMs based on the votes received.
  - *Potential for validation*: Comparison between human and simulated prompts characteristics. If human votes are available for the simulated prompts, a comparison of performances can be made.
  - *Example of validation*: **WizardArena** (Luo et al., 2024a) (validation of judge simulators)

$\Rightarrow$ Other benchmarks comprising in-the-wild prompts/conversations and human votes exist, such as **WildBench** (Lin et al., 2024) and **CRSArena-Dial** (Bernard et al., 2025a).

Summary of Resources for Validation

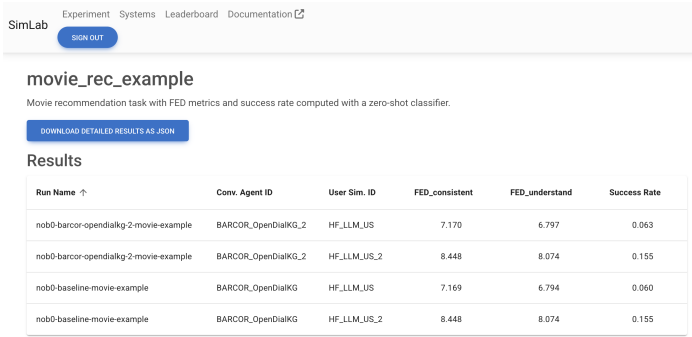
Resource	Interactions	Validation approach	
		Characteristics	Performance
UQV100 (Bailey et al., 2016)	Search	✓	✓
TREC Session Track (Carterette et al., 2016)	Search	✓	✓
ORCAS (Craswell et al., 2020)	Search	✓	×
Archive Query Log (Reimer et al., 2023)	Search	✓	×
MovieLens-1M (Harper and Konstan, 2015)	Recommendation	✓	×
MIND (Wu et al., 2020)	Recommendation	✓	×
OTTO (Normann et al., 2023)	Recommendation	✓	×
TREC iKAT (Aliannejadi et al., 2024a,b)	Conversational	✓	✓
LAPS (Joko et al., 2024)	Conversational	✓	×
BIDD-1k (Trippas et al., 2024)	Conversational	✓	×
Chatbot Arena (Chiang et al., 2024)	Conversational	✓	✓*

* Possible only if votes are available for simulated prompts.

Table: Overview of presented resources for validation.

Benchmarking Conversational Simulators

- **SimLab** (Bernard et al., 2025b)
  - Centralized platform for benchmarking conversational simulators. It offers the possibility of testing and comparing different simulators on a set of tasks.
  - Automatic validation of simulators is envisioned, where tasks associated with conversational datasets involving human users and validation metrics are used to validate simulators.



Conclusion and Future Challenges

Summary

- User simulation is an essential step toward AGI with three major applications
  - Automated large-scale reproducible evaluation of interactive AI systems
  - Complete formal user modeling, required for effective personalization of any AI system
  - Large amounts of useful synthetic interaction data for training AI algorithms
- Most work on user simulation has been done for search engines, less so for recommender systems, but increasingly more common for conversational assistants
- LLMS accelerated research on user simulation and broadened the scope of applications (e.g., social system simulation)
- Many opportunities and challenges in the future
  - New opportunities for an exciting interdisciplinary research community on user simulation
  - Many difficult challenges (e.g., validation of user simulation, interpretable simulation approaches, and industry-academia collaboration)

Future Direction: Embracing Simulation-based Evaluation

- Simulation-based evaluation has not been widely adopted in the IR and RecSys communities
- Could be due to several factors:
  - Complexity of creating realistic simulations
  - Lack of consensus on simulation-based evaluation methodology
  - Open questions regarding the validity of simulations
  - Resources required to develop and run simulations
- Next steps
  - Leverage existing test collections and turn them into user simulators
  - Organize evaluation activities regularly (e.g., at TREC) for evaluating both user simulators and using simulation to evaluate IR systems

Future Direction: Fostering Industry-academia Collaboration

- User simulation is a technology that can help to foster collaboration between academia and industry
- Academia: Access to realistic datasets for evaluation is always a major challenge
- Industry: It is difficult to release datasets (e.g., due to privacy concerns)
- Releasing user simulators trained/estimated using commercial search log data should have much less privacy concerns than releasing any log data (directly)
- Self-sustainable innovation ecosystem
  - Academic researchers develop models/algorithms for user simulation and make them available as open source
  - Commercial service providers train and validate user simulators against their logs, and publish the trained simulators (without having to share any actual user data)
  - Academic researchers can develop and validate new search and recommendation algorithms against published simulators
  - Service providers get access to the most advanced algorithms developed by (external)

Key Technical Challenge: Realisticity

- Informally, it is easy to understand what it means to simulate a user computationally
- Mathematically defining the problem remains a major open challenge (e.g., behavior similarity vs. model similarity)
- From psychology perspective, how to implement Theory of Mind (Premack and Woodruff, 1978) is a major challenge

"It remains an open question as to how realistic (i.e. human-like) simulators can be, or indeed should be. It is important to note that simulators do not need to be perfect mirrors of human behaviour, but instead simply need to be "good enough." By this, we mean that output from simulations should correlate well with human assessments on a given task with respect to some evaluation metric. The main requirement is reproducibility." – *Sim4IR workshop (Balog et al., 2022)*

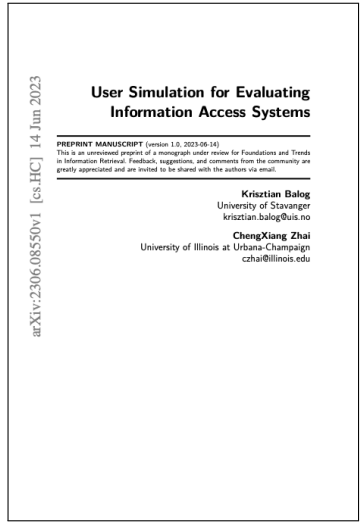
Opportunities for Interdisciplinary Research

- User Simulation overlaps with multiple related areas
  - Information Retrieval: Conversational Search
  - Recommender Systems: Conversational recommendation
  - Agent Systems: Conversational task assistants
  - Machine Learning: Reinforcement Learning
  - HCI and Psychology: Simulators as Testable Hypotheses about Users
  - Natural Language Processing: User Simulation and Large Language Models
- How can we establish and grow an interdisciplinary research community around user simulation?

Discussion

Resources

- FnTIR book:
  `https://arxiv.org/abs/2306.08550`
- Website: `https://usersim.ai`
  - Tutorials and slides
  - Annotated bibliography
  - List of toolkits
- Mailing list: usersim@googlegroups.com
- Slack channel: ACM SIGIR / #usersim



References

Jafar Afzali, Aleksander Mark Drzewiecki, Krisztian Balog, and Shuo Zhang. 2023. UserSimCRS: A User Simulation Toolkit for Evaluating Conversational Recommender Systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23)*. 1160–1163. https://doi.org/10.1145/3539597.3573029

Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer, and Paul Thomas. 2023. Can Generative LLMs Create Query Variants for Test Collections? An Exploratory Study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. 1869–1873.

Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024a. TREC iKAT 2023: A Test Collection for Evaluating Conversational and Interactive Knowledge Assistants. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. 819–829. https://doi.org/10.1145/3626772.3657860

Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffery Dalton, and Leif Azzopardi. 2024b. TREC iKAT 2023: The Interactive Knowledge Assistance Track Overview.

Mohammad Aliannejadi, Leif Azzopardi, Hamed Zamani, Evangelos Kanoulas, Paul Thomas, and Nick Craswell. 2021. Analysing Mixed Initiatives and Search Strategies during Conversational Search. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21)*. 16–26. https://doi.org/10.1145/3459637.3482231

References

Leif Azzopardi, Timo Breuer, Björn Engelmann, Christin Kreutz, Sean MacAvaney, David Maxwell, Andrew Parry, Adam Roegiest, Xi Wang, and Saber Zerhoudi. 2024. SimIIR 3: A Framework for the Simulation of Interactive and Conversational Information Retrieval. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 197–202.

Leif Azzopardi and Maarten de Rijke. 2006. Automatic Construction of Known-Item Finding Test Beds. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. 603–604. https://doi.org/10.1145/1148170.1148276

Leif Azzopardi, Maarten de Rijke, and Krisztian Balog. 2007. Building Simulated Queries for Known-Item Topics: An Analysis Using Six European Languages. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*. 455–462. https://doi.org/10.1145/1277741.1277820

Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffery Dalton. 2018. Conceptualizing Agent-Human Interactions During the Conversational Search Process. In *Proceedings of the 2nd International Workshop on Conversational Approaches to Information Retrieval (CAIR '18)*.

Leif Azzopardi, Paul Thomas, and Alistair Moffat. 2019. Cwl_eval: An Evaluation Tool for Information Retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*. 1321–1324. https://doi.org/10.1145/3331184.3331398

References

Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. 725–728. https://doi.org/10.1145/2911451.2914671

Krisztian Balog, David Maxwell, Paul Thomas, and Shuo Zhang. 2022. Report on the 1st Simulation for Information Retrieval Workshop (Sim4IR 2021) at SIGIR 2021. *SIGIR Forum* 55, 2, Article 10 (mar 2022). https://doi.org/10.1145/3527546.3527559

Krisztian Balog and ChengXiang Zhai. 2024. User Simulation for Evaluating Information Access Systems. *Foundations and Trends® in Information Retrieval* 18, 1-2 (2024), 1–261. https://doi.org/10.1561/1500000098

Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. 2012. Time Drives Interaction: Simulating Sessions in Diverse Searching Environments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. 105–114. https://doi.org/10.1145/2348283.2348301

Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. 2013. Modeling Behavioral Factors in Interactive Information Retrieval. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM '13)*. 2297–2302. https://doi.org/10.1145/2505515.2505660

Marcia J. Bates. 1989. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review* 13 (1989), 407–424. Issue 5. https://doi.org/10.1108/eb024320

Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. 1982. ASK for Information Retrieval: Part I. Background and Theory. *Journal of Documentation* (1982).

References

Nolwenn Bernard and Krisztian Balog. 2024. Towards a Formal Characterization of User Simulation Objectives in Conversational Information Access. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '24)*. 185–193.

Nolwenn Bernard, Hideaki Joko, Faegheh Hasibi, and Krisztian Balog. 2025a. CRS Arena: Crowdsourced Benchmarking of Conversational Recommender Systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25)*. 1028–1031. https://doi.org/10.1145/3701551.3704120

Nolwenn Bernard, Sharath Chandra Etagi Suresh, Krisztian Balog, and ChengXiang Zhai. 2025b. SimLab: A Platform for Simulation-based Evaluation of Conversational Information Access Systems. arXiv:2507.04888 [cs.IR]

Christine L. Borgman. 1996. Why are Online Catalogs Still Hard to Use? *Journal of the American Society for Information Science* 47, 7 (1996), 493–503. https://doi.org/10.1002/(SICI)1097-4571(199607)47:7<493::AID-ASI3>3.0.CO;2-P

Timo Breuer, Norbert Fuhr, and Philipp Schaer. 2022. Validating Simulations of User Query Variants. In *Proceedings of the 44th European Conference on IR Research (ECIR '22)*. 80–94. https://doi.org/10.1007/978-3-030-99736-6_6

Ben Carterette, Ashraf Bah, and Mustafa Zengin. 2015. Dynamic Test Collections for Retrieval Evaluation. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15)*. 91–100. https://doi.org/10.1145/2808194.2809470

References

Ben Carterette, Paul Clough, Mark Hall, Evangelos Kanoulas, and Mark Sanderson. 2016. Evaluating Retrieval over Sessions: The TREC Session Track 2011-2014. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. 685–688. https://doi.org/10.1145/2911451.2914675

Qinyuan Cheng, Linyang Li, Guofeng Quan, Feng Gao, Xiaofeng Mou, and Xipeng Qiu. 2022. Is MultiWOZ a Solved Task? An Interactive TOD Evaluation Framework with User Simulator. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 1248–1259. https://doi.org/10.18653/v1/2022.findings-emnlp.90

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*. 8359–8388. https://doi.org/10.5555/3692070.3692401

C. Cleverdon and M. Kean. 1968. Factors Determining the Performance of Indexing Systems. Aslib Cranfield Research Project, Cranfield, England.

Michael D. Cooper. 1973. A Simulation Model of an Information Retrieval System. *Information Storage and Retrieval* 9, 1 (1973), 13–32. https://doi.org/10.1016/0020-0271(73)90004-1

Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 18 Million Clicked Query-Document Pairs for Analyzing Search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. 2983–2989. https://doi.org/10.1145/3340531.3412779

References

Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-Bias Models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08)*. 87–94. https://doi.org/10.1145/1341531.1341545

Paul Crook and Alex Marin. 2017. Sequence to Sequence Modeling for User Simulation in Dialog Systems. In *Proceedings of Interspeech 2017*. 1706–1710. https://doi.org/10.21437/Interspeech.2017-161

Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. CAsT-19: A Dataset for Conversational Information Seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 1985–1988. https://doi.org/10.1145/3397271.3401206

Sam Davidson, Salvatore Romeo, Raphael Shu, James Gung, Arshit Gupta, Saab Mansour, and Yi Zhang. 2023. User Simulation with Large Language Models for Evaluating Task-Oriented Dialogue. arXiv:2309.13233 [cs.CL]

Georges E. Dupret and Benjamin Piwowarski. 2008. A User Browsing Model to Predict Search Engine Click Data from Past Observations. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. 331–338. https://doi.org/10.1145/1390334.1390392

Danial Ebrat, Eli Paradalis, and Luis Rueda. 2024. Lusifer: LLM-based user simulated feedback environment for online recommender systems. *arXiv preprint arXiv:2405.13362* (2024).

References

W. Eckert, E. Levin, and R. Pieraccini. 1997. User Modeling for Spoken Dialogue System Evaluation. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. 80–87. https://doi.org/10.1109/ASRU.1997.658991

Layla El Asri, Jing He, and Kaheer Suleman. 2016. A Sequence-to-Sequence Model for User Simulation in Spoken Dialogue Systems. In *Proceedings of Interspeech 2016*. 1151–1155. https://doi.org/10.21437/Interspeech.2016-1175

Björn Engelmann, Timo Breuer, Jana Isabelle Friese, Philipp Schaer, and Norbert Fuhr. 2024. Context-Driven Interactive Query Simulations Based on Generative Large Language Models. In *Proceedings of the 46th European Conference on Information Retrieval (ECIR '24)*. 173–188.

Björn Engelmann, Timo Breuer, and Philipp Schaer. 2023. Simulating Users in Interactive Web Table Retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. 3875–3879.

Sebastian Günther and Matthias Hagen. 2021. Assessing Query Suggestions for Search Session Simulation. In *Joint Proceedings of the Causality in Search and Recommendation (CSR) and Simulation of Information Retrieval Evaluation (Sim4IR) Workshops 2021*. 38–45.

Izzeddin Gür, Dilek Hakkani-Tür, Gokhan Tür, and Pararth Shah. 2018. User Modeling for Task Oriented Dialogues. In *2018 IEEE Spoken Language Technology Workshop (SLT '18)*. 900–906. https://doi.org/10.1109/SLT.2018.8639652

References

Matthias Hagen, Maximilian Michel, and Benno Stein. 2016. Simulating Ideal and Average Users. In *Proceedings of the 12th Asia Information Retrieval Societies Conference (AIRS '16)*. 138–154. https://doi.org/10.1007/978-3-319-48051-0_11

Donna Harman. 1992. Relevance Feedback Revisited. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '92)*. 1–10. https://doi.org/10.1145/133160.133167

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems* 5, 4 (2015), 1–19. https://doi.org/10.1145/2827872

Yifan He, To Eun Kim, Fernando Diaz, Jaime Arguello, and Bhaskar Mitra. 2025. Tip of the Tongue Query Elicitation for Simulated Evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*.

Zhiyuan Hu, Yue Feng, Anh Tuan Luu, Bryan Hooi, and Aldo Lipani. 2023. Unlocking the Potential of User Feedback: Leveraging Large Language Model as User Simulators to Enhance Dialogue System. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. 3953–3957. https://doi.org/10.1145/3583780.3615220

Eugene Ie, Chih wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. 2019. RecSim: A Configurable Simulation Platform for Recommender Systems. arXiv:1909.04847 [cs.LG]

Peter Ingwersen and Kalervo Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*. The Information Retrieval Series, Vol. 18. Springer. https://doi.org/10.1007/1-4020-3851-8

References

Anthony Jameson, Bettina Berendt, Silvia Gabrielli, Federica Cena, Cristina Gena, Fabiana Vernero, and Katharina Reinecke. 2014. Choice Architecture for Human-Computer Interaction. *Foundations and Trends in Human-Computer Interaction* 7, 1–2 (oct 2014), 1–235. https://doi.org/10.1561/1100000028

Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P. de Vries, Jeff Dalton, and Faegheh Hasibi. 2024. Doing Personal LAPS: LLM-Augmented Dialogue Construction for Personalized Multi-Session Conversational Search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. 796–806. https://doi.org/10.1145/3626772.3657815

Heikki Keskustalo, Kalervo Järvelin, and Ari Pirkola. 2008. Evaluating the Effectiveness of Relevance Feedback Based on a User Simulation Model: Effects of a User Scenario on Cumulated Gain Value. *Information Retrieval* 11, 3 (June 2008), 209–228. https://doi.org/10.1007/s10791-007-9043-7

Johannes Kiesel, Marcel Gohsen, Nailia Mirzakhmedova, Matthias Hagen, and Benno Stein. 2024. Who Will Evaluate the Evaluators? Exploring the Gen-IR User Simulation Space. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF '24)*. 166–171.

To Eun Kim and Aldo Lipani. 2022. A Multi-Task Based Neural Model to Simulate Users in Goal Oriented Dialogue Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. 2115–2119. https://doi.org/10.1145/3477495.3531814

Florian Kreyssig, Iñigo Casanueva, Paweł Budzianowski, and Milica Gašić. 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 (SIGDIAL '18)*. 60–69. https://doi.org/10.18653/v1/W18-5007

References

Carol C. Kuhlthau. 1991. Inside the Search Process: Information Seeking from the User's Perspective. *Journal of the American Society for Information Science* 42, 5 (1991), 361–371. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-%23

Sahiti Labhishetty and Chengxiang Zhai. 2021. An Exploration of Tester-Based Evaluation of User Simulators for Comparing Interactive Retrieval Systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. 1598–1602. https://doi.org/10.1145/3404835.3463091

Sahiti Labhishetty and ChengXiang Zhai. 2022. RATE: A Reliability-Aware Tester-Based Evaluation Framework of User Simulators. In *Proceedings of the 44th European Conference on IR Research (ECIR '22)*. 336–350. https://doi.org/10.1007/978-3-030-99736-6_23

Anton Leuski. 2000. Relevance and Reinforcement in Interactive Browsing. In *Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM '00)*. 119–126. https://doi.org/10.1145/354756.354809

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. A Stochastic Model of Human-machine Interaction for Learning Dialog Strategies. *IEEE Trans. Speech Audio Process.* 8, 1 (2000), 11–23. https://doi.org/10.1109/89.817450

Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2022. Controllable Dialogue Simulation with In-context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 4330–4347. https://doi.org/10.18653/v1/2022.findings-emnlp.318

References

Tingting Liang, Chenxin Jin, Lingzhi Wang, Wenqi Fan, Congying Xia, Kai Chen, and Yuyu Yin. 2024. LLM-REDIAL: A Large-Scale Dataset for Conversational Recommender Systems Created from User Behaviors with LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*. 8926–8939.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild. arXiv:2406.04770 [cs.CL]

Hsien-chin Lin, Christian Geishauser, Shutong Feng, Nurul Lubis, Carel van Niekerk, Michael Heck, and Milica Gasic. 2022. GenTUS: Simulating User Behaviour and Language in Task-oriented Dialogues with Generative Transformers. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL '22)*. 270–282.

Hsien-chin Lin, Nurul Lubis, Songbo Hu, Carel van Niekerk, Christian Geishauser, Michael Heck, Shutong Feng, and Milica Gasic. 2021. Domain-independent User Simulation with Transformers for Task-oriented Dialogue Systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL '21)*. 445–456.

Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2021. How Am I Doing?: Evaluating Conversational Search Systems Offline. *ACM Trans. Inf. Syst.* 39, 4, Article 51 (Aug. 2021), 22 pages. `https://doi.org/10.1145/3451160`

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Qingwei Lin, Jianguang Lou, Shifeng Chen, Yansong Tang, and Weizhu Chen. 2024a. WizardArena: Post-training Large Language Models via Simulated Offline Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS '24, Vol. 37)*. 111544–111570.

References

Xufang Luo, Zheng Liu, Shitao Xiao, Xing Xie, and Dongsheng Li. 2022. MINDSim: User Simulator for News Recommenders. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. 2067–2077. https://doi.org/10.1145/3485447.3512080

Xiang Luo, Zhiwen Tang, Jin Wang, and Xuejie Zhang. 2024b. DuetSim: Building User Simulator with Dual Large Language Models for Task-Oriented Dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 5414–5424.

Shengnan Lyu, Arpit Rana, Scott Sanner, and Mohamed Reda Bouadjenek. 2021. A Workflow Analysis of Context-Driven Conversational Recommendation. In *Proceedings of the Web Conference 2021 (WWW '21)*. 866–877. https://doi.org/10.1145/3442381.3450123

Gary Marchionini. 1995. *Information Seeking in Electronic Environments*. Cambridge University Press. https://doi.org/10.1017/CBO9780511626388

David Maxwell and Leif Azzopardi. 2016a. Agents, Simulated Users and Humans: An Analysis of Performance and Behaviour. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM '16)*. 731–740. https://doi.org/10.1145/2983323.2983805

David Maxwell and Leif Azzopardi. 2016b. Simulating Interactive Information Retrieval: SimIIR: A Framework for the Simulation of Interaction. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. 1141–1144. https://doi.org/10.1145/2911451.2911469

References

David Maxwell and Leif Azzopardi. 2018. Information Scent, Searching and Stopping: Modelling SERP Level Stopping Behaviour. In *Proceedings of the 40th European Conference on IR Research (ECIR '18)*. 210–222. https://doi.org/10.1007/978-3-319-76941-7_16

David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. 2015. Searching and Stopping: An Analysis of Stopping Rules and Strategies. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. 313–322. https://doi.org/10.1145/2806416.2806476

Martin Mladenov, Chih-Wei Hsu, Vihan Jain, Eugene Ie, Christopher Colby, Nicolas Mayoraz, Hubert Pham, Dustin Tran, Ivan Vendrov, and Craig Boutilier. 2021. RecSim NG: Toward Principled Uncertainty Modeling for Recommender Ecosystems. arXiv:2103.08057 [cs.LG]

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. arXiv:1904.08375 [cs.IR]

Philipp Normann, Sophie Baumeister, and Timo Wilm. 2023. OTTO Recommender Systems Dataset. https://doi.org/10.34740/kaggle/dsv/4991874

Vicki L. O'Day and Robin Jeffries. 1993. Orienteering in an Information Landscape: How Information Seekers Get from Here to There. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (CHI '93)*. 438–445. https://doi.org/10.1145/169059.169365

References

Paul Owoicho, Ivan Sekulic, Mohammad Alianejadi, Jeffery Dalton, and Fabio Crestani. 2023. Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. https://doi.org/10.1145/3539618.3591683

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems (InfoScale '06)*. https://doi.org/10.1145/1146847.1146848

Olivier Pietquin. 2004. *A Framework for Unsupervised Learning of Dialogue Strategies*. Ph. D. Dissertation. Faculté Polytechnique de Mons, Belgium.

Peter Pirolli and Stuart Card. 1999. Information Foraging. *Psychological Review* 106 (1999), 643–675. Issue 4. https://doi.org/10.1037/0033-295X.106.4.643

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.

Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and Characterizing User Intent in Information-Seeking Conversations. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '18)*. 989–992. https://doi.org/10.1145/3209978.3210124

Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. 117–126. https://doi.org/10.1145/3020165.3020183

References

Jan Heinrich Reimer, Sebastian Schmidt, Maik Fröbe, Lukas Gienapp, Harrisen Scells, Benno Stein, Matthias Hagen, and Martin Potthast. 2023. The Archive Query Log: Mining Millions of Search Result Pages of Hundreds of Search Engines from 25 Years of Web Archives. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. 2848–2860. https://doi.org/10.1145/3539618.3591890

Alexandre Salle, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. 2021. Studying the Effectiveness of Conversational Search Refinement Through User Simulation. In *Proceedings of the 43rd European Conference on IR Research (ECIR '21)*. 587–602. https://doi.org/10.1007/978-3-030-72113-8_39

G. Salton. 1970. Evaluation problems in Interactive Information Retrieval. *Information Storage and Retrieval* 6, 1 (1970), 29–44. https://doi.org/10.1016/0020-0271(70)90011-2

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers (NAACL-HLT '07)*. 149–152.

Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. 2006. A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies. *The Knowledge Engineering Review* 21, 2 (June 2006), 97–126.

Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating Mixed-Initiative Conversational Search Systems via User Simulation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22)*. 888–896. https://doi.org/10.1145/3488560.3498440

References

Catherine L. Smith and Paul B. Kantor. 2008. User Adaptation: Good Results from Poor Systems. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. 147–154. https://doi.org/10.1145/1390334.1390362

Mark D. Smucker and Charles L.A. Clarke. 2012. Time-Based Calibration of Effectiveness Measures. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. 95–104. https://doi.org/10.1145/2348283.2348300

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2023. Data Augmentation for Conversational AI. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. 5220–5223.

Karen Spärck Jones. 1979. Search Term Relevance Weighting given Little Relevance Information. *Journal of Documentation* 35, 1 (1979), 30–48. https://doi.org/10.1108/eb026672

Louise T. Su. 1992. Evaluation measures for interactive information retrieval. *Information Processing & Management* 28, 4 (1992), 503–516. https://doi.org/10.1016/0306-4573(92)90007-M

Weiwei Sun, Shuyu Guo, Shuo Zhang, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2023. Metaphorical User Simulators for Evaluating Task-oriented Dialogue Systems. *ACM Transactions on Information Systems* 42, 1, Article 17 (aug 2023). https://doi.org/10.1145/3596510

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating User Satisfaction for the Evaluation of Task-oriented Dialogue Systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. 2499–2506. https://doi.org/10.1145/3404835.3463241

References

Silvia Terragni, Modestas Filipavicius, Nghia Khau, Bruna Guedes, André Manso, and Roland Mathis. 2023. In-Context Learning User Simulators for Task-Oriented Dialog Systems. arXiv:2306.00774 [cs.CL]

Johanne R. Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. 2024. What do Users Really Ask Large Language Models? An Initial Log Analysis of Google Bard Interactions in the Wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. 2703–2707. https://doi.org/10.1145/3626772.3657914

Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval (CHIIR '18)*. 32–41. https://doi.org/10.1145/3176349.3176387

Bo-Hsiang Tseng, Yinpei Dai, Florian Kreyssig, and Bill Byrne. 2021. Transferable Dialogue Systems and User Simulators. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL '21)*. 152–166. https://doi.org/10.18653/v1/2021.acl-long.13

Andrew Turpin and Falk Scholer. 2006. User Performance versus Precision Measures for Simple Search Tasks. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. 11–18. https://doi.org/10.1145/1148170.1148176

References

Andrew Turpin, Falk Scholer, Kalvero Jarvelin, Mingfang Wu, and J. Shane Culpepper. 2009. Including Summaries in System Evaluation. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. 508–515. https://doi.org/10.1145/1571941.1572029

Stefan Ultes, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Iñigo Casanueva, Paweł Budzianowski, Nikola Mrkšić, Tsung-Hsien Wen, Milica Gašić, and Steve Young. 2017. PyDial: A Multi-domain Statistical Dialogue System Toolkit. In *Proceedings of ACL 2017, System Demonstrations (ACL '17)*. 73–78.

Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. 2019. QRFA: A Data-Driven Model of Information-Seeking Dialogues. In *Proceedings of the 41st European Conference on IR Research (ECIR '19)*. 541–557. https://doi.org/10.1007/978-3-030-15712-8_35

Anna Volodkevich, Veronika Ivanova, Alexey Vasilev, Dmitry Bugaychenko, and Maxim Savchenko. 2025. Sim4Rec: Flexible and Extensible Simulator for Recommender Systems for Large-Scale Data. In *European Conference on Information Retrieval*. Springer, 425–430.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2025. User Behavior Simulation with Large Language Model-based Agents. *ACM Transactions on Information Systems* 43, 2 (2025), 1–37. https://doi.org/10.1145/3708985

References

Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP '23)*. 10052–10065. https://doi.org/10.18653/v1/2023.emnlp-main.621

Zhenduo Wang, Zhichao Xu, Vivek Srikumar, and Qingyao Ai. 2024. An In-depth Investigation of User Response Simulation for Conversational Search. In *Proceedings of the ACM on Web Conference 2024 (WWW '24)*. 1407–1418. https://doi.org/10.1145/3589334.3645447

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20)*. 3597–3606. https://doi.org/10.18653/v1/2020.acl-main.331

Ke Yang and ChengXiang Zhai. 2025. Ten Principles of AI Agent Economics. *arXiv preprint arXiv:2505.20273* (2025).

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. 2024. Evaluating Large Language Models as Generative User Simulators for Conversational Recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 1490–1504.

Steve Young. 1999. Probabilistic Methods in Spoken Dialogue Systems. *Philosophical Transactions of the Royal Society (Series A)* 358 (1999), 1389–1402. Issue 1769.

References

Steve Young, Milica Gašić, Simon Keizer, FranÁois Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The Hidden Information State Model: A Practical Framework for POMDP-based Spoken Dialogue Management. *Computer Speech & Language* 24, 2 (2010), 150–174. https://doi.org/10.1016/j.csl.2009.04.001

Saber Zerhoudi, Sebastian Günther, Kim Plassmeier, Timo Borst, Christin Seifert, Matthias Hagen, and Michael Granitzer. 2022. The SimIIR 2.0 Framework: User Types, Markov Model-Based Interaction Simulation, and Advanced Query Generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*. 4661–4666. https://doi.org/10.1145/3511808.3557711

Shuo Zhang and Krisztian Balog. 2020. Evaluating Conversational Recommender Systems via User Simulation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '20)*. 1512–1520. https://doi.org/10.1145/3394486.3403202

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. 177–186. https://doi.org/10.1145/3269206.3271776

Yinan Zhang, Xueqing Liu, and ChengXiang Zhai. 2017. Information Retrieval Evaluation as Search Simulation: A General Formal Framework for IR Evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '17)*. 193–200. https://doi.org/10.1145/3121050.3121070

Yuye Zhang and Alistair Moffat. 2006. Some Observations on User Search Behaviour. *Australian Journal of Intelligent Information Processing Systems* 9, 2 (2006), 1–8.

References

Yinan Zhang and Chengxiang Zhai. 2015. Information retrieval as card playing: A formal model for optimizing interactive retrieval interface. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 685–694.

Kesen Zhao, Shuchang Liu, Qingpeng Cai, Xiangyu Zhao, Ziru Liu, Dong Zheng, Peng Jiang, and Kun Gai. 2023. KuaiSim: A comprehensive simulator for recommender systems. *Advances in Neural Information Processing Systems* 36 (2023), 44880–44897.

Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. How Reliable is Your Simulator? Analysis on the Limitations of Current LLM-based User Simulators for Conversational Recommendation. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. 1726–1732.

Qi Zhu, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, et al. 2022. Convlab-3: A flexible dialogue system toolkit based on a unified data format. *arXiv preprint arXiv:2211.17148* (2022).