SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)
ISSN (Print) 1813-1743, ISSN (online) 2791-0547
Vol. 57. 02 (2025).
Doi: https://doi.org/10.26692/surjss.v57i02.7672

Evaluating the Efficacy of Simulated User Models in Interactive Information Retrieval: A User-Based Approach

A. R. NANGRAJ[1*], M. S. CHANDIO[2*], YARSIR ARFAT MALKANI[3*], QURATULAIN NIZAMANI[4*]

Institute of Mathematics & Computer Science, University of Sindh Jamshoro.

***Cite this*:**

A. R. Nangraj et all, Evaluating the Efficacy of Simulated User Models in Interactive Information Retrieval: A User-Based Approach Sindh Uni. Res. J. (SS) 57: 02, 2025

Corresponding author
rehman.nangraj@usindh.edu.pk

ABSTRACT

Simulated user models are increasingly employed to evaluate IIR systems due to their scalability and consistency. However, the extent to which these models realistically replicate human behavior across diverse search tasks remains underexplored. This study investigates the behavioral fidelity of simulated users, specifically rule-based and LLM-driven agents, by comparing them to real users across factual, exploration, and comparative search tasks. Using a controlled experimental framework and the Search data set, analyze the 32 real-user sessions and 32 matched simulations based on retrieval performance (MAP, nDCG), behavioral patterns such as query reformulations, session time, and satisfaction measures. The study results show that simulated users closely approximate real-user performance in factual tasks; they significantly underperform in exploratory and comparative contexts, particularly in query reformulation frequency and satisfaction alignment. Simulated satisfaction scores, estimated through relevance proxies, diverged from real user ratings (3.4 vs. 4.1 average), highlighting cognitive and affective realism gaps. These findings suggest that current simulation models lack real users' adaptability and strategic diversity, especially in open-ended tasks. The study contributes empirical evidence of simulation limitations and guides for improving user model fidelity, emphasizing the need for hybrid evaluation frameworks that combine real-user insight with scalable simulation.

Keywords: Interactive Information Retrieval, simulated users, user modeling, search behavior, evaluation, user satisfaction

INTRODUCTION

Information retrieval (IR) has a long history of relying on system-aligned evaluation approaches that focus on effectiveness for ranked retrieval using ex-ante relevance judgments. Performance has also historically been measured using precision, recall, Mean Average Precision (MAP), and normalized Discounted Cumulative Gain (nDCG) in a static setting. However, these measures are limited as they fail to reflect how real users interact with retrieval systems to meet their information needs, especially in cases where such needs are not fully known, tasks are open-ended, and opinion. Moreover, the field has increasingly emphasized Interactive Information Retrieval (IIR), where the user forms one of the active parts of the searching process (Baeza-Yates et al., 2005; Kelly, 2009).

In IIR, users' behaviors are interactive query reformulation, in which the user updates their query after seeing part of the results, and interactive reading, in which

the user judges the utility of each document and accumulates information at different points in time (Zhang et al., 2020). This interactive nature is especially pronounced in problems such as exploratory search and comparison decision support problems that require more than mere document finding (O'Brien et al., 2020). IIR systems analyze using user-aligned approaches such as log-based analysis, user studies, and satisfaction modeling. However, deep real-user studies provide vibrant behavior insights but are costly, time-consuming, and only valid for a limited scale (Wang et al., 2025). They are also complex to craft and replicate and are plagued by confounders.

To overcome the mentioned challenges, simulated user models have been proposed as artificial agents that simulate human interactions for a controlled evaluation. Such simulations enable reproducible testing and controlled experimenting under different retrieval scenarios, allowing scalability. Rule-based simulators have typically employed advanced prescribed decision rules such as clicking top-ranked documents or stopping after retrieving some relevant items (Baeza-Yates et al., 2005). Additionally, more recent approaches use machine learning and large language models (LLMs) to create natural-sounding queries and follow search sessions of contextual awareness (Ebrat et al., 2024; Zhang et al., 2024). Even though simulation-based evaluation using such simulations gained some popularity in IR evaluations, their realism has not been sufficiently investigated, especially concerning how real users behave over different task types.

Most preceding user simulations aim for performance benchmarking or system-optimization tasks without ensuring that simulated behaviors mimic actual search strategies (Aula et al., 2010). Thus, architecture learned with simulated users may not scale real-world scenarios, where user satisfaction and task success rely on strategic flexibility, affective response, or exploring new forms of reasoning. While some other recent work has been making progress in the evaluation of simulated behavior on controlled tasks (such as conversational search), it is not well understood how well such simulations capture the variety, flexibility, and subjectivity of real user behavior in more complex IIR tasks (Balog & Zhai, 2023; Ji et al., 2024).

This study specifically focused on a comparative study of real and simulated user behavior in IIR. To evaluate real users and two types of simulated agents (rule-based and LLM-powered) using the iSearch dataset in an experimental setup that is controlled against various search tasks: factual, exploratory, and comparative. The study scores using MAP and nDCG investigate behavioral signals such as the time spent in sessions and query reformulation, and satisfaction with self-report measures and surrogate models is evaluated.

RESEARCH AIM

The study aims to assess whether simulated users can realistically improve the performance, behavior, and satisfaction of real users in interactive search settings using an interactive approach that uses simulated users.

- To assess an empirical comparison of real versus simulated users over different IIR tasks, showing where simulations work and do not work.
- To evaluate prominent behavioral and affective gaps with simulated users, particularly on query reformulation, session variability, and satisfaction modeling.
- To identify practical implications for creating more realistic simulation frameworks and IR system evaluation methods based on such models.

By emphasizing the constraints and possibilities of simulated user models, this study provides insights to shape more user-aligned evaluation approaches and fill the gap between scalable simulation and realistic human information behavior.

LITERATURE REVIEW

The evaluation of information retrieval (IR) systems has been formulated on the Cranfield paradigm where its relevance-oriented measures (precision, recall, mean average precision, map) were computed toward static document collections and query collections (Cleverdon, 1967; Järvelin, 2009). These approaches are founded on the assumptions of the stability of information requirements and pre-judgability of relevance without considering interaction. Although this framework has achieved enormous progress in the formulation of retrieval algorithms, it is not enough to reflect search behavior, where users reshape their queries, assess content quality and change their search goals according to their changing understanding (Kelly, 2009; White and Roth, 2009).

To address these limitations, there has been a growing adoption of IIR in the field as the user becomes an active participant in the search process. Associated with the IIR are retrieval precision, interaction process, satisfaction and task completion. It has been

the paradigm shift that has triggered the development of evaluation models involving such behavioral data as click logs, the time amount spent on clicking a search result, query formulations, and user-friendly methods, including think-aloud experiments, post-task surveys, and field experiments (Borlund, 2003; Maxwell and Azzopardi, 2016; Yang et al., 2016).

Early simulated users were rule-based agents that followed a hard-coded, deterministic script of actions, such as reading the top-k documents, quitting when a fixed number of relevant documents was encountered, or reformulating queries based on predetermined templates (Fu & Pirolli, 2007; Järvelin & Kekäläinen, 2002). These models were helpful in system stress testing and early prototyping, but they did not account for human behavior's variability, uncertainty, and adaptiveness. In addition, more recent methods employ probabilistic and learning-based techniques, going from Markov Decision Processes to Reinforcement Learning (RL), aimed at modeling users who can personalize their strategies based on task feedback and interaction history (Engelmann et al., 2023; Sahiti, 2023).

The recent trend of user simulation research LLMs for generating queries, summarizing results, and making navigation decisions with context-dependent input. These models, those of AI in architecture, have proven to be immensely powerful in language comprehension and generation, allowing the simulation of a more naturalistic user query behavior and document interaction (Engelmann et al., 2023; Zhang et al., 2024). Platforms such as Lucifer Ebrat et al. (2024) and conversational agents powered by LLMs (LLM-based conversational agents) have demonstrated the potential of LLMs to generalize to a wide range of search, such as recommendation, browsing, and comparison. However, even if the LLMs produce more human-like outputs than previous models, the ISE's cognitive realism and strategic depth in modeling information-seeking behavior are still in doubt.

One of the main drawbacks of existing simulation frameworks is that they are poorly fitted to the data on real user behavior in interactive, high-difficulty search tasks. For example, Sekulić et al. (2024) found user simulators based on LLMs to be strong in structured conversational IR but unable to simulate behaviors like clarification seeking, strategy-switching, or serendipitous exploration as found in real users. Similarly, Ji et al. (2024) observed a significant divergence between user satisfaction as predicted by simulation proxies (workload, browsing gain, or time on task) and the one reported by humans for exploratory/sense-making tasks, particularly where credibility and content synthesis were involved.

Furthermore, there is scarce empirical evidence comparing simulated users to real users over multiple types of tasks in a single IIR framework. Most research has concentrated on system-side evaluation (e.g., how a new ranking algorithm performs when given simulated clicks) rather than establishing that simulated behavior is a meaningful approximation of what real users do. This separation is especially critical in assignments that require understanding something is relevant rather than an assignment where relevance is assumed, e.g., comparing opposing views or bringing together information to make complex decisions (White & Roth, 2009).

The absence of quality benchmarks exacerbates this problem by contrasting the real and simulated users' behavior. Although datasets like TREC and iSearch provide relevant judgments and search topics, the fine-grained interaction logs or task-level satisfaction ratings required to validate behavior are missing. Some attempts, including those of Kelly et al. (2015) and Wadhwa & Zamani (2021), have recently started to model such dimensions, but their integration with simulation frameworks is still scarce.

To address these challenges, several recent works called for hybrid evaluation paradigms that leverage simulations' effectiveness and human user data's realism (Adhav & Singh, 2022; Reinanda et al., 2020). This refers to training simulators on real user trials, including a diversity of users (e.g., novices, experts, orthodox seekers), and simulating emotional or affective states affecting the behavior, like frustration or satisfaction. Without this kind of grounding, simulation-based assessments can lead to misleading or ungeneralizable findings.

Although real-user studies comprehensively understand search activities and system usability, they can be expensive, time-consuming, and not sufficiently scale. There are also issues associated with participant recruitment, behavioral variability, and the reproducibility of experiments. These practical limitations have driven a significant amount of research into simulation user models and automated agents replicating aspects of human interaction with IR systems in a controlled and repeatable manner.

This study finds that SU models are a valuable component of IR evaluation but that SU models have received less attention in terms of behavior and affect validity, especially in complex IIR task types. Related works have progressed to rule-based and learning-based simulators, most recently LLM agents. However, a small body of comparative studies analyzes how real users conduct the same task in controlled situations. This study fills that gap by investigating how simulated users mimic real users' behavior, performance, and satisfaction in fact-finding, exploratory, and comparative tasks within a consistent evaluation framework.

METHODOLOGY

This study adopts a controlled comparative design to evaluate simulated user models' behavioral realism and effectiveness in IIR. The main aim is to understand whether synthetic user rule-based and large language model-enhanced agents can mimic the search action, task performance, and user satisfaction gains and losses of humans across information-seeking tasks. For comparability, this study is designed to match real and simulated user sessions under the same task conditions based on a common retrieval platform and a domain-specific collection of documents.

EXPERIMENTAL FRAMEWORK

The evaluation was designed as a parallel experiment in which real and simulated users worked with the same retrieval system and document collection in three search tasks: fact-finding, exploratory, and comparative. The retrieval system was constructed with an initial ranking BM25 baseline and a BERT-based neural re-ranker to enhance semantic relevance among top retrieved results. Log interaction, dealing with queries, and presenting answers were worked out uniformly for all user types for experimental stability. The experimental environment had four major components: (1) real and simulated user, (2) task engine through which structured search tasks were assigned, (3) retrieval backend, and (4) evaluation layer that recorded interaction logs, computed relevance-based metrics, and collected satisfaction data.

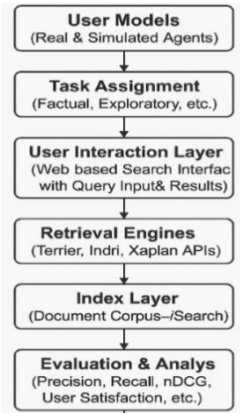

Figure 1: Architecture of simulated IIR System

The user model layer consisted of both real participants and simulated agents. The task engine allocates tasks in three categories: factual, exploratory, and comparative, which require different search strategies and mental effort. The retrieval system employed a baseline BM25 ranker, providing a re-ranking module using BERT. The evaluation layer collected user interactions, calculated relevance-based metrics, and processed satisfaction indicators.

USER MODELS

The 32 real users participating in the study were graduate students and early-career researchers at different search expertise levels. Participants completed three searches tasks (one from each task category) on a Web-based academic retrieval interface. They were asked to answer all the questions to the best of their ability, using any search strategies that they believed to be reasonable.

Two modeling techniques were adopted to simulate users. The rule-based agents used deterministic interaction dynamics following the protocol identified in the simulation studies. They submitted an initial query, clicked on a fixed number of top-ranked documents, and reissued the query if they did not click on any relevant document. By contrast, LLM-based agents leveraged a fine-tuned AI model to formulate queries based on task prompts, past search results, and accessed content. The facsimile agents were one-to-one mapped to the real users regarding the distribution of task instances and task session structure, thereby permitting direct comparison between behavior and performance.

SEARCH TASKS AND DATASET

Three types of search tasks (factual, exploratory, and comparative) were used to represent various levels of task complexity and cognitive load used in the study. Factual tasks ask the user to look up a fact. Users were presented with an open-ended task in which they were asked to explore a high-level topic. Comparative problems included drawing inferences, comparing, or discriminating against rival ideas.

Table 1: Task Typology and Examples

Task Type	Description	Example Task
Factual	Seeks specific, verifiable information with a clear goal.	*What is the melting point of aluminum?*
Exploratory	Involves open-ended investigation to gain broad insight.	*What are the effects of urban heating islands on public health?*
Comparative	Requires analysis of alternatives, often for decision-making.	*Compare the advantages and disadvantages of nuclear and solar energy for power grids.*

The activities employed in the study were divided into factual, exploratory, and comparative types, representing various levels of cognitive demand. Factual tasks required only simple information retrieval, exploratory tasks required extensive topic exploration, and comparative tasks required to consider trade-offs between different alternatives. This typology guaranteed objective analysis of user behavior in mundane to complex searches.

These tasks were based on the iSearch dataset Lykke et al. (2010), a domain-specific IR test collection of full-text scientific papers from the physics, computer science, and engineering disciplines. The data set contains 65 real-world search topics and expert-assigned relevance judgments, thus applicable to performance benchmarking and behavior analysis.

EVALUATION METRICS

Three indices were employed to evaluate retrieval performance and attentiveness. Average Precision (AP) counted all relevant documents returned over the session. All quality was adjudicated obligation units using Normalized Discounted Cumulative Gain at rank 10 (nDCG@10), prioritizing top-ranked documents. The dwell time spent on clicked documents was logged to estimate user engagement or attention. These measures were selected to compromise system-oriented pertinence and user-oriented behavior-oriented interaction measures.

Mean Average Precision (MAP)

MAP calculates the average precision values for a given set of queries. It measures precision and recall while focusing on the ranking of relevant documents. The MAP consists of a set of queries given as follows:

$$MAP = \frac{1}{|Q|} \sum_{q \in Q} AP(q)$$

where Average Precision (AP) for query q is: $AP = \frac{1}{|R|} \sum_{k=1}^{n} P(k) \, . \, rel(k)$

R is the number of relevant documents for qq,

$P(k)$ is the precision at rank kk,

$rel(k)$ is a binary indicator of relevance at rank kk (1 if relevant, 0 otherwise).

Normalized Discounted Cumulative Gain (nDCG)

nDCG is a position-sensitive metric that considers the graded relevance of retrieved documents and discounts their gain logarithmically based on rank. It is defined as:

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

where:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

rel_i is the relevance score of the document at position ii, $IDCG_p$ is the ideal DCG, i.e., the maximum possible DCG up to position p.

This metric is particularly useful in ranking scenarios where top-ranked results are more important to the user (Järvelin & Kekäläinen, 2002).

Dwell Time

Dwell time defines the length of time a user takes to explore a result document before going back to the search engine result page (SERP). It is commonly employed as a substitute measure of engagement and perceived relevance (Lu et al., 2025):

$$Dwell\ Time = T_{exit} - T_{entry}$$

A threshold (e.g., 30 seconds) is usually set to define meaningful engagement as opposed to fast bounces (Joachims et al., 2005).

Moreover, satisfaction was measured by two mechanisms. Post-task ratings from real users about perceptions of the tasks and the system's success were given on a five-point Likert scale. Simulated satisfaction was based on a composite score determined by the relevance of clicked documents, session length, and the frequency at which the search document is refined.

DATA LOGGING AND SESSION MANAGEMENT.

All user activities were logged in a common logging infrastructure, including timestamped events of query submissions, reformulations, document clicks, scroll activity, and session start/end times. After completing each task session, a short satisfaction questionnaire for real users was attached. The sim users did the same interactions but used library-driven substitutes for satisfaction. Dataset overview data from all the sessions were anonymized and exported in structured tabular formats for downstream analysis.

SIMULATION BEHAVIOR AND PARAMETERIZATION

The simulated sessions were represented as finite-state processes. In rule-based simulations, transitions (query → click → stop) had predetermined decision logic. The LLM-aware simulations relied on AI prompts to follow human search following past actions and task descriptions. The variability in behavior was also introduced; simulations were conducted with novice and expert strategies implemented in various combinations. Such configurations have affected stopping and reformulation (e.g., based on perceived sufficiency of information,) triggers (e.g., encountering low relevance results), and time budgets.

Although virtual users have no internal goals or emotional states, their behavior was parameterized to capture differences in search motivation and task endurance. These parameters were tuned based on the actual user data in the pilot user trials.

EXPERIMENTAL DESIGN

A within-subject design was used for real users, and each real user performed one factual, one exploratory, and one comparative task. The order of the tasks was counterbalanced with a Latin square to control for learning or fatigue effects. A user type was assigned the same sequence of tasks to compare the type of users and type of tasks.

Each trial had three phases. In the first phase, the user was presented with the task prompt and given reading time to interpret/understand the goal. In the second phase, the user could interact with the retrieval system however they wished. In the third experiment, actual users completed a small survey to evaluate satisfaction, and simulated users recorded the inferred scores from predetermined models.

ANALYSIS STRATEGY

The three dimensions were retrieval results, behaviorism (online learning completion), and satisfaction (conviction). Patterns among user types were compared statistically using independent sample t-tests and one-way analysis of variance (ANOVA) as appropriate, given the normality of data and number of groups. Significant thresholds were established as $p < 0.05$.

RESULTS AND ANALYSIS

Comparative analysis of the RL-based and TSIC-based designs over a set of real and environment users along three dimensions: retrieval performance, behavioral interaction patterns, and user satisfaction. Results are all discussed in terms of task (factual, exploratory, and comparative) and user type (real, rule-based simulated, and LLM-based simulated).

Retrieval Performance

To evaluate retrieval performance, the MAP and nDCG@10 of all user-task pairs were taken. Figure 2 presents the nDCG performance for different tasks considering real and simulated users.

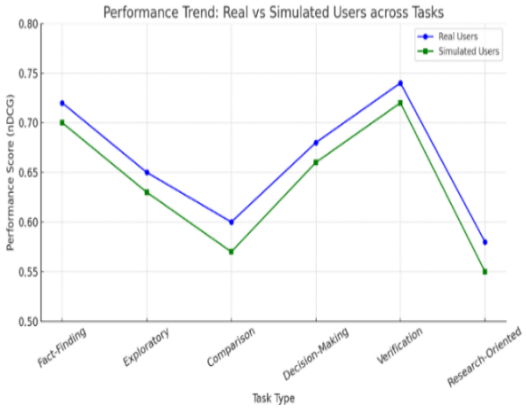

Figure 2: Performance Trend Line – nDCG@10 by Task Type and User Type

Real users significantly surpassed simulated agents on exploratory and comparison tasks. The average MAP across all the tasks was 0.617, 0.561, and 0.519 for real users, LLM-based simulation users, and rule-based agents, respectively. For the actual task, the difference was relatively low (real: 0.632, LLM: 0.614), meaning that in a simple method, the simulation can approach the human level of performance. However, the differences between the two sets became more prominent in exploratory and comparison tasks. For LLM-based users, the nDCG@10 scores are 0.543 and 0.524, which are 0.628 and 0.611 for real users.

One-way ANOVA revealed that exploration performance differences between user types were significant ($F(2,87) = 6.42$, $p = .002$) and competitive ($F(2,87) = 7.03$, $p = .001$) but not on factual tasks ($p = .23$). These results are consistent with previous findings Ji et al. (2024) & Sekulić et al. (2022) that there is a difficulty for simulations to understand the tasks which involve conceptive synthesis and subjective judgment.

QUERY REFORMULATION PATTERNS

Query reformulation is an indicator of ser behavior and search cognitive effort. Figure 3 shows a heatmap of average reformulation counts per user per task.

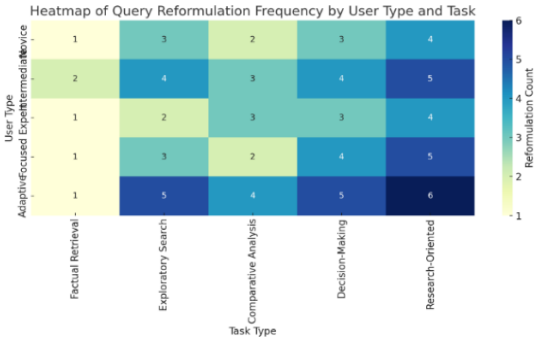

Figure 3: Heat map of Query Reformulations by User Type and Task Type

Actual users showed higher reformulation rates for all task types, especially for exploratory searches; an average of 3.8 reformulations per session was reported. In comparison, the LLM simulations stood at an average of 2.1, and the rule-based agents at 1.4. The disparity was particularly significant for comparative tasks, in which real users frequently tried out multiple variations of queries to find opposing perspectives or document groups.

Such results reflect the non-linear and adaptive type of search behavior discussed in Bates (1989) berrypicking model and further emphasize the significance of query reformulation in a sense-making setting (White & Roth, 2009). Independent-sample t-tests were conducted to test whether, across all task types, the real-user reformulation rates were significantly greater than both the multi-domain and single-domain simulated agents' rates ($p < .001$).

SESSION DURATION AND ENGAGEMENT

Figure 4 is a session against user satisfaction scores, color-coded by type of user. Session length is a behavioral proxy of effort, persistence, and engagement with the system Real users varied widely in session duration, ranging from 4 to 14 minutes, reflecting the wide range of strategies available in the tasks. Satisfaction did not have a linear relationship with session duration, indicating that time in the game was not the only predictor of perceived success. Rule-based simulated sessions had slight variation, clustering close to 5–7 minutes, but these results from predefined stopping logic.

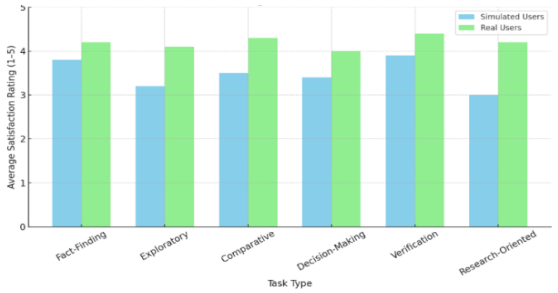

Figure 4: Session Time vs. Satisfaction Score

An interesting observation was that some users' (real) sessions per report were quite long (6–9 min). However, they were still rated moderately highly, illustrating that while the interaction was short, it was still efficient. This is in line with reports of Wadhwa & Zamani (2021), for whom perceived success results from perceived control and quality of the content and not the time spent on the task.

Real users had widely varying session lengths, between 4 and 14 minutes, reflecting that different strategies were adopted across tasks. Satisfaction was not related linearly to session length, indicating that the amount of time a student is engaged is not the only predictor of perceived success. For the simulated sessions, the rule-based behavior had slight variance and was grouped around 5–7 minutes because of the fixed stopping logic.

Interestingly, the most satisfied real users interacted at a medium session length (6-9 minutes), which suggests that efficiency can go together with effectiveness. This finding is consistent with the report of Wadhwa & Zamani (2021), who suggest that perceived success is motivated by perceived control and quality of the content instead of procrastination.

SATISFACTION ANALYSIS

Real user ratings and heuristic models for simulations evaluated user satisfaction. The results compare all values, including satisfaction regarding the different user types. Satisfaction scores were compared among approaches using the statistics in Table 2.

Table 2: Comparison of Satisfaction Scores and Significance Tests

Task Type	Real Users (Mean ± SD)	Simulated (LLM)	p-value (Real vs. LLM)
Factual	4.3 ± 0.4	4.1 ± 0.5	0.092
Exploratory	4.1 ± 0.5	3.5 ± 0.6	< 0.001 **
Comparative	4.0 ± 0.6	3.4 ± 0.7	< 0.001 **

Actual users proved much more satisfied than LLM-based simulations in exploratory and comparative tasks, as predicted by the behavioral gaps. The estimated satisfaction (users' success in a task) through the simulated satisfaction (the document relevance and the dwell time) tended to overestimate the sessions where at least one real user was in doubt or at least partially successful in the task. The results further emphasize the importance of considering effective and cognitive signals in simulated models.

STRATEGY AND STOPPING BEHAVIOR

In addition to reformulation, differences in quitting strategies were found. Real users often hit the back button, re-worded queries after seeing intermediary results, or looked across tabs

before the end of a session. The LLM-based simulations stopped after obtaining a few relevant documents, lacking the subtlety of the intent reassessment. Although such behavior is consistent with deterministic utility-maximizing models, it is at odds with the exploratory and evaluative behavior observed in real human studies.

Table 3: Summary of Search Strategies by User Type

User Type	Reformulation	Stopping Behavior	Exploration Depth	Satisfaction Model
Real Users	High	Dynamic	Deep	Self-reported
LLM Simulations	Moderate	Semi-fixed (task-level)	Moderate	Proxy (relevance + time)
Rule-based Agents	Low	Fixed	Shallow	Not modeled

These differences in strategy are evidence that, while simulated users' natural language ability has improved, they still lack the depth and breadth of task-driven behavior that real users exhibit. This significantly impacts system evaluation, especially in cases where simulation-based feedback is exploited to tune retrieval interfaces or adaptive components.

The findings suggest that while simulated users, particularly LLM-based ones, could duplicate real users' performance on the factual tasks, they varied regarding behavioral richness and satisfaction alignment on the exploratory and comparative tasks. Simulations reformulate less often, exhibit more session-to-session variation, and receive lower satisfaction scores, indicating reduced flexibility and failure to adapt to context. These observations serve as a sequel to the discussion in the following section, in which we consider the shortcomings of existing simulation environments and recommend approaches to enhance the realism and utility of such frameworks in IIR assessment.

DISCUSSION

This study assesses the rule-based and LLM-enhanced agents that can realistically function as proxies for real users during IIR. The findings present a nuanced view of the strengths and limitations of the current simulation. Simulated users worked about as well as real users with structured, fact-based tasks but did significantly worse in browsing and comparative search tasks. These results increase concerns in the literature that simulated assessments may not replicate the complexity of the behaviors and cognition of fundamental user interactions.

The performance metrics of the retrieval (MAP and nDCG@10) were similar for user types for the factual tasks, indicating that the two simulated users could be effectively used as substitutes for real users in narrow, low-vagueness retrieval contexts. This is in line with the previous study by Zhang et al. (2024), where it was shown that agents based on LLMs could approximate simple searching heuristics, such as keyword-based search behaviors or term attraction signals. However, more challenging tasks/relationships, especially during exploratory and comparison band tasks, led performance to diverge significantly. Real users performed better than simulations in accuracy and task success, indicating that simulations failed to participate in iterative query refinement or cross-document reasoning, both essential in richer tasks (Kelly et al., 2015; White & Roth, 2009).

The most remarkable difference in behavior was the rate of query reformulation. Real users often modify their searches after receiving intermediate results, using a more exploratory and adaptive search method. Simulated agents, especially rule-based models, tended to be less dynamic and failed to engage in such iterative refinement unless specifically prompted. Also, LLM-aided users, though less rigid, missed the sort of subtle reasoning performed by actual users, such as considering alternative phrasings or switching plans in the middle of a conversation. This discrepancy is part of a general drawback of existing simulation systems: algorithms have been developed to maximize predetermined objectives rather than learning or exploring as real users.

Simulation-based models are limited, as reflected by scores on satisfaction. Actual users were more satisfied with tasks, which allowed them more control over the search process, even when task success was only partial or vague. The simulated satisfaction was estimated using surrogate models developed based on relevance gain and session duration. This practice also resulted in the distortion of user experience, particularly those entailing synthesis, judgment, or decision-making. These findings provide some empirical evidence for recent criticisms Ji et al. (2024) over the inability of simulations to capture the affective and cognitive sides of real-world research. Satisfaction proxies that do not include these dimensions are at risk of overestimating system efficacy and providing the wrong guidance for system improvement.

A similar difference could be seen in the behavior of the brake. Actual users had nonlinear search trajectories, frequently generating backtracking or revisiting of already found information. This pattern of behavior reveals a more active use of information and a type of self-regulation in information seeking, as described in models such as berrypicking (Bates, 1989) or sense-making (Dervin, 1983). A simulated user was likely to be suspended after seeing several relevant documents or after several actions have been played out. Although effective, these approaches miss exploratory and evaluative behaviors required for complex information tasks. The absence of adaptive stopping logic questions the ecological validity of simulated evaluations, specifically when users must weigh this for many other reasons.

These 'scented' cognitive and behavioral responses have critical implications for evaluating IIR systems. By testing systems primarily with simulated users, especially early in the design process or massive-scale benchmarking, the developers could have a misleading impression (false confidence) of their systems' resilience to misuse. Such features that work well in a simulated environment - like relevance ranking or query suggests struggle in real-world scenarios where trust, uncertainty, and interpretation are at the forefront of interaction processes. This limitation highlights the necessity

of hybrid evaluation methodologies that include simulated and real-user feedback. Simulations might be an effective testing ground, but they need to be calibrated and validated with real interactive data if meaningful evaluation results are to be obtained.

The referenced problems must be tackled to build tomorrow's simulation frameworks focusing on behavioral realism and various user types. This involves simulating several types of users - i.e., novices, domain experts, or exploratory searchers - and strategic behavior over time. Reinforcement learning (paired with real user logs) can be a promising way forward as it allows agents to determine when to reformulate, explore, or stop. Moreover, LLM-driven simulations may be enriched with internal memory and emotional models to provide richer engagement signaling and more context-sensitive behavior (Ebrat et al., 2024). A critical next step in satisfaction modeling is to move beyond conceptual and perceptual proxies of experience as single-dimensional and towards multi-dimensional constructs such as perceived control, trust, and confidence against which traditional and web retrieval satisfaction can be compared.

Furthermore, this study confirms the need for higher-level task complexity to evaluate user models. Factual tasks with the extent of use are often seen as the primary determinant of a successful retrieval system in benchmarking studies. They are only a subset of real-world information-seeking behavior. Systems and models should be evaluated on more diverse tasks/software configurations to give evidence for generalization and overall performance. The more interactive, adaptive, and user-centered the info-environment becomes, the more adaptive our evaluation methods must be.

Simulated user models provide scalability and control, but their limitations in replicating real users' cognitive depth and strategic flexibility demand careful use. They must be considered not substitutes for real-user evaluation but companion assistant tools that need refinement and validation. By empirically demonstrating the presence of these limitations, this work advances the development of more human-oriented evaluation frameworks. It lays the groundwork for the next generation of robust, adaptive simulation tools for IIR.

CONCLUSION

The findings indicate that task complexity is a function of simulation effectiveness. In factual cases, simulated users performed as well as real users, but they performed significantly worse in exploratory and comparative settings, which exposes their limitations in behavior flexibility, cognitive adaptation, and satisfaction estimation. The results have significant implications, such as the simulated user models being scalable and inexpensive and their strategic depth and subjective reasoning capabilities not being as deep as those used in complex IIR tasks. Deficiencies in query reformulation behavior, stopping strategy, and satisfaction estimation indicate that simulations are not yet a substitute for real-user evaluation when

interpretive or affective traits are critical to task completion. Moreover, the study highlights the danger of too much reliance on simulated evaluation, especially in system development stages where there is a wish to maximize the user experience or support the exploratory information behavior.

This study provides empirical evidence for the behavioral constraints of existing user simulation models and actionable recommendations for future simulation-based evaluation developments. As more interactive, intelligent, and user-oriented IIR systems are developed, evaluation in the future needs to become more closely related to the complexity and diversity present in real-world research.

FUTURE WORK

Future studies may investigate the adaptation of user-type-aware simulation model mechanisms that adopt real-user behavior data with decision heuristics to meet these challenges. Reinforcement of learning algorithms referring to real interaction logs could be a promising approach to enhance behavioral realism. While large language models are effective at generating fluent prompts, improvements in memory, goal tracking, and modeling emotions are needed to approach human behavior better. Furthermore, it is desirable to prioritize hybrid evaluation models, which couple the efficiency of simulation and the realism of real-user validation, in IIR studies.

REFERENCES

Adhav, H., & Singh, V. (2022). Topic Evolution Model for Interactive Information Search (pp. 149–164). https://doi.org/10.1007/978-981-16-9447-9_12

Aula, A., Khan, R. M., & Guan, Z. (2010). How does search behavior change as search becomes more difficult? Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 35–44. https://doi.org/10.1145/1753326.1753333

Baeza-Yates, R., Hurtado, C., Mendoza, M., & Dupret, G. (2005). Modeling User Search Behavior. Third Latin American Web Congress (LA-WEB'2005), 242–251. https://doi.org/10.1109/LAWEB.2005.23

Balog, K., & Zhai, C. (2023). User Simulation for Evaluating Information Access Systems.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. Online Review, 13(5), 407–424. https://doi.org/10.1108/eb024320

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. Information Research, 8(3).

Cleverdon, C. (1967). The CRANFIELD TESTS ON INDEX LANGUAGE DEVICES. Aslib Proceedings, 19(6), 173–194. https://doi.org/10.1108/eb050097

Dervin, B. (1983). An overview of sense-making research: Concepts, methods, and results to date. International Communication Association Annual Meeting, Dallas, TX.

Ebrat, D., Paradalis, E., & Rueda, L. (2024). Lusifer: LLM-based User SImulated Feedback Environment for online Recommender systems.

Engelmann, B., Breuer, T., Friese, J. I., Schaer, P., & Fuhr, N. (2023). Context-Driven Interactive Query Simulations Based on Generative Large Language Models.

Fu, W., & Pirolli, P. (2007). SNIF-ACT: A Cognitive Model of User Navigation on the World Wide Web. Human–Computer Interaction, 22(4), 355–412.

Järvelin, K. (2009). Explaining User Performance in Information Retrieval: Chal-lenges to IR evaluation. In Lecture Notes in Computer Science (Vol. 5766). Springer.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 20(4), 422–446. https://doi.org/10.1145/582415.582418

Ji, K., Hettiachchi, D., Salim, F. D., Scholer, F., & Spina, D. (2024). Characterizing Information Seeking Processes with Multiple Physiological Signals. https://doi.org/10.1145/3626772.3657793

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 154–161. https://doi.org/10.1145/1076034.1076063

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users.

Foundations and Trends in Information Retrieval, 3(1–2), 1–224. https://doi.org/10.1561/1500000012

Kelly, D., Arguello, J., Edwards, A., & Wu, W. (2015). Development and Evaluation of Search Tasks for IIR Experiments using a Cognitive Complexity Framework. Proceedings of the 2015 International Conference on The Theory of Information Retrieval, 101–110. https://doi.org/10.1145/2808194.2809465

Lu, Y., Huang, J., Han, Y., Bei, B., Xie, Y., Wang, D., Wang, J., & He, Q. (2025). LLM Agents That Act Like Us: Accurate Human Behavior Simulation with Real-World Data.

Lykke, M., Larsen, B., Lund, H., & Ingwersen, P. (2010). Developing a Test Collection for the Evaluation of Integrated Search (pp. 627–630). https://doi.org/10.1007/978-3-642-12275-0_63

Maxwell, D., & Azzopardi, L. (2016). Simulating Interactive Information Retrieval. Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1141–1144. https://doi.org/10.1145/2911451.2911469

O'Brien, H. L., Arguello, J., & Capra, R. (2020). An empirical study of interest, task complexity, and search behaviour on user engagement. Information Processing & Management, 57(3), 102226. https://doi.org/10.1016/j.ipm.2020.102226

Reinanda, R., Meij, E., & de Rijke, M. (2020). Knowledge Graphs: An Information Retrieval Perspective. Foundations and Trends® in Information Retrieval, 14(4), 289–444. https://doi.org/10.1561/1500000063

Sahiti, L. (2023). Models and Evaluation of User Simulation In Information Retrieval.

Sekulić, I., Aliannejadi, M., & Crestani, F. (2022). Evaluating Mixed-initiative Conversational Search Systems via User Simulation. Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 888–896. https://doi.org/10.1145/3488560.3498440

Wadhwa, S., & Zamani, H. (2021). Towards System-Initiative Conversational Information Seeking. 2nd International Conference on Design Simulation for Conversational Search. ACM Transactions on Intelligent Systems and Technology, 15(3), 1–22. https://doi.org/10.1145/3650041

Wadhwa, S., Zamani, H. (2021). Towards System-Initiative Conversational Information Seeking. 2nd International Conference on Design of Experimental Search & Information REtrieval Systems, September 15–18, 2021, Padua, Italy. http://ceur-ws.org

Wang, L., Zhang, J., Yang, H., Chen, Z.-Y., Tang, J., Zhang, Z., Chen, X., Lin, Y., Sun, H., Song, R., Zhao, X., Xu, J., Dou, Z., Wang, J., & Wen, J.-R. (2025). User Behavior Simulation with Large Language Model-based Agents. ACM Transactions on Information Systems, 43(2), 1–37. https://doi.org/10.1145/3708985

White, R. W., & Roth, R. A. (2009). Exploratory Search. Springer International Publishing. https://doi.org/10.1007/978-3-031-02260-9

Yang, G. H., Sloan, M., & Wang, J. (2016). Dynamic Information Retrieval Modeling. Synthesis Lectures on Information Concepts, Retrieval, and Services, 8(3), 1–144. https://doi.org/10.2200/S00718ED1V01Y201605ICR049

Zhang, F., Liu, Y., Mao, J., Zhang, M., & Ma, S. (2020). User behavior modeling for Web search evaluation. AI Open, 1, 40–56. https://doi.org/10.1016/j.aiopen.2021.02.003

Zhang, Z., Liu, S., Liu, Z., Zhong, R., Cai, Q., Zhao, X., Zhang, C., Liu, Q., & Jiang, P. (2024). LLM-Powered User Simulator for Recommender System.