Do Simulated Users Need to Remember? Analyzing the Impact of Memory Models in Conversational Search Evaluation

Nailia Mirzakhmedova
Bauhaus-Universität Weimar
Weimar, Germany

Marcel Gohsen
Bauhaus-Universität Weimar
Weimar, Germany

Johannes Kiesel
GESIS - Leibniz Institute for the Social Sciences
Cologne, Germany

Matthias Hagen
Friedrich-Schiller Universität Jena
Jena, Germany

Benno Stein
Bauhaus-Universität Weimar
Weimar, Germany

Abstract

Conversational search systems are typically evaluated using a fixed reference collection of conversations or through user studies with a live system. However, fixed-reference conversations can cover only a few plausible conversations, and user studies are costly, time-consuming, and often hard to reproduce. A promising alternative that avoids coverage and cost issues is user simulation, in which a computer program takes on the role of a user and interacts with the system under evaluation. But the complexity of human search behavior raises the question of how "realistic" the simulations actually need to be for reliable evaluations of conversational search systems. In this paper, we ask: Do simulated users need to remember? While real users may learn and forget information during conversational search sessions, which inspired previous research to also model memory capabilities in simulations, it remains unclear whether this actually influences the results of system evaluations. To investigate the impact of memory modeling, we analyze conversations of simulated users and of humans with four conversational search systems. Our results suggest that incorporating long-term memory into simulators can help reproduce system effectiveness rankings obtained from human conversations, whereas incorporating short-term memory can diminish the reproduction. We also find that simulators are generally valid and reproducible—and memory modeling even increases run-to-run reproducibility of system rankings—but overall, simulations approximate human evaluation scores better when "helpful" assistants are evaluated than when assistants with deteriorated response quality are assessed. Our code and data are available at https://github.com/webis-de/SIGIR-26.

CCS Concepts

• **Information systems → Users and interactive retrieval**; • **Computing methodologies → Modeling and simulation**.

Keywords

Conversational information access; Conversational search; User simulation; Conversational search evaluation

This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGIR '26, Melbourne, VIC, Australia
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809556

ACM Reference Format:
Nailia Mirzakhmedova, Marcel Gohsen, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2026. Do Simulated Users Need to Remember? Analyzing the Impact of Memory Models in Conversational Search Evaluation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3805712.3809556

1 Introduction

Conversational search allows a user to iteratively deepen their knowledge of some topic through a natural-language dialogue with multiple turns [59]. But as different system responses can be acceptable for a turn and as each of them may elicit a different user reaction, the robust evaluation of conversational search is not that straightforward and in fact is still not "solved" [15, 26]. For instance, the widely used Cranfield-style evaluation setup [47] with a static set of fixed queries is not suited to the conversational/generative multi-turn domain, as only a limited range of plausible conversations on a topic can be covered. As an alternative, analyzing system behavior in user studies can mitigate the coverage issue, but user studies are often costly and face reproducibility challenges [9].

As a third option, simulating user interactions with conversational search systems has recently gained attention with the development of large language models (LLMs) that can generate human-like utterances [17, 21, 32, 33, 42, 50, 51]. For such a simulation to be realistic, Balog [7] formulated key requirements, including modeling users' personas, preferences, and knowledge states. Notably, a user's knowledge is not static but evolves over time as they learn new information and forget old information across multiple interactions with a conversational search system, which might necessitate some memory component in simulated users to store and retrieve acquired "internal" knowledge relevant to a given turn for follow-up interactions. However, even though several studies have explored LLM-based simulators with respect to specific user profiles or goals [28, 58] and with respect to some initial steps towards tracking knowledge and interaction consistency [17, 56, 64], the question of how "realistic" or simply "good enough" a simulator (and its memory modeling) needs to be for reliable and reproducible evaluations of conversational search systems is still open [8].

In this paper, we investigate whether and how modeling short- and long-term memory affects the robustness of simulation-based conversational search evaluation. By 'short-term memory', we refer to a simulated user's ability to retain and use information from

recent turns of a current conversation, while 'long-term memory' rather is a simulated user's ability to store and retrieve general facts, preferences, and knowledge accumulated from previous dialogues. Figure 1 illustrates these two types of memory in a simulated user interacting with a conversational search system. As for the evaluation robustness, we analyze three key aspects based on the following research questions (RQs):

RQ1: (*Validity*) How well do relative rankings of conversational search systems obtained from memory-equipped user simulations reflect the rankings obtained from real user studies?
RQ2: (*Reproducibility*) Do memory components improve the reproducibility of relative system rankings across multiple simulation runs?
RQ3: (*Approximation*) How well can absolute effectiveness scores from real user studies be approximated by memory-equipped simulations?

Interestingly, our results indicate that incorporating a short-term memory component does not consistently improve the evaluation robustness and can even introduce variability in system rankings across simulation runs. In contrast, equipping simulators with a long-term memory component improves the validity and reproducibility of conversational search system evaluations, both in automatic and human assessments, and it enhances the approximation of absolute effectiveness scores from real user studies.

2 Related Work

The term conversational search (or conversational information seeking) describes a task-oriented sequence of exchanges between one or more users and an information system [19, 43, 59]. In contrast to "traditional" ad hoc search systems, a system that allows for such an exchange, called a conversational search system, can actively request information from the user [34]. Such systems are also known as mixed-initiative systems [5]. In the following, we briefly review the literature on the evaluation of conversational search systems in general, on evaluation via user simulation in particular, and on memory models for both humans and simulated users.

Evaluation of Conversational Search Systems. Typically, conversational search systems are evaluated in online, offline, or simulated settings [7]. Online evaluation refers to deriving some measure of user satisfaction from observing how real users interact with a system in production to fulfill their information needs. This means that academic research that applies online evaluation is typically restricted to industry collaborations, for example, in living-lab shared tasks [2, 30, 49]. In contrast, offline evaluation refers to observing how systems respond to a fixed set of pre-collected user queries; for instance, collected from a reference system in TREC shared tasks [3, 4, 20]. The evaluation itself then typically relies on relevance judgments for some fixed document set (can be reused for other systems' responses based on the same set) with the judgments stemming from (expensive) human judging or from automated LLM-based judging, which has not only been investigated in theory [24] but also already in practice [25, 46]. While offline evaluation is more reproducible than online evaluation, the restriction to a few pre-collected conversations from one or a few reference systems is a clear bottleneck. As a way out, evaluation via user simulation establishes only the start of a conversation—the first user turn or the information need—and then proceeds by observing how simulated users interact with a system [7]. Fixed (human) relevance judgments for the responses to not-fixed turns are then no longer feasible, so that LLMs are often employed for judgments [33].

User Simulation. The main goal of user simulation in IR is to predict the effectiveness and/or efficiency of a search system by observing interactions with simulated users instead of (or in addition to) real users. In this context, Balog [7] calls a simulation "realistic" if an evaluation of systems via the simulator leads to the same system ranking as an evaluation via real users; as important modeling aspects, he lists persona, interests and preferences, multimodal interactions, learning and forgetting, and the mental model of the simulator. Similarly, Bignold et al. [12] stress that simulating users with diverse search behaviors, such as the accuracy of the information they provide, how quickly their goals change, biases towards certain types of information, or varying knowledge levels, helps to get broader insights into system effectiveness.

Historically, simulation for evaluation has been first suggested for traditional "ten blue links" search systems with models for clicks / result interactions [18], query formulation [14], and stopping [39] to, for instance, assess the effectiveness of personalized query suggestions [54] or to compare real users' effectiveness with upper and lower bounds [10, 29]. Later, similar questions and studies also appeared in the recommender systems literature (e.g., [13, 61]).

As for conversational search, most simulation studies nowadays employ LLMs for generating "user" utterances, often enriched by persona information [55, 63], emotional states [36], retrieval-augmented generation to integrate specific knowledge [22, 31], information on similar conversations with real users as behavior reference [52], or balancing between uniform and boosted user adaptation [37]. Available frameworks for user simulation include UserSimCRS [1] (focus: conversational recommender systems), SimIIR [6, 38, 60] (focus: complex and conversational search model), and GenIRSim [33] (focus: web interface for running and analyzing search simulations). Among these, only SimIIR mentions simulating memory, but only as a promising idea for future work.

To ensure that user simulators for retrieval evaluation fulfill the intended purposes, they themselves need to be evaluated. In this respect, Balog and Zhai [9] describe three essential angles to assess the credibility of a simulation. First and foremost, the retrieval performance (i.e., effectiveness and/or efficiency) predicted by a simulator should be comparable to the actual retrieval performance observed in real-world scenarios with real users. Second, a user simulator's behavioral characteristics, such as interaction patterns, should align with the user group they are intended to mimic. And third and finally, human evaluations of simulated outputs should be conducted to assess characteristics such as naturalness or usefulness; for instance, as pairwise comparisons of judgments in the style of the Turing test (bot versus human) [61].

However, the choice of a specific metric to evaluate simulations in some particular scenario is often unclear, as there are no respective "standard" metrics yet. Shared tasks can lay the foundation for establishing such standard metrics, and the micro-shared task at Sim4IA 2025 [48] was among the first shared tasks in which the goal of the participants was to build user simulators. The TREC User

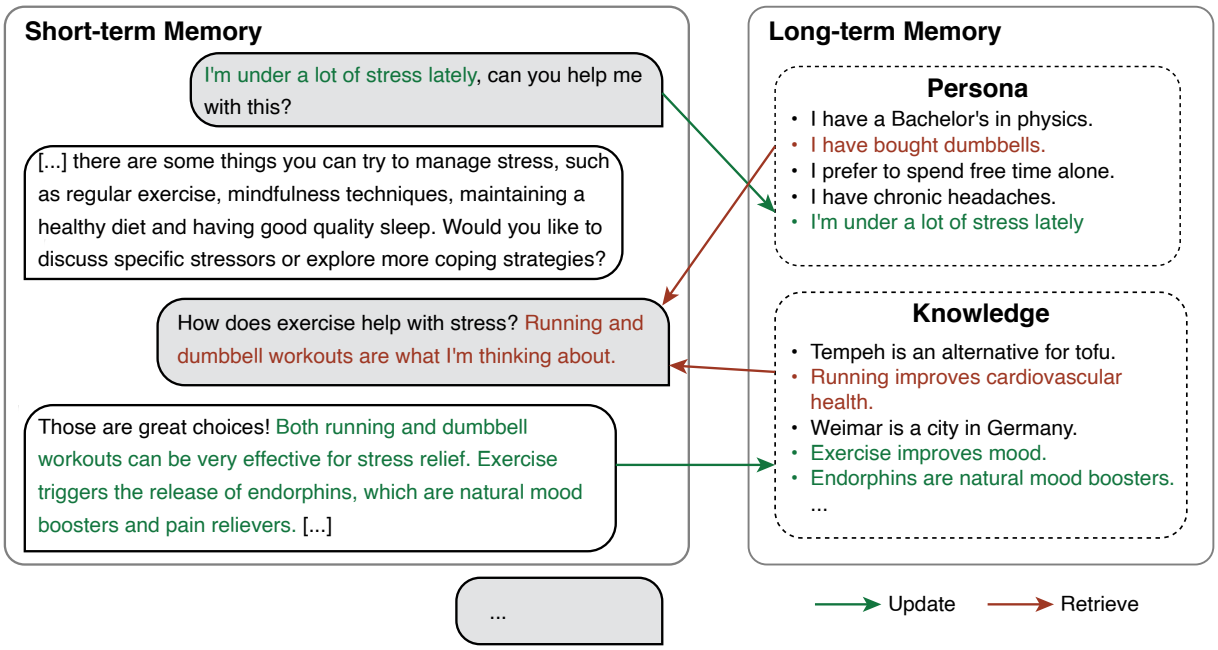

Figure 1: Example conversational search dialogue of a simulated user (gray bubbles) that has a short-term and long-term memory component for interacting with a conversational search system (white bubbles). The long-term memory is divided into components for Persona (top right) and Knowledge (bottom right), which are updated with new information from the dialogue (green arrows) and from which relevant facts are retrieved (red arrows) to inform the ongoing conversation.

Simulation track,[1] which emerged from this shared task, continues this line of research to formulate standard metrics and methods to assess user simulation. Interestingly, a byproduct of the collaborative shared-task efforts was the development of best practices for user simulators that can serve as reusable research resources [27].

Models of Memory. So far, memory has been modeled in LLMs mostly via the context window, but Zhong et al. [62] integrate a model of long-term memory into a conversational system using a storage of past interactions with hierarchical summaries and a forgetting mechanism employing the Ebbinghaus Forgetting Curve theory [23]. Ebbinghaus suggested modelling the time passed since information was learned,[2] and he empirically determined the relative time savings b (expressed as a percentage) of relearning information after passing time t as $b(t) = (100 \cdot k)/(\log(t)^c + k)$, with the constants c and k approximated to $c = 1.25$ and $k = 1.84$. If $b(t) = 100$ at some point, 100% of the time to relearn some knowledge is saved, as the knowledge is retained in the memory.

More recently, Woźniak et al. [57] proposed another model based on the optimum repetition intervals for learning, and they defined the retrievability r_s (i.e., the probability with which a memory trace can be found) as a function of time

$$r_s(t) = e^{-\frac{t}{s}}, \tag{1}$$

where s describes the retrievability decline by forgetting.

Although various other mathematical models of memory have been proposed over the last 100 years [45], the ones by Ebbinghaus and Woźniak et al. are among the most common due to their simplicity. We use Woźniak et al.'s r_s in our simulations as it provides a straightforward way to model users with more or less stable (i.e., "better" or "worse") memory via the parameter s.

Particularly in the field of conversational recommendation, user memory is often explicitly modeled in simulators. Wang et al. [55] attempt to mimic human memory as sensory, short-term, and long-term memory and define the transitions of knowledge between these stages. Knowledge that reaches long-term memory is forgotten according to a custom function that models the probability of retention. In contrast to this universal modeling, Cai et al. [16] and Chen et al. [17] have rather recommendation-specific approaches to modeling long-term memory. While Cai et al. model memory as past recommendations and reasons for recommendation and rejection, Chen et al. extract uttered preferences from past interactions as the users' memory.

Although not all user simulators explicitly describe a memory component, all simulators that interact with conversational systems have a form of short-term memory, represented as past interaction turns or current conversation context. While many implementations of user simulations rely on some form of memory model, to the best of our knowledge, we are the first to investigate whether these memory mechanisms are relevant for evaluating conversational search systems.

3 Problem Statement

Our goal is to investigate whether user simulators require a memory model for reliable evaluation of conversational search systems. To

[1] https://trec.usersim.ai/
[2] Ebbinghaus used sequences of letters in his experiments.

this end, we compare user simulators with and without dynamic memory mechanisms for (1) knowledge acquisition, (2) knowledge forgetting, and (3) persona development (i.e., changing preferences and interests). If these memory processes were essential for robust evaluation of conversational search using simulation, one would expect that system evaluations derived from memory-equipped simulations are closer to human-based evaluations than those from user simulators without memory. Such a result would validate the importance of simulating memory processes for evaluating conversational search systems, whereas an opposite result (i.e., evaluations with and without memory are similar) would suggest that user simulators may not need to model memory capabilities.

3.1 User Simulator Architecture

To investigate the impact of memory variants, we use a streamlined text-only version of the modular user simulation architecture of Balog [7] (overview in Figure 2): (1) Natural language understanding module (NLU) processes the dialogue content and triggers memory updates with new knowledge gained from system responses, as well as persona updates based on explicit user statements about preferences and interests. (2) Memory module stores the user's knowledge and persona characteristics. This includes knowledge learned from interactions and information on user preferences and interests. When triggered to update, the module also manages the potential forgetting of unused information. An integrated retrieval component allows for the fetching of information from memory that is relevant to the current dialogue. (3) Natural language generation module (NLG) creates user turns based on the current dialogue context and relevant knowledge retrieved from the memory module, which ensures that the simulated user interactions are consistent with the user's persona and knowledge state.

3.2 Simulation Objectives

To assess the robustness of user simulators for evaluating conversational search systems, we follow the methodology of Bernard and Balog [11] and assess a simulator's ability to reproduce human-based system rankings. More formally, let $M(S, U)$ be some metric for assessing the utility of a conversational system S measured on conversations of a real user population U. We then call a user simulator $\hat{U}$ *valid* iff for two arbitrary conversational systems, S_1 and S_2, the system ranking according to $M(\cdot, \hat{U})$ is consistent with the ranking obtained for $M(\cdot, U)$:

$$M(S_1, U) > M(S_2, U) \quad \Leftrightarrow \quad M(S_1, \hat{U}) > M(S_2, \hat{U}) \tag{2}$$

Moreover, a simulator $\hat{U}$ is *reproducible* iff the ranking of two arbitrary conversational systems, S_1 and S_2, according to $M(\cdot, \hat{U})$ is consistent across multiple simulation runs i, j:

$$M(S_1, \hat{U}^{(i)}) > M(S_2, \hat{U}^{(i)}) \quad \Leftrightarrow \quad M(S_1, \hat{U}^{(j)}) > M(S_2, \hat{U}^{(j)}) \tag{3}$$

Finally, a user simulator $\hat{U}_1$ *approximates* human evaluation *better* than a user simulator $\hat{U}_2$ (notation: $\hat{U}_1 \succ \hat{U}_2$) iff $M(S, U)$ is closer to $M(S, \hat{U}_1)$ than to $M(S, \hat{U}_2)$:

$$\hat{U}_1 \succ \hat{U}_2 \quad \Leftrightarrow \quad |M(S, U) - M(S, \hat{U}_1)| < |M(S, U) - M(S, \hat{U}_2)| \tag{4}$$

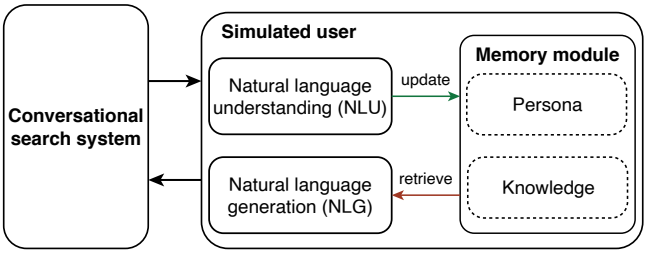

Figure 2: Our setup of a simulated user interacting with a conversational search system. The three user modules are responsible for processing the dialogue content (NLU), for updating and retrieving information (memory module), and for generating user turns (NLG) in the conversation.

4 Experimental Setup

In this section, we describe our experimental setup, including the dataset, search systems, user simulators, and evaluation metrics.

4.1 Dataset

For our experiments, we use the TREC-iKAT collection [3], as it features diverse topics, detailed persona descriptions, a comprehensive document collection, and an established conversational search system for comparison. The dataset contains 20 topics with 36 associated dialogues that have 35 unique first user utterances. We employ these first utterances as conversation starters for both the user simulators in our experiments and the human participants in our study. Each dialogue has a distinct set of persona descriptions (6–17 statements). The associated document collection is a subset of 116 million passages from ClueWeb22-B [41].

4.2 Conversational Search Systems

Due to the lack of publicly available conversational search systems, we implement a baseline system and we use the existing iKAT system that was developed for the TREC iKAT track [3]. Both systems follow a standard retrieve-and-generate architecture, which is commonly used in conversational search [40].

Retrieval Backend. Our baseline system uses standard BM25, while the iKAT system enhances BM25 by incorporating user persona information into the queries. Both backends are set to return the top-5 passages for each turn of a conversation.

Response Generation. From the retrieved passages, a Gemma-2-27B model [44] generates system responses. To simulate different levels of system helpfulness, we use prompting to create two variants of each system: helpful and unhelpful. The helpful variants are prompted to provide relevant and informative responses based on the retrieved passages, while the unhelpful variants are prompted to generate less relevant and less informative responses. This allows us to evaluate the user simulators' ability to differentiate between systems that provide varying levels of assistance to users. Thus, we include four conversational search systems in our experiments: Baseline helpful/unhelpful and iKAT helpful/unhelpful.

```
You are interacting with a conversational information access
system.
Here are some of the things you already know:
{retrieved knowledge}
You have the following personal attributes: {persona}
Continue the conversation using relevant information where
appropriate.

{conversation context}
```

Figure 3: Prompt template employed to simulate a user in a conversational search setting. Text with a red background is used in user models with memory module and is always adapted to the respective conversation. Text with a blue background are user persona statements relevant to the current conversation. Text with a purple background is the context of the conversation at the current turn.

4.3 User Simulators

As for the response generation, we use the open-source Gemma-2-27B model [44] also as the backbone of our user simulators, because the model better followed instructions to act as a user (and not an assistant)[3] compared to various Llama, Mistral, and Phi models that we tested in a pilot study. Figure 3 shows the prompt template, highlighting the parts that are adjusted when using the various memory models described below.

Given the capabilities of modern LLMs to generate and understand natural language, we focus on the memory modeling aspects of the user simulator. To isolate the effects of memory modeling on the evaluation of conversational search systems, we keep the natural language understanding (NLU) and natural language generation (NLG) modules unchanged across all user simulator variants. We only replace the memory module that manages the user's knowledge and persona information and consider the six user simulators given in Table 1 that differ in their employment of short-term and long-term memory, and of their ability to update the state of knowledge (i.e., learning and forgetting).

Persona. The persona component models the user's personal preferences and interests, which are represented as a set of natural language statements. To ensure fairness across user simulator variants, the initial set of persona statements was randomly selected from the TREC iKAT dataset and remained consistent across all user simulators throughout the experiments. We consider persona as a static component for the users with no learning capabilities (i.e., No-Memory and STM). For others, the new persona information is extracted from the *user's utterances* by prompting the Gemma 2-27B model. An example of the persona update process is shown in Figure 1 (the first utterance, in green).

Short-term Memory. The short-term memory (STM) component stores the current dialogue context. An example of the dialogue context is shown in Figure 1 (left). For the users with no STM, namely No-Memory, LTM, and $\text{LTM}_{\text{Forget}}$, only the most recent system response is maintained as context.

[3]The issue of LLMs assuming the role of an assistant rather than a user in conversational settings has been reported in prior work [53].

Table 1: Memory modeling capabilities of the tested simulators; STM/LTM: short-/long-term memory.

Simulator	Knowledge			Learn	Forget
	Persona	STM	LTM		
No-Memory	✓	✗	✗	✗	✗
STM	✓	✓	✗	✗	✗
LTM	✓	✗	✓	✓	✗
$\text{LTM}_{\text{Forget}}$	✓	✗	✓	✓	✓
STM + LTM	✓	✓	✓	✓	✗
STM + $\text{LTM}_{\text{Forget}}$	✓	✓	✓	✓	✓

Long-term Memory. The long-term memory (LTM) stores information gained from previous dialogues with the system. Users without LTM only have access to their persona information and the current dialogue. For users with long-term memory, new knowledge is extracted from the *system utterances* using the same approach as for persona updates. An example of this learning process is shown in Figure 1 (in green, last utterance). The user simulators with forgetting capabilities ($\text{LTM}_{\text{Forget}}$, STM+$\text{LTM}_{\text{Forget}}$) additionally employ a forgetting mechanism to avoid retrieving outdated information from the long-term memory. We implement a forgetting mechanism using Equation 1 (Section 2) to adjust retrieved information scores based on the number of turns since the information was learned as a proxy for time t and the stability level $s = 100$ as a proxy for the forgetting rate.

Memory Retrieval. For simulators with a memory module, relevant persona and knowledge information is retrieved from the memory storage at each turn using semantic textual similarity.[4] To limit the number of retrieved statements, we use a top-10 retrieval and apply various similarity thresholds $\theta \in \{0.0, 0.1, 0.2\}$ to model different levels of memory recall. The retrieved information is then used in the response generation process to ensure that the user's responses are consistent with their persona and knowledge state. The process of retrieval is highlighted in Figure 1 (in red).

In this study, we treat simulated users with different memory retrieval similarity thresholds but the same memory model as a single user simulator variant and treat them as different random seeds for the simulation runs. We leave the exploration of this aspect for future work.

4.4 Dialogue Simulation

For each combination of user simulator and conversational search system, we generated synthetic dialogues using the conversation starters from the TREC iKAT dataset (Section 4.1). Each dialogue began with a conversation starter, followed by alternating turns between the user simulator and the conversational search system. The user simulator generated messages based on the available context and, if applicable, retrieved information from its memory module. The conversational search system responded to the user's messages by retrieving relevant passages and generating responses as described in Section 4.2. Each dialogue continued for a fixed number of turns (5 turns) to ensure comparability across all experiments.

[4]https://huggingface.co/NovaSearch/stella_en_400M_v5

To account for the inherent randomness in LLM-based generation, we conducted three independent simulation runs for each user simulator and conversational search system combination. This resulted in a total of 72 sets of simulated dialogues (6 user simulators × 4 conversational search systems × 3 runs), each containing dialogues for all 35 conversation starters from the TREC iKAT dataset. Ergo, we obtained a total of 2 520 simulated dialogues.

4.5 Human Reference Dialogues

As a point of reference, we collected human dialogues using the four conversational search systems on the conversation starters from the TREC iKAT dataset. We developed a simple web interface that shows instructions, the current topic, and the ongoing conversation, where users interact with the system through a typical chat interface. Each participant was randomly assigned to one system and was instructed to interact with it in three to five turns to explore the assigned topic. The participants were encouraged to ask follow-up or clarifying questions or to provide feedback. To prevent participant fatigue and ensure natural interactions, users could skip topics they found uninteresting or outside their comfort zone. Participants saw the same conversation starters as the simulators, presented in the same order to maintain comparable learning conditions.

With this procedure, we collected 252 dialogues for 35 conversation starters from 21 participants, totaling 1 753 conversation turns (6.9 turns per dialogue). Participants were recruited via convenience sampling and were computer science students and researchers familiar with conversational search technologies.

4.6 Evaluation Metrics

We use automatic and manual evaluation methods on a per-dialogue level as the two measures of utility to assess how well the user simulators meet the objectives described in Section 3.2.

Automatic Evaluation. Without ground-truth answers available, we evaluate the conversational search systems using the EXAM++ answerability metric [25], which measures how well system responses cover relevant information. Following their methodology, we manually created 10 questions per conversation starter to assess the relevance and completeness of system responses. We employed an LLM-based grader to automatically evaluate the dialogues based on the generated questions. The grader assesses how well the dialogue addresses each question on a scale from 0 to 5, where 0 indicates no relevant information provided, and 5 indicates a comprehensive and accurate answer.

To account for LLM variability and inherent bias, we employed three different LLM graders from different families: Llama-3-70B-Instruct[5], Gemma-2-27B[6], and Phi-4[7]. Each dialogue was graded independently by each grader with greedy decoding, resulting in a total of over 80 thousand grades (2 772 dialogues × 10 questions × 3 graders). The final grade for each question was computed by taking the median of the three grades for each question.

For each dialogue, we compute the EXAM++ coverage score, which quantifies what proportion of exam questions $r \in R_c$ for conversation starter c are adequately addressed in dialogue d, using a minimum grade threshold τ:

$$\text{EXAM++}(d) = \frac{1}{|R_c|} |\{r | \text{grade}(d, r) \geq \tau, \forall r \in R_c\}| \tag{5}$$

In our experiments, we use a minimum grade threshold $\tau = 4$. This threshold was chosen based on preliminary experiments that showed it had the highest correlation with human-based evaluations across different runs of the simulators. The final score for each conversational search system is then computed by averaging the EXAM++ coverage scores across all dialogues D for a given system S and user population U:

$$\text{Coverage}(S, U) = \frac{1}{|D|} \sum_{d \in D} \text{EXAM++}(d) \tag{6}$$

Manual Evaluation. As an alternative to automatic evaluation, we also conducted a manual evaluation of the dialogues using Acute-Eval [35]. This method involves pairwise comparisons of a user's dialogues with two different systems for the same conversation starter. The annotators were presented with two dialogues side-by-side and were asked to choose which assistant, assistant A or assistant B, they would prefer to interact with on the given topic, with an option to provide explanations for their choice.

Five annotators, all being researchers or computer science students, participated in the evaluation of a subset of the human and simulator dialogues (490 dialogue pairs, 98 pairs per annotator). Following Li et al. [35], each specific dialogue was used only once. The results were aggregated using the Acute-Eval framework to compute win rates and overall rankings of the conversational search systems.

5 Results

We analyze the user simulators against the three formal objectives from Section 3.2: validity (i.e., consistency with system rankings from human interactions), reproducibility (i.e., stability across multiple runs), and approximation (i.e., closeness to human scores).

5.1 Validity: Consistency with Human Rankings

As discussed in Section 3, a key objective of user simulation for conversational search evaluation is the ability to yield the same evaluation outcomes when substituting human users with simulated ones. Valid simulators thus should replicate the system ranking of conversational search systems obtained from human interactions (Equation 2). We analyze how different memory variants in the simulator affect the rankings of the four conversational search systems using the metrics defined in Section 4.6.

Ranking by Answer Coverage (EXAM++). Table 2 presents the rankings of conversational search systems according to EXAM++ coverage scores, calculated across all dialogues from both human users and various simulated user configurations. Based on human dialogues, the ground truth system ranking is:

$$\text{Baseline}_{\text{helpful}} \succ \text{iKAT}_{\text{helpful}} \succ \text{Baseline}_{\text{unhelpful}} \succ \text{iKAT}_{\text{unhelpful}}$$

For each conversation, we calculate EXAM++ coverage scores for a simulator based on grades obtained across all runs. With this aggregation, all simulators reproduce the human ranking. This

[5] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[6] https://huggingface.co/google/gemma-2-27b
[7] https://huggingface.co/microsoft/phi-4

Table 2: Rankings of the conversational search systems based on EXAM++ coverage scores across different user simulators and runs. The Spearman's ρ and Kendall's τ correlation coefficients quantify the agreement between the coverage scores resulting from simulated and human dialogues. The 'all' rows are the aggregated results across all runs. The colors indicate the different conversational search systems: $\text{Baseline}_{\text{helpful}}$, $\text{iKAT}_{\text{helpful}}$, $\text{Baseline}_{\text{unhelpful}}$, $\text{iKAT}_{\text{unhelpful}}$.

User	Run	ρ	τ	Ranking (EXAM++ score)
Human	–	–	–	Base_{h} (0.2710) ≻ iKAT_{h} (0.2416) ≻ Base_{u} (0.0763) ≻ iKAT_{u} (0.0428)
No-Memory	1	0.659	0.527	Base_{h} (0.3118) ≻ iKAT_{h} (0.3086) ≻ Base_{u} (0.1118) ≻ iKAT_{u} (0.0771)
	2	0.526	0.411	Base_{h} (0.3118) ≻ iKAT_{h} (0.2765) ≻ iKAT_{u} (0.1824) ≻ Base_{u} (0.1588)
	3	0.581	0.450	Base_{h} (0.3412) ≻ iKAT_{h} (0.2676) ≻ Base_{u} (0.1969) ≻ iKAT_{u} (0.1471)
	all	0.742	0.573	Base_{h} (0.3217) ≻ iKAT_{h} (0.2845) ≻ Base_{u} (0.1650) ≻ iKAT_{u} (0.1372)
STM	1	0.634	0.503	Base_{h} (0.3114) ≻ iKAT_{h} (0.2794) ≻ Base_{u} (0.1000) ≻ iKAT_{u} (0.0914)
	2	0.621	0.494	Base_{h} (0.3500) ≻ iKAT_{h} (0.2882) ≻ Base_{u} (0.1735) ≻ iKAT_{u} (0.1429)
	3	0.544	0.421	iKAT_{h} (0.3294) ≻ Base_{h} (0.3156) ≻ Base_{u} (0.2118) ≻ iKAT_{u} (0.1853)
	all	0.702	0.542	Base_{h} (0.3273) ≻ iKAT_{h} (0.2968) ≻ Base_{u} (0.1588) ≻ iKAT_{u} (0.1437)
LTM	1	0.660	0.520	Base_{h} (0.3371) ≻ iKAT_{h} (0.3000) ≻ Base_{u} (0.1200) ≻ iKAT_{u} (0.0829)
	2	0.705	0.566	Base_{h} (0.3400) ≻ iKAT_{h} (0.3147) ≻ iKAT_{u} (0.0971) ≻ Base_{u} (0.0943)
	3	0.581	0.455	Base_{h} (0.3206) ≻ iKAT_{h} (0.2824) ≻ Base_{u} (0.1324) ≻ iKAT_{u} (0.0647)
	all	0.768	0.596	Base_{h} (0.3327) ≻ iKAT_{h} (0.2990) ≻ Base_{u} (0.1154) ≻ iKAT_{u} (0.0817)
$\text{LTM}_{\text{Forget}}$	1	0.620	0.490	Base_{h} (0.3265) ≻ iKAT_{h} (0.3257) ≻ Base_{u} (0.1429) ≻ iKAT_{u} (0.0882)
	2	0.619	0.494	Base_{h} (0.3059) ≻ iKAT_{h} (0.2765) ≻ Base_{u} (0.1118) ≻ iKAT_{u} (0.1029)
	3	0.696	0.567	Base_{h} (0.3118) ≻ iKAT_{h} (0.3000) ≻ Base_{u} (0.0971) ≻ iKAT_{u} (0.0853)
	all	0.759	0.597	Base_{h} (0.3147) ≻ iKAT_{h} (0.2990) ≻ Base_{u} (0.1175) ≻ iKAT_{u} (0.0922)
STM+LTM	1	0.606	0.492	Base_{h} (0.3257) ≻ iKAT_{h} (0.2314) ≻ Base_{u} (0.1176) ≻ iKAT_{u} (0.0971)
	2	0.578	0.460	Base_{h} (0.3471) ≻ iKAT_{h} (0.2971) ≻ Base_{u} (0.1412) ≻ iKAT_{u} (0.0824)
	3	0.623	0.498	Base_{h} (0.3324) ≻ iKAT_{h} (0.2265) ≻ Base_{u} (0.1147) ≻ iKAT_{u} (0.0971)
	all	0.729	0.570	Base_{h} (0.3350) ≻ iKAT_{h} (0.2519) ≻ Base_{u} (0.1227) ≻ iKAT_{u} (0.0922)
$\text{STM+LTM}_{\text{Forget}}$	1	0.708	0.572	iKAT_{h} (0.3143) ≻ Base_{h} (0.3086) ≻ Base_{u} (0.1114) ≻ iKAT_{u} (0.0853)
	2	0.584	0.480	Base_{h} (0.2657) ≻ iKAT_{h} (0.2265) ≻ Base_{u} (0.1088) ≻ iKAT_{u} (0.0824)
	3	0.585	0.458	Base_{h} (0.3000) ≻ iKAT_{h} (0.2824) ≻ Base_{u} (0.1294) ≻ iKAT_{u} (0.0853)
	all	0.757	0.591	Base_{h} (0.3913) ≻ iKAT_{h} (0.2748) ≻ Base_{u} (0.1165) ≻ iKAT_{u} (0.0843)

shows that aggregating across multiple runs can mitigate the variance in the rankings obtained from LLM-based simulators, even those without memory. However, correlation coefficient analysis of per-dialogue coverage scores demonstrates that simulators equipped with long-term memory (LTM) show higher and more consistent agreement with human rankings (Spearman's ρ between 0.759 and 0.768, Kendall's τ between 0.596 and 0.597) compared to short-term memory (STM) ($\rho = 0.702$, $\tau = 0.542$) or No-Memory simulators ($\rho = 0.742$, $\tau = 0.573$). However, the combination of long-term memory with short-term memory (STM + LTM) (without forgetting mechanisms) correlates less with human rankings ($\rho = 0.729$, $\tau = 0.570$) than the No-Memory simulator.

*Ranking by User Preference (*Acute-Eval*).* The results in Table 3 summarize the rankings of the conversational search systems based on Acute-Eval preferences derived from the dialogues. In contrast to the automatic EXAM++ evaluation, the human annotators preferred the personalized system ($\text{iKAT}_{\text{helpful}}$) over the

Table 3: Rankings of the conversational search systems based on Acute-Eval preferences derived from dialogues with different user groups. The rankings are based on the number of wins in pairwise comparisons per user group. The colors indicate the different systems: $\text{Baseline}_{\text{helpful}}$, $\text{iKAT}_{\text{helpful}}$, $\text{Baseline}_{\text{unhelpful}}$, $\text{iKAT}_{\text{unhelpful}}$.

User	Ranking (Acute-Eval wins)
Human	iKAT_{h} ≻ Base_{h} ≻ Base_{u} ≻ iKAT_{u}
No-Memory	Base_{h} ≻ iKAT_{h} ≻ iKAT_{u} ≻ Base_{u}
STM	Base_{h} ≈ iKAT_{h} ≻ iKAT_{u} ≈ Base_{u}
LTM	iKAT_{h} ≻ Base_{h} ≻ Base_{u} ≻ iKAT_{u}
$\text{LTM}_{\text{Forget}}$	iKAT_{h} ≻ Base_{h} ≻ Base_{u} ≻ iKAT_{u}
STM+LTM	Base_{h} ≻ iKAT_{h} ≻ Base_{u} ≻ iKAT_{u}
$\text{STM+LTM}_{\text{Forget}}$	iKAT_{h} ≻ Base_{h} ≻ iKAT_{u} ≻ Base_{u}

non-personalized one ($\text{Baseline}_{\text{helpful}}$), while still preferring both "helpful" systems over the "unhelpful" ones:

$$\text{iKAT}_{\text{helpful}} \succ \text{Baseline}_{\text{helpful}} \succ \text{Baseline}_{\text{unhelpful}} \succ \text{iKAT}_{\text{unhelpful}}$$

The No-Memory and STM simulators do not yield that ranking; interestingly, for the STM-simulated dialogues, the annotators showed no significant preference within the "helpful" and within the "unhelpful" system groups, resulting in ties. In contrast, the results obtained from LTM-only simulators (both with and without forgetting) correctly reproduce the human preferences for the personalized system ($\text{iKAT}_{\text{helpful}} \succ \text{Baseline}_{\text{helpful}}$), while the combined short-term and long-term memory variant (STM+LTM) reverses the order of the two "helpful" systems.

RQ1. *How well do relative rankings of conversational search systems obtained from memory-equipped user simulations reflect the rankings obtained from real user studies?* Our results indicate that all simulators reproduce the system ranking based on EXAM++ coverage scores, but for the human preference rankings from ACUTE-EVAL, only those with long-term memory are able to do so. Interestingly, including the current conversation context (STM) does not consistently improve the validity of the simulator and can even lead to less agreement with human per-dialogue scores compared to using long-term memory alone (LTM) or not using memory at all (No-Memory).

5.2 Reproducibility: Robustness across Runs

A reliable evaluation requires that simulators produce consistent rankings across different stochastic runs (Equation 3). As shown in Table 2, most of the simulators allow for successful distinction between "helpful" and "unhelpful" systems across individual runs, however the specific rankings can vary. In particular, the unhelpful assistants are ranked differently across runs for No-Memory and LTM simulators. Similarly, helpful assistants are ranked differently across runs for STM and $\text{STM+LTM}_{\text{Forget}}$ simulators.

Here, we do not observe a clear pattern of variability across runs for different memory configurations. The highest run-to-run variance in correlation with human rankings is achieved by STM + $\text{LTM}_{\text{Forget}}$ with correlation coefficients ranging from $\rho = 0.584$ to $\rho = 0.708$. In second place is the No-Memory simulator with generally lower and highly varying correlation coefficients ($0.526 \leq \rho \leq 659$). The most consistent correlation is demonstrated by STM+LTM, yielding coefficients between $\rho = 0.578$ and $\rho = 0.623$.

RQ2. *Does a memory component improve the reproducibility of relative system rankings across multiple simulation runs?* While our findings do not reveal a systematic pattern of which memory model is generally the best to improve reproducibility of system rankings, having no memory at all results in one of the highest run-to-run variances of correlation with human rankings. Therefore, modeling memory in any way, be it short-term or long-term memory, is recommended to keep these variances under control.

5.3 Approximation of Human-based Scores

A third objective of good simulators is to approximate the absolute effectiveness scores derived from human interactions (Equation 4).

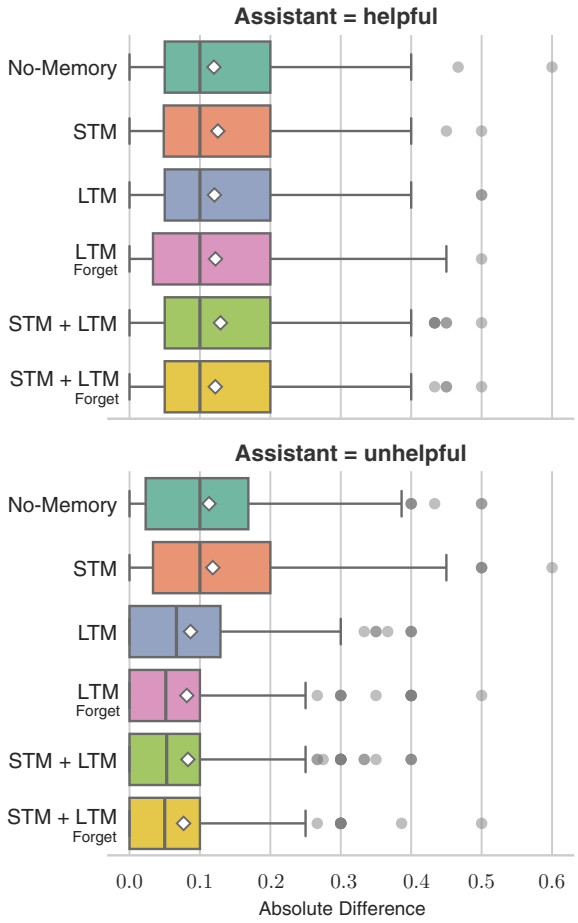

Figure 4: Mean absolute differences in per-dialogue EXAM++ coverage scores across dialogues from human interactions compared to those from simulated interactions, grouped by memory model and assistant type (helpful vs. unhelpful). Diamond markers indicate the distributions' means.

EXAM++ Coverage and Grades. The plots in Figure 4 show the mean absolute differences in per-dialogue EXAM++ coverage scores (Equation 5) between evaluations based on human dialogues and those based on simulated dialogues across all runs grouped by assistant type (helpful and unhelpful). For "helpful" assistants, we see that the distribution of absolute differences is relatively similar across all memory models, with all simulators showing a mean absolute difference of around 0.1. While there are variations in the spread of the data, adding complex memory structures does not drastically improve the simulator's ability to mimic human scoring compared to having no memory at all. Interestingly, in the "unhelpful" assistant group, there is a clearer distinction between the effectiveness of different memory models. Here, models with long-term memory (LTM, $\text{LTM}_{\text{Forget}}$, STM+LTM, $\text{STM+LTM}_{\text{Forget}}$) generally outperform those without it (No-Memory, STM).

More details on the distributions of EXAM++ coverage scores and grades (raw values obtained from question-answering evaluations)

Table 4: Mean and standard deviations of per-conversation EXAM++ Coverage ($\tau = 4$) and EXAM++ Grades across all dialogues of the simulators. The p-values come from a Wilcoxon signed-rank test with Holm-Bonferroni correction (paired per-dialogue for coverage scores) that compares a simulator's scores with that of the human users.

	EXAM++ Coverage			EXAM++ Grades		
User	Mean	Std	p-value	Mean	Std	p-value
Baseline helpful						
Human	0.27	0.13	–	2.19	1.73	–
No-Memory	0.32	0.15	1.00	2.35	1.82	0.81
STM	0.33	0.17	1.00	2.30	1.89	1.00
LTM	0.33	0.16	1.00	2.35	1.83	0.73
$\text{LTM}_{\text{Forget}}$	0.31	0.13	1.00	2.33	1.84	1.00
STM+LTM	0.34	0.15	1.00	2.32	1.90	1.00
$\text{STM+LTM}_{\text{Forget}}$	0.29	0.15	1.00	2.17	1.89	1.00
iKAT helpful						
Human	0.25	0.12	–	2.06	1.72	–
No-Memory	0.28	0.15	0.90	2.15	1.82	1.00
STM	0.30	0.16	0.46	2.15	1.89	1.00
LTM	0.30	0.15	0.40	2.27	1.82	0.16
$\text{LTM}_{\text{Forget}}$	0.30	0.17	0.46	2.23	1.83	0.44
STM+LTM	0.25	0.17	1.00	1.97	1.87	1.00
$\text{STM+LTM}_{\text{Forget}}$	0.27	0.16	1.00	2.08	1.85	1.00
Baseline unhelpful						
Human	0.08	0.11	–	1.10	1.41	–
No-Memory	0.15	0.15	**<0.01**	1.61	1.63	**<0.01**
STM	0.16	0.15	**<0.01**	1.51	1.66	**<0.01**
LTM	0.12	0.13	0.46	1.38	1.54	**<0.02**
$\text{LTM}_{\text{Forget}}$	0.12	0.14	0.90	1.42	1.56	**<0.01**
STM+LTM	0.13	0.14	0.22	1.32	1.58	0.09
$\text{STM+LTM}_{\text{Forget}}$	0.12	0.13	0.46	1.34	1.58	0.20
iKAT unhelpful						
Human	0.03	0.06	–	0.91	1.31	–
No-Memory	0.13	0.14	**<0.01**	1.43	1.58	**<0.01**
STM	0.14	0.15	**<0.01**	1.39	1.60	**<0.01**
LTM	0.08	0.11	**<0.02**	1.11	1.45	0.21
$\text{LTM}_{\text{Forget}}$	0.09	0.13	**<0.03**	1.19	1.47	**<0.01**
STM+LTM	0.09	0.12	**<0.02**	1.15	1.47	0.08
$\text{STM+LTM}_{\text{Forget}}$	0.08	0.12	**<0.04**	1.19	1.45	**<0.01**

can be found in Table 4. We can observe that the standard deviations across all pairs of users and conversational search systems are quite similar. To test if the mean of the distributions of EXAM++ scores obtained by simulated users is significantly different from the mean of scores obtained from human interactions, we compute statistical significance tests as described in Section 3. To select an appropriate test, we first test for normality with the Shapiro-Wilk test, which indicates that most of the distributions are not normally distributed ($p < 0.05$). Thus, we employ a nonparametric paired Wilcoxon signed-rank test with the Holm-Bonferroni correction for multiple comparisons to compare the means of the EXAM++ scores and grades between human and simulated users.

Table 5: ACUTE-EVAL results of pairwise system preferences based on dialogues with (a) humans, (b) LTM, and (c) $\text{LTM}_{\text{Forget}}$ (these simulators reproduced the human ranking). Each cell shows the percentage of wins for the system in the row compared to the system in the column. Bold win percentages indicate significant differences ($p < 0.05$).

(a) Human-based Results

Losses % matches

Wins % matches	Base_h	Base_u	iKAT_h	iKAT_u
Base_h		**100**	44	**100**
Base_u	**0**		**0**	62
iKAT_h	56	**100**		**100**
iKAT_u	**0**	38	**0**	

(b) LTM-based Results

Losses % matches

Wins % matches	Base_h	Base_u	iKAT_h	iKAT_u
Base_h		**100**	42	**100**
Base_u	**0**		**0**	57
iKAT_h	57	**100**		**89**
iKAT_u	**0**	43	**11**	

(c) $\text{LTM}_{\text{Forget}}$-based Results

Losses % matches

Wins % matches	Base_h	Base_u	iKAT_h	iKAT_u
Base_h		**100**	33	**100**
Base_u	**0**		**9**	75
iKAT_h	67	**91**		**100**
iKAT_u	**0**	25	**0**	

The p-values show that using simulated users instead of humans doesn't significantly affect EXAM++ coverage in the case of "helpful" assistants ($p > 0.05$). However, for the "unhelpful" assistant, using simulated users inflated the coverage scores and grades, resulting in significantly different score distributions compared to evaluations of human interactions. In particular, the scores for the $\text{Baseline}_{\text{unhelpful}}$ demonstrate that the score inflation is worse for the simulators with no memory or short-term memory only. Our observation of score inflation caused by user simulation aligns with observations by Balog and Zhai [9, p. 154], who identified a potential "superuser" phenomenon when using large language models for user simulation. These models often possess broader knowledge than typical human users, which may explain the systematically higher evaluation scores of simulated interactions. This suggests that while the relative rankings are informative, the absolute scores from simulated users should be interpreted with caution.

ACUTE-EVAL *Win Rates.* Table 5 presents the pairwise win rates from the ACUTE-EVAL manual evaluation comparing conversational search systems based on dialogues from both human and simulated users. In human dialogues, the personalized system ($\text{iKAT}_{\text{helpful}}$) achieved 56% wins against the non-personalized $\text{Baseline}_{\text{helpful}}$ system. Out of the simulated user configurations that correctly reproduced the human ranking (cf. Table 3), simulated dialogues with the LTM user achieved a win rate of 57%, and its forgetting variant resulted in 67%. As for "unhelpful", human-based dialogues

showed a win rate of 62% for $\text{Baseline}_{\text{unhelpful}}$ over $\text{iKAT}_{\text{unhelpful}}$. Simulated dialogues with LTM user achieved a win rate of 57%, and its forgetting variant resulted in 75%.

RQ3. *How well can absolute effectiveness scores from real user studies be approximated by memory-equipped simulations?* Our findings show that the use of long-term memory (LTM) reduces the absolute difference in effectiveness scores compared to non-memory models. While all simulators tend to score higher than humans due to the "superuser" effect, LTM models consistently provide the closest approximation to human scores and avoid the significant divergence of No-Memory on unhelpful systems.

6 Conclusion

In this paper, we have analyzed the impact of various memory models on the evaluation of conversational search systems with simulated users. Our results show that even simulators without a memory component can be used to distinguish helpful conversational search systems from unhelpful ones. However, the memoryless simulators fail to capture nuanced effectiveness differences as observed by humans, especially preferences for personalized interactions. With respect to capturing such differences, our experiments indicate that simulators equipped with long-term memory can successfully reproduce human-derived system rankings and preferences for personalization. However, further adding a short-term memory component does not necessarily yield more reliable results. We hypothesize that this is due to the complexity of managing and integrating information from both memory types, which may introduce inconsistencies in the simulated behavior.

Even though our findings indicate the potential of memory modeling in user simulators for evaluating conversational search systems, our study still has limitations. To isolate the effect of memory components, we kept other aspects of the user model constant, which may limit the generalizability of our results. Additionally, our evaluation focused on a specific set of conversational search systems and user tasks, which may not cover the full spectrum of real-world scenarios. In future work, we thus plan to explore more diverse system architectures and user goals to validate the robustness of our findings. Furthermore, we also plan to dig deeper into individual aspects of memory, such as forgetting rates and personas, and to explore new learning and forgetting mechanisms. Although we touched on these aspects in this paper, further research is needed to examine the impact on the reliability of evaluation results.

Finally, we believe that future research on user simulation for conversational search evaluation, but also for general retrieval evaluation, should shift from just creating more complex user models to focusing more on validation and reproducibility. It is essential to set strict standards for ensuring that simulators yield consistent evaluation results across runs and that they accurately reflect human judgments and preferences before we can actually rely on them to evaluate conversational or other search systems.

Acknowledgments

We would like to thank Sahand Majdabadi Farahani for his support of our experiments and the data collection, and we thank the anonymous reviewers for their helpful feedback.

We also acknowledge the peoples of the Woi Wurrung and Boon Wurrung language groups of the eastern Kulin Nation on whose unceded lands ACM SIGIR 2026 was hosted. We pay our respects to their Elders past and present, and extend that respect to all Aboriginal and Torres Strait Islander peoples today and their continuing connection to land, sea, sky, and community.

References

[1] Jafar Afzali, Aleksander Mark Drzewiecki, Krisztian Balog, and Shuo Zhang. 2023. UserSimCRS: A User Simulation Toolkit for Evaluating Conversational Recommender Systems. In *Proceedings of WSDM 2023*. 1160–1163. doi:10.1145/3539597.3573029

[2] Eugene Agichtein, Michael Johnston, Anna Gottardi, Lavina Vaz, Cris Flagg, Yao Lu, Shaohua Liu, Sattvik Sahai, Giuseppe Castellucci, Jason Ingyu Choi, Prasoon Goyal, Di Jin, Saar Kuzi, Nikhita Vedula, Lucy Hu, Samyuth Sagi, Luke Dai, Hangjie Shi, Zhejia Yang, Desheng Zhang, Chao Zhang, Daniel Pressel, Heather Rocker, Leslie Ball, Osman Ipek, Kate Bland, James Jeun, Oleg Rokhlenko, Akshaya Iyengar, Arindam Mandal, Yoelle Maarek, and Reza Ghanadan. 2023. Advancing Conversational Task Assistance: The Second Alexa Prize TaskBot Challenge. https://www.amazon.science/alexa-prize/proceedings/alexa-lets-work-together-introducing-the-second-alexa-prize-taskbot-challenge

[3] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffery Dalton, and Leif Azzopardi. 2024. TREC iKAT 2023: A Test Collection for Evaluating Conversational and Interactive Knowledge Assistants. In *Proceedings of SIGIR 2024*. 819–829. doi:10.1145/3626772.3657860

[4] Mohammad Aliannejadi, Simon Lupart, Marcel Gohsen, Zahra Abbasiantaeb, Nailia Mirzakhmedova, Johannes Kiesel, and Jeffrey Dalton. 2025. TREC iKAT 2025: The Interactive Knowledge Assistance Track Overview. In *Proceedings of TREC 2025*. https://pages.nist.gov/trec-browser/trec34/ikat/proceedings/#trec-ikat-2025-the-interactive-knowledge-assistance-track-overview

[5] James Allen, Curry Guinn, and Eric Horvitz. 1999. Mixed-Initiative Interaction. *IEEE Intelligent Systems and Their Applications* 14, 5 (1999), 14–23. doi:10.1109/5254.796083

[6] 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 SIGIR-AP 2024*. 197–202. doi:10.1145/3673791.3698427

[7] Krisztian Balog. 2021. Conversational AI from an Information Retrieval Perspective: Remaining Challenges and a Case for User Simulation. In *Proceedings of DESIRES 2021*. 80–90. https://ceur-ws.org/Vol-2950/paper-03.pdf

[8] Krisztian Balog, David Maxwell, Paul Thomas, and Shuo Zhang. 2021. Report on the 1st Simulation for Information Retrieval Workshop (Sim4IR 2021) at SIGIR 2021. *SIGIR Forum* 55, 2 (2021), 16 pages. doi:10.1145/3527546.3527559

[9] 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. doi:10.1561/1500000098

[10] Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. 2013. Modeling Behavioral Factors in Interactive Information Retrieval. In *Proceedings of CIKM 2013*. 2297–2302. doi:10.1145/2505515.2505660

[11] Nolwenn Bernard and Krisztian Balog. 2024. Towards a Formal Characterization of User Simulation Objectives in Conversational Information Access. In *Proceedings of ICTIR 2024*. 185–193. doi:10.1145/3664190.3672529

[12] Adam Bignold, Francisco Cruz, Richard Dazeley, Peter Vamplew, and Cameron Foale. 2021. An Evaluation Methodology for Interactive Reinforcement Learning with Simulated Users. *Biomimetics* 6, 1 (2021), 13:1–13:16. doi:10.3390/biomimetics6010013

[13] Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, and Claudia Hauff. 2019. SIREN: A Simulation Framework for Understanding the Effects of Recommender Systems in Online News Environments. In *Proceedings of FAT* 2019*. 150–159. doi:10.1145/3287560.3287583

[14] Timo Breuer, Norbert Fuhr, and Philipp Schaer. 2022. Validating Simulations of User Query Variants. In *Proceedings of ECIR 2022*. 80–94. doi:10.1007/978-3-030-99736-6_6

[15] Timo Breuer, Christin Katharina Kreutz, Norbert Fuhr, Krisztian Balog, Philipp Schaer, Nolwenn Bernard, Ingo Frommholz, Marcel Gohsen, Kaixin Ji, Gareth J. F. Jones, Jüri Keller, Jiqun Liu, Martin Mladenov, Gabriella Pasi, Johanne R. Trippas, Xi Wang, Saber Zerhoudi, and ChengXiang Zhai. 2024. Report on the 1st Workshop on Simulations for Information Access (Sim4IA 2024) at SIGIR 2024. *SIGIR Forum* 58, 2 (2024), 14 pages. doi:10.1145/3722449.3722460

[16] Shihao Cai, Jizhi Zhang, Keqin Bao, Chongming Gao, Qifan Wang, Fuli Feng, and Xiangnan He. 2025. Agentic Feedback Loop Modeling Improves Recommendation and User Simulation. In *Proceedings of SIGIR 2025*. 2235–2244. doi:10.1145/3726302.3729893

[17] Luyu Chen, Quanyu Dai, Zeyu Zhang, Xueyang Feng, Mingyu Zhang, Pengcheng Tang, Xu Chen, Yue Zhu, and Zhenhua Dong. 2025. RecUserSim: A Realistic and Diverse User Simulator for Evaluating Conversational Recommender Systems. In *Proceedings of WWW 2025 (Companion)*. 133–142. doi:10.1145/3701716.3715258

[18] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. *Click Models for Web Search*. Springer. doi:10.1007/978-3-031-02294-4

[19] J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. 2018. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum* 52, 1 (2018), 34–90. doi:10.1145/3274784.3274788

[20] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. TREC CAsT 2019: The Conversational Assistance Track Overview. In *Proceedings of TREC 2019*. https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.CAsT.pdf

[21] 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. doi:10.48550/arXiv.2309.13233

[22] Kaustubh D. Dhole. 2024. KAUCUS: Knowledge Augmented User Simulators for Training Language Model Assistants. arXiv:2401.16454. doi:10.48550/arXiv.2401.16454

[23] Hermann Ebbinghaus. 1885. *Über Das Gedächtnis: Untersuchungen Zur Experimentellen Psychologie*. Duncker and Humblot, Berlin, Germany.

[24] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of ICTIR 2023*. 39–50. doi:10.1145/3578337.3605136

[25] Naghmeh Farzi and Laura Dietz. 2024. Pencils Down! Automatic Rubric-based Evaluation of Retrieve/Generate Systems. In *Proceedings of ICTIR 2024*. 175–184. doi:10.1145/3664190.3672511

[26] Alexander Frummet, Andrea Papenmeier, Maik Fröbe, Johannes Kiesel, Vaibhav Adlakha, Norbert Braunschweiler, Mateusz Dubiel, Satanu Ghosh, Marcel Gohsen, Christin Kreutz, Milad Momeni, Markus Nilles, Sachin Pathiyan Cherumanal, Abbas Pirmoradi, Paul Thomas, Johanne R. Trippas, Ines Zelch, and Oleg Zendel. 2024. Report on the 8th Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI 2024) at CHIIR 2024. *SIGIR Forum* 58, 1 (2024), 12 pages. doi:10.1145/3687273.3687282

[27] Marcel Gohsen, Zahra Abbasiantaeb, Mohammad Aliannejadi, Krisztian Balog, Timo Breuer, Jeffrey Dalton, Maik Fröbe, Christin Kreutz, Andreas Kruff, Simon Lupart, Nailia Mirzakhmedova, Harrisen Scells, Philipp Schaer, Benno Stein, and Johannes Kiesel. 2025. User Simulation in Practice: Lessons Learned from Three Shared Tasks. *SIGIR Forum* 59, 2 (2025), 18 pages. doi:10.1145/3799914.3799917

[28] Justyna Gromada, Alicja Kasicka, Ewa Komkowska, Lukasz Krajewski, Natalia Krawczyk, Morgan Veyret, Bartosz Przybył, Lina M. Rojas-Barahona, and Michał K. Szczerbak. 2025. Evaluating Conversational Agents with Persona-driven User Simulations based on Large Language Models: A Sales Bot Case Study. In *Proceedings of EMNLP 2025 (Industry Track)*. 230–245. doi:10.18653/v1/2025.emnlp-industry.16

[29] Matthias Hagen, Maximilian Michel, and Benno Stein. 2016. Simulating Ideal and Average Users. In *Proceedings of AIRS 2016*. 138–154. doi:10.1007/978-3-319-48051-0_11

[30] Rolf Jagerman, Krisztian Balog, Philipp Schaer, Johann Schaible, Narges Tavakolpoursaleh, and Maarten de Rijke. 2017. Overview of TREC OpenSearch 2017. In *Proceedings of TREC 2017*. https://trec.nist.gov/pubs/trec26/papers/Overview-O.pdf

[31] Johannes Kiesel, Çagri Çöltekin, Marcel Gohsen, Sebastian Heineking, Maximilian Heinrich, Maik Fröbe, Tim Hagen, Mohammad Aliannejadi, Sharat Anand, Tomaz Erjavec, Matthias Hagen, Matyás Kopp, Nikola Ljubesic, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevicius, Harrisen Scells, Moritz Wolter, Ines Zelch, Martin Potthast, and Benno Stein. 2025. Overview of Touché 2025: Argumentation Systems. In *Working Notes of CLEF 2025*. 4536–4562. https://ceur-ws.org/Vol-4038/paper_378.pdf

[32] Johannes Kiesel, Marcel Gohsen, Nailia Mirzakhmedova, Matthias Hagen, and Benno Stein. 2024. Simulating Follow-up Questions in Conversational Search. In *Proceedings of ECIR 2024*. 382–398. doi:10.1007/978-3-031-56060-6_25

[33] 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 *Proceedings of CLEF 2024*. 166–171. doi:10.1007/978-3-031-71736-9_11

[34] Johannes Kiesel, Lars Meyer, Martin Potthast, and Benno Stein. 2021. Meta-Information in Conversational Search. *ACM Transactions on Information Systems* 39, 4 (2021), 50:1–50:44. doi:10.1145/3468868

[35] Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons. arXiv:1909.03087. doi:10.48550/arXiv.1909.03087

[36] Hsien-Chin Lin, Shutong Feng, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gasić. 2023. EmoUS: Simulating User Emotions in Task-Oriented Dialogues. In *Proceedings of SIGIR 2023*. 2526–2531. doi:10.1145/3539618.3592092

[37] Yajiao Liu, Xin Jiang, Yichun Yin, Yasheng Wang, Fei Mi, Qun Liu, Xiang Wan, and Benyou Wang. 2023. One Cannot Stand for Everyone! Leveraging Multiple User Simulators to Train Task-oriented Dialogue Systems. In *Proceedings of ACL 2023*. 1–21. doi:10.18653/v1/2023.acl-long.1

[38] David Maxwell and Leif Azzopardi. 2016. Simulating Interactive Information Retrieval: SimIIR: A Framework for the Simulation of Interaction. In *Proceedings of SIGIR 2016*. 1141–1144. doi:10.1145/2911451.2911469

[39] David Maxwell and Leif Azzopardi. 2018. Information Scent, Searching and Stopping: Modelling SERP Level Stopping Behaviour. In *Proceedings of ECIR 2018*. 210–222. doi:10.1007/978-3-319-76941-7_16

[40] Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2025. A Survey of Conversational Search. *ACM Transactions on Information Systems* 43, 6 (2025), 167:1–167:50. doi:10.1145/3759453

[41] Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. ClueWeb22: 10 Billion Web Documents with Rich Information. In *Proceedings of SIGIR 2022*. 3360–3362. doi:10.1145/3477495.3536321

[42] Paul Owoicho, Ivan Sekulić, Mohammad Aliannejadi, Jeffrey Dalton, and Fabio Crestani. 2023. Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond. In *Proceedings of SIGIR 2023*. 632–642. doi:10.1145/3539618.3591683

[43] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of CHIIR 2017*. 117–126. doi:10.1145/3020165.3020183

[44] Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Gï¿½rner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118. doi:10.48550/arXiv.2408.00118

[45] David C. Rubin and Amy E. Wenzel. 1996. One Hundred Years of Forgetting: A Quantitative Description of Retention. *Psychological Review* 103 (1996), 734–760. doi:10.1037/0033-295X.103.4.734

[46] David P. Sander and Laura Dietz. 2021. EXAM: How to Evaluate Retrieve-and-Generate Systems for Users Who Do Not (Yet) Know What They Want. In *Proceedings of DESIRES 2021*. 136–146. https://ceur-ws.org/Vol-2950/paper-16.pdf

[47] Mark Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval* 4, 4 (2010), 247–375. doi:10.1561/1500000009

[48] Philipp Schaer, Christin Katharina Kreutz, Krisztian Balog, Timo Breuer, Andreas Kruff, Mohammad Aliannejadi, Christine Bauer, Nolwenn Bernard, Nicola Ferro, Marcel Gohsen, Nurul Lubis, and Saber Zerhoudi. 2025. Report on the Second Workshop on Simulations for Information Access (Sim4IA 2025) at SIGIR 2025. *SIGIR Forum* 59, 2 (2025), 15 pages. doi:10.1145/3799914.3799927

[49] Anne Schuth, Krisztian Balog, and Liadh Kelly. 2015. Extended Overview of the Living Labs for Information Retrieval Evaluation (LL4IR) CLEF Lab 2015. In *Working Notes of CLEF 2015*. https://ceur-ws.org/Vol-1391/inv-pap8-CR.pdf

[50] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating Mixed-initiative Conversational Search Systems via User Simulation. In *Proceedings of WSDM 2022*. 888–896. doi:10.1145/3488560.3498440

[51] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2024. Analysing Utterances in LLM-Based User Simulation for Conversational Search. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (2024), 62:1–62:22. doi:10.1145/3650041

[52] 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 (2023), 17:1–17:29. doi:10.1145/3596510

[53] Silvia Terragni, Modestas Filipavicius, Nghia Khau, Bruna Guedes, André Ferreira Manso, and Roland Mathis. 2023. In-Context Learning User Simulators for Task-Oriented Dialog Systems. arXiv:2306.00774. doi:10.48550/arXiv.2306.00774

[54] Suzan Verberne, Maya Sappelli, Kalervo Järvelin, and Wessel Kraaij. 2015. User Simulations for Interactive Search: Evaluating Personalized Query Suggestion. In *Proceedings of ECIR 2015*. 678–690. doi:10.1007/978-3-319-16354-3_75

[55] Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. 2025. Know You First and Be You Better: Modeling Human-Like User Simulators via Implicit Profiles. In *Proceedings of ACL 2025*. 21082–21107. https://aclanthology.org/2025.acl-long.1025/

[56] Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan 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), 55:1–55:37. doi:10.1145/3708985

[57] Piotr A. Woźniak, Edward J. Gorzelańczyk, and Janusz A. Murakowski. 1995. Two Components of Long-Term Memory. *Acta Neurobiologiae Experimentalis* 55, 4 (1995), 301–305. doi:10.55782/ane-1995-1090

[58] 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 NAACL 2024*. 1490–1504. doi:10.18653/v1/2024.naacl-long.83

[59] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2023. Conversational Information Seeking. *Foundations and Trends in Information Retrieval* 17, 3-4 (2023), 244–456. doi:10.1561/1500000081

[60] 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 CIKM 2022*. 4661–4666. doi:10.1145/3511808.3557711

[61] Shuo Zhang and Krisztian Balog. 2020. Evaluating Conversational Recommender Systems via User Simulation. In *Proceedings of KDD 2020*. 1512–1520. doi:10.1145/3394486.3403202

[62] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. MemoryBank: Enhancing Large Language Models with Long-Term Memory. In *Proceedings of AAAI 2024*. 19724–19731. doi:10.1609/aaai.v38i17.29946

[63] Junkai Zhou, Liang Pang, Ya Jing, Jia Gu, Huawei Shen, and Xueqi Cheng. 2024. Knowledge Boundary and Persona Dynamic Shape A Better Social Media Agent. arXiv:2403.19275. doi:10.48550/arXiv.2403.19275

[64] Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. A LLM-based Controllable, Scalable, Human-Involved User Simulator Framework for Conversational Recommender Systems. arXiv:2405.08035. doi:10.48550/arXiv.2405.08035