Analysing Utterances in LLM-Based User Simulation for Conversational Search

IVAN SEKULIĆ, Universitá della Svizzera italiana, Lugano, Switzerland
MOHAMMAD ALIANNEJADI, University of Amsterdam, Amsterdam, The Netherlands
FABIO CRESTANI, Universitá della Svizzera italiana, Lugano, Switzerland

Clarifying underlying user information needs by asking clarifying questions is an important feature of modern conversational search systems. However, evaluation of such systems through answering prompted clarifying questions requires significant human effort, which can be time-consuming and expensive. In our recent work, we proposed an approach to tackle these issues with a user simulator, *USi*. Given a description of an information need, *USi* is capable of automatically answering clarifying questions about the topic throughout the search session. However, while the answers generated by *USi* are both in line with the underlying information need and in natural language, a deeper understanding of such utterances is lacking. Thus, in this work, we explore utterance formulation of large language model (LLM)–based user simulators. To this end, we first analyze the differences between *USi*, based on GPT-2, and the next generation of generative LLMs, such as GPT-3. Then, to gain a deeper understanding of LLM-based utterance generation, we compare the generated answers to the recently proposed set of patterns of human-based query reformulations. Finally, we discuss potential applications as well as limitations of LLM-based user simulators and outline promising directions for future work on the topic.

CCS Concepts: • **Information systems** → *Evaluation of retrieval results*; **Information retrieval;**

Additional Key Words and Phrases: User simulation, conversational search, mixed-initiative

ACM Reference Format:
Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2024. Analysing Utterances in LLM-Based User Simulation for Conversational Search. *ACM Trans. Intell. Syst. Technol.* 15, 3, Article 62 (May 2024), 22 pages.
https://doi.org/10.1145/3650041

1 INTRODUCTION

Conversational information retrieval, also known as *conversational search*, refers to the process of retrieving relevant information in response to a natural language conversation or query. The primary goal of a conversational search system is to satisfy the user's information need by retrieving relevant information from a given collection. To successfully do so, the system needs to have a clear understanding of the underlying user need. Since users' queries are often under-specified and vague, a mixed-initiative paradigm of conversational search allows the system to take the initiative in the conversation and ask the user clarifying questions or issue other requests. Clarifying the

Authors' addresses: I. Sekulić and F. Crestani, Universita della Svizzera italiana, Lugano, Switzerland; e-mails: ivan.sekulic@usi.ch, fabio.crestani@usi.ch; M. Aliannejadi, University of Amsterdam, Amsterdam, The Netherlands; e-mail: m.aliannejadi@uva.nl.

ACM 2157-6904/2024/05-ART62
https://doi.org/10.1145/3650041

user information need has benefited both the user and the conversational search system [4, 32, 77], providing a solid motivation for such mixed-initiative systems.

However, evaluating the described mixed-initiative conversational search systems takes considerable work [47]. The challenge arises from expensive and time-consuming user studies required for holistic evaluation of conversational systems [20]. Such studies require real users to interact with the search system for several conversational turns and provide answers to potential clarifying questions prompted by the system. A relatively simple solution is to conduct offline corpus-based evaluation [4]. However, this limits the system to selecting clarifying questions from a pre-defined set of questions, which only transfers well to the real-world scenario. Moreover, such offline evaluation remains limited to single-turn interaction, as the pre-defined questions are associated with corresponding answers and unaware of previous interactions. User simulation has been proposed to tackle the shortcomings of corpus-based and user-based evaluation methodologies. A simulated user aims to capture the behaviour of a real user, i.e., being capable of having multi-turn interactions on unseen data, while still being scalable and inexpensive like other offline evaluation methods [7, 56, 78].

In this article, we extend our conversational **User Simulator (*USi*)**, proposed in Sekulić et al. [61], to explore utterance formulation of **large language model (LLM)**–based user simulators. Given an initial information need, *USi* interacts with the conversational system by accurately answering clarifying questions prompted by the system. The answers align with the underlying information needed and help elucidate the intent. Moreover, *USi* generates answers in fluent and coherent natural language, making its responses comparable to real users.

We experiment with two LLM-based approaches to simulate users. First, we base our proposed user simulator on a large-scale transformer-based language model. We fine-tune GPT-2 [49] to generate answers to posed clarifying questions. This method was presented in our recent paper [61]. Second, we use in-context learning, that is, prompting, a few-shot technique made possible with the next generation of LLMs, such as GPT-3 [12], LLaMa [71], and Chinchilla [28]. A GPT-3-based method, *ConvSim*, was recently proposed by Owoicho et al. [44]. Both methods generate answers to clarifying questions in line with the initial information needed, simulating the behaviour of a real user. In the first case, we ensure that through a specific training procedure, resulting in a semantically controlled language model. With a GPT-3-based simulator, we utilise in-context learning (i.e., prompting) to guide the model into following specific steps to answer posed questions.

We evaluate the feasibility of our approaches with an exhaustive set of experiments, including automated metrics and human judgements. First, we compare the quality of the answers generated by our methods and several competitive sequence-to-sequence baselines by computing several automated **natural language generation (NLG)** metrics. In Sekulić et al. [61], we found that the GPT-2-based model significantly outperforms the baselines. This work extends the experiments to *ConvSim* [44], a GPT-3-based model, and finds an even more robust performance. However, as automated NLG metrics often yield unrealistic evaluations, we further analyse a crowdsourcing study conducted in Owoicho et al. [44] to assess how honest and accurate the generated answers are compared with solutions caused by humans. Furthermore, we extend this evaluation setting in a multi-turn conversational scenario. The crowdsourcing judgements show significant differences both in the *naturalness* and *usefulness* of answers generated by *USi* and ones developed by *ConvSim*. The*ConvSim* model outperforms the other, especially in the multi-turn setting. Performance compared with human responses remains similar. Next, we perform a qualitative analysis of utterance reformulations generated by our LLM-based approaches in response to clarifying questions. We map our findings to recently proposed patterns for conversational recommender systems [79] and

find that user simulators tend to rewrite the original query to explain the underlying information need further. However, we note that types of such reformulations highly depend on the training data and the prompts given to the models. Finally, we discuss the applications and future work of LLM-based user simulators.

In summary, our contributions are the following.

— We compare two streams of LLM-based user simulators by the automated NLG metrics.
— We analyse the type of utterances generated by LLM-based methods.
— We discuss in detail the potential applications of LLM-based user simulators, their cost, and their limitations. Moreover, we outline potential future work in the space of user simulation, aimed at going beyond answering clarifying questions.

The rest of the article is organised as follows. Section 2 reviews related work on the topic. Section 3 describes a user's role in conversational search system evaluation and the desirable characteristics of a simulated user. In Section 4, we motivate and describe in detail the implementation of the two approaches to user simulation, covering both *USi* [61] and *ConvSim* [44]. In Section 5, we construct several experiments to answer key research questions on the feasibility of the proposed methods. In Section 6, we then analyse patterns identified in the simulator's responses, compare them to human-generated answers, and extend an existing set of practices for utterance reformulations. We present the results in Section 7. In Section 8, we discuss the advantages versus shortcomings of the approaches and outline our future work aspirations. In Section 9, we present our conclusions.

2 RELATED WORK

Our work is part of a broad area of conversational information retrieval and user simulation. In this section, we present an overview of the relevant work on the topics.

2.1 Conversational Search

Recent advancements in conversational agents have stimulated research in conversational information access [16, 67, 75] that started many years earlier [18]. The report from the Dagstuhl Seminar N. 19461 [5] identifies conversational search as one of the essential areas of **information retrieval (IR)** in the upcoming years. Radlinski and Craswell [50] propose a theoretical framework for conversational search, highlighting the multi-turn user–system interactions as one of the desirable properties of modern conversational search. This property is tied with a mixed-initiative paradigm in IR [30], where the system is passive and prompts the user with engaging content, such as clarifying questions.

Clarification has attracted considerable attention from the research community, including studies on human-generated dialogues on **question-answering (QA)** forums, utterance intent analysis, and asking clarifying questions [11]. Asking clarifying questions is beneficial for the conversational search system and the user. For example, Kiesel et al. [32] studied the impact of voice query clarification on user satisfaction and found that users like to be prompted for clarification. Moreover, Aliannejadi et al. [4] proposed an offline evaluation methodology for the asking clarifying questions and showed the benefits of clarification in terms of improved performance in document retrieval once the question is answered. Hashemi et al. [27] proposed a Guided Transformer model for document retrieval and next clarifying question selection in a conversational search setting. Zamani et al. [77] proposed reinforcement learning–based models for generating clarifying questions and the corresponding candidate answers from weak supervision data. Sekulić et al. [59] proposed a GPT-2-based model for generating facet-driven clarifying questions.

Although extensive work related to clarification in search exists, effective and efficient evaluation methodologies of mixed-initiative approaches are still being determined.

Another research direction in the conversational search area is multi-turn passage retrieval, led by the TREC **Conversational Assistant Track (CAsT)** [19] and **Interactive Knowledge Assistance Tack (iKAT)** [2]. The system needs to understand the conversational context and retrieve appropriate passages from the collection. As a further improvement, Ren et al. [54] introduced the task of conversations with search engines, where the system generates a short, summarised response of the retrieved passages. Other studies in the area of conversational search include user intent classification [48], response ranking [19, 58, 63], document features for clarifying questions [62], user engagement prediction [39, 60], and query rewriting [46, 62, 72].

In the field of **natural language processing (NLP)**, researchers have studied question ranking [51] and generation [52, 74] in dialogue. These studies usually rely on large amounts of data from query logs [53], industrial chatbots [74], and QA websites [51, 52, 70]. For example, Rao and Daumé [51] developed a neural model for question selection on an artificial dataset of clarifying questions and answers extracted from QA forums. Their later study proposed an adversarial training mechanism for generating clarifying questions given a product description from Amazon [52]. Unlike these studies, we study user–system interaction in an IR setting. The user's information need is presented in short queries (versus a long detailed post on StackOverflow), resulting in a ranked list of relevant documents. Furthermore, the IR system can ask clarifying questions to elucidate the user's information need, which needs to be answered.

2.2 User Simulation in Information Retrieval

Given the complexity of human–computer interactions and natural language, there has been an ongoing discussion in the NLP community about the credibility of automatic evaluation metrics based on text overlap [43]. Metrics such as BLEU and ROGUE, which try to judge a system's output solely based on how much overlap it has with a reference utterance, cannot capture the performance of the system accurately [8]. Hence, human annotation should be done to evaluate a system's performance when a generative model is used in summarisation and machine translation tasks. Moreover, the evaluation of a system becomes even more complex if an ongoing interaction between the user and the system exists. Not only must the system evaluate the generated utterance, but it should also be able to incorporate a human response. For this reason, researchers adopt human-in-the-loop techniques to mimic human–computer interactions and further perform human annotation to evaluate the whole system's performance (in response to humans). Recent work of Lipani et al. [38] proposes a metric for offline evaluation of conversational search systems based on a user interaction model.

To alleviate the need for time-consuming and expensive human evaluation, researchers proposed replacing the user with a user simulation system [56, 66]. Simulation in IR has long been studied (1973) [17] to generate pseudo-docs and pseudo-queries to analyse literature search system performance. The work was then followed by Griffiths [26], proposing a general simulation framework for IR systems. Tague et al. [69] later studied the problems for user simulation in bibliographic retrieval systems. User simulation for evaluation was first proposed in 1990 by Gordon [25], who proposed a framework for generating simulated queries. This work has been long followed in the literature to study various hypothetical user and system actions (e.g., issuing 100 queries in a session) that cannot be done in a real system [6]. In particular, Azzopardi [6] proposed to study the cost and gain of user and system actions and studied the effect of different strategies using simulated queries and actions of users (e.g., clicking on relevant documents). Mostafa et al. [42] studied different dimensions of users' interests and their impact on user modelling and

information filtering. Diaz and Arguello [21] adapted an offline vertical selection prediction model in the presence of user feedback for user simulation.

More recently, there has been research on simulating users to evaluate the effectiveness of systems [15, 56, 66, 76, 78]. For example, Carterette et al. [15] proposed a conceptual framework for investigating various aspects of simulations: system effectiveness, user models, and user utility. With the recent developments of conversational systems, more attention has been drawn towards simulating users in a conversation. Sun et al. [66] proposed a simulated user for evaluating conversational recommender systems based on predefined actions and structured response types. Kim and Lipani [33] extend their work by offering a multi-task neural model that predicts user action, satisfaction, and an utterance in conversational recommender systems.

Closer to our work, Salle et al. [56] proposed a user simulator for information-seeking conversations in which the simulator takes an information need and responds to the system accordingly. However, we would like to draw attention to the various limitations of their work. Even though their proposed simulator takes an information need as input and aims to answer the system's request according to the need, it fails to generate responses. In other words, the approach is limited to predicting the relevance of the system's utterance to the user's information need and selecting an appropriate answer from a list of human-generated answers. The simulator becomes valid only if predefined pools of clarifying questions and their answers are available. In this work, we take one step further and generate human-like answers in natural language. Also, the work by Zhang and Balog [78], which simulates users for recommender system evaluation, uses structured data and response types. This work proposes a simulator that generates natural language responses based on unstructured data.

Finally, Zhang et al. [79] study query reformulations in conversational recommender systems. They identify several types of query reformulations and find that users often reformulate their query by repeating the previous utterance by rephrasing it or further expressing their information needs. In this work, we analyse reformulations of user utterances and utterances generated by our simulators in mixed-initiative conversational search systems.

3 USER SIMULATION

In this section, we explain a user's role in evaluating conversational search systems. We also discuss several desired characteristics of a user simulator and propose two simulation methods, with a focus on answering clarifying questions.

3.1 User's Role in Conversational Search System Evaluation

Previous work in task-oriented dialogue systems and conversational search systems mainly evaluate the performance of the systems in an offline setting using a corpus-based approach [20]. The offline evaluation must accurately reflect the nature of conversational systems, as the evaluation is possible only at a single-turn level. Thus, to adequately capture the nature of the conversational search task, it is necessary to involve users in the evaluation procedure [10, 36]. User involvement allows proper evaluation of multi-turn conversational systems, in which users and systems take turns in a conversation. Even with such an approach, which most precisely captures the performance of the systems in a real-world scenario, the involvement of users in the evaluation is tiresome, expensive, and unscalable. To alleviate the evaluation of dialogue systems while still accurately capturing the overall performance, a simulated user approach has been proposed [66, 78]. The simulated user is intended to provide a substitute for real users [7], as it is easily scalable, cheap, fast, and consistent. Next, we formally describe the characteristics of a simulated user for conversational search system evaluation.

3.2 Problem Definition

As mentioned, evaluating conversational search systems is challenging due to the necessity of human judgements at each turn of interaction with the search system. In this work, we aim to alleviate the procedure of evaluating certain types of mixed-initiative conversational search systems. Specifically, we provide a simulated user with an information need capable of answering various clarifying questions prompted by any modern conversational search system.

Our simulated user U is at first initialised with a given information need in. Simulated user U formulates its need in the form of the initial query q, which is then given to the general mixed-initiative conversational system S. The system S elucidates the information needed in through a series of clarifying questions cq. We do not go into details of the implementation of such a system, but different approaches have been proposed in recent literature [4, 27]. Next, the simulated user U needs to provide an answer a to the system's question. The answer a needs to align with the user's information need in.

3.2.1 Single-turn responses. Formally, user U needs to generate an answer a to the system's clarifying question cq, conditioned on the initial query q and the original user's intent in:

$$a = f(cq|in, q). \tag{1}$$

The user U is expected to answer the question in line with its information need, not just based on a potentially vague and under-specified query, as traditional chatbots would be inclined to do.

3.2.2 Conversation history-aware user. The system can take further initiative and ask additional clarifying questions. Thus, our simulated user U needs also to track the conversation flow. Formally, at the conversational turn i, U generates an answer given by

$$a_i = f(cq_i|in, q, H), \tag{2}$$

where H is conversational history, consisting of the interaction between the user and the system up until the current turn: $H = \{(cq_j, a_j)\}$, where $j \in [1 \dots i-1]$. The following section explains how we modelled the described simulated user.

4 SIMULATION METHODOLOGY

In this section, we motivate and describe the two approaches to user simulators for answering clarifying questions. The first one is based on semantically controlled text generation via fine-tuning the LLM model, specifically GPT-2 [49], which is proposed in Sekulić et al. [61]. The other approach is based on in-context learning (prompting) and the next generation of LLMs, GPT-3 [12], which is proposed by Owoicho et al. [44].

4.1 Semantically Controlled Text Generation

We define generating answers to clarifying questions as a sequence generation task. Thus, we employ language modelling as our primary tool for generating sequences. The goal of a **language model (LM)** is to learn the probability distribution $p_\theta(x)$ of a sequence of length n: $x = [x_1, x_2, \dots, x_n]$, where θ are the parameters of the LM. Current state-of-the-art language models, such as GPT-2, learn the distribution in an auto-regressive manner, i.e., formulating the task as a next-word prediction task:

$$p_\theta(\boldsymbol{x}) = \prod_{i=1}^{n} p_\theta(x_i|x_{<i}). \tag{3}$$

However, recent research showed that transformer-based LLMs, although generating text of near-human quality, are prone to "hallucination" [22] and generally lack semantic guidance [55].

Thus, we fine-tune a semantically conditioned LM with a specific fine-tuning technique and careful input arrangement. As mentioned in the previous section, answer generation must be conditioned on the underlying information need. To this aim, we learn the probability distribution of generating an answer a:

$$p_\theta(\boldsymbol{a}|in, q, cq) = \prod_{i=1}^{n} p_\theta(a_i|a_{<i}, in, q, cq), \tag{4}$$

where a_i is the current token of the answer and $a_{<i}$ are all the previous ones, whereas in, q, and cq correspond to the information needed, the initial query, and the recent clarifying question from Equation (1), respectively.

4.1.1 GPT-2-based simulated user. GPT-2 is a large-scale, transformer-based LM trained on a dataset of 8 million web pages capable of synthesising text of near human quality [49]. As trained on a highly diverse dataset, it can generate text on various topics, which can be primed with an input sequence. GPT-2 has previously been used for various text-generation tasks, including dialogue systems and chatbots [13]. Therefore, it suits our task of simulating users by generating answers to clarifying questions in a conversational search system.

We base our proposed user simulator *USi* on the GPT-2 model with language modelling and classification losses, i.e., DoubleHead GPT-2. In this variant, the model learns to generate the appropriate sequence through the language modelling loss and how to distinguish a correct answer to the "distractor". This has been shown to improve the sequence generation [49] and has demonstrated superior performance over only-language loss GPT-2 in the initial stage of experiments. The two losses are linearly combined.

Single-turn responses. We formulate the input to the GPT-2 model, based on Equation (4), as

$$input\_seq = in[SEP]q[SEP]cq[bos]a[eos], \tag{5}$$

where $[bos]$, $[eos]$, and $[SEP]$ are unique tokens indicating the beginning of the sequence, the end of the sequence, and a separation token, respectively. Information needs in, initial query q, clarifying question cq, and a target answer a are tokenised before constructing the entire input sequence to the model. Additionally, we construct segment embeddings, which indicate different segments of the input sequence, namely, in, q, cq, and a.

When training the DoubleHead variation of the model, we formulate the first part of the input as described above. Additionally, we sample the ClariQ dataset for distractor answers and process them like the original answer, based on Equation (5). Therefore, the DoubleHead GPT-2 variant accepts as input two sequences, one with the original target answer in the end and the other with the distractor answer. It then needs not only to learn to model the target answer but also to distinguish between original and distractor answers and provide a binary label indicating which of the two solutions is desirable. We sample the distractor answers from the datasets above. When possible, we ensure that if the target answer starts with "Yes", the distractor answer starts with "No" to enforce the connection between the solution, the clarifying question, and the information needed. Likewise, if the answer starts with "No", we sample a distractor answer that begins with "Yes". Note that *USi* does not generate answers that begin strictly with a "yes" or a "no".

Conversation history-aware model. The conversation history-aware model calls for a different input and training formulation. The input to history-aware GPT-2 is constructed as

$$input\_seq = in[user]q[system]cq_{<i}[user]a_{<i}[system]cq_i[bos]a_i[eos],$$

where $[user]$ and $[system]$ are additional unique tokens indicating the conversational turns between the (simulated) user and the conversational system, respectively.

Inference. During inference, we omit the answer a from the input sequence, as our goal is to generate this answer to a previously unseen question. To generate answers, we use a combination of state-of-the-art sampling techniques to develop a textual sequence from the trained model. We utilise temperature-controlled stochastic sampling with top-k [23] and top-p (nucleus) filtering [29]. After the initial experiments and consultation with previous work, we fix the temperature parameters to 0.7, k to 0, and p to 0.9.

4.2 Prompt-Based Text Generation

In this section, we describe the prompt-based generation method. We follow Owoicho et al. [44] and utilise recently developed GPT-3 [12] to answer posed clarifying questions. To this end, we use prompting [24] — a method to describe the task for the LLM to perform without requiring further fine-tuning. The prompt is a chunk of text that describes a task we are interested in, preferably giving several examples of such a study being executed. We want to generate the answer a to the clarifying question cq. As mentioned, the answer must align with the underlying information needs description.

Prompt-based generation has several potential advantages over the previously introduced fine-tuning-based approach. For example, only a couple of examples must be given to the model, thus mitigating the need to create task-specific datasets. As such, prompt-based few-show learning can adapt to various tasks. While we focus solely on utilising such methods for answering clarifying questions in this work, they can be used for other user-specific utterance-generation tasks, such as providing explicit feedback [44]. Nonetheless, prompting only became a recently valid method due to significant advancements in LLMs. However, the next generation of LLMs requires significantly more processing power and is not feasible to run on single-compute nodes. This consequently raises the cost of such methods. Thus, fine-tuning medium-sized LLMs, such as GPT-2, might still be a potentially desirable path. In this article, we compare the two methods across several aspects and discuss the potential advantages of one over the other.

5 EVALUATION METHODOLOGY

In this section, we describe our methodology for evaluating the proposed user simulation methods. We compare text generated by our simulators to human-generated text with regard to multiple aspects. First, we use automated NLG metrics to assess the differences between the two simulation methods. Second, we employ crowdsourcing to evaluate the *usefulness* and *naturalness* of the generated answers. Finally, we perform a qualitative analysis of the simulator's utterance reformulations and map them into recently identified patterns (see [79]). All of the comparisons are performed both in single- and multi-turn settings.

5.1 Single-Turn Conversational Data

For training and evaluating our proposed simulated user *USi*, we utilise two publicly available datasets, Qulac [4], and ClariQ [3]. Both datasets aim to foster research in asking clarifying questions in open-domain conversational search. Qulac was created on top of the TREC Web Track 2009-12 collection. The Web Track collection contains ambiguous and faceted queries, often requiring clarification when addressed in a conversational setting. Given a topic from the dataset, clarifying questions were collected via crowdsourcing. Then, given a topic and a specific facet of the case, workers were employed to gather answers to these clarifying questions. This results in a tuple of ($topic$, $facet$, $clarifying\_question$, $answer$). Most of the topics in the dataset are multi-faceted and ambiguous, meaning that the clarifying questions and answers must align with the actual facet. ClariQ is an extension of Qulac created for the ConvAI3 challenge [3] and contains additional non-ambiguous topics. Relevant statistics of the datasets are presented in Table 1.

Table 1. Statistics for Qulac and ClariQ Datasets

	Qulac	ClariQ
Number of topics	198	237
Number of facets	762	891
Number of questions	2,639	3,304
Number of question–answer pairs	10,277	11,489

We utilise these datasets by feeding the corresponding elements to Equation (4). Specifically, *facet* from Qulac and ClariQ represents the underlying information need, as it describes in detail the intent behind the issued *query*. Q represents the current asked question, whereas *answer* is our language modelling target.

5.2 Multi-turn Conversational Data

A significant drawback of Qulac and ClariQ is that they are both built for single-turn offline evaluation. A conversational search system will likely engage in a multi-turn dialogue to elucidate user needs. To bridge the gap between single- and multi-turn interactions, we construct multi-turn data that resembles a more realistic interaction between a user and the system. Our user simulator *USi* is then further fine-tuned on this data.

To acquire the multi-turn data, we construct a crowdsourcing-based human-to-human interaction. At each conversational turn, the crowdsourcing worker is tasked to behave as a search system by asking a clarifying question on the topic of the conversation. Then, another worker is tasked to provide the answer to that question, considering the underlying information needed and the conversation history, imitating the actual user's behaviour. We construct in 500 conversations up to a depth of three, i.e., we have three sequential question–answer pairs for a topic and its facet.

We construct several edge cases to further study the effects of specific clarifying questions on the search experience. In such cases, the clarifying question prompted by the search system is considered faulty, as it is either a repetition, off-topic, unnecessary, or completely ignores the previous user's answers. We obtain answers to these questions to provide more realistic data for the training of our model, making our simulated user as human-like as possible. These clarifying questions are intended to simulate a conversational search system of poor quality and provide insight into users' responses to such questions. We employ workers to provide answers to an additional 500 clarifying questions of poor quality, up to the depth of two. The specific edge cases and their descriptions with examples are presented in Table 2. We publicly release the acquired multi-turn datasets in Sekulić et al. [61]. In this work, we use the multi-turn data to evaluate both fine-tuning-based and prompting-based approaches to generating answers to clarifying questions in a conversational setting.

5.3 Research Questions

We aim to evaluate whether our proposed simulated user can replace real users in answering clarifying questions of conversational search systems, which would make evaluating such systems significantly less troublesome. Overall, we aim to answer four main research questions, extending the list from Sekulić et al. [61]:

RQ1: To what extent are the answers generated by the two simulation methods in line with the underlying information need?
RQ2: How coherent and natural is the language of the generated answers?
RQ3: How do LLM-based simulators behave in multi-turn interactions?
RQ4: What are the advantages and disadvantages of either simulation methodology?

Table 2. Multi-turn Dataset Acquired Through Crowdsourcing

Question case	Description	Sample conversation	N
Normal	A good system naturally continues the conversation.	U: I'm looking for information on dieting S: Are you looking for dieting tips? U: Yes and exercise tips as well. S: Do you need anything specific in relation to counting calories you consume daily? U: Yes, I would like to know more about that topic.	500
Repeat	System repeats the previous question.	U: Find information on raised gardens. S: Do you need information on materials needed? U: No, I want to find plans. S: Do you need information on materials needed? U: I want what I previously asked for.	50
Off-topic	System asks the user an off-topic question.	U: I'm looking for an online world atlas. S: Are you interested in satellite maps? U: No, I want an online world atlas. S: Which mountain ski resort would you like information around the Pocono area? U: I am not interested in this topic.	50
Similar	System asks a question similar to the previous one, ignoring the user's answer.	U: I'm looking for information about Mayo Clinic Jacksonville FL. S: Would you like to request an appointment? U: Yes. S: Are you looking for the address of Mayo Clinic Jacksonville FL? U: I just want to request an appointment.	400

Sample conversations of depth three are omitted due to space limitations.

In order to address these questions, we first compute several NLG metrics to compare the generated answers to the oracle human answers from ClariQ. As several NLG metrics received criticism from the NLP community, significantly since they do not correlate well with the text's coherence, we perform a crowdsourcing study to evaluate the *naturalness* of generated answers. To evaluate whether the generated answers align with the information needed, we conduct an additional crowdsourcing study and assess the *usefulness* of answers. Finally, we perform a qualitative analysis of generated answers by identifying certain patterns in utterance formulations.

As it was done in Sekulić et al. [61], we compare our LLM-based user simulators with two sequence-to-sequence baselines. The first baseline is a multi-layer bidirectional **long short-term memory (LSTM)** encoder–decoder network for sequence-to-sequence tasks [68].[1] The second baseline is a transformer-based encoder–decoder network, based on Vaswani et al. [73]. We perform a hyperparameter search to select the models' learning rate, number of layers, and hidden dimension. Both baselines are trained with the same input as our primary model.

[1]We use the IBM implementation for our experiments: https://github.com/IBM/pytorch-seq2seq

5.4 Automated NLG Metrics

We first study the language-generation ability of *USi* and the previous baselines. We compute several standard metrics for evaluating the generated language. We use two widely adopted metrics based on n-gram overlap between the generated and the reference text. These are BLEU [45] and ROUGE [37]. Next, we compute the EmbeddingAverage and SkipThought metrics to capture the semantics of the generated text, as they are based on the word embeddings of each token in the developed and the target text. The metric is then defined as a cosine similarity between the means of the word embeddings in the two texts [35]. The models are trained on the ClariQ training set and evaluated on the unseen ClariQ development set. We evaluate ClariQ's development set since the test set does not contain question–answer pairs. We take a small portion of the training set for our development. The answers generated by *USi* and the baselines are compared against oracle answers from ClariQ, generated by humans.

5.5 Response Naturalness and Usefulness

To simulate a real user, the generated responses by our model need to be fluent and coherent. Thus, we study the *naturalness* of the generated answers. We define *naturalness* as an answer being natural, fluent, and likely caused by a human. Similarly, fluency [14] and humanness [57] have been used for evaluating generated text. We also assess the *usefulness* of the answers generated by our simulated user. We define *usefulness* as an answer that aligns with the underlying information need and guides the conversation towards the topic of the information need. This definition of usefulness can be related to similar metrics in previous work, such as adequacy [65] and informativeness [16].

We perform a crowdsourcing study to assess the *naturalness* and *usefulness* of generated answers to clarifying questions. We use Amazon Mechanical Turk to acquire workers based in the United States with at least a 95% task approval rate. The study was done in a pair-wise setting, i.e., each worker was presented with several answer pairs. Our model generated one of the answers, and the other was by a human, taken from the ClariQ collection. Their task was to judge which answers were more natural or valuable, depending on the study. The workers have been provided with the context, i.e., the initial query, facet description, and clarifying question.

We annotate 230 answer pairs for *naturalness* and 230 answer pairs for *usefulness*, each judged by two crowdsource workers. We would define a *win* for our model if both annotators voted our generated answer as more natural/useful and a *loss* for our model if both voted the human-generated answer as more natural/useful. If the two workers voted differently on a single answer pair, we define that as a *tie*. With this study, we aim to shed light on research questions RQ1 and RQ2, i.e., whether the generated answers are natural and in line with the underlying information need compared with human-generated answers. Additionally, we compare Transformer-seq2seq to *USi*.

We also compare the two LLM-based simulation approaches in a multi-turn casual setting. The results of both single- and multi-turn comparisons are presented in Section 7.2.

6 RESPONSES TO CLARIFYING QUESTIONS

In this section, we analyse human- and simulator-generated answers to posed clarifying questions. Specifically, we conduct expert annotation to identify patterns in the given answers, grounding our findings in prior work. To this end, we analyse the answers in light of patterns identified by Krasakis et al. [34], focusing on the Qulac dataset [4]. Krasakis et al. [34] find that users' answers vary in polarity in length. For example, the user can answer with a negative short answer, such as *"No"*, but also potentially provide a longer answer, e.g., *"No, I'm looking for X instead"*. Naturally, the answer can also be positive polarity depending on the information needed and prompted clarifying

questions. Furthermore, we compare the generated answers to patterns identified by Zhang et al. [79]. Although Zhang et al. [79] focus on query reformulations in conversational recommender systems, we find the overlap of the findings to be high. Thus, we map their proposed query reformulation types to answers in a mixed-initiative conversational search. Finally, we analyse answers to faulty clarifying questions proposed by Sekulić et al. [61].

6.1 Responses Patterns

We analyse answers to prompted clarifying questions in light of previously identified utterance reformulation types [79]. In other words, we map and expand the existing utterance reformulation ontology for conversational recommender systems to answer formulations in conversational search. While specific differences exist between recommender and search systems, our initial analysis suggested that the common conversational setting incites similar user behaviours. In their study, Zhang et al. [79] analyse how users reformulate their utterances in subsequent turns given a prompt from the conversational recommendation agent about its lack of understanding of user's needs. Similarly, in conversational search, we have the user's initial query, the clarifying question prompted by the search system, and the user's answer. Thus, we analyse these answers through the lens of reformulations from the user's initial query.

Zhang et al. [79] identify seven utterance reformulation behaviours: (1) *start/restart* – users start to present their need; (2) *repeat* – user repeats previous utterance without significant change; (3) *repeat/rephrase* – user repeats last turn with different wording; (4) *repeat/simplify* – user repeats the word with a more straightforward expression, reducing complexity; (5) *clarify/refine* – user clarifies or refines the expression of an information need; (6) *change* – user changes the information need (topic shift); (7) *stop* – user ends the search session. We encourage an interested reader to refer to Zhang et al. [79] for a more elaborate explanation of the reformulation types. In our analysis, we focus specifically on answers to clarifying questions. Thus, some user utterances must be observed and not discussed in other sections. Mainly by the design of our research setting, described in Section 4, we do not deal with utterance types (1) *start/restart*, (6) *change*, or (7) *stop*. However, we add two additional categories, mostly to deal with edge cases: (8) *hallucination* – when the provided answer is not in line with the underlying information need and (9) *short answer* – when the answer is just "no" or "yes". Examples of the observed utterance types are presented in Table 4.

6.2 Responses to Faulty Clarifying Questions

In order to gain further insight into designing a reliable user simulator for conversational search evaluation, we must adapt it to be resilient to unexpected system responses. For example, if a conversational search system responds with an off-topic clarifying question or an unrelated passage, the simulated user needs to react in a natural, human-like manner. However, to design such a simulator, we first need to learn how real users would react to incorrect responses from the search system. To this end, we acquired a dataset of human responses when prompted with faulty clarifying questions. The published dataset is multi-turn and can thus be used to improve our multi-turn user simulator model.

Examples from the acquired dataset are presented in Table 2. The dataset contains several scenarios in which a conversational search system asks follow-up clarifying questions. We acquired a dataset of 1,000 conversations, with crowd workers assuming the user role and responding to clarifying questions. Initial analysis of the crowd workers' answers offers several insights. In the case of appropriate clarifying questions (*Natural*), users tend to respond naturally by refining their information needs, as expected. However, in the case of faulty clarifying questions (*repeat*, *off-topic*, or *similar*), users either repeat their previous answer (20% of analysed answers), expand their last

Table 3. Performance of Different Answer Generation Methods, Measured by Automated NLG Metrics on the ClariQ Development Set

Model	BLEU-1	BLEU-2	BLEU-3	ROUGE_L	SkipThoughtCS	EmbeddingAvgCS
LSTM-seq2seq	0.1989	0.1401	0.0988	0.2210	0.3158	0.7012
Transformer-seq2seq	0.2041	0.1352	0.0936	0.2067	0.3666	0.7077
USi [61]	**0.3029**	**0.2404**	**0.2054**	**0.2359**	**0.4025**	**0.7322**
ConvSim [44]	0.1949	0.1394	0.1014	0.1898	0.3911	0.6766

reply with more details on their information need (23%), or rephrase the previous answer with different wording (37%). Next, we aim to evaluate the resilience of our proposed *USi* to such faulty questions by analysing its correspondence to human-generated answers.

7 RESULTS

In this section, we present the results of the evaluation methods described above. First, we show the performance of user simulation approaches as measured by automated NLG metrics, followed by a crowdsourcing-based study on response usefulness and naturalness. Finally, we qualitatively analyse the generated utterances.

7.1 Automated NLG Metrics

Performance of the baseline models and our simulated user models, as evaluated by automated NLG metrics described in Section 5.4, is presented in Table 3. *USi* significantly outperforms all baselines by all computed metrics on the ClariQ data. Even though LSTM-seq2seq showed strong performance in various sequence-to-sequence tasks, such as translation [68] and dialogue generation [64], it performs relatively poorly on our task. A similar outcome is observed for Transformer-seq2seq. We hypothesise that the poor performance in this task is due to limited training data, as the success of these seq2seq models on various tasks was conditioned on large training sets. Our GPT-2-based model does not suffer from the same problem, as it has been pre-trained on a large body of text, making the fine-tuning enough to capture the essence of the task, which is generating answers to clarifying questions.

An interesting observation is the fact that the GPT-3-based model, *ConvSim* [44] performs worse than the GPT-2-based *USi*. We attribute this result to the aforementioned issues with the unreliableness of the automated NLG metrics. As such, they capture solely exact matching of the wordings of the generated answer and the gold answers, largely failing to adjust to differences in vocabulary between the two answers, although they might be conveying the same message. Thus, in the next section, we report more reliable crowdsourcing-based annotations of the

7.2 Naturalness and Usefulness

Table 6 presents the results of the crowdsourcing study on *usefulness* and *naturalness*, comparing answers generated by *USi* and humans, as described in Section 5.5. Both in terms of *naturalness* and *usefulness*, we observe a large number of *ties*, i.e., the two workers annotating the answer pair did not agree on which one is more natural/beneficial. Since we are comparing answers generated by our GPT-2-based simulated user with the answers written entirely by humans, this result favours our proposed model. Moreover, the difference between *losses* and *wins* for our model is relatively tiny (38% losses, 17% wins) for *naturalness* and even smaller in terms of *usefulness* (32% losses, 23% wins). We conduct a trinomial test for statistical significance [9], an alternative to the Sign and binomial tests that consider ties.[2] Regarding *naturalness*, we reject the null hypothesis of equal

[2]Another point-of-view would be to test for equivalent effectiveness, as in Jayasinghe et al. [31]. However, we refrain from it since it does not consider ties.

Table 4. Identified Reformulation Patterns in Responses Generated by Our Proposed User Simulator(s)

	Explanation	Example
Repeat	Repeat the last utterance without significant change.	U: What are treatments for carotid cavernous fistula? S: Would you like to search for a doctor for a carotid cavernous fistula treatment? U: No, I want to know what treatments are available for carotid cavernous fistula.
Repeat/rephrase	Repeat the last utterance by using different wording.	U: I want to learn about rock art. S: Do you want pictures of rock art? U: Yes, I want information on rock art.
Repeat/simplify	Repeat the last utterance in a simpler language.	U: Find me information about the sales tax in Illinois. S: Do you want to know what the Illinois state tax rate is? U: Yes, I want to know about the taxes in the state.
Clarify/refine	Refine the information need expressed in last utterance.	IN: Find the homepage for the NASA Voyager mission. U: Tell me about Voyager. S: Are you interested in where the Voyager has been? U: No, I am interested in the website for the mission.
Other	Short response, none above applicable.	U: "All men are created equal" S: Are you looking for Declaration of Independence at the National Archives? U: No.
Hallucination	Response not grounded in information need.	IN: How is workplace diversity achieved and managed? U: Tell me about diversity. S: Would you like the legal definition of diversity? U: Yes, I would like to know about the definition of diversity.

The patterns are extended from the list proposed by Zhang et al. [79].

Table 5. Results on *Naturalness* and *Usefulness* of Responses, *USi* vs. Transformer-Seq2Seq Baseline [61] vs. Human-Generated Answers

	USi [61]	Baseline [61]	Ties	*USi* [61]	Human	Ties
Naturalness	50%†	3%	47%	17%	38%	45%
Usefulness	66%†	3%	31%	22%	27%	51%

The † sign indicates a statistically significant difference under the trinomial test with p-value < 0.05.

performance with $p < 0.05$, i.e., human-generated answers are more natural than those generated by *USi*. Nonetheless, 45% of ties between *USi*- and human-generated answers suggests the high quality of the generated text. Regarding *usefulness*, we accept the null hypothesis of equal performance with $p = 0.43$, i.e., there is no statistically significant difference between the performance of human annotators and *USi*.

Table 5 presents the comparison results between the Transformer-seq2seq and *USi*. We observe a win of the proposed *USi* over the baseline by a large margin. Our GPT-2-based model significantly outperforms the baseline ($p < 0.05$) both in terms of *naturalness* (50% wins and 3% losses) and *usefulness* (66% wins and 3% losses). This finding is in line with the automated evaluation of generated answers.

Regarding the research questions RQ1 and RQ2, i.e., whether the responses generated by our model align with the underlying information need and, at the same time, coherent and fluent, we arrive at the satisfactory performance of the simulated user. The generated answers to clarifying questions seem to be able to compete with the solutions produced by humans both in terms of *naturalness* and *usefulness*. Moreover, the strong performance of *USi* over Transformer-seq2seq additionally motivates the use of large-scale pre-trained language models, such as GPT-2, for the

Table 6. Results of a Crowdsourcing Study Assessing Naturalness and Usefulness of Generated Answers Between *ConvSim*, *USi*, and Human-Constructed Answers

		ConvSim [44]	*USi* [61]	Ties	ConvSim [44]	Human	Ties
MultiSingle	*Naturalness*	37%†	22%	41%	36%	25%	39%
	Usefulness	44%†	19%	37%	36%†	20%	44%
	Naturalness	45%†	18%	37%	25%	28%	47%
	Usefulness	62%†	12%	26%	26%	16%	58%

Each value indicates the percentage of pairwise comparisons won by the specific model and the percentage of ties. Sign † indicates a statistically significant difference as computed by a trinomial test with p-value < 0.05.

Table 7. Qualitative Analysis of Answers Generated by User Simulator *USi*

Information need	Clarifying question	Generated answer
How do you treat severe angular cheilitis?	Are you looking for the definition of angular cheilitis?	No, I want to know about treatment.
What home remedies are there for angular cheilitis?	Are you looking for the definition of angular cheilitis?	No, I am looking for home remedies.
Find a list of hobby stores in Maryland.	Is this for a school craft or play?	No, I want to find a list of hobby stores in Maryland.
What hobby stores carry trains?	What location do you want to find a hobby store in?	I want to know where I can purchase trains.
What hobby stores carry trains?	Is this a hobby for an adult?	Yes.

task. These results make a strong case for using a user simulator for mixed-initiative conversational search system evaluation.

7.3 Qualitative Analysis

7.3.1 Single-Turn Analysis. In this section, we analyse several conversation samples of our user simulator with a hypothetical conversational search system. Table 7 shows five interaction examples. The user simulator *USi* is initialised with the information that needs description text. Given an initial query (omitted in the table for space), the conversational search system asks a clarifying question to elucidate *USi*'s intent. Then, *USi* generates the answer to the prompted question. The information needed and the questions for these examples are taken from the ClariQ development set. Most TREC-style datasets contain the information need (facet/topic) description alongside the initial query. Thus, our simulated user can help evaluate conversational search systems on any of such datasets, as it only requires a description for initialisation. Then, the system we aim to evaluate can produce clarifying questions and receive answers from *USi*.

The first two examples in Table 7 initialise *USi* with different information needs. However, given the same initial query "How to cure angular cheilitis" and the same prompted clarifying question, *USi* answers differently, in line with the basic information needed for each case. In the table's last three rows, we have different information needs for one broad topic of hobby stores. Given the initial query "I'm looking for information on hobby stores", *USi* again answers questions in line with the underlying information need. We notice that the text produced by our GPT-2-based user simulator is coherent and fluent and, in the given examples, indeed in line with the underlying information need. Moreover, *USi* is not bound by answering the question in a "yes" or "no" fashion. Instead, it can produce various answers and even express its uncertainty (e.g., "I don't know").

Table 8. Prevalence of Utterance Reformulation Types for Answers to Clarifying Questions

	Human	*USi* [61]	*ConvSim* [44]
Repeat	2%	0%	3%
Repeat/rephrase	4%	7%	6%
Repeat/simplify	4%	8%	5%
Clarify/refine	63%	37%	83%
Other	25%	40%	3%
Hallucination	2%	7%	0%

Table 8 shows the prevalence of the aforementioned types of utterance reformulations on the ClariQ development set. We expertly annotated 150 answers generated by both generative approaches as well as human answers taken directly from the ClariQ dataset. As indicated in Table 8, *USi* hallucinates in 7% of analysed cases. The hallucination accounted for in the table is limited to cases when a long answer is generated. However, we observed that *USi* often needs a better short answer. For example, with an information need related to finding the list of dinosaurs with pictures, when prompted with a clarifying question "Are you looking for pictures of dinosaurs?", *USi* answers "No". Such short answers are mapped under *Other* in Table 8, as the focus of the analysis was to capture the extent of the short-versus-long answers. Moreover, we observe the hallucination phenomena in several answers taken from ClariQ, constructed by crowd workers. We attribute this to the potentially swift manner in which the answers were written rather than to crowd workers not understanding that their answers were not in line with the given information need. On the contrary, the prompt- and GPT-3-based *ConvSim* method does not suffer from the mentioned issue.

The prevalence of different utterance reformulations differs between human-generated answers and the answers generated by *USi* and *ConvSim*. Specifically, we observe a greater frequency of short answers (e.g., "yes", "no") in answers generated by GPT-2-based *USi*. On the other hand, GPT-3-based *ConvSim* tends to refine and clarify the given information need in the majority of the cases. While both long and short types of answers to clarifying questions are acceptable, as long as they are in line with the information need, certain users have a slight preference towards one or the other. Thus, as discussed in the last section, as a step towards more realistic user simulators, we aim to model users according to their cooperativeness level. In other words, the simulator would be able to generate either concise or long and elaborate answers depending on the cooperativeness parameter for a specific underlying user model.

7.3.2 Multi-turn Analysis. We perform an initial case study on the multi-turn variant of *USi*. While the initial analysis of multi-turn conversations suggests that *usefulness* and *naturalness* of single-turn interactions transfer into a multi-turn setting, additional evaluation is needed to support that claim strongly. Thus, future work includes a pair-wise comparison of multi-turn conversations inspired by ACUTE-Eval [36].

We also aim to observe user simulator behaviour in unexpected, edge-case scenarios. For example, initial analysis of the created multi-turn dataset showed that humans tend to repeat their previous answers when the clarifying question is off-topic or repeated. Similarly, our multi-turn *USi* has been observed to generate answers such as "I already told you what I'm looking for" when prompted with a repeated question. However, such edge cases need to be clarified for the multi-turn model, which leads to a higher presence of hallucination than in the single-turn variation. This means that the user simulator drifts off the topic of the conversation and starts generating answers outside the basic information needed. This effect is well documented in recent literature

on text generation [22] and should be approached carefully. Although edge cases are also present in the acquired dataset, the GPT-2-based model needs additional mechanisms to simulate the behaviour of users in such cases. We leave a deeper analysis of the topic for future research.

8 DISCUSSION AND FUTURE WORK

In this section, we discuss the advantages and shortcomings of both simulation approaches, their applicability in evaluation, and topics for future work.

8.1 Performance versus Cost

While both GPT-2- and GPT-3-based user simulators can generate natural and valuable answers to clarifying questions, as demonstrated by our experiments presented in Section 7.2, GPT-3 is still significantly better. The difference in performance becomes wider in the multi-turn setting, indicating the overall superiority of the GPT-3-based simulator for the task. This was expected, as GPT-3 was trained on a significantly larger dataset (570 GB of text) than GPT-2 (40 GB of text) and is much more significant in terms of parameters (175 billion for GPT-3 vs. 1.5 billion for GPT-2) [12]. However, the increase in performance comes with a rise in cost. The *Davinci* model used in our experiments costs $0.0200 per 1K tokens.[3] When the cost of pre-training such models is considered, it extends well beyond the cost of pre-training and GPT-2-based methods. For example, to run GPT-3, we need at least 5 80 GB A100 GPUs,[4] whereas GPT-2 runs smoothly on a single 12 GB GPU. As such, it would be incredibly beneficial if smaller-scale LLMs could be used on specific tasks with reasonable success. Achieving such performances with smaller-scale LLMs could be the direction towards sustainable **artificial intelligence (AI)** [41].

8.2 Beyond Answering Clarifying Questions

In this article, we focused on simulators for answering clarifying questions posed by the system. However, as indicated in Section 3, a user's role in conversational search system evaluation extends beyond answering clarifying questions. Thus, a well-designed user simulator should have different properties, such as information-seeking behaviour and explicit feedback generation. We hypothesise that GPT-2- and GPT-3-based methods could be used to a reasonable extent for such purposes. However, an essential distinction between the two methods is that GPT-2 would require fine-tuning. This entails the need for appropriate datasets for each property we intend to include in the simulator. For example, if we aim to include an explicit feedback generation module in our simulator, that is, the simulator's ability to generate positive or negative feedback in a natural language to the system's responses, we would require a substantially large dataset of such feedback, together with initial queries and the system's responses. Then, GPT-2, and models alike, could be fine-tuned for the task.

On the other hand, the next generation of LLMs, such as GPT-3, allows the use of an in-context learning approach — *prompting*, explained in Section 4.2. This eliminates the need for task-specific datasets, as the prompt contains a brief task description and a few examples. The description and examples of the task being carried out correctly can be designed directly by the researcher and typically do not require external annotation. Thus, such LLMs are highly adaptable to other simulator-related tasks, making the extensions to the aforementioned desired properties significantly easier. While some of the properties have been recently explored in *ConvSim* [44], namely, providing explicit feedback, future work includes the expansion of current simulators to others.

[3]Cost at the time of submission: https://openai.com/pricing

[4]The exact compute required is not available as the model is closed source. This is the estimate by the AI research community.

However, we note that prompt design is not a silver bullet compared with the fine-tuning approach, as minor adjustments to the prompt text can result in the generation of vastly different utterances.

8.3 Parameterised Simulator

We analysed the types of utterance formulations given by our generative LLM-based simulators. As reported in Section 7.3, we observe certain patterns in the formulations and discuss specific differences between the GPT-2- and GPT-3-based models. Our line of thought is that a fine-tuned model exhibits behaviours of the data it was trained on. As such, the distributions and varieties of the answers are similar to the data it was trained on. On the other hand, prompting allows for easier tweaking of the reformulations we want to exhibit. However, a realistic user simulator should closely follow the behaviours of real users [7]. To achieve that, we need both the underlying user model we design our simulators by and more control over the types of utterances generated by the simulators. A solution towards this goal lies in *parameterised* user simulators.

Parameterised user simulators would allow for adjustment towards certain types of users. For example, Salle et al. [56] model cooperativeness – how lengthy and informative the simulator responses are, and patience – how many turns of answering clarifying questions is the simulator willing to partake in before giving up. Similarly, Owoicho et al. [44] model *ConvSim*'s patience. However, many other parameters could be introduced, allowing for fine-grade user models. Such parameters are demandingness – how precisely does the system's response need to be for the user to provide positive feedback to it and chattiness – is the user simulator simply answering questions and giving feedback or does it include certain conversational elements (i.e., "That's interesting", "I didn't know that"). Another important aspect of the user simulators is their ability to develop and change their information need throughout the conversation, including shifting the topic of the conversation completely.

8.4 Applications of the Simulator

The first application of our proposed simulators, discussed throughout this article, is *evaluation* of conversational search systems. This can be done by allowing the search system to interact with the simulator, which then assumes a user's role and exhibits the implemented behaviours. The conversational search system's primary goal is to satisfy the underlying information need, which can be evaluated using a standard Cranfield paradigm [38]. In other words, starting from the initial query, the system must provide a ranked list of documents, which are evaluated against query relevance judgements. Evaluation via simulation enables quicker, more cost-effective, and more robust comparisons of conversational search systems without the potential loss in quality.

User simulators can be used for evaluating conversational search systems, but also for creating synthetic data to be used for downstream tasks [40]. Moreover, we can identify break points in search systems by purposely generating faulty utterances, thus probing the search system's robustness. For example, we might design a simulator that performs a sudden topic shift and indicate what kind of actions would be expected of the search system.

9 CONCLUSIONS

In this article, we have examined two recently proposed approaches for simulating users to alleviate the evaluation of mixed-initiative conversational search systems. More specifically, we demonstrated the feasibility of substituting expensive and time-consuming user studies with scalable and inexpensive user simulators. Through several experiments, including automated metrics and crowdsourcing studies, we showed the simulator's capabilities in generating fluent and accurate answers to clarifying questions prompted by the search system. A crowdsourcing study of answer

usefulness and *naturalness* showed that answers generated by *USi* tied with human-generated solutions in 51% and 45% of cases, respectively. Moreover, we confirmed the even stronger performance of the GPT3-based simulator, *ConvSim*, especially in the multi-turn setting. We also performed a qualitative analysis of the utterance reformulation types, finding that the majority of the answers aim to clarify and refine the underlying information need. Based on all of the results and observations, we discussed further steps towards a more realistic simulator by introducing parameters that would allow for modelling different types of users, such as cooperativeness and patience. Moreover, we plan to investigate the application of *ConvSim* for proactive user simulation [1].

REFERENCES

[1] Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the LLMs talk: Simulating human-to-human conversational QA via zero-shot LLM-to-LLM interactions. In *WSDM*. ACM.

[2] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffery Dalton, and Leif Azzopardi. 2024. TREC iKAT 2023: The interactive knowledge assistance track overview. *arXiv preprint arXiv:2401.01330* (2024).

[3] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2020. ConvAI3: Generating clarifying questions for open-domain dialogue systems (ClariQ). (2020).

[4] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *SIGIR*. 475–484.

[5] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational Search (Dagstuhl Seminar 19461). In *Dagstuhl Reports*, Vol. 9. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[6] Leif Azzopardi. 2011. The economics in interactive information retrieval. In *SIGIR*. ACM, 15–24.

[7] Krisztian Balog. 2021. Conversational AI from an information retrieval perspective: Remaining challenges and a case for user simulation. (2021).

[8] Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *ACL*.

[9] Guorui Bian, Michael McAleer, and Wing-Keung Wong. 2011. A trinomial test for paired data when there are many ties. *Mathematics and Computers in Simulation* 81, 6 (2011), 1153–1160.

[10] Alan W. Black, Susanne Burger, Alistair Conkie, Helen Hastie, Simon Keizer, Oliver Lemon, Nicolas Merigaud, Gabriel Parent, Gabriel Schubiner, Blaise Thomson, et al. 2011. Spoken dialog challenge 2010: Comparison of live and control test results. In *SIGDIAL*. 2–7.

[11] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What do you mean exactly?: Analyzing clarification questions in CQA. In *CHIIR*.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[13] Paweł Budzianowski and Ivan Vulic. 2019. Hello, it's GPT-2 — how can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. *EMNLP-IJCNLP 2019* (2019), 15.

[14] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *ACL*.

[15] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. 2011. Simulating simple user behavior for system effectiveness evaluation. In *CIKM*. 611–620.

[16] Aleksandr Chuklin, Aliaksei Severyn, Johanne R. Trippas, Enrique Alfonseca, Hanna Silen, and Damiano Spina. 2019. Using audio transformations to improve comprehension in voice question answering. In *CLEF*. Springer, 164–170.

[17] Michael D. Cooper. 1973. A simulation model of an information retrieval system. *Information Storage and Retrieval* 9, 1 (1973), 13–32.

[18] Fabio Crestani and Heather Du. 2006. Written versus spoken queries: A qualitative and quantitative comparative analysis. *Journal of the American Society for Information Science and Technology* 57, 7 (2006), 881–890.

[19] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624* (2020).

[20] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* 54, 1 (2021), 755–810.

[21] Fernando Diaz and Jaime Arguello. 2009. Adaptation of offline vertical selection predictions in the presence of user feedback. In *SIGIR*. 323–330.

[22] Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455* (2021).

[23] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833* (2018).
[24] Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723* (2020).
[25] Michael D. Gordon. 1990. Evaluating the effectiveness of information retrieval systems using simulated queries. *Journal of the American Society for Information Science* 41, 5 (1990), 313–323.
[26] José-Marie Griffiths. 1976. *The Computer Simulation of Information Retrieval Systems*. Ph. D. Dissertation. University of London (University College).
[27] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2020. Guided Transformer: Leveraging multiple external sources for representation learning in conversational search. In *SIGIR*. 1131–1140.
[28] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* (2022).
[29] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751* (2019).
[30] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *CHI*. 159–166.
[31] Gaya K. Jayasinghe, William Webber, Mark Sanderson, Lasitha S. Dharmasena, and J. Shane Culpepper. 2015. Statistical comparisons of non-deterministic IR systems using two dimensional variance. *Information Processing & Management* 51, 5 (2015), 677–694.
[32] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2018. Toward voice query clarification. In *SIGIR*. 1257–1260.
[33] To Eun Kim and Aldo Lipani. 2022. A multi-task based neural model to simulate users in goal oriented dialogue systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2115–2119.
[34] Antonios Minas Krasakis, Mohammad Aliannejadi, Nikos Voskarides, and Evangelos Kanoulas. 2020. Analysing the effect of clarifying questions on document ranking in conversational search. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval (ICTIR'20)*. 129–132.
[35] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *EMNLP-IJCNLP*. 540–551.
[36] Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087* (2019).
[37] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. 74–81.
[38] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2021. How am I doing?: Evaluating conversational search systems offline. *ACM TOIS* (2021).
[39] Tom Lotze, Stefan Klut, Mohammad Aliannejadi, and Evangelos Kanoulas. 2021. Ranking clarifying questions based on predicted user engagement. *CoRR* abs/2103.06192 (2021).
[40] Selina Meyer, David Elsweiler, Bernd Ludwig, Marcos Fernandez-Pichel, and David E. Losada. 2022. Do we still need human assessors? Prompt-based GPT-3 user simulation in conversational AI. In *Proceedings of the 4th Conference on Conversational User Interfaces (CUI'22)*. 6 pages.
[41] Nafise Sadat Moosavi, Angela Fan, Vered Shwartz, Goran Glavaš, Shafiq Joty, Alex Wang, and Thomas Wolf (Eds.). 2020. *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*. Association for Computational Linguistics.
[42] Javed Mostafa, Snehasis Mukhopadhyay, and Mathew Palakal. 2003. Simulation studies of different dimensions of users' interests and their impact on user modeling and information filtering. *Information Retrieval* 6, 2 (2003), 199–223.
[43] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *EMNLP*. 2241–2252.
[44] Paul Owoicho, Ivan Sekulić, Mohammad Aliannejadi, Jeff Dalton, and Fabio Crestani. 2023. Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. In *SIGIR*.
[45] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL*. 311–318.
[46] Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. *arXiv preprint arXiv:2002.12328* (2020).
[47] Gustavo Penha and Claudia Hauff. 2020. Challenges in the evaluation of conversational search systems. *KDD Workshop on Conversational Systems Towards Mainstream Adoption* (2020).

[48] Chen Qu, Liu Yang, W. Bruce Croft, Yongfeng Zhang, Johanne R. Trippas, and Minghui Qiu. 2019. User intent prediction in information-seeking conversations. In *CHIIR*. 25–33.
[49] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. (2019).
[50] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *CHIIR*. 117–126.
[51] Sudha Rao and Hal Daumé. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *ACL (1)*. 2736–2745.
[52] Sudha Rao and Hal Daumé III. 2019. Answer-based adversarial training for generating clarification questions. *arXiv:1904.02281* (2019).
[53] Gary Ren, Xiaochuan Ni, Manish Malik, and Qifa Ke. 2018. Conversational query understanding using sequence to sequence modeling. In *WWW*. 1715–1724.
[54] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten de Rijke. 2020. Conversations with search engines. *ACM Transactions on Information Systems* 1, 1 (2020).
[55] Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. 2020. Leading conversational search by suggesting useful questions. In *The Web Conference*. 1160–1170.
[56] Alexandre Salle, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. 2021. Studying the effectiveness of conversational search refinement through user simulation. In *ECIR*. 587–602.
[57] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *NAACL*. 1702–1723.
[58] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2020. Extending the use of previous relevant utterances for response ranking in conversational search. In *Proceedings of the 29th Text REtrieval Conference, TREC*.
[59] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2021. Towards facet-driven generation of clarifying questions for conversational search. In *ICTIR*.
[60] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2021. User engagement prediction for clarification in search. In *ECIR (1)*. 619–633.
[61] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating mixed-initiative conversational search systems via user simulation. In *WSDM'22: International Conference on Web Search and Data Mining* (Phoenix, AZ).
[62] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Exploiting document-based features for clarification in conversational search. In *ECIR*.
[63] Ivan Sekulić, Amir Soleimani, Mohammad Aliannejadi, and Fabio Crestani. 2020. Longformer for MS MARCO document re-ranking task. *arXiv preprint arXiv:2009.09392* (2020).
[64] Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *EMNLP*. 2210–2219.
[65] Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *CICLing*. 341–351.
[66] Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. *arXiv preprint arXiv:2105.03748* (2021).
[67] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *SIGIR*. 235–244.
[68] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *NeurIPS* 27 (2014), 3104–3112.
[69] Jean Tague, Michael Nelson, and Harry Wu. 1980. Problems in the simulation of bibliographic retrieval systems. In *SIGIR*. 236–255.
[70] Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to make context more useful? An empirical study on context-aware neural conversational models. In *ACL (2)*. 231–236.
[71] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[72] Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, and Shayne Longpre. 2021. A comparison of question rewriting methods for conversational passage retrieval. In *ECIR*.
[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).
[74] Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. 2018. Learning to ask questions in open-domain conversational systems with typed decoders. In *ACL (1)*. 2193–2203.
[75] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *SIGIR*. 55–64.
[76] Grace Hui Yang and Ian Soboroff. 2016. TREC 2016 dynamic domain track overview. In *TREC*.

[77] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *The Web Conference*. 418–428.
[78] Shuo Zhang and Krisztian Balog. 2020. Evaluating conversational recommender systems via user simulation. In *KDD*. 1512–1520.
[79] Shuo Zhang, Mu-Chun Wang, and Krisztian Balog. 2022. Analyzing and simulating user utterance reformulation in conversational recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 133–143.

Received 14 May 2023; accepted 7 February 2024