Simulating Follow-up Questions in Conversational Search

Johannes Kiesel[1], Marcel Gohsen[1], Nailia Mirzakhmedova[1], Matthias Hagen[2], and Benno Stein[1]

[1] Bauhaus-Universität Weimar, Bauhausstr. 9a, 99423 Weimar, Germany
<first-name>.<last-name>@uni-weimar.de
[2] Friedrich-Schiller Universität Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany
matthias.hagen@uni-jena.de

Abstract Evaluating conversational search systems based on simulated user interactions is a potential approach to overcome one of the main problems of static conversational search test collections: the collections contain only very few of all the plausible conversations on a topic. Still, one of the challenges of user simulation is generating realistic follow-up questions on given outputs of a conversational system. We propose to address this challenge by using state-of-the-art language models and find that: (1) on two conversational search datasets, the tested models generate questions that are semantically similar to those in the datasets, especially when tuned for follow-up questions; (2) the generated questions are mostly valid, related, informative, and specific according to human assessment; and (3) for influencing the characteristics of the simulated questions, small changes to the prompt are insufficient.

Keywords: Conversational search · Follow-up questions · Simulation

1 Introduction

Conversational search has been a focus of information retrieval research for more than a decade [4], yet many challenges remain, particularly in system evaluation. In conversational search, users iteratively deepen their knowledge through dialog with the search system. Unfortunately, such highly interactive conversations pose a challenge for system evaluation: The many plausible utterances from which the user and system choose each turn can lead to completely different conversations, but traditional test collections can only cover a few such conversation branches.

To acquire sufficient interaction data with a conversational search system for its evaluation—without expensive testing with real users—one can simulate users at a large scale [8]. With this evaluation methodology, a simulation system interacts with a search system to mimic a real user as closely as possible. Simulation parameters affect which actions the simulated user takes, enabling the simulation of a wide variety of users. Moreover, it allows for the repetition of a simulation with a different search system, facilitating A/B testing.

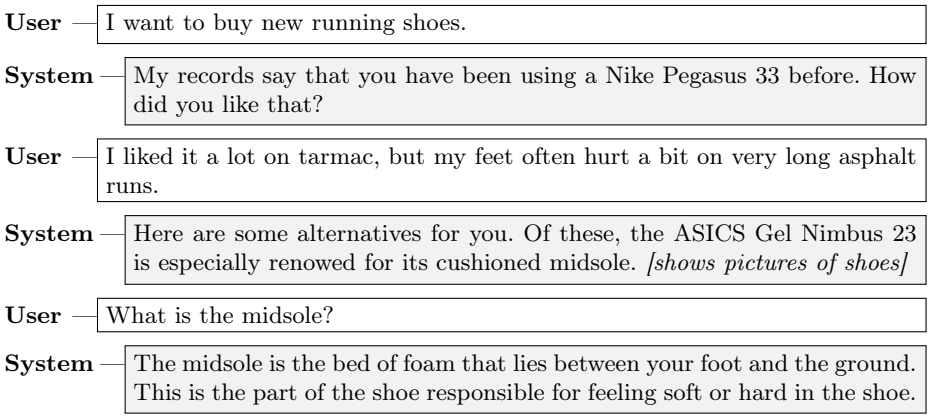


Figure 1. An envisioned dialogue between a user and a conversational search system, taken from Balog [8], including the follow-up question "What is the midsole?"

Particularly important for user simulation is to have the simulator pose realistic follow-up questions, which are questions about something the search system said earlier (cf. Figure 1). Follow-up questions are key for both conversational search [4] and conversational question answering [22], and are frequently contained in conversational search datasets: In TREC CAsT 2022 [27], we found that 54% of the user utterances after the initial one contain follow-up questions.

In this paper, we explore the potential of using language models to simulate user follow-up questions. Language models have recently reached an impressive level in mimicking human language use, especially in continuing conversations. Moreover, they can be tuned and/or instructed to conduct the discussions in specific manners, making them a natural choice for user simulation.

Specifically, we pose and address the following research questions on follow-up question generation through state-of-the-art language models:

1. Are the generated questions lexically or semantically similar to human-generated questions in conversational search datasets? (cf. Section 4)
2. According to human judgments, are the generated questions appropriate follow-up questions for the respective conversation? (cf. Section 5)
3. Can one adapt the prompt to steer the language model to generate questions according to different user profiles? (cf. Section 6)

In the following, Section 3 covers the task of follow-up question simulation, the selection of suitable datasets to analyze simulators, and the models that we explore for simulation. Section 4 compares the follow-up questions our models generate to those in the datasets using standard automated metrics, showing that they are semantically similar. Section 5 presents a human assessment of the simulated questions, showing that experts judge them as valid, related to the context, informative, and mostly specific. Section 6 investigates whether the models can be set to simulate specific users (e.g., naive or savvy ones) through small intuitive prompt modifications, and presents a negative result.

However, our results are mostly positive and highlight the promise of large language models for user simulation, even if the simulation of specific users requires further research (e.g., exploring fine-tuning or few-shot prompts).

2 Related Work

Our work is heavily inspired by the user simulator for conversational search proposed by Balog [8]. His proposed architecture is modular, featuring a personalized knowledge graph to implement a user model, an interaction model, and a mental model, as well as modules for response generation (planning, execution, learning) and both natural language understanding (for input) and generation (for output, alongside clicks). Our work fits within the generation module, but can, for example, also be integrated as a module in generic AI assistant platforms like DeepPavlov [42]. These use several independent modules to generate candidate responses for an input and one control model to select from these candidates. Simulators of both kinds can then interact with conversational search frameworks like Macaw [41] or DECAF [2] to evaluate specific conversational retrieval models. Similar to our work, Kim and Lipani [23] simulate user utterances for conversational search, but do not perform a human assessment of the simulation, nor can their model integrate with the aforementioned frameworks.

Concepts related to follow-up question simulation. Several concepts related to user simulation exist in the literature. Simulating user utterances is different to both simulating user feedback [28] and generating (for the system) clarification questions [3, 19, 30]. Our simulation is similar to question suggestion [33]. However, question suggestion attempts to find or generate useful next questions for a user to ask based on their previous question ("People also ask"), not the system's answer. Moreover, follow-up questions are different to question reformulations [9]: the former correspond to new questions based on what was said, whereas the latter are (small) changes to the previous question, usually intended to improve the question if the answer was not satisfactory. Conceptually, follow-up questions open a new context space in the conversation [31], whereas reformulations do not. Finally, follow-up questions seem related to but are actually independent to conversational action types. For example, each of the user actions defined by Azzopardi et al. [6] (reveal, inquire, navigate, interrupt, interrogate) can be performed using a follow-up question, but also other types of questions.

Recent approaches to evaluation in information retrieval. Ideas to replace the Cranfield paradigm of evaluation based on document collection, topics, and relevance judgments date back over two decades [35], but are only recently gaining more attention. Faggioli et al. [15] discuss pros and cons of automating relevance judgments for IR along a human–machine collaboration spectrum and provide an overview of existing ideas. Dietz et al. [12] and Dietz and Dalton [13] present a method for generating (non-conversational) test collections for various tasks from Wikipedia, and show that models are ranked similarly for a manual and their

generated collection. Another approach is to create test collections that contain questions along with the queries instead of relevance judgments. A document is considered more relevant to the query the more of the questions a language model can answer using the information from the document [34]. This approach could be paired with simulated users to evaluate conversational systems.

3 Simulating Follow-up Questions via Language Models

This section defines the task at hand (Section 3.1), selects suitable existing datasets (Section 3.2), and selects language models (Section 3.3) for our empirical investigations on simulating follow-up questions in the later sections.

3.1 Defining Follow-up Question Simulation

For user simulation in conversational search, as described by Balog [8], we define the task of follow-up question simulation as follows:

> **Task**: Given an informative textual response to a user's query, generate a question that the user might ask based on the provided information.

A deliberately vague point of our definition is the restriction to follow-up questions *that the user might ask*. Balog [8] provides a list of several factors that influence what a user might ask: current personal interests and preferences, persona (personality, educational and socio-economical background, etc.), current knowledge, and current understanding of the system's capabilities. In this initial paper, we simplify matters by concentrating our main experiments on a generic user, but make a first attempt at user modeling in Section 6. We also refrain from instructing the simulator on which point exactly to follow up on, both for simplicity and as it is currenlty unclear whether user simulators should do so.

3.2 Selecting Datasets for Follow-up Question Simulation

We select TREC CAsT 2022 [27] and Webis-Nudged-Questions-23 [16] for our experiments from a larger pool of datasets that we reviewed. Neither the MS MARCO conversational [7] nor the Webis-Conversational-Query-Reformulations-21 [21] contain system responses. The questions in the SCAI-QReCC dataset [5, 38] are on texts (generated from search results) that are unavailable. The Wizard of Wikipedia dataset [14] is not task-oriented. The MultiWOZ dataset [10, 40] focuses heavily on transactions (e.g., hotel booking). The Webis-Exhibition-Questions-21 [20] contains questions on elements displayed in a virtual environment, which goes beyond the scope of our study.

The two selected datasets focus on conversational search and contain system responses and subsequent user utterances. The TREC CAsT 2022 dataset is the newest in a series of four datasets created for the TREC CAsT shared task [27]. It includes 205 unique turns over 50 conversations. We manually identified all follow-up questions in the user utterances, with 26 targeting a specific

system response preceding them; these questions were matched with their respective system responses for our experiments. We utilize 100 system responses that received follow-up questions. The Webis-Nudged-Questions-23 [16] contains 8376 crowdsourced questions to 30 short informative texts on three topics (argument search, exhibition, and product search). We employ all 30 texts as system responses and a random sample of 30 questions per response as user utterances.

3.3 Selecting and Tuning Models for Follow-up Question Simulation

We've chosen GPT-4 [26] as the latest successor to Chat-GPT, and the Llama models as robust open-source alternatives with strong performance [37]. For Llama, we compare models of different sizes (Llama2-7B with 7 billion parameters and Llama2-13B) and versions (Alpaca-7B, a tuned Llama1 [36]). We found Llama2-chat models to perform similarly and thus discarded them.

Figure 2 shows the prompt that we use for all models, with variable texts highlighted. The prompt reflects the instructions given to crowdworkers to generate follow-up questions for the Inquisitive dataset [24] (there called questions-under-discussion and used for monological texts, but otherwise equivalent).

To further adapt the models to follow-up question simulation, we employ instruction fine-tuning for Llama models using the same prompt. Fine-tuning GPT-4 is not available at the time of writing. We fine-tune on three datasets: TREC CAsT 2022 and Webis-Nudged-Questions-23, also used for evaluation, and Inquisitive (see above) as a generic non-conversational dataset. For the first two datasets, we ensure that the model is not tuned on the same conversation it's generating questions for, using a 3-fold cross-validation approach.

For fine-tuning and generation we use the HuggingFace Transformers library [39] on a single NVIDIA A100 GPU (40 GB). We use standard low-rank adaptation [17] for efficient fine-tuning, applying low-rank updates ($r = 64$) with a scaling parameter of $\alpha = 16$ to all linear layers. We fine-tune each model for one epoch, as further epochs did not reduce the loss significantly. We used a batch size of 4 and a learning rate of $2 \cdot 10^{-4}$.

In our evaluation, we assess zero-shot generation of both the original and fine-tuned models.[3] To illustrate the simulated follow-up questions, the Appendix shows examples and the most common leading bigrams for some models.

4 Comparing Simulated and Original Human Questions

As a first approach to check whether the questions that large language models generate are questions a user might ask (as per our task definition in Section 3.1), we analyze whether the questions are similar to the ones that humans asked in the respective dataset. This corresponds to our first research question: Are the generated questions lexically or semantically similar to human-generated questions in conversational search datasets?

[3] Code, data, and models are available at https://github.com/webis-de/ECIR-24.

```
### Instruction: Follow-up questions are the questions elicited from
readers as they naturally read through text. You are a savvy user. You
ask elaborate questions about the implications of what was being said.
Given the text below, write follow-up questions that you would ask if you
were reading this text for the first time.

### Text: The nation's largest gun-rights group is taking some Texans to
task over their headline-generating demonstrations advocating the legal,
open carrying of weapons.

### Follow-up questions:
```

Figure 2. Example of a prompt we employ to simulate a user asking questions. Text with a red background is always adapted to the respective conversation. Text with a purple background is used to model different users (cf. Section 6) and is blank for the other experiments.

4.1 Similarity Computation

We assess both lexical (i.e., same word sequence) and semantic similarity (similar meaning) between simulated and human questions. For both we use standard metrics from the machine translation and paraphrasing literature.

For lexical similarity, we employ the standard metric in machine translation, BLEU [29]. BLEU computes word n-gram precision between a candidate text and a set of reference texts. We compute BLEU up to 4-grams and apply Smoothing 4 from Chen and Cherry [11] to prevent inflation of BLEU for short questions.

For semantic similarity, we employ a nowadays commonly used metric in paraphrasing, Sentence-BERT [32]. This method embeds both simulated and human generated sentences with Sentence-BERT along with TinyBERT [18], as TinyBERT embeddings are tuned for natural language understanding and the embedding process is reasonably efficient. The semantic similarity is then the cosine similarity of the embedding vectors, as it is standard for embeddings.

However, many follow-up questions can be imagined for each system response, and we thus generate several questions per model and compare these to all human-generated questions in the datasets (cf. Section 3.2). We repeat the process if no question is generated up to five times. We then calcutate the overall score for a set of questions simulated by a model for a dataset, $\hat{Q}$, and the human questions in the same dataset, Q, for a similarity measure φ (BLEU or Sentence-BERT). Let $Q_{\mathrm{s}(\hat{q})}$ be the set of questions within the human questions Q that are asked for the same system response as a simulated question $\hat{q}$. Then, the overall score is the average similarity across each simulated question $\hat{q}$ to its most similar question $q \in Q_{\mathrm{s}(\hat{q})}$, with similarity measured according to φ:

$$\mathrm{score}_{\varphi}(\hat{Q}, Q) = \frac{\sum_{\hat{q} \in \hat{Q}} \left(\max_{q \in Q_{\mathrm{s}(\hat{q})}} \varphi(\hat{q}, q) \right)}{|\hat{Q}|}$$

To reduce randomness, we report mean scores for 10 generation runs each.

Table 1. Scores for BLEU and Sentence-BERT for simulated questions on TREC CAsT 22 (CAsT) and Webis-Nudged-Questions-23 (WNQ). Reported values are the means of 10 runs. All values from 0 (no similarity) to 1 (identical), with best values in each column marked bold.

Model		BLEU		Sent.-BERT	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	0.02	0.11	0.22	0.68
Alpaca-7B	none	0.03	0.14	**0.23**	0.70
Alpaca-7B	CAsT	0.03	0.08	0.20	0.46
Alpaca-7B	Inquisitive	0.03	0.13	0.21	0.63
Alpaca-7B	WNQ	0.03	0.13	0.22	0.66
Llama2-7B	none	0.03	0.18	0.18	0.63
Llama2-7B	CAsT	**0.04**	0.09	0.19	0.45
Llama2-7B	Inquisitive	0.03	0.20	0.20	0.70
Llama2-7B	WNQ	0.03	0.21	0.20	**0.71**
Llama2-13B	none	0.03	0.19	0.21	0.66
Llama2-13B	CAsT	**0.04**	0.07	0.20	0.41
Llama2-13B	Inquisitive	**0.04**	**0.23**	0.22	0.68
Llama2-13B	WNQ	0.03	0.22	0.20	0.70

4.2 Results

For lexical similarity ($\varphi =$ BLEU), Table 1 shows that scores are very low for the TREC CAsT dataset for all models (between 0.02 and 0.04), but higher for Webis-Nudged-Questions-23 (between 0.07 and 0.23). Higher scores for Webis-Nudged-Questions-23 are expected, since this dataset has on average 30 human-generated questions per system response that are matched with each simulated question, whereas for TREC CAsT there are only 1.3 on average. However, we still conclude that the lexical similarity of the simulated questions to the human questions is quite low for both datasets. Moreover, we find that models fine-tuned on the TREC CAsT dataset generate questions for Webis-Nudged-Questions-23 that are even more dissimilar than those generated without fine-tuning. Interestingly, fine-tuning on the Inquisitive dataset seems to have no strong negative effect, and in some cases even leads to the most similar questions.

For semantic similarity ($\varphi =$ Sentence-BERT), Table 1 shows much higher scores. Llama2-7B tuned on Webis-Nudged-Questions-23 achieves a Sentence-BERT value of 0.71 on the same dataset. Moreover, we observe the same performance decrease when fine-tuning on TREC CAsT as when fine-tuning on Webis-Nudged-Questions-23, indicating that the questions in the datasets are both lexically and semantically dissimilar.

In summary, we find that the simulated questions rarely match human questions when it comes to lexical similarity, but much more so for semantic similarity. However, one has to consider that there are on average only 1.3 human-generated questions per system response for TREC CAsT, yet many follow-up questions are plausible for each system response, which naturally reduces the score. Therefore, we conclude that the simulated questions are relatively similar, at least in their meaning, to the questions that humans ask.

5 Judging Simulated Questions

As a second approach to determine whether the questions generated by large language models are questions users might ask (as per our task definition in Section 3.1), we employ human experts to evaluate the questions. This corresponds to our second research question: According to human judgments, are the generated questions appropriate follow-up questions for the respective conversation?

5.1 Human Judgment of Simulated Questions

To analyze whether a simulated utterance is an appropriate follow-up question, we adopt three criteria employed by Ko et al. [24] to evaluate the questions in the Inquisitive dataset (cf. Section 3.2) and extend it with the criterion of specificity suggested by Adiwardana et al. [1] for conversational systems. Specifically, we ask human judges the following for each simulated utterance: (1) Is it a valid question? An invalid question is incomplete, incomprehensible, or not even a question at all; (2) If it is a valid question, is it also related to the context, i.e., the system response?; (3) If it is a valid and related question, is it also an informative question? An answer for an informative question is not contained in the system response; and (4) If it is a valid, related, and informative question, is it also a specific question, i.e., not a question that could be asked for any system response? For example, a question to the effect of "tell me more" is not specific.

These binary questions are answered by three of the authors as experts on the matter. Each expert has a background in natural language processing and having inspected the Inquisitive dataset to gain an exact understanding of the task. However, to avoid annotation biases, the experts did not know which model produced an utterance or whether it was taken from the dataset, and they received utterances in a random order. We selected one question from each of the 13 models used in the previous experiment and paired it with the original questions in the dataset, yielding 1820 simulated responses and 7280 judgments. Since there is some subjectivity to the judgments, we ensured consistency that the same expert judged all utterances that were simulated for a system response. Finally, for assessing agreement, the three expert annotators independently evaluated all simulated utterances for one single TREC CAsT conversation. The agreement, measured by Fleiss κ, was found to be moderate for each rating, which we deemed acceptable for this task: 0.51 for "valid," 0.52 for "related," 0.53 for "informative," and 0.57 for "specific."

5.2 Results

Table 2 summarizes the evaluation results for different datasets and criteria. Notably, GPT-4 consistently outperforms other models in all dataset-criterion combinations. Nonetheless, all models produce valid questions—even more often than the crowdworkers employed for generating the Webis-Nudged-Questions-23. The drop in performance is minimal when considering relatedness, suggesting that current language models excel in these tasks. However, GPT-4 stands

Table 2. Ratio of simulated questions judged as valid, related, informative, and specific for the TREC CAsT 22 (CAsT) and Webis-Nudged-Questions-23 (WNQ). Each judgment implies the judgments to its left. Highest ratio in each column marked bold.

Model		Valid		Related		Informative		Specific	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	**0.98**	0.97	0.97	**0.97**	**0.84**	**0.87**	**0.84**	**0.87**
Alpaca-7B	none	0.93	0.97	0.93	0.93	0.73	0.63	0.72	0.63
Alpaca-7B	CAsT	0.92	0.87	0.85	0.80	0.80	0.80	0.72	0.70
Alpaca-7B	Inquisitive	0.94	0.80	0.92	0.80	0.77	0.67	0.75	0.67
Alpaca-7B	WNQ	0.96	0.77	0.94	0.77	0.75	0.67	0.75	0.67
Llama2-7B	none	0.92	0.80	0.84	0.77	0.60	0.50	0.57	0.47
Llama2-7B	CAsT	0.94	0.93	0.84	0.70	0.76	0.57	0.73	0.43
Llama2-7B	Inquisitive	0.95	0.87	0.94	0.83	0.73	0.77	0.72	0.77
Llama2-7B	WNQ	0.96	**1.00**	0.94	0.93	0.65	0.63	0.65	0.63
Llama2-13B	none	0.90	0.93	0.88	0.90	0.57	0.50	0.52	0.43
Llama2-13B	CAsT	0.87	0.90	0.79	0.77	0.71	0.73	0.63	0.57
Llama2-13B	Inquisitive	**0.98**	0.90	**0.98**	0.83	0.77	0.67	0.75	0.67
Llama2-13B	WNQ	0.94	0.97	0.89	0.93	0.58	0.60	0.58	0.57
Original questions	-	0.95	0.60	0.91	0.50	0.87	0.40	0.77	0.40

out in terms of informativeness. It is worth noting that the models without fine-tuning consistently perform worse than their tuned counterparts. The utterances, judged as informative, can be considered valid follow-up questions. As the last two columns show, most such questions are also specific, although the models fine-tuned on TREC CAsT occasionally produce unspecific questions. This result is not surprising, given that TREC CAsT dataset itself contains unspecific questions (cf. original questions).

In summary, we find that, for the best models, the simulated utterances are often valid follow-up questions. GPT-4 is the best model for simulation and close to human performance (better than crowdworkers), but also Llama models perform well, especially when fine-tuned. Moreover, fine-tuning can be used to adapt the model to ask fewer specific questions. Therefore, we conclude that simulated questions are often valid follow-up questions as per human judgment.

6 Modeling Specific Users through Prompt Modifications

As a third approach to assess whether the questions generated by large language models align with what users might ask (outlined in Section 3.1), we simulate distinct user profiles by modifying the model prompt. Human experts are then asked to evaluate whether the generated questions accurately reflect these modifications. We purposefully attempt to simulate specific users by only changing the model prompt, as prompt modifications are a cost-effective strategy and necessitate no additional training data. Thus, if prompt adjustments prove effective in representing various user perspectives, language models can readily simulate a wide array of users with minimal effort. This corresponds to our third research question: Can one adapt the prompt to steer the language model to generate questions according to different user profiles?

Table 3. Ratio of simulated questions judged as asked by a naive, savvy, implication-focused, or reasons-focused user for the TREC CAsT 22 (CAsT) and Webis-Nudged-Questions-23 (WNQ). Ratios higher than for the same model without prompt modification marked bold.

Model			Naive		Savvy		Implications		Reasons	
Base	Tuning	Prompt	CAsT	WNQ	CAsT	WNQ	CAsT	WNQ	CAsT	WNQ
GPT-4	none		0.07	0.23	0.29	0.33	0.14	0.17	0.13	0.27
GPT-4	none	Naive+Implic.	**0.11**	0.20	0.29	**0.37**	**0.15**	**0.23**	**0.16**	0.17
GPT-4	none	Naive+Reasons	**0.20**	0.23	0.23	0.33	0.12	**0.40**	**0.34**	0.27
GPT-4	none	Savvy+Implic.	0.01	0.03	**0.54**	**0.77**	**0.25**	**0.50**	0.12	0.17
GPT-4	none	Savvy+Reasons	**0.08**	0.17	**0.34**	**0.60**	**0.19**	**0.53**	**0.23**	0.20
Alpaca-7B	none		0.16	0.23	0.22	0.20	0.19	0.30	0.11	0.10
Alpaca-7B	none	Naive+Implic.	0.15	0.13	0.15	**0.40**	**0.22**	0.30	0.09	0.10
Alpaca-7B	none	Naive+Reasons	0.11	0.20	0.18	**0.27**	0.19	**0.33**	0.07	0.10
Alpaca-7B	none	Savvy+Implic.	**0.18**	0.13	**0.27**	**0.53**	**0.20**	**0.47**	**0.15**	0.07
Alpaca-7B	none	Savvy+Reasons	0.07	0.17	**0.24**	**0.27**	0.19	0.30	**0.16**	0.07
Llama2-13B	none		0.52	0.40	0.01	0.03	0.06	0.07	0.12	0.13
Llama2-13B	none	Naive+Implic.	0.45	**0.57**	0.01	0.00	0.06	**0.20**	0.09	0.10
Llama2-13B	none	Naive+Reasons	0.47	**0.47**	**0.03**	0.03	0.06	**0.13**	**0.22**	**0.23**
Llama2-13B	none	Savvy+Implic.	0.38	0.33	**0.06**	0.03	**0.19**	**0.13**	0.06	0.03
Llama2-13B	none	Savvy+Reasons	0.43	0.20	**0.03**	**0.07**	**0.07**	0.07	**0.19**	0.13
Llama2-13B	Inquisitive		0.50	0.40	0.02	0.10	0.03	0.20	0.19	0.23
Llama2-13B	Inquisitive	Naive+Implic.	0.40	**0.70**	0.01	0.07	**0.06**	**0.27**	**0.30**	**0.27**
Llama2-13B	Inquisitive	Naive+Reasons	0.43	**0.70**	**0.07**	0.00	**0.06**	0.17	**0.38**	0.20
Llama2-13B	Inquisitive	Savvy+Implic.	0.45	**0.70**	0.02	0.03	**0.04**	**0.23**	**0.36**	**0.30**
Llama2-13B	Inquisitive	Savvy+Reasons	0.42	**0.57**	**0.03**	0.10	**0.05**	0.20	**0.30**	0.17
Original questions			0.42	0.63	0.05	0.13	0.14	0.20	0.12	0.07

6.1 Prompt Modifications

From the vast number of possible user attributes we selected two dimensions for our experiment: (1) naive vs. savvy user, corresponding to terms used in classical user simulation [25]; and (2) users focusing on questions about implications vs. reasons of something in the system response. We modify the prompt (Figure 2) by adding the text: "You are a [savvy/naive] user. You ask [simple/elaborate] questions about the [implications/reasons] of what was being said." We use the same evaluation setup and experts as in Section 5 to have the experts judge if a given utterance aligns with the specified user type. Measured in the same way, Fleiss κ shows substantial agreement for rating questions as from a user that is focused on "implications" ($\kappa = 0.75$) and "savvy" ($\kappa = 0.63$), moderate agreement for focused on "reasons" ($\kappa = 0.46$), and fair agreement for "naive" ($\kappa = 0.37$).

6.2 Results

Due to time constraints, our experts could only evaluate 4 out of the 14 different models for each combination of savvy/naive and implications/reasons. As shown in Table 3, our attempts to enhance the simulation through minor prompt adjustments were not successful. Although the prompt modifications did clearly affect the simulation, the observed effects are not consistent with our hypothesis. Especially for Llama2-13B fine-tuned on the Inquisitive dataset, there is a

significant increase in certain ratios, but this increase was limited to only one of the datasets. It appears that GPT-4 may be the most effective model for identifying savvy or naive users, as it generated more responses aligned with the prompt, particularly on the TREC CAsT dataset. However, this effect was not as pronounced for users focusing on implications or reasons.

In summary, we find that small modifications to the prompt are insufficient to steer the simulation towards specific user attributes. Of course, our experimental setup is limited: different modifications or different attributes could yield improved results. However, the attractiveness of small prompt modifications lies in their simplicity of implementation. Our results indicate that, at least with the tested models, this straightforward way of modeling users is not yet feasible.

7 Conclusion

User simulation is a promising yet hypothetical approach to the evaluation of conversational search systems, addressing the drawbacks of static test collections for a highly interactive task. This paper presents another step towards a complete user simulation—the simulation of follow-up questions to system responses. As per the literature, follow-up questions are frequent and of key importance in conversational search. We showed that large language models are capable of simulating users asking follow-up questions. The semantic similarity (Sentence-BERT) to human-generated questions reaches as high as 0.71 for one of the two conversational search datasets we tested on. Moreover, human experts judged the simulated questions in blind evaluation to be mostly valid, related to the system response, and informative. Furthermore, we found that fine-tuning models to datasets, even if they are out-of-domain, can improve the simulation—more so than using larger models. While GPT-4 is ahead of the open models in our benchmark, nearly matching human performance, the gap is not excessive. However, we also presented a negative result: although the prompt interface to language models suggests that modifications to the prompt could be used to alter the simulation to represent different users, we found that our slight modifications were insufficient and failed to control the simulation as intended.

Nonetheless, our results are mostly positive and highlight the promise of large language models for user simulation, even if the simulation of specific users requires further research. For example, instead of prompt-modification one could explore fine-tuning or few-shot prompts. Both methods attempt to mimic a user based on a few example questions. The latter adds these examples to the prompt, which is more direct, but limited to only a few examples. Another venue for research is to create a dataset of users to be then used in simulation, possibly by extracting user attributes from existing conversations [43].

Although many questions remain open, our results provide further evidence of the potential of user simulation to evaluate conversational search systems. Furthermore, our method is not restricted to the simulation of follow-up questions, and can be adapted to simulate other user interactions in the future.

8 Limitations

The question of how to simulate users of a conversational search system has many facets, many of which we could not address in this paper. Even for the method of language models, which we focused on in this paper, we could not explore the entire parameter space for the simulation. We approached the task with both zero-shot and fine-tuning, but not with the middle ground of few-shot learning (also called in-context learning). In terms of modeling specific users, this work has barely scratched the surface of what is possible. Going back to the idea of the personal knowledge graph from Balog [8], one could also use methods that integrate such knowledge graphs into language models to model different users. Furthermore, we only used two different datasets, which naturally cannot represent the many different scenarios in which a user might search—for many of which no dataset currently exists. Finally, we did not test our simulator with an actual retrieval system, but evaluated the simulation as it continues a human conversation (for TREC CAsT). Ideally, the language model picked up the conversation and continued it naturally, but we have not evaluated whether it actually did so, nor do we know of any evaluation metrics for checking this.

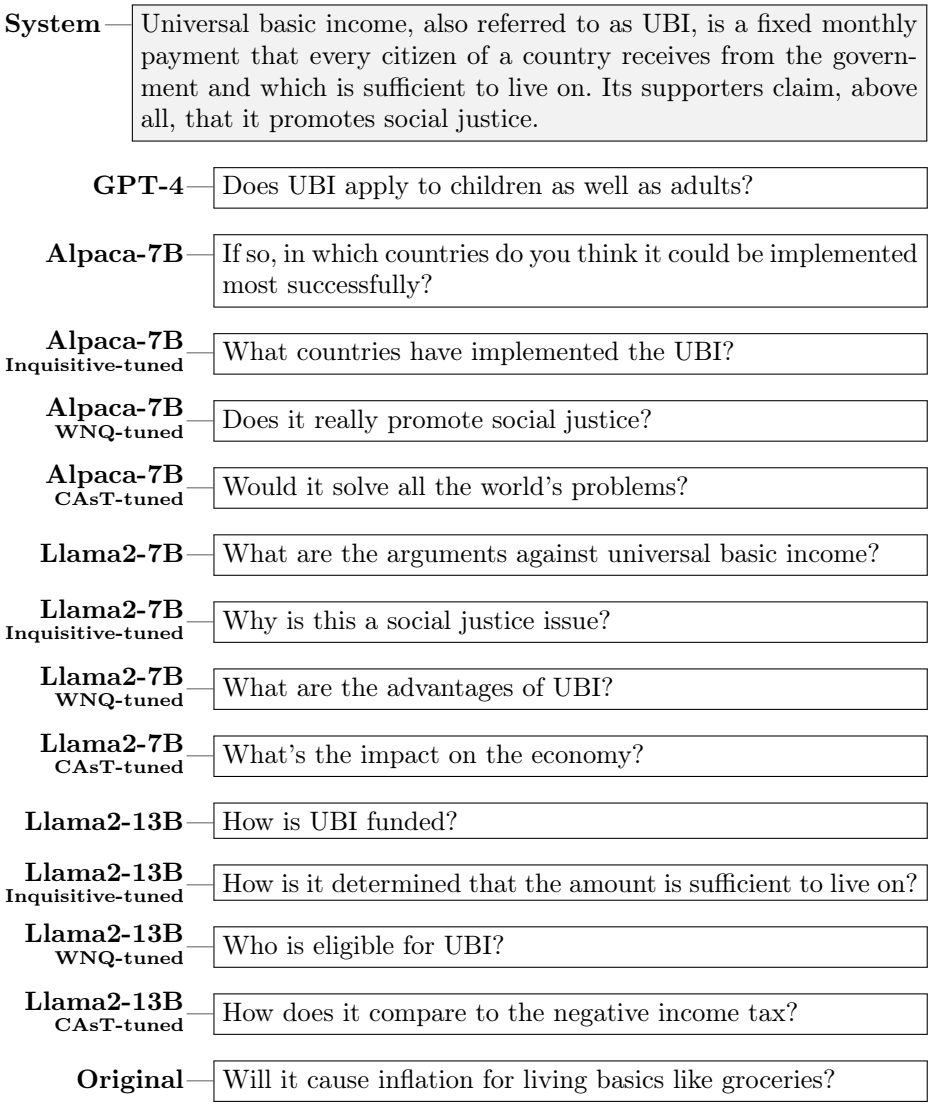
Acknowledgements This work was partially supported by the European Commission under grant agreement GA 101070014 (https://openwebsearch.eu)

Appendix

Most frequent leading bigrams (lemmatized) and their frequency for original questions and questions simulated by selected models (IT=Inquitisitive-tuned).

Rank	Original		Model					
			Llama2-7B		Llama2-7B (IT)		Llama2-13B	
For TREC CAsT 22								
1	what [be]	0.12	what [be]	0.62	what [be]	0.38	what [be]	0.70
2	tell [i]	0.11	what [do]	0.15	why [be]	0.17	what [do]	0.06
3	how [do]	0.06	how [do]	0.04	how [do]	0.14	how [do]	0.04
4	what [make]	0.05	why [do]	0.03	why [do]	0.07	how [can]	0.02
5	so [what]	0.03	why [be]	0.02	how [be]	0.05	why [be]	0.02
6	can [you]	0.02	who [be]	0.02	how [can]	0.03	who [be]	0.02
7	how [be]	0.02	how [many]	0.01	why [would]	0.02	why [do]	0.02
8	let [we]	0.02	how [can]	0.01	how [much]	0.02	how [be]	0.01
9	how [can]	0.02	do [you]	0.01	what [do]	0.02	when [be]	0.01
10	what [about]	0.02	what [would]	0.01	how [many]	0.01	how [many]	0.01
For Webis-Nudged-Questions-23								
1	what [be]	0.21	what [be]	0.61	what [be]	0.40	what [be]	0.69
2	why [be]	0.03	what [do]	0.17	how [do]	0.15	what [do]	0.05
3	how [do]	0.03	how [do]	0.04	why [be]	0.14	who [be]	0.04
4	how [much]	0.03	who [be]	0.03	how [be]	0.06	how [do]	0.04
5	who [be]	0.03	why [do]	0.02	why [do]	0.06	why [be]	0.02
6	why [do]	0.03	why [be]	0.02	how [would]	0.04	why [do]	0.02
7	how [many]	0.02	how [much]	0.01	why [would]	0.04	how [much]	0.01
8	how [be]	0.01	how [can]	0.01	how [much]	0.03	when [be]	0.01
9	what [do]	0.01	how [be]	0.01	what [do]	0.01	how [can]	0.01
10	what [other]	0.01	why [would]	0.01	how [many]	0.01	how [would]	0.01

Original follow-up question from the dataset and simulated questions by model.



Bibliography

[1] Adiwardana, D., Luong, M., So, D.R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., Le, Q.V.: Towards a human-like open-domain chatbot. CoRR **abs/2001.09977** (2020), URL https://arxiv.org/abs/2001.09977

[2] Alessio, M., Faggioli, G., Ferro, N.: DECAF: A modular and extensible conversational search framework. In: 46th International ACM SIGIR Conference

on Research and Development in Information Retrieval (SIGIR 2023), ACM (2023), https://doi.org/10.1145/3539618.3591913
[3] Aliannejadi, M., Zamani, H., Crestani, F., Croft, W.B.: Asking clarifying questions in open-domain information-seeking conversations. In: Piwowarski, B., Chevalier, M., Gaussier, É., Maarek, Y., Nie, J., Scholer, F. (eds.) 42th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019), pp. 475–484, ACM (2019), https://doi.org/10.1145/3331184.3331265
[4] Allan, J., Croft, W.B., Moffat, A., Sanderson, M.: Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne. SIGIR Forum **46**(1), 2–32 (2012), https://doi.org/10.1145/2215676.2215678
[5] Anantha, R., Vakulenko, S., Tu, Z., Longpre, S., Pulman, S., Chappidi, S.: Open-domain question answering goes conversational via question rewriting. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tür, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021), pp. 520–534, Association for Computational Linguistics (2021), https://doi.org/10.18653/v1/2021.naacl-main.44
[6] Azzopardi, L., Dubiel, M., Halvey, M., Dalton, J.: Conceptualizing agent-human interactions during the conversational search process. In: Spina, D., Arguello, J., Joho, H., Kiseleva, J., Radlinski, F. (eds.) 2nd International Workshop on Conversational Approaches to Information Retrieval (CAIR 2018), ACM (Jul 2018)
[7] Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., Wang, T.: MS MARCO: A human generated machine reading comprehension dataset. CoRR **abs/1611.09268** (2016), https://doi.org/10.48550/arXiv.1611.09268, URL http://arxiv.org/abs/1611.09268
[8] Balog, K.: Conversational AI from an information retrieval perspective: Remaining challenges and a case for user simulation. In: Alonso, O., Marchesin, S., Najork, M., Silvello, G. (eds.) 2nd International Conference on Design of Experimental Search & Information REtrieval Systems (DESIRES 2021), CEUR Workshop Proceedings, vol. 2950, pp. 80–90, CEUR-WS.org (2021)
[9] Boldi, P., Bonchi, F., Castillo, C., Vigna, S.: Query reformulation mining: models, patterns, and applications. Information Retrieval **14**(3), 257–289 (2011), https://doi.org/10.1007/S10791-010-9155-3
[10] Budzianowski, P., Wen, T.H., Tseng, B.H., Casanueva, I., Ultes, S., Ramadan, O., Gasic, M.: MultiWOZ - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), pp. 5016–5026, Association for Computational Linguistics (2018)
[11] Chen, B., Cherry, C.: A systematic comparison of smoothing techniques for sentence-level BLEU. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 362–367, Association for Computational Linguistics, Baltimore, Maryland, USA (Jun 2014), https://doi.org/10.3115/v1/W14-3346

[12] Dietz, L., Chatterjee, S., Lennox, C., Kashyapi, S., Oza, P., Gamari, B.: Wikimarks: Harvesting relevance benchmarks from wikipedia. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022), pp. 3003–3012, ACM (2022), https://doi.org/10.1145/3477495.3531731

[13] Dietz, L., Dalton, J.: Humans optional? Automatic large-scale test collections for entity, passage, and entity-passage retrieval. Datenbank-Spektrum **20**(1), 17–28 (2020), https://doi.org/10.1007/s13222-020-00334-y

[14] Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., Weston, J.: Wizard of wikipedia: Knowledge-powered conversational agents. In: 7th International Conference on Learning Representations (ICLR 2019), OpenReview.net (2019)

[15] Faggioli, G., Dietz, L., Clarke, C.L.A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., Wachsmuth, H.: Perspectives on large language models for relevance judgment. CoRR **abs/2304.09161** (2023), https://doi.org/10.48550/arXiv.2304.09161

[16] Gohsen, M., Kiesel, J., Korashi, M., Ehlers, J., Stein, B.: Guiding oral conversations: How to nudge users towards asking questions? In: ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2023), pp. 34–42, ACM, New York (Mar 2023), https://doi.org/10.1145/3576840.3578291

[17] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: 10th International Conference on Learning Representations (ICLR 2022), OpenReview.net (2022)

[18] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: TinyBERT: Distilling BERT for natural language understanding. CoRR **abs/1909.10351** (2019)

[19] Kiesel, J., Bahrami, A., Stein, B., Anand, A., Hagen, M.: Toward Voice Query Clarification. In: 41st International ACM Conference on Research and Development in Information Retrieval (SIGIR 2018), pp. 1257–1260, ACM (Jul 2018), https://doi.org/10.1145/3209978.3210160, URL https://dl.acm.org/doi/10.1145/3209978.3210160

[20] Kiesel, J., Bernhard, V., Gohsen, M., Roth, J., Stein, B.: What is that? Crowdsourcing questions to a virtual exhibition. In: Elsweiler, D. (ed.) 2022 Conference on Human Information Interaction & Retrieval (CHIIR 2022), pp. 358–362, ACM (Mar 2022), https://doi.org/10.1145/3498366.3505836

[21] Kiesel, J., Cai, X., Baff, R.E., Stein, B., Hagen, M.: Toward conversational query reformulation. In: Alonso, O., Najork, M., Silvello, G. (eds.) 2nd International Conference on Design of Experimental Search & Information Retrieval Systems (DESIRES 2021), CEUR Workshop Proceedings, vol. 2950, pp. 91–101 (Sep 2021)

[22] Kim, G., Kim, H., Park, J., Kang, J.: Learn to resolve conversational dependency: A consistency training framework for conversational question answering. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP 2021), pp. 6130–6141, Association for Computational Linguistics (2021), https://doi.org/10.18653/v1/2021.acl-long.478

[23] Kim, T.E., Lipani, A.: A multi-task based neural model to simulate users in goal oriented dialogue systems. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) 45th International ACM SIGIR Conference

on Research and Development in Information Retrieval (SIGIR 2022), pp. 2115–2119, ACM (2022), https://doi.org/10.1145/3477495.3531814
[24] Ko, W.J., Chen, T.Y., Huang, Y., Durrett, G., Li, J.J.: Inquisitive question generation for high level text comprehension. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), pp. 6544–6555, Association for Computational Linguistics (2020), https://doi.org/10.18653/v1/2020.emnlp-main.530
[25] Maxwell, D., Azzopardi, L.: Information scent, searching and stopping - modelling SERP level stopping behaviour. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) Advances in Information Retrieval - 40th European Conference on IR Research (ECIR 2018), Lecture Notes in Computer Science, vol. 10772, pp. 210–222, Springer (2018), https://doi.org/10.1007/978-3-319-76941-7_16
[26] OpenAI: GPT-4 technical report (2023), https://doi.org/10.48550/arXiv.2303.08774
[27] Owoicho, P., Dalton, J., Aliannejadi, M., Azzopardi, L., Trippas, J., Vakulenko, S.: TREC CAsT 2022: Going beyond user ask and system retrieve with initiative and response generation. In: Voorhees, E.M., Ellis, A. (eds.) 31st Text REtrieval Conference (TREC 2022), NIST Special Publication, National Institute of Standards and Technology (2022)
[28] Owoicho, P., Sekulic, I., Aliannejadi, M., Dalton, J., Crestani, F.: Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. In: Chen, H.H., Duh, W.J.E., Huang, H.H., Kato, M.P., Mothe, J., Poblete, B. (eds.) 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023), pp. 632–642, ACM (2023), https://doi.org/10.1145/3539618.3591683
[29] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002), https://doi.org/10.3115/1073083.1073135
[30] Rao, S., III, H.D.: Answer-based adversarial training for generating clarification questions. In: Burstein, J., Doran, C., Solorio, T. (eds.) 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), pp. 143–155, Association for Computational Linguistics (2019), https://doi.org/10.18653/V1/N19-1013
[31] Reichman, R.: Getting Computers to Talk Like You and Me: Discourse Context, Focus, and Semantics:(an ATN Model). MIT press (1985)
[32] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (Aug 2019)
[33] Rosset, C., Xiong, C., Song, X., Campos, D., Craswell, N., Tiwary, S., Bennett, P.N.: Leading conversational search by suggesting useful questions. In: Huang, Y., King, I., Liu, T., van Steen, M. (eds.) The Web Conference 2020 (WebConf 2020), pp. 1160–1170, ACM / IW3C2 (2020), https://doi.org/10.1145/3366423.3380193
[34] Sander, D.P., Dietz, L.: EXAM: How to evaluate retrieve-and-generate systems for users who do not (yet) know what they want. In: Alonso, O., Marchesin, S., Najork, M., Silvello, G. (eds.) 2nd International Conference on Design of Experimental Search & Information REtrieval Systems (DESIRES 2021), CEUR Workshop Proceedings, vol. 2950, pp. 136–146, CEUR-WS.org (2021)

[35] Soboroff, I., Nicholas, C.K., Cahan, P.: Ranking retrieval systems without relevance judgments. In: Croft, W.B., Harper, D.J., Kraft, D.H., Zobel, J. (eds.) Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001), pp. 66–73, ACM (2001), https://doi.org/10.1145/383952.383961

[36] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023)

[37] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and efficient foundation language models. CoRR **abs/2302.13971** (2023), https://doi.org/10.48550/arXiv.2302.13971

[38] Vakulenko, S., Kiesel, J., Fröbe, M.: SCAI-QReCC shared task on conversational question answering. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., Piperidis, S. (eds.) 14th Language Resources and Evaluation Conference (LREC 2022), pp. 4913–4922, European Language Resources Association (ELRA), Paris, France (Jun 2022)

[39] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing. CoRR **abs/1910.03771** (2019), https://doi.org/10.48550/arXiv.1910.03771, URL http://arxiv.org/abs/1910.03771

[40] Ye, F., Manotumruksa, J., Yilmaz, E.: MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In: Lemon, O., Hakkani-Tür, D., Li, J.J., Ashrafzadeh, A., García, D.H., Alikhani, M., Vandyke, D., Dusek, O. (eds.) 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2022), pp. 351–360, Association for Computational Linguistics (2022)

[41] Zamani, H., Craswell, N.: Macaw: An extensible conversational information seeking platform. In: Huang, J.X., Chang, Y., Cheng, X., Kamps, J., Murdock, V., Wen, J.R., Liu, Y. (eds.) 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020), pp. 2193–2196, ACM (2020), https://doi.org/10.1145/3397271.3401415

[42] Zharikova, D., Kornev, D., Ignatov, F., Talimanchuk, M., Evseev, D., Petukhova, K., Smilga, V., Karpov, D., Shishkina, Y., Kosenko, D., Burtsev, M.: DeepPavlov dream: Platform for building generative AI assistants. In: Bollegala, D., Huang, R., Ritter, A. (eds.) 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), pp. 599–607, Association for Computational Linguistics (2023), https://doi.org/10.18653/v1/2023.acl-demo.58

[43] Zhu, L., Li, W., Mao, R., Pandelea, V., Cambria, E.: PAED: Zero-shot persona attribute extraction in dialogues. In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), pp. 9771–9787, Association for Computational Linguistics (2023), https://doi.org/10.18653/v1/2023.acl-long.544